PUCRS

FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

RAFAEL CAUDURO OLIVEIRA MACEDO

ON THE ANALYSIS OF REMD PROTEIN STRUCTURE PREDICTION SIMULATIONS FOR REDUCING VOLUME OF ANALYTICAL DATA

Porto Alegre

2017

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ON THE ANALYSIS OF REMD PROTEIN STRUCTURE PREDICTION SIMULATIONS FOR REDUCING VOLUME OF ANALYTICAL DATA

RAFAEL CAUDURO OLIVEIRA MACEDO

Dissertation presented as a partial requisite to obtain the Master degree in Computer Science in the Pontifícia Universidade Católica do Rio Grande do Sul.

Adviser: Professor Dr. Osmar Norberto de Souza

Porto Alegre
2017

Ficha Catalográfica

M141 Macedo, Rafael Cauduro Oliveira

On the Analysis of REMD Protein Structure Prediction Simulations for Reducing Volume of Analytical Data / Rafael Cauduro Oliveira Macedo . – 2017.
146 f.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Osmar Norberto de Souza.

1. Replica Exchange Molecular Dynamics. 2. Predição de Estruturas de Proteínas. 3. Filtragem. 4. Métrica de Qualidade. I. Souza, Osmar Norberto de. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).

Rafael Cauduro Oliveira Macedo

On the Analysis of REMD Protein Structure Prediction Simulations for Reducing Volume of Analytical Data

This Dissertation/Thesis has been submitted in partial fulfillment of the requirements for the degree of Doctor/Master of Computer Science, of the Graduate Program in Computer Science, School of Computer Science of the Pontifícia Universidade Católica do Rio Grande do Sul.

Sanctioned on August 30th, 2017.

COMMITTEE MEMBERS:

Prof. Dr. Rafael Andrade Caceres (PPG-CS/UFCSPA)

Prof. Dr. Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS)

Prof. Dr. Osmar Norberto de Souza (PPGCC/PUCRS - Advisor)

DEDICATION

I dedicate this study to my mother Carmen Sylvia Cauduro Oliveira Macedo, to whom I own all my creativity; my step-mother Maria Valesca Azevedo Macedo, who despite all the adversities of life still believed in me and embraced me as a son; and lastly to my father Roberto Manoel Juckowsky Macedo, for being my role-model and teaching me to be whom I am today.

To the apple tree that gestated the apple, it had no idea that it would result in the discovery of the laws of gravity. We are constantly changing the world without knowing. The greatest enigma of life is to find out if it was for the best or worst.

Rafael C. O. Macedo

ACKNOWLEDGEMENTS

Various people and organizations contributed to the realization of this work. I would like to profoundly thanks everyone who directly or indirectly helped in this journey so far.

This dissertation was achieved in cooperation with Dell Inc. using incentives of Brazilian Informatics Law (Law nº 8.2.48 of 1991) and also using the incentives of the PROSUP program of CAPES. Thank you for the vote of confidence.

To my colleagues in the LABIO group, whose work would pose almost impossible without, especially Carlos Sequeiros, Michele Tanus, Thiago Lipinski Paes and Vanessa Paixão Côrtes. To all the healthy discussions we had, and the unimaginable enlightenment this provided me. To all the laughter we had, which made all hard work worth it. To your disposition to help me (regardless of the trivial problems I had), which became more frequent than I would have liked it.

To my other colleagues in FARMINF group, whose I unfortunately do not share the same daily conviviality, but are nevertheless dear to me.

To all my colleagues in the master and doctorate degree, which provided me goals to aim for and for the marvelous projects I could take part in.

To all the teachers and professors I had in life, whose passionate work made me climb where I am today. For the warm welcome and opportunities provided, which led me step by step into my master's degree, and for teaching me the required tools to sew my wings, so that I could soar high in the skies to reach my objectives and ambitions.

To the members of the examination board, whose critics and helpful suggestion most surely will improve this dissertation.

To my orienting professor, who lit the path when things were dark and all the learning patiently passed on.

And finally, but not least, to my family and friends, who made life worth living. You will always have a special place in my heart.

Thank you!

ANALISE DE SIMULAÇÕES REMD PARA PREDIÇÃO DE ESTRUTURAS DE PROTEÍNAS PARA REDUZIR O VOLUME DE DADOS ANALÍTICOS

RESUMO

Proteínas executam um papel vital em todos os seres vivos, mediando uma série de processos necessários para a vida. Apesar de existirem maneiras de determinar a composição dessas moléculas, ainda falta-nos conhecimentos suficiente para determinar de uma maneira rápida e barata a sua estrutura 3D, que desempenha um papel importante na suas funções. Um dos principais métodos computacionais aplicados ao estudo das proteínas e o seu processo de enovelamento, o qual determina a sua estrutura, é Dinâmica Molecular. Um aprimoramento deste método, conhecido como Replica Exchange Molecular Dynamics (ou REMD), é capaz de produzir resultados muito melhores, com o revés de significativamente aumentar o seu custo computacional e gerar um volume muito maior de dados. Esta dissertação apresenta um novo método de otimização deste método, intitulado Filtragem de Dados Analíticos, que tem como objetivo otimizar a análise pós-simulação filtrando as estruturas preditas insatisfatórias através do uso de métricas de qualidade absolutas. A metodologia proposta tem o potencial de operar em conjunto com outras abordagens de otimização e também cobrir uma área ainda não abordada por elas. Adiante, a ferramenta SnapFi é apresentada, a qual foi designada especialmente para o propósito de filtrar estruturas preditas insatisfatórias e ainda operar em conjunto com as diferentes abordagens de otimização do método REMD. Um estudo foi então conduzido sobre um conjunto teste de simulações REMD de predição de estruturas de proteínas afim de elucidar uma séries de hipóteses formuladas sobre o impacto das diferentes temperaturas na qualidade final do conjunto de estruturas preditas do processo REMD, a eficiência das diferentes métricas de qualidade absolutas e uma possível configuração de filtragem que utiliza essas métricas. Foi observado que as temperaturas mais altas do método REMD para predição de estruturas de proteínas podem ser descartadas de forma segura da análise posterior ao seu término e também que as métricas de qualidade absolutas possuem uma alta variância (em termos de qualidade) entre diferentes simulações de predições de estruturas de proteínas. Além disso, foi observado que diferentes configurações de filtragem que utilize tais métricas carrega consigo esta variância.

Palavras-Chave: Replica Exchange Molecular Dynamics, Predição de Estruturas de Proteínas, Filtragem, Métrica de Qualidade

ON THE ANALYSIS OF REMD PROTEIN STRUCTURE PREDICTION SIMULATIONS FOR REDUCING VOLUME OF ANALYTICAL DATA

ABSTRACT

Proteins perform a vital role in all living beings, mediating a series of processes necessary to life. Although we have ways to determine the composition of such molecules, we lack sufficient knowledge regarding the determination of their 3D structure in a cheap and fast manner, which plays an important role in their functions. One of the main computational methods applied to the study of proteins and their folding process, which determine its structure, is Molecular Dynamics. An enhancement of this method, known as Replica-Exchange Molecular Dynamics (or REMD) is capable of producing much better results, at the expense of a significant increase in computational costs and volume of raw data generated. This dissertation presents a novel optimization for this method, titled Analytical Data Filtering, which aims to optimize post-simulation analysis by filtering unsatisfactory predicted structures via the use of different absolute quality metrics. The proposed methodology has the potential of working together with other optimization approaches as well as covering an area still untouched at large by them to the best of the author knowledge. Further on, the SnapFi tool is presented, a tool designed specially for the purpose of filtering unsatisfactory structure predictions and also being able to work with the different optimization approaches of the Replica-Exchange Molecular Dynamics method. A study was then conducted on a test dataset of REMD protein structure prediction simulations aiming to elucidate a series of formulated hypothesis regarding the impact of the different temperatures of the REMD process in the final quality of the predicted structures, the efficiency of the different absolute quality metrics and a possible filtering configuration that take advantage of such metrics. It was observed that high temperatures may be safely discarded from post-simulation analysis of REMD protein structure prediction simulations, that absolute quality metrics posses a high variance of efficiency (regarding quality terms) between different protein structure prediction simulations and that different filtering configurations composed of such quality metrics carry on this inconvenient variance.

Keywords: Replica-Exchange Molecular Dynamics, Protein Structure Prediction, Filtering, Quality Metric

LIST OF FIGURES

Figure 1 - Chemical Structure of an Amino Acid.....26
Figure 2 - The 20 Different Types of Amino Acids.....27
Figure 3 - Two Amino Acids Joining Together to Form a Simple Peptide Via a Peptide Bond.....27
Figure 4 - A Polypeptide Structure Diagram Divided According to its Amino Acid Residues.....28
Figure 5 - Hierarchical Levels of Proteins.....28
Figure 6 - Schematic Representing the Torsion Angles of a Peptide.....29
Figure 7 - Schematic of an α Helix.....30
Figure 8 - Schematic of Both Types of β Sheets.....31
Figure 9 - Tertiary Structure Representation of a Computationally Designed Peptide.....32
Figure 10 - Cartoon Drawing and Partial Atom Representation of the Structure of Hemoglobin.....33
Figure 11 - Energy Landscape of a Protein Folding.....38
Figure 12 - Diagram Illustrating a One-Dimensional Minima Problem.....41
Figure 13 - Illustration of the Metropolis Criterion for the Acceptance or Rejection of Movements in the Monte Carlo Method.....43
Figure 14 - Schematic of a REMD simulation.....44
Figure 15 - Graphical Representation of Several Aspects of the Boltzmann Distribution.....45
Figure 16 - A Ramachandran Plot of the Torsion Angles that Specify Protein Backbone Conformation.....50
Figure 17 - Illustration of Gradient Descent on a Series of Level Sets.....59
Figure 18 - Pie Chart of the Optimization Categories Assigned to Related Works.....72
Figure 19 - Current Workflow Diagram of Optimizing and Running a REMD PSP Simulation.....78
Figure 20 - Proposed Workflow Diagram of Optimizing and Running a REMD PSP Simulation.....79

Figure 21 - Example of a Filtering Configuration File.....81
Figure 22 - Workflow Diagram of the SnapFi Filtration Process.....83
Figure 23 - Top 1% Scored Structures Clustered by Temperature According to Each Absolute Quality Metric Versus GDT_TS.....85
Figure 24 - Worst 1% Scored Structures Clustered by Temperature According to Each Absolute Quality Metric Versus GDT_TS.....86
Figure 25 - Cumulative Distribution of the Top Scored Structures Predicted According to GDT_TS Versus the Temperature in Which They were Extracted (R1).....87
Figure 26 - Cumulative Distribution of the Top Scored Structures Predicted According to GDT_TS Versus the Temperature in Which They were Extracted (R2).....87
Figure 27 - Cumulative Distribution of the Top Scored Structures Predicted According to GDT_TS Versus the Temperature in Which They were Extracted (R3).....88
Figure 28 - Quartile Distribution of GDT_TS Plot for the Top 1% Best Scored Structures According to Each Absolute Quality Metrics and for Each Temperature of the Simulation (R1).....89
Figure 29 - Quartile Distribution of GDT_TS Plot for the Top 1% Best Scored Structures According to Each Absolute Quality Metrics and for Each Temperature of the Simulation (R2).....90
Figure 30 - Quartile Distribution of GDT_TS Plot for the Top 1% Best Scored Structures According to Each Absolute Quality Metrics and for Each Temperature of the Simulation (R3).....90
Figure 31 - Overlapping Cartoon Drawings of the Top Predicted Structure of Protein 1L2Y Versus the Experimentally Determined Structure.....97
Figure 32 - Overlapping Cartoon Drawings of the Top Predicted Structure of Protein 1UAO Versus the Experimentally Determined Structure.....97
Figure 33 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1UAO REMD PSP Simulation.....99
Figure 34 - GDT_TS Histograms of the REMD PSP Simulation of Protein 1L2Y (20%).....102

Figure 35 - GDT_TS Histograms of the REMD PSP Simulation of Protein 1L2Y (10%).....103
Figure 36 - GDT_TS Histograms of the REMD PSP Simulation of Protein 1E0L (20%).....104
Figure 37 - GDT_TS Histograms of the REMD PSP Simulation of Protein 1E0L (25%).....104
Figure 38 - Filtering Configuration Input File for the SnapFi Tool.....105
Figure 39 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1E0L and the Remaining Structures After the Proposed Filtering Method is Applied.....106
Figure 40 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1FME and the Remaining Structures After the Proposed Filtering Method is Applied.....106
Figure 41 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1L2Y and the Remaining Structures After the Proposed Filtering Method is Applied.....107
Figure 42 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1LE1 and the Remaining Structures After the Proposed Filtering Method is Applied.....107
Figure 43 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1PSV and the Remaining Structures After the Proposed Filtering Method is Applied.....108
Figure 44 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1RIJ and the Remaining Structures After the Proposed Filtering Method is Applied.....108
Figure 45 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1UAO and the Remaining Structures After the Proposed Filtering Method is Applied.....109
Figure 46 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 1VII and the Remaining Structures After the Proposed Filtering Method is Applied.....109

Figure 47 - GDT_TS Histograms of All Predicted Structures Generated by the REMD PSP Simulation of the Protein 2WXC and the Remaining Structures After the Proposed Filtering Method is Applied.....110
Figure A.1 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1E0L REMD PSP Simulation.....142
Figure A.2 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1FME REMD PSP Simulation.....142
Figure A.3 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1L2Y REMD PSP Simulation.....143
Figure A.4 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1LE1 REMD PSP Simulation.....143
Figure A.5 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1PSV REMD PSP Simulation.....144
Figure A.6 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1RIJ REMD PSP Simulation.....144
Figure A.7 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1UAO REMD PSP Simulation.....145
Figure A.8 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 1VII REMD PSP Simulation.....145
Figure A.9 - Cumulative Distribution Plot of the Top GDT_TS Bands for the Protein 2WXC REMD PSP Simulation.....146

LIST OF TABLES

Table 1 - Functionality Abridgment of Relative Quality Metrics.....51
Table 2 - Functionality Abridgment of Absolute Quality Metrics.....54
Table 3 - Pilot Search Results.....63
Table 4 - Number of Studies Found on Each Database.....65
Table 5 - Related Works Found for Each Database.....66
Table 6 - Related Works with Assigned Optimization Category.....66
Table 7 - Proteins Simulated in the REMD Dataset.....73
Table 8 - Number of Replicas for Each Protein.....74
Table 9 - Simulations Protocols Used for Testing the Additional Hypothesis.....76
Table 10 - RMSD Calculation Intervals for Each Protein.....77
Table 11 - Analysis of Clusters' Centroids According to RMSD & Population.....91
Table 12 - RMSD Scores According to Each Top Scored Structure.....93
Table 13 - Temperatures in which 50% Convergence of the Top GDT_TS Bands are Attained.....96
Table 14 - Temperatures in which 80% Convergence of the Top GDT_TS Bands are Attained.....96
Table 15 - Number of Predicted Structures Within the Achieved GDT_TS Bands of the 1PSV PSP Simulation.....99
Table 16 - Number of Predicted Structures Within the Achieved GDT_TS Bands of the 1VII PSP Simulation.....100
Table 17 - Number of Predicted Structures Within the Achieved GDT_TS Bands of the 1LE1 PSP Simulation.....100
Table 18 - Top RMSD Score From the Entire REMD PSP Simulation of Each Protein Versus the RMSD Score Extracted by the Top Scored Structure According to Each Absolute Quality Metric.....101
Table 19 - Number of Structures Contained in the Initial and Filtered Ensembles of Predicted Structures for Each Simulation.....110

LIST OF ACRONYMS

3D - Tridimensional

BC - Balance Condition

BLAST - Basic Local Alignment Search Tool

CASP - Critical Assessment of Structure Prediction

Cut-REMD - Cutoff Temperature Replica-Exchange Molecular Dynamics

Cu-MD - Cutoff Molecular Dynamics

dDFIRE - dipole Distance-scaled, Finite Ideal Gas Reference

DFIRE - Distance-scaled, Finite Ideal Gas Reference

DNA - Deoxyribonucleic Acid

DOPE - Discrete Optimized Protein Energy

EAF - Exchange Attempt Frequency

GDT_TS - Global Distance Test - Total Score

GOAP - Generalized Orientation-dependent All-atom Potential

GOAP_AG - Generalized Orientation-dependent All-atom Potential (Angular)

ID - Identifier

LABIO - Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas

MD - Molecular Dynamics

MMCM - Multiple Markov Chain Method

NMR - Nuclear Magnetic Resonance

NP - Non-Polynomial

PDB - Protein Data Bank

PSP - Protein Structure Prediction

QCS - Quality Control Score

REMD - Replica-Exchange Molecular Dynamics

RMSD - Root-Mean-Square Deviation

SB - Structural Bioinformatics

SEE - Secondary Structure Element

LIST OF SYMBOLS

Å - Ångström.....31
α - Alpha.....23
β - Beta.....23
χ - Chi.....28
∆ - Delta.....42
e - Euler.....42
ω - Omega.....28
φ - Phi.....28
ψ - Psi.....28

TABLE OF CONTENTS

1. INTRODUCTION....21
1.1. Organization....24
2. THEORETICAL BASE....25
2.1. Proteins and their composition....25
2.1.1. Primary structure of proteins....28
2.1.2. Secondary structure of proteins....29
2.1.2.1. Alpha helix....29
2.1.2.2. Beta sheets....30
2.1.3. Tertiary structure of proteins....31
2.1.4. Quaternary structure of proteins....32
2.2. The proteins structure prediction (PSP) problem....33
2.3. Computational methods for predicting the 3D structure of proteins....34
2.3.1. Comparative modeling....34
2.3.2. Fold recognition....34
2.3.3. First principle methods with database information....35
2.3.4. First principle methods without database information....36
2.4. The Levinthal paradox....37
2.5. Molecular dynamics....39
2.6. Monte Carlo....40
2.7. Replica-Exchange Molecular Dynamics....43
2.8. AMBER14....47
2.9. CASP: Critical Assessment of Structure Prediction....48
2.10. Ramachandran Plot....49
2.11. Quality metrics....50
2.11.1. Relative quality metrics....50
2.11.1.1. RMSD....51
2.11.1.2. GDT_TS....52
2.11.1.3. QCS....52
2.11.2. Absolute quality metrics....53
2.11.2.1. DFIRE....55
2.11.2.2. dDFIRE....55
2.11.2.3. DOPE....56
2.11.2.4. GFactor....56
2.11.2.5. GOAP....56
2.11.2.6. OPUS-PSP....57
2.11.2.7. Probscore....57
2.11.2.8. RWplus....58
2.11.2.9. Minimized Energy....58
3. MOTIVATION AND OBJECTIVES....61
3.1. Motivation....61
3.2. Broad Objectives....62
3.3. Specific Objectives....62
4. RELATED WORKS....63
5. METHODOLOGY....73

5.1. REMD PSP Simulations Test Dataset....73
5.2. Case Study of a REMD PSP Simulation: Protein 1UNC....75
5.3. Relative Metrics Calculation....77
6. DISCUSSION AND RESULTS....78
6.1. Analytical Data Filtering Methodology....78
6.2. Introducing the SnapFi Tool....79
6.3. Discussion & Results: Case Study of the Protein 1UNC....83
6.3.1. First Formulated Hypothesis....84
6.3.2. Second Formulated Hypothesis....86
6.3.3. Third Formulated Hypothesis....88
6.4. Testing Additional Hypothesis Using the Protein 1UNC....91
6.4.1. First Additional Formulated Hypothesis....91
6.4.2. Second Additional Formulated Hypothesis....93
6.5. Preliminary Filtering Configuration Proposed....94
6.6. Discussion & Results: REMD PSP Simulations Test Dataset....95
6.6.1. Verification of the Second Formulated Hypothesis....96
6.6.2. Verification of the Third Formulated Hypothesis....98
6.7. Final Filtering Configuration Proposed....102
7. CONCLUSIONS....112
7.1. SnapFi Tool....112
7.2. Formulated Hypothesis....112
7.3. Proposed Filtering Methodology....113
7.4. Final Considerations....114
REFERENCES....116
APPENDIX A - CUMULATIVE DISTRIBUTION PLOTS OF THE TOP GDT_TS BANDS FOR EACH REMD SIMULATION OF TEST DATASET....142

1. INTRODUCTION

Understanding biological mechanisms and processes has a major impact on the scientific community, enabling the research of new drugs, early prediction of diseases related to genetic disorders and novel treatments for previously cryptic diseases and resilient infections. Biological macromolecules known as proteins are the primary components of cell structures and have a pivotal role in their function, executing and intermediating a myriad of biological processes. Knowledge of the structure, dynamic and functions of such molecules can greatly enhance our understanding of living beings, from small to large, some of which posses a much higher control of natural phenomenons such as aging, illnesses, etc. Although there are experimental ways to determine the biological function of a biological molecule, this functional analysis cannot describe the physical and chemical behavior native to a molecule. Some times the molecule chemical composition itself (such as electrical charge, pH, etc) does not fully determine its biological function. Studying the tridimensional structure of such molecules has a lot of advantages therefore. In fact, there were 12 Nobel Prizes in chemistry and physiology or medicine awarded for work in this field from 1956 to 2006. Almost one in four chemistry prizes since 1956 have been for structure work. [Mic07].

Every protein is formed by a sequence of 20 different amino-acid residues that, by interacting physicochemically, form its unique 3D structure. These sequences, translated from the DNA, may be retrieved from the GenBank [Ben17] at not cost over the Internet. The three-dimensional structures of these sequences, on the other hand, are not so abundant. While some of them may be obtained from a few databases around the world, most notably the Protein Data Bank (or PDB) [Ber00], the majority of these structures are still unknown to us. Currently, there are around 96 million non-redundant unique proteins sequences. (BLAST, accessed September 15th, 2016) [Alt90]. However, in PDB, we only find around 123,837 proteins' 3D structures. By eliminating redundancy by filtering similar structures, only 1,375 different protein folds are obtained (PDB Statistics, accessed Sept 11th, 2017).

There is a large gap between our capability of generating new protein sequences and our capability of solving the 3D structure of new proteins, with different protein folding from those already known [Pav11]. An important tool to reduce this gap is Structural Bioinformatics. Structural Bioinformatics, as defined by Luscombe et al., is "conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale." [Lus01]. Structural Bioinformatics, an important subarea of Bioinformatics, aims to understand how a protein reaches its 3D structure from just its amino-acid sequence. This problem is called the Protein Structure Prediction problem (or PSP), which emerged more than 60 years ago. Although progress has been made, it still

unsolved at large. Given its biological importance and NP-Complete complexity, the protein structure prediction (or PSP) problem is one of the big challenges of modern science [Cre09, Dil12].

There are experimental techniques available to determine a protein structure with a high degree of precision. Among these are Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography. These techniques, however, are costly both in matters of time and money. They also require an specialist who known the techniques in length. Faster and cheaper ways to achieve the same, or at least similar results, are in great need for the scientific community.

Seeking to solve the PSP problem, a series of computational methods are being proposed and/or enhanced. Among these are the homology modeling [Mar00], fold recognition [Ale95, Söd05, Käl12], *de novo* methods [Sim99, Zha05, Lam16] and the *ab initio* methods [Che05, Jay06]. Between those cited, the *ab initio* methods, while not the most precise, are the only ones that are able to result in novel protein folding patterns [Flo06]. That is because they use only the primary amino acid sequence of a protein and attempt to simulate its folding process in complex software packages. Molecular Dynamic (MD) methods are used for this, which calculate every atom movement and interaction between other atoms in the system with a series of variables such as the atoms force fields and interatomic potentials. These methods result in one or several predictions for the native structure of the simulated protein, that is, the 3D structure of the protein found in nature and usually the most physically stable.

The main problem with common MD methods is that the energy landscape of proteins is far from being a perfect parabola. The true nature of the energy landscape of the proteins is very rough [Fra91], due to this and aided by the fact that in conventional MD methods the system temperature remain constant, the predicted protein conformations tend to get trapped in one of several local minimums. This impair the capacity of the method to have the desired sampling efficiency. Beyond that, the temperature chosen for the simulation has a large impact on the results, as in high temperature a larger degree of movement is present in the atoms, making them less prone to converge into a stable structure. On the other hand, low temperatures cause the simulations to converge too rapidly in local minimums, never reaching better structural conformations.

The REMD (Replica-Exchange Molecular Dynamics) was designed to solve this problem with a few modifications. Firstly, several different replicas would be simulated in parallel at different temperatures, and these replicas can exchange their system temperatures at fixes intervals. Secondly, the Monte Carlo method is used to enable detrimental movements in the energy landscape as a mean to overcome peaks separating local minimums from the global minimum [Han97, Sug99]. The REMD method is being increasingly more used over the last few years, specially in studies of specific protein's folding and dynamic [Sue03, Sei05, Lei07, Xu08, Lei08, Bec07, Kou10].

A significant drawback for this method, however, is that even with powerful computers, the execution of this protein fold simulation algorithm and the posterior analysis of the resulting data can take up to several weeks.

If this time can be shortened, it could reduce both the computational and personnel costs for future projects significantly.

Although the REMD method guarantee a wider sampling in the energy landscape over conventional MD methods, which is generally a good thing to have, the extra predicted structures are not necessarily closer to the native structure of the target protein. Having this issue in sight, we decided to develop a tool capable of receiving a large volume of predicted structures (packed together in an ensemble file) from a finished REMD PSP simulation aiming to predict a protein with a still undefined 3D structure and filter the unsatisfactory predictions, that is, predicted structures distant from the native structure of the protein. In order to do so, several absolute quality metrics are used. Further information about these metrics are discussed in chapter 2.10. Additionally, a study directed to determine the impact of the different temperatures on the final quality results in each REMD replica was performed as part of the test to find good filtering methodologies.

Although the idea is rather simple, no such tool was found in the literature search performed, containing several different virtual databases with extensive use by the scientific community. The proposed tool was also modeled with the specific intent of working together with other optimization techniques of the REMD method, proposed by several different authors. The final output of the tool is an ensemble of predicted structures containing a far lower ratio of unsatisfactory predictions. This ensemble can then be further analyzed by specialized professionals to extract the best predicted structure, that is, the prediction closer to the native structure of the target protein. The smaller and filtered ensemble can make this process considerably faster, and possibly with a higher degree of precision as well.

This dissertation presents the Snapshot Filtration (SnapFi) tool, which purpose is to reduce the massive volume of data generated from a REMD PSP simulation by means of filtering unsatisfactory structure predictions, along with a filtering configuration composed by different absolute quality metrics. It is also presented other techniques that can effectively filter unsatisfactory predictions with an acceptable degree of precision. Although this feature could not have being tested in full, it is also expected that the proposed tool is capable of working together with several other REMD optimization techniques. The tool is also capable of working with conventional MD methods, given that it is properly configurated first.

In order to test possible filters for the SnapFi tool, a series of REMD PSP simulations were used, kindly granted by Lipinksi-Paes and Norberto de Souza [Lip17] through personal communication. The tested set of proteins contained α, β and αβ proteins conformations. All simulations were performed under 50 ns each. The solution structure of the human villin c-terminal headpiece subdomain was used to test several preliminary hypothesis, some even outside the proposed work spectrum of this dissertation, but whose results are nevertheless worth reporting to future colleagues interested in further studying it.

To verify the quality of the filters found and the efficiency of the SnapFi tool, several different comparisons were performed, evaluating the divergence between the final quality results (using both the RMSD and the GDT_TS score)

between the initial ensemble and the final ensemble of predicted structures. A comparison of the resulting quality of filtering the initial predicted structures ensemble with each of the absolute quality metrics alone was performed as well. Overall, the tested filtering methods could not achieve the desired reduction of computational cost, but are capable of providing more reliable results when comparing to simple filtering means (e.g. using a single absolute quality metric to filter out unsatisfactory structure predictions).

1.1. Organization

This dissertation is organized in 8 chapters:

- The first chapter presents the introduction to the problem and an overall description of the study performed.

- The second chapter contain the theoretical base required for the understanding the performed study. Firstly, the proteins and their composition are explained, followed shortly by a brief explanation of the PSP problem. The computational methods for predicting the 3D structures of proteins are then summarized. The Levinthal paradox is then introduced followed by the explanation of Molecular Dynamics. The Monte Carlo and replica exchange are then described. The AMBER14 software in then introduced. The biannual CASP conference is also briefly described. Finally, different ways of evaluating a predicted protein structures are presented.

- In the third chapter, the motivation and the objectives of this dissertation are presented.

- Chapter 4 presents a detailed analysis of the related works pertaining REMD PSP simulations optimizations.

- Chapter 5 contains the methodology used in this study, describing the proteins tested, the REMD PSP test dataset used, and so on.

- The sixth chapter finally present the SnapFi tool along with the proposed filter methodology and the results obtained from this study.

- Finally, chapter 7 presents the final consideration and the conclusions extracted for this work.

- The last chapter contains the bibliography used in this study.

2. THEORETICAL BASE

This chapter addresses a theoretical base on the main concepts that will be used on this dissertation. Firstly, the concepts of proteins are addressed, followed shortly by the PSP problem. In the sequel, the methods of PSP are displayed with further details. Lastly, the protein structure quality metrics used in this work are explained in an overall manner.

2.1. Proteins and their composition

Proteins are vital for all biological processes in our body. They mediate virtually every process that takes place in a cell, exhibiting an almost endless diversity of functions. They are the most abundant biological macromolecules, occurring in all cells and all parts of cells. Not only abundant in quantity, they are also abundant in variety, where thousands of different kinds may be found in a single cell. All proteins of every organism, from the simplest bacteria to human beings, are constructed from the same ubiquitous set of 20 amino acids. Since each amino acid has a side chain with distinctive chemical properties, this group of 20 precursor molecules may be regarded as the alphabet in which the language of protein structure is written [Leh12]. All 20 of the common amino acids are α amino acids. They have a carboxyl group and an amino group bonded to the same carbon atom (the central α carbon). They differ from each other in their side chains, or R groups, which vary in structure, size, and electric charge, and which influence the solubility of the amino acids in water. Figure 1 represents this common chemical structure of amino acids. While there are more than these 20 types of amino acids, the other existing amino acids are less common. Some of them are residues modified after a protein has been synthesized and others are amino acids present in living organisms but not as constituents of proteins [Leh12]. Figure 2 represents the set of 20 different types of amino acids, grouped by electrical charge and polarity.

To generate a particular protein, amino acids are covalently linked in a characteristic linear sequence. The most remarkable aspect is that cells can produce proteins with strikingly different properties and activities by joining the same 20 amino acids in many different combinations and sequences. From these basic building blocks of life, different organisms can make a wide diversity of products as enzymes, hormones, antibodies, transporters, muscle fibers, the lens protein of the eye, feathers, spider webs, rhinoceros horn, milk proteins, antibiotics, mushroom poisons, and an infinite number of other substances having distinct biological activities. Figure 3 depicts two amino acids (Glycines) joining together to form a simple peptide (a dipeptide) via a peptide bond. When many amino acids are joined together, the resulting structure is called a polypeptide (Figure 4). A polypeptide with enough molecular weights can be called a protein. Although there is no clear

boundary to when a polypeptide is not a protein, a polypeptide is usually addressed as a protein when its molecular weight is 10.000 or higher [Leh12].

Regarding the structure of the proteins, there are four defined hierarchy levels (Figure 5). A description of all covalent bonds linking amino acids residues in a polypeptide chain is the primary structure of a protein. In other words, the primary structure of a protein can be defined as its amino-acid sequence. The secondary structure of a protein refers to particularly stable arrangements of amino acids residues that form recurring structural patterns. The most prominent of those are α helixes and β sheets. Where a regular pattern is not found, the secondary structure is sometimes referred to as undefined or as a random coil. The tertiary structure describes all aspects of the 3D folding of a polypeptide. Whereas the secondary structure refers to the spatial arrangement of amino acid residues that are adjacent in a segment of a polypeptide, the tertiary structure includes longer-range aspects of amino acid sequence, i.e. amino acids that are far apart in the polypeptide sequence and in other secondary structures. Finally, when a protein has two or more polypeptide subunits, the arrangement of these protein subunits in 3D complexes is referred to as quaternary structure [Leh12].

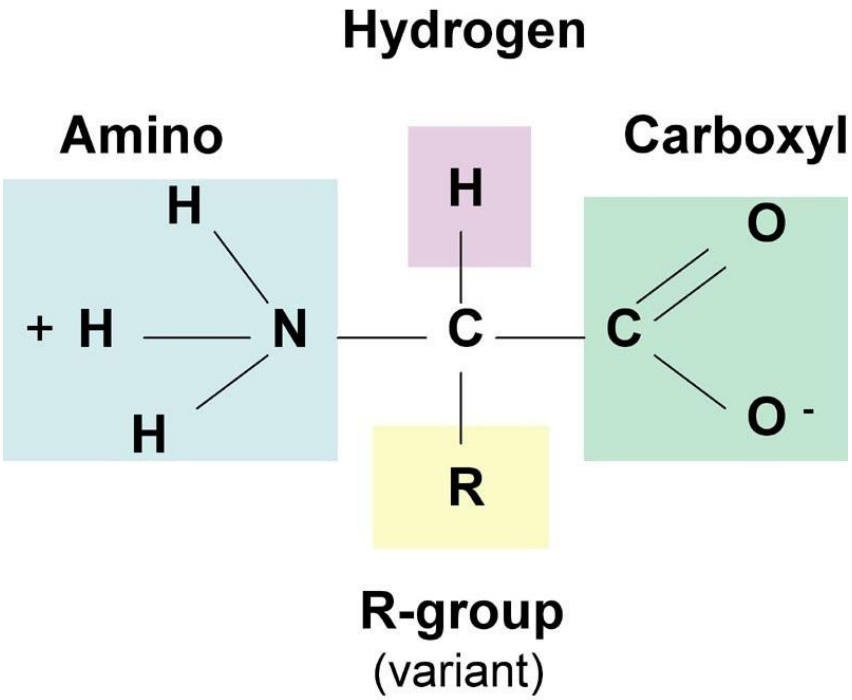

Figure 1 - Chemical structure of an amino acid, where *R* represents the side chains. Regardless of the amino acid type, they all have the same base structure: when un-ionized, a carboboxyl group (-COOH) and an amino group ($-NH_2^+$) bond to the same carbon atom (the α carbon). The amino acid can also assume a zwitterion (or hybrid) form, where the carboxyl group lose its hydrogen atom (becoming COO-) and the amino group gain one (becoming NH_3^+). As the α-carboxylic acid group is a weak acid and the α-amine group is a weak base, the first form is rarely found in nature, being its zwitterion form the most common of the two. The amino acids differ among themselves solely by their side chains, also knows as the *R* group, which also bond to the same central α carbon. Adapted figure from [Joh18].

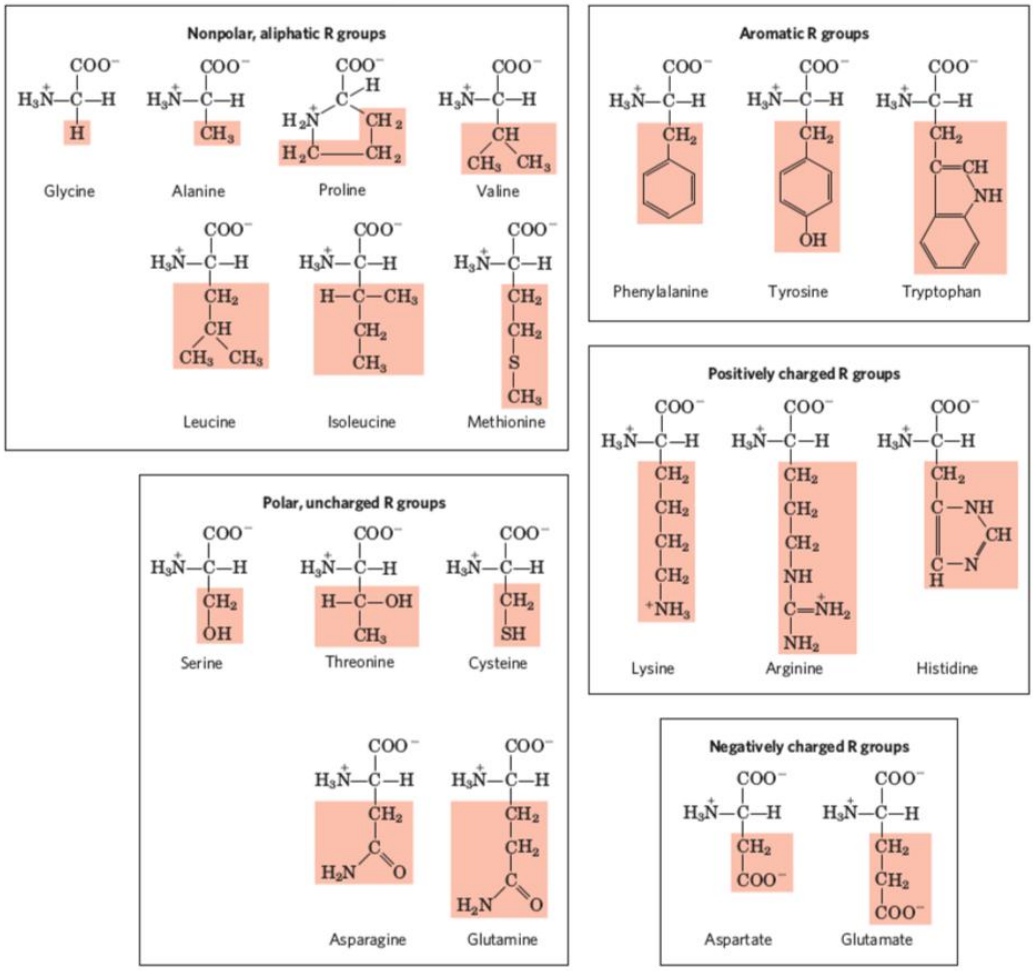

Figure 2 - The 20 different types of amino acids, classified according to charge and polarity. Figure obtained from [Leh12].

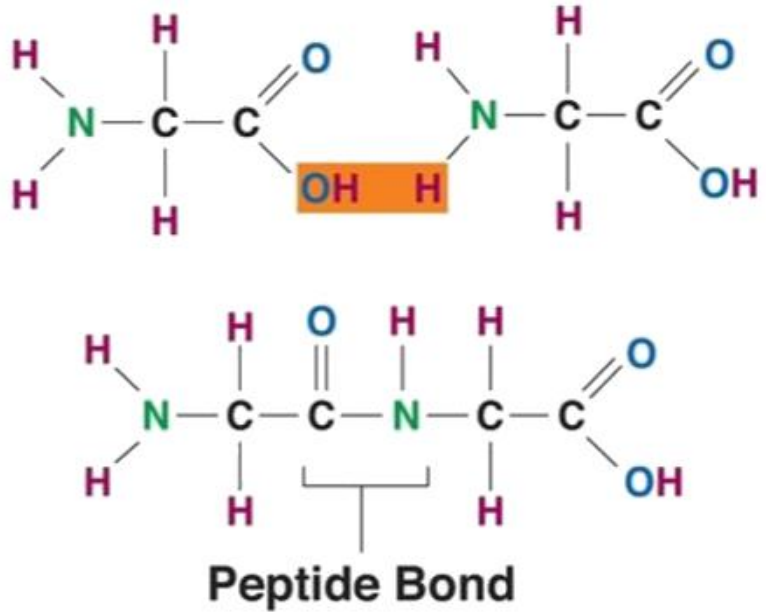

Figure 3 - Two amino acids (Glycines) joining together to form a simple peptide (a dipeptide) via a peptide bond. The excess atoms (H_2O) join together to form a water molecule, unbinding from their original amino acids. Adapted figure from [Zim17].

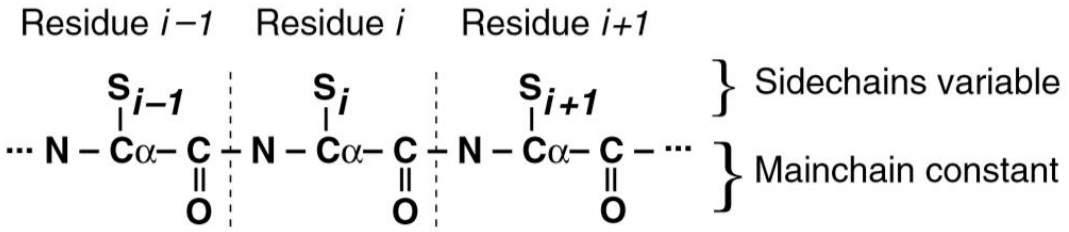

Figure 4 - A polypeptide structure diagram divided according to its amino acid residues. Figure extracted from [Les05].

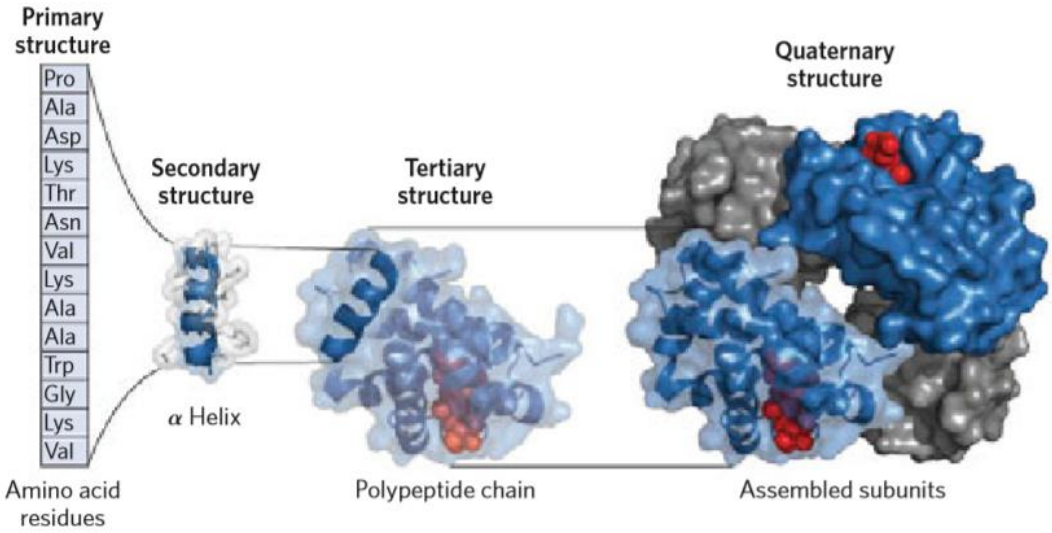

Figure 5 - Hierarchical levels of proteins, ranging from the simple primary structure composed solely by amino acid sequence of the protein to the quaternary structure, containing the arrangement of the subunits of the protein and the position of all atoms in a 3D spectrum. Figure extracted from [Leh12].

2.1.1. Primary structure of proteins

The primary structure of a protein is the description of all covalent bonds (mainly peptide bonds and disulfide bonds) linking its amino acid residues in a polypeptide chain [Leh12]. This amino acid sequence that define peptides and proteins can be obtained through different methods such as mass spectrometry [Ast19], Edman degradation (Edm50) or through the DNA nucleotide sequence that code the protein [Leh12]. By convention, the primary structure of a protein is reported starting from the amino-terminus (N) end to the carboxyl-terminus (C) end.

Peptide conformation are also defined by three dihedral angles (also known as torsion angles) called φ (phi), ψ (psi), and ω (omega), reflecting rotation about each of the three repeating bonds in the peptide backbone [Leh12]. The first two revolves around the N-Cα and the Cα-C, while the latter revolves around the peptide bond. The side chain also posses torsion angles that vary from residue to residue. These are called χ_1 (chi1), χ_2 (chi2), etc. Figure 6 depict these torsion angles.

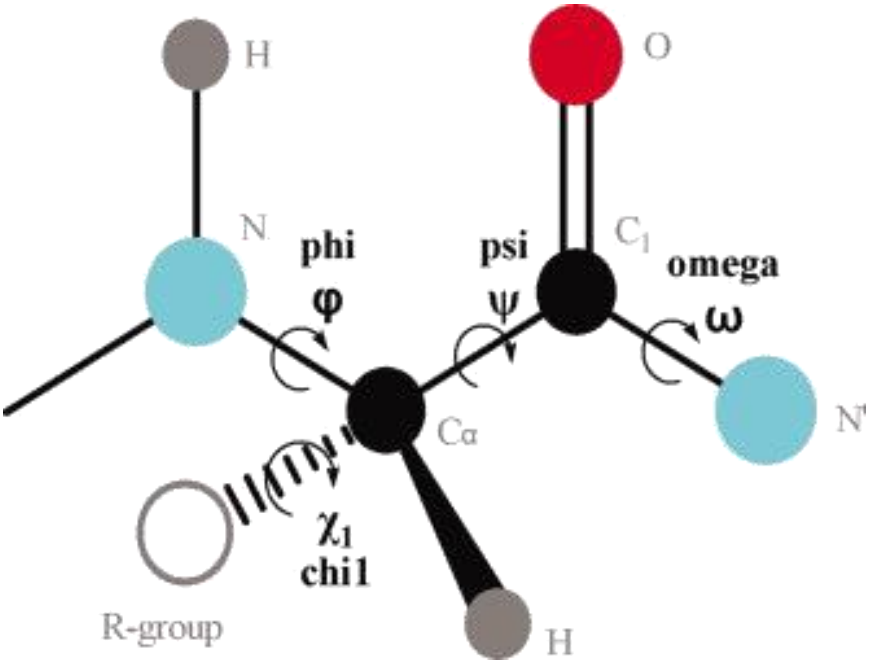

Figure 6 - Schematic representing the torsion angles of a peptide. The *φ* and *ψ* angles revolve around the N-Cα and the Cα-C atoms, while the ω angle revolve around the peptide bond between the two amino acid residues. Figure extracted from [Tie07].

2. 1. 2. Secondary structure of proteins

The secondary structure of proteins refers to the local spatial arrangement of main-chain atoms, without regard to the positioning of its side chains or its relationship to other segments [Leh12]. The secondary structure also includes other forms of bonding between residues beyond the peptide bonds, such as the hydrogen bonds. Among all the possible conformations that the peptide structure can adopt, some are particularly stable and occur on a regular basis in a wide gamma of different proteins. The most prominent are the α helix and the β sheets conformations. These are called regular structures (or regular patterns). These two patterns are particularly common because they result from hydrogen bonds between the nitrogen and the oxygen from the -NH and CO groups of peptide units, without involving the side chain of the peptides. Although the hydrogen bond is a weak bond, their high quantity grant great stability to these structures [Mac13]. These two structures will be further discussed in the following sections. Where a regular pattern is not found, the secondary structure is sometimes referred to as undefined or as a random coil.

2. 1. 2. 1. Alpha helix

The α helix (Figure 7) was the first regular structure theorized by Pauling in 1948 and based on x-ray studies of animal hair, porcupine quills, silk, wool and other materials performed by William Astbury and colleagues in the 1930 decade [Ast31, Ast33, Ast34, Ast35]. The model he theorized, and later

confirmed in work with Corey and coworker Herman Branson [Pau51], was the simplest arrangement the polypeptide chain can assume that maximizes the use of internal hydrogen bonding. In this structure, the polypeptide backbone is tightly wound around an imaginary axis drawn longitudinally through the middle of the helix, and the R groups of the amino acid residues protrude outward from the helical backbone [Leh12].

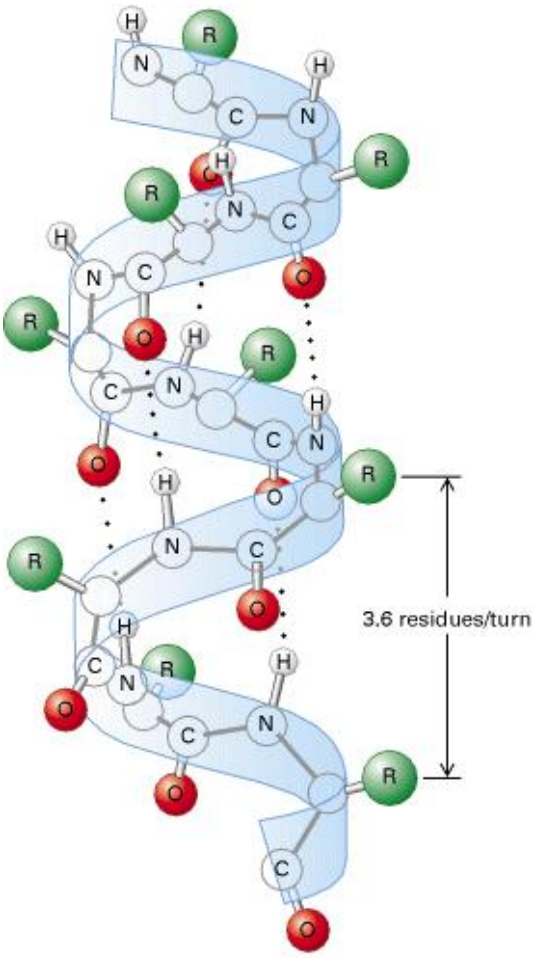

Figure 7 - Schematic of an α helix. The polypeptide chain adopts an helical pattern around an imaginary axis with 3.6 amino acids per turn. The hydrogen bonds between the main chain of the peptide units stabilize the structure. The side chain protrude outward from the helical backbone. Figure extracted from [Was17a].

2.1.2.2. Beta sheets

On their paper published in 1951 [Pau51], Pauling and Corey also predicted another regular structure beyond the α helix. This structure was the β sheet, which is also supported by hydrogen bonds created between the main chain of the peptide units, but it adopts an extended conformations that zigzags instead of forming an helix. The arrangement of several segments side by side, all of which are part of the β sheet, resemble a pleated sheet.

Hydrogen bonds form between adjacent segments of polypeptide chain within the sheet, forming a planar structure. The individual segments that form a β sheet are usually nearby on the polypeptide chain but can also be quite distant from each other in the linear sequence of the polypeptide; they may even be in different polypeptide chains. This fact alone make the β sheet a difficult structure to be predicted using *in silico* methods.

The *R* groups of adjacent amino acids protrude from the zigzag structure in opposite directions, creating the alternating pattern, never interacting between themselves. Depending on the relative orientation of the β sheet segments, it can be classified as either parallel or antiparallel (having the same or opposite amino-to-carboxyl orientations, respectively). Their structure is very similar, with minor changes regarding the repeating period of the segments, where a parallel β sheet has a repeating period of 6.5 Å and the antiparallel β sheet has a repeating period of 7 Å (The Ångström, Å, named after the physicist Anders J. Ångström, is equal to 0.1 nm. Although not an SI unit, it is used universally by structural biologists to describe atomic distances—it is approximately the length of a typical CH or OH bond). The hydrogen bonding patterns are also different, whereas the interstrand hydrogen bond are essentially in line in the antiparallel β sheet and slightly distorted or not in-line for the parallel variant [Leh12]. Figure 8 depicts both variants of the β sheets.

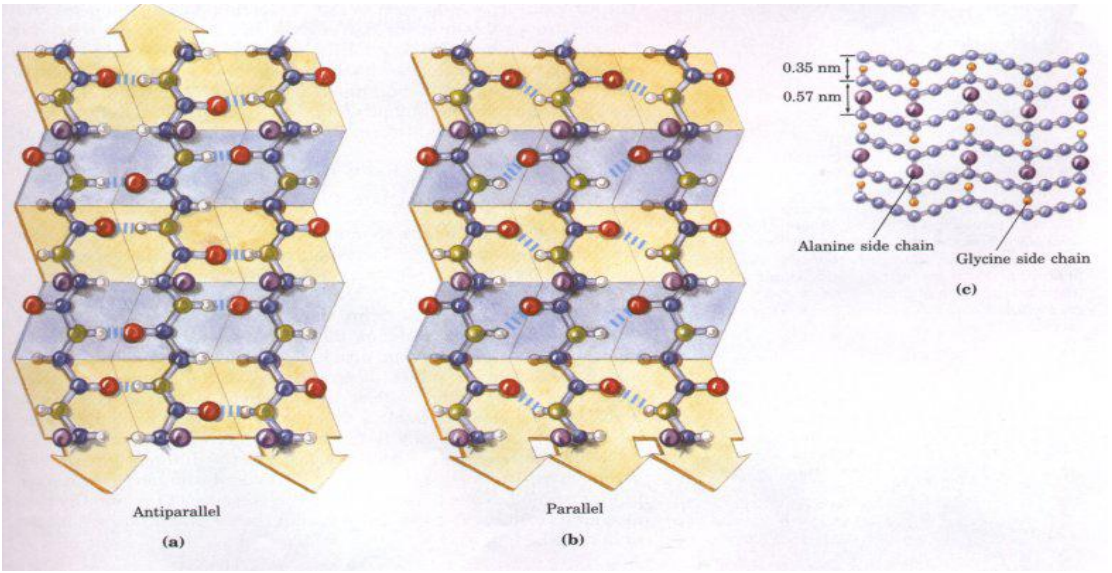

Figure 8 - Schematic of both types of β sheets. (a) antiparallel β sheets. (b) parallel β sheets. Figure adapted from [Leh93].

2. 1. 3. Tertiary structure of proteins

The secondary structure elements are in most cases too simple to execute the complex functions of proteins. This requires a higher degree of organization, which is achieved by the folding of the entire polypeptide chain. [Kes10]. The result of this wide scale folding is called the tertiary structure of a protein.

The tertiary structure of a protein is, therefore, the representation of the arrangements of secondary structures distributed in a 3D space. In short, the protein tertiary structure is defined by its atomic coordinates. Whereas secondary structure includes solely spatial arrangement of amino acid residues that are adjacent in a segment of a polypeptide, tertiary structure includes longer-range aspects of amino acid sequence such as the interaction of atoms in different secondary structures. This structure level is also called the native structure or functional structure of the protein, since the function a protein fulfills is a direct consequence of its 3D structure [Kes10].

While secondary structures are similar in all proteins, there are countless of different types of tertiary structures. The two most common forms of experimentally determining the tertiary structure of a protein is using Nuclear Magnetic Resonance spectroscopy and X-ray crystallography. As already cited previously in Chapter 1, these processes are both time consuming and very costly. Figure 9 depicts an example of the tertiary structure of a protein.

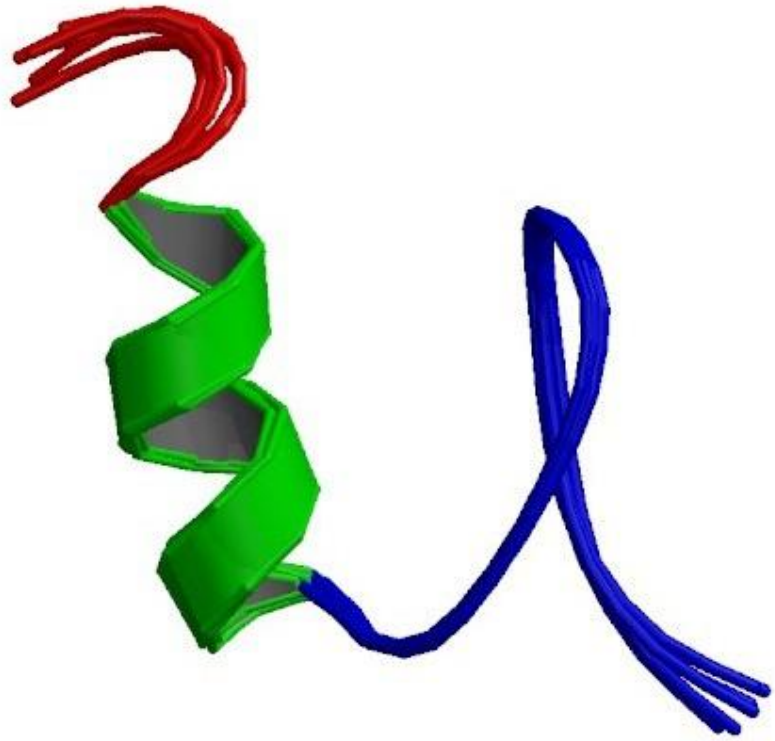

Figure 9 - Tertiary structure representation of a computationally designed peptide extracted from the Protein Data Bank (PDB ID: 1PSV). The green helix is an α helix, while the blue loop contains a β sheet.

2.1.4. Quaternary structure of proteins

Some proteins contain two or more separate polypeptide chains, or subunits, which may be identical or different. The arrangement of these protein subunits in 3D complexes constitutes quaternary structure [Leh12]. These subunits are kept together by means of non-covalent interactions between the subunits, the same forces that maintain the tertiary structure of the proteins stable [Mac13].

An important example of a protein with a quaternary structure is the Hemoglobin present in our blood (Figure 10), which consists of a tetramer, that is, a protein structure containing four subunits proteins. In this case, four globular proteins.

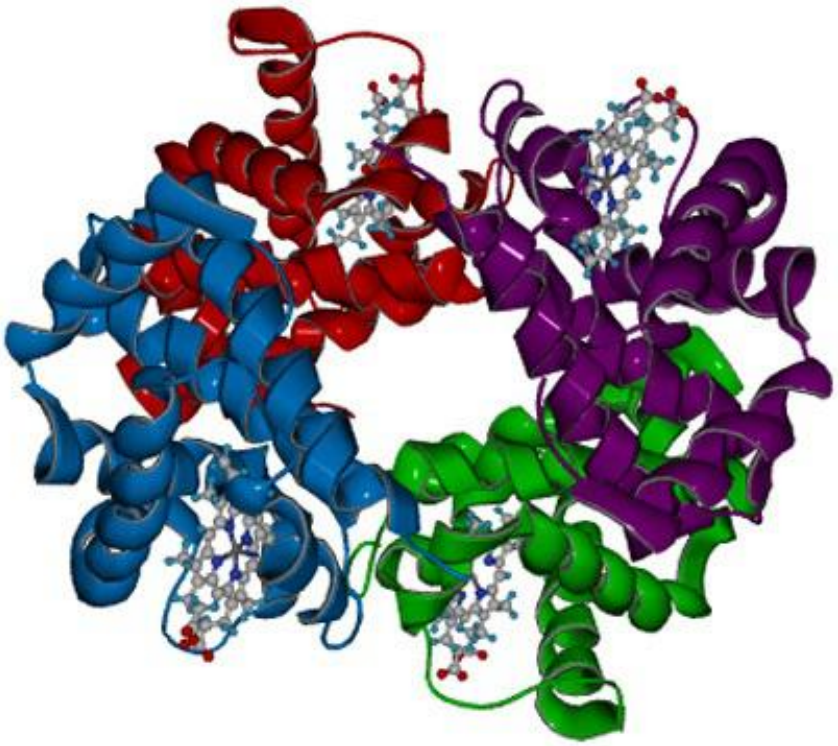

Figure 10 - Cartoon drawing and partial atom representation of the structure of hemoglobin, composed by four equal globular protein subunits. Figure extracted from [Was17b].

2.2. The proteins structure prediction (PSP) problem

The PSP problem is that of finding the native-like structure of a given protein based on its amino-acid sequence only. The amino-acid sequence and its resulting 3D structure are directly linked to the protein function. “[...] each type of protein has a unique amino acid sequence that confers a particular 3D structure. This structure, in turn, confers a unique function.” [Leh12]. The exact amino-acid sequence of a protein is so important that a single change in one of them can result in harmful and often severe consequences. Thousands different human genetic diseases have been traced to the production of defective proteins. The defect can range from a single change in the amino acid sequence (as in sickle cell anemia) to larger missing portions of the polypeptide chain (as in most cases of Duchenne muscular dystrophy).

However, the prediction of the 3D structure of proteins and even simplified approaches are, as shown by Crescenzi et al. [Cre09], an NP-Complete problem. Throughout the past six decades, many different algorithmic approaches have been attempted, and although progress has

been made, the definitive solution to this problem remains a mystery. While the general objective is to predict the 3D structure from the primary sequence of a protein, our current knowledge, and computational power are simply insufficient to solve a problem of such high complexity. Besides the efforts of many different research groups, this problem, even after almost 70 years, remains unsolved and still very pertinent to the society [Hel08]. Perhaps even more now with the recent advances in pharmaceutical industry and the identification of several new genetic related diseases.

2.3. Computational methods for predicting the 3D structure of proteins

Computational methods for protein structure prediction can be classified into four groups according to Floudas et al. [Flo07]:

1. comparative modeling [Bia14, Mar00, Lyr14, Bra12];
2. fold recognition [Bow91, Jon92];
3. first principles methods with database information [Roh04, Sri95] and
4. first principles methods without database information [Osg00].

2.3.1. Comparative modeling

Comparative Modeling (or Homology Modeling) relies on the principle that sequences which are related evolutionarily exhibit similar 3D folded structures, that its sequence similarity suggests structural similarity [Flo07].

This approach is, currently, the method that yields the best results in the PSP area for proteins with a reasonable evolutionary relation. It also confers a significantly high accuracy compared to the other methods. The drawback of Comparative Modeling is that, in order to use this approach, one must have a good structural model, evolutionarily related to the target protein. Many times it is not possible to have such structure at hand, making this approach limited to a certain degree of proteins. Furthermore, this approach cannot result in any novel protein folding as it relies on already existing folding methods. It also does not confer the possibility to further study the folding process of the given protein.

This method, in summary, is an excellent choice when the target protein has a similar protein model, but a poor choice when targeting a novel protein without any similar models. Examples of Comparative Modeling methods are: SWISS-MODEL [Bia14], MODELLER [Mar00], ReformAlign [Lyr14] and PyMOD [Bra12].

2.3.2. Fold recognition

Fold Recognition or Protein Threading, is relatively similar to the Homology Modeling. It differs to the Homology Modeling because it does not use a single homologous model protein, but rather a series of proteins with similar sequence on particular points. That is, even if the target protein does

not posses another model protein with a highly similar amino acid sequence and resolved 3D structure, it could still have specific parts of its amino acid sequence similar to specific parts of other proteins, which fold in a similar form [Lev76, Flo06]. In such cases, the proteins are said to be remotely homologous [Lip17].

It is based on the principle that the number of different folded proteins structures is significantly more limited than the vast number of different sequences generated out of genome projects [Lev76]. The method works by trying to identify remotely homologous proteins within a collection of candidates. When such proteins are identified, the sequence alignment process begin, similar to the Comparative Modeling method. When it is not possible to identify homologies by aligning pairwise sequences, the protein threading technique is used [Jon92]. It then uses statistical data to predict the correct fold of the protein according to the folding of others proteins with a similar sequence that it was aligned.

The drawback, similarly to the Homology Modeling, is that this approach relies on having proteins with similar amino-acid chains in some part [Lip17]. It also relies on statistics and probability that do not guarantee to find the best matches. Examples of Fold Recognition methods are: GENTHREADER [Jon99], 123D [Ale95], ORFEUS [Gin03], PROSPECT (Protein structure prediction and evaluation computer toolkit) [Xu00], Bio Shell-Threading [Gni14], FFAS03server [Jar05], RaptorX server [Käl12], Phyre server [Kel09], HHpred [Söd05], LOOPP server [Teo04], SPARKS-X [Yan11].

2.3.3. First principle methods with database information

The first principle methods (or *ab initio* methods) with database information do not compare a target to a known protein directly, but rather compare fragments, that is, short amino acid subsequences. These short structures can be obtained from the Protein Data Bank [Ber00]. Once appropriate fragments have been identified, they are assembled to a structure, often with the aid of scoring functions and optimization algorithms.

The *ab initio* term refer to methods for structure prediction that do not use experimentally known structures. The scoring functions resemble energy functions, and the fragment assembly with optimization algorithms resembles free energy optimization, therefore this type of method was given its own distinct classification. This classification, however, is somewhat vague, as already cited by Floudas and colleagues in [Lev76], and could (with a certain debate) be incorporated as a specific type of Fold Recognition as fragment assembly methods cannot be considered *ab initio* structure prediction methods in the same strict sense as *ab initio* methods that are based on free energy minimization.

Although the *ab initio* methods are significantly slower and less precise than the other two categories presented, its main advantage is due to the fact that they are capable of predicting novel folding as they are not bound by known protein structures [Flo07]. Due to the fact that fragments are used for comparison, when the first principle methods with database information result

in novel foldings, these are direct results of the composition of motifs or fragments of supersecondary structures of known structure [Tra07].

Examples of first principle methods with database information are: TASSER e ITASSER [Roy10, Zha04], ROSETTA e ROSETTA@home [Roh04, Sim99], FRAGFOLD [Jon01], CABS-Fold [Bla13], SIMFOLD [Chi03], PROFESY (PROFile Enumerating SYstem) [Lee04], A3N (Artificial neural network N-gram-based method) [Dor10a], CReF (Central-residue-fragment-based method) [Dor10b], PEP-FOLD [Lam16], BHAGEERATH [Jay06, Nar06] e QUARK [Xu12].

2. 3. 4. First principle methods without database information

The first principle methods (or *ab initio* methods) that do not rely on database information attempt to predict protein structure make directly use of the Anfinsen's thermodynamic hypothesis [Anf73], which states that the native structure of a given protein corresponds to its lowest free energy state. "It is generally assumed that a protein folding sequence folds to a native conformation or ensemble of conformations that is at or near the global free-energy minimum" [Bra12].

These methods attempt to identify this lowest free energy state of the target protein in its environment based on complex computational simulations that employ the use of the laws of physics and using only its amino acid sequence. The only additional information required by these methods beyond the target protein's primary structure is a suitable potential energy (or force field) function. These functions describe the internal energy of the protein and its interactions with the surroundings. Since predicting protein structures generally involve many atoms, it is not yet feasible to treat these systems using quantum mechanics. The problem, therefore, become much more tractable when turning to empirical potential energy functions, which are much less computationally demanding than quantum mechanics. These potential energy functions posses a cutoff radius which measures in which distance (in Ångströms) the interactions between atoms is calculated. The greater the cutoff value is, the larger the number of interactions is calculated, making the result more precise but significantly more costly in computational terms.

Using such approaches comes at a cost however. As the developed force fields distance from the real world physics (with yet unknown formula), the results of the simulation do not converge to a single accurate result [Sto17]. Thus finding such functions can be considered one of the two main subproblems for finding the correct native-like conformation for a given protein. "[...] the problem of finding native-like conformations for a given sequence can be decomposed into two subproblems: (a) developing an accurate potential and (b) developing an efficient protocol for searching the resultant energy landscape" [Bra12].

A number of laboratories across the world still research better potential energy functions, which in turn are (generally) incorporated in the different softwares solutions to this method that are usually developed by the same

authors. Examples of developed force fields functions are: AMBER [Cor95], CHARMM [Bro83], GROMOS [Chr05] e OPLS [Jor96].

Differently from the method with database information, the first principle method without database information can result in truly unique protein folds. Moreover, this method can be applied to any given target sequence using only physically meaningful potentials and atom representations. These advantages alone are enough reason to motivate its use. The main drawback of this method, on the other hand, is that due to the large degree of freedom to fold a protein, it must consider a massive number of possible conformations. Even the simplest of proteins can demand a large computing power, as the PSP problem was proved to be a NP-Complete problem [Cre09]. This will be further discusses in section 2.4. With such a broad range of targets and the inability to directly or indirectly apply database information, these methods are the most difficult of the protein structure prediction methods [Lev76].

Even though these methods are computationally demanding, first principle structure prediction is an indispensable complementary approach to any knowledge-based approach for several reasons:

1. In some cases, even a remotely related structural homologue may not be available. In these cases, first principle methods are the only alternative;

2. New structures continue to be discovered which could not have been identified by methods which rely on comparison to known structures.

3. Knowledge-based methods have been criticized for predicting protein structures without having to obtain a fundamental understanding of the mechanisms and driving forces of structure formation. First principle structure prediction methods, in contrast, base their predictions on physical models for these mechanisms. As such, they can therefore help to discriminate correct from incorrect modeling assumptions, and to deepen the understanding of the mechanisms of protein folding.

The filtering tool proposed in this work will focus on methods of this group.

2.4. The Levinthal paradox

The free energy landscape of large molecules like proteins is vast and complex. There are many degrees of freedom and a myriad of possible conformations it could adopt. According to Tramontano [Tra04]: “The number of possible conformational states of a protein is enormous (at least 2^{100} for a chain of 100 amino acids) and is therefore computationally intractable. In fact, this observation is also relevant for the development of a folding theory as it is obvious that a protein cannot explore such a large number of states in a reasonable time frame, as required by the hypothesis that the native structure is thermodynamically the most stable”.

To better illustrate this, for a protein to sequentially sample all its possible conformational states at a rate similar to the observed protein folding rate,

which is about 1 picosecond per state, it would take about 10^{38} seconds. To have a perspective, the age of the universe is currently estimated to be 10^{17} seconds. This observation remain true even if the protein fold at a rapid rate of nanoseconds. Contradicting the theory is that most small proteins fold spontaneously within millisecond or even microsecond. Cyrus Levinthal raised this problem first in 1968, in an attempt to explain that, rather than sampling possible conformations randomly, nature search for "folding pathways" to find the native state of a protein [Lev68].

This paradox was solved in 1992 by the folding funnel theory, which explain that the loss of entropy of the energy chain is immediately compensated by an energy gain [Leo92, Loc01, Nym98, Onu97, Soc98, Wol97, Fin97]. The theory also allows to estimate a protein folding time in agreement with experimental observations [Gal00]. Figure 6 depicts the proposed energy landscape funnel.

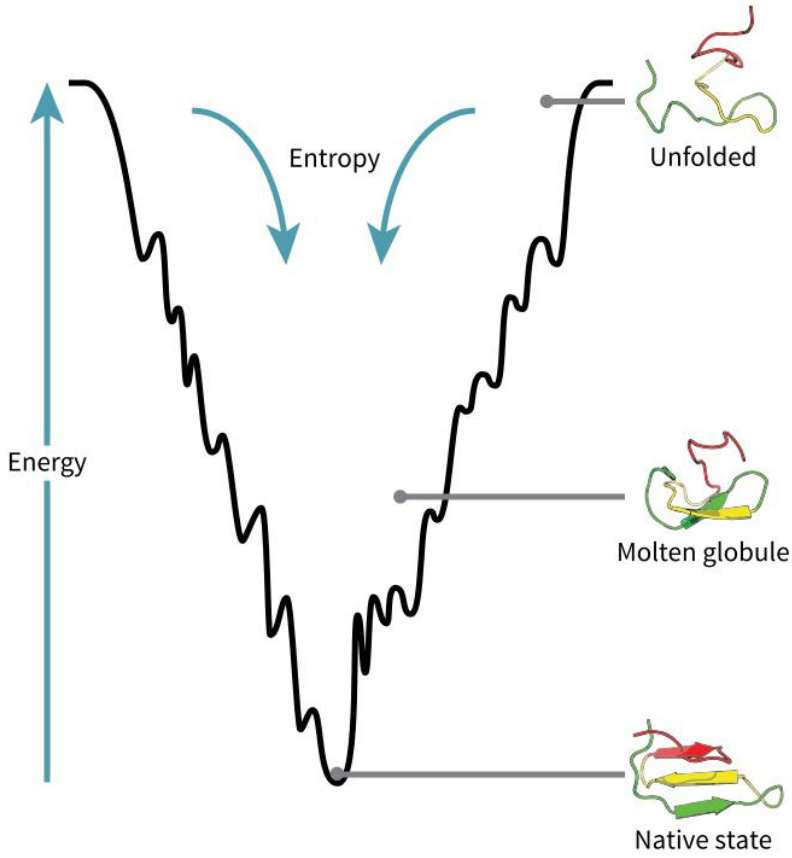

Figure 11 - Energy landscape of a protein folding. The energy landscape is "rough", with several peaks and non-native local minimums in which partially folded proteins (also called a Molten globule) can become trapped. According to the Anfinsen's thermodynamic hypothesis [Anf73], the native state of the protein correspond to the global minimum of the energy landscape. Figure obtained from [Spl17].

The funnel theory helped us better understand the pathway taken during a protein folding process. However, simulating this folding process with *ab initio* methods, even with simplified approaches, is a NP-Complete problem as proved by Pierluigi Crescenzi and colleagues [Cre09]. Many researches

around the world are still attempting to solve this problem or at least find an alternative viable solution to it. While the general objective is to predict the 3D structure from the primary sequence of a protein, our current knowledge, and computational power are simply insufficient to solve a problem of such high complexity. Besides the efforts of many different research groups, this problem, even after almost 70 years, remains unsolved and still very pertinent to the society [Hel08]. Perhaps even more now with the recent advances in pharmaceutical industry and the identification of several new genetic related diseases.

2.5. Molecular dynamics

The first Molecular Dynamic (or MD) method was published on 1959 by Alder and Wainwright [Ald59] and later by Rahman [Rah64] in 1964. In most MD simulations, the trajectories of atoms and molecules in the system are determined by solving Newton's equations of motion for a system of interacting particles using potential energy (or force field) functions. Unfortunately, as our capacity to simulate quantum mechanics is still out of our computational capabilities, these *in silico* simulations are not 100% accurate. Long MD simulations can generate errors in numerical integration that accumulate as the simulation continues. There are way to minimize this, but not eliminate them entirely.

MD simulations also require a simulation medium (i.e., a solvent). A series of different solvation models were then developed. They not only enable MD simulations, but also enable thermodynamic calculations applicable to reactions and processes which take place in solution, including biological, chemical and environmental processes [Sky15]. There are 3 different types of solvation models, each with own pros and cons:

1. **Explicit Solvation Models** provide the most descriptive and realistic models for the solvent, where its molecules are explicitly described in the simulation system with determined position, rotation, charge, etc. This impacts in a large increase to the degrees of freedom of the system and thus significantly impact the computational costs of the (already costly) simulation.

2. **Implicit Solvation Models** provide the system with a reasonable description of the solvent behavior and are generally computational efficient. They consider the solvent as a continuous isotropic medium with the underlying assumption that the solvent molecules themselves may be removed from the system if the continuous medium replacing them sufficiently represents equivalent properties. These models, however, are not as precise as the explicit solvation and fail to account for local fluctuations in the solvent density.

3. **Hybrid Solvation Models** incorporate aspects of implicit and explicit solvation models, aiming to minimize computational cost while retaining at least some of the precise resolution of the solvent. Contrary to the other

two solvation methods, these are specialized models built by extensive research and vary considerably from one to another.

Despite high the computational costs and its inexact accuracy, MD methods are still the most versatile *in silico* technique to study biological macromolecules, and thus predict a protein 3D structure, to date. Molecular Dynamics is what makes the *ab initio* methods possible. It also enables otherwise impossible experiments, such as simulating a protein fold at extreme pressures or temperatures. Beyond that, some proteins do not form crystals or dense solutions, which are necessary for experimental protein structure resolution. Additionally, experiments *in vitro* or *in vivo* can pose a hazardous risk to health if poorly performed, adding to their already high costs of time and money. The *in silico* approaches are an attractive solution to these problems.

In 1977, McCammon and colleagues performed the first Molecular Dynamic simulation involving proteins [Mcc77]. The team simulated the bovine pancreatic trypsin inhibitor protein on a vacuum environment for $8{,}8e^{-12}$ seconds. Since then, MD techniques are being enhanced and, as a consequence, the target proteins are both larger and more complex. The advances in computers and parallel architectures also greatly contributed to enable such simulations to be performed on a feasible time. They also made possible for more realistic force fields functions to be used, which require a longer processing time.

Molecular dynamics (MD) “has had a long history and has evolved into an important and widely used theoretical tool that allows researchers in chemistry, physics, and biology to model the detailed microscopic dynamical behavior of many different types of systems, including gases, liquids, solids, surfaces, and clusters.” [Tuc99]. While the majority of works still use MD methods only as a mean of refining models [Dal12, Dor13, Jag08, Kri04, Mar12, Mel12, Mir14, Par12, Lee02], it has gained more attention by the scientific community after successful works which used MD to predict the native structure of proteins such as [Dua98, Bow09, Zag02, Lei08, Lei09].

2.6. Monte Carlo

The simulation of a protein folding is a optimization problem. However, simple optimization algorithms fail to solve it due to the rough nature of a energy landscape of proteins. The local optimization methods identify in a series of steps one path to the local minimum, and do not allow the function value to increase at any stage [Zve08] (i.e., the free energy of the simulated protein cannot increase at any given point). This means that energy peaks cannot be crossed over. Therefore, when local running optimization algorithms on an energy landscape, without means to overcome energy peaks, the quality of the results depend solely on a favorable topology and a “lucky start” of the starting point, when it is located inside the global minimum energy valley. Observing Figure 12, it is possible to infer that only the P_1 starting point will ever reach the global minimum A of the energy landscape due to the energy peaks separating the starting points P_2 and P_3 from it.

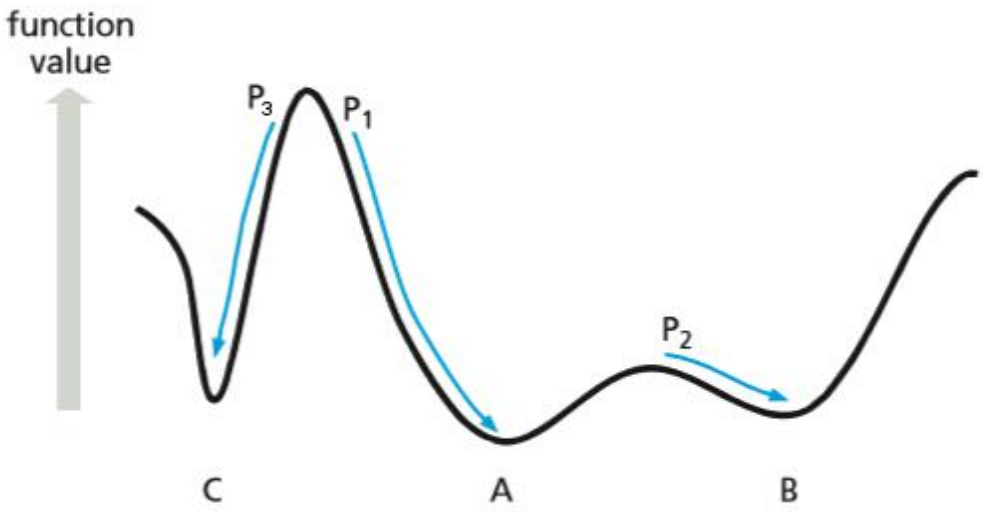

Figure 12 - Diagram illustrating a one-dimensional minima problem. The black line represent the energy landscape of a protein obtained through an hypothetical energy function. The points *A*, *B* and *C* represent local minima. P_1, P_2 and P_3 represent three different starting points on the energy landscape. Figure adapted from [Zve08].

While it is possible to locate the global minimum with such methods, this would require the method to be run several times using different random starting points. Eventually one starting point would be withing the global minimum energy valley and the function would correctly identify it. This however, is intractable for a problem of such high complexity as the protein structure prediction problem [Zve08]. As described in chapter 2.4, the number of possible conformations a protein can adopt in its energy landscape is simply overwhelming.

The Monte Carlo method address this issue by allowing movements considered detrimental in the energy landscape. This freedom of movement must be controlled however, as not to impair the ability of the method to identify the global minimum in a reasonable time frame. Otherwise it would behave exactly like running several local optimization methods hoping for one instance to identify the global minimum, or in the worst case scenario randomly walk in the energy landscape without direction. Detrimental movements are then controlled by a probability function.

A folding state of a protein can be defined by the position of all atoms in the system. It is possible to assign an energy value to this state by using a suitable force field. Therefore we can determine an energy value to all possible folding states of a proteins. When the system is in equilibrium, the relative probability of any given state occurring is given by the Boltzmann weighting $e^{-E_1/kT}$, where E_1 is the assigned energy value for that state, k is the Boltzmann's constant and T is the absolute temperature in Kelvins (K) [Zve08]. The exact probability of state 1 occurring is given by dividing this by the partition function Z (see EQ. 1), where M denotes the number of all states accessible by the system.

$$Z = \sum_{i=1}^{M} e^{-E_i / kT} \qquad \text{(EQ. 1)}$$

The probability of an state p_i occurring is then defined by EQ. 2.

$$p_i = \frac{1}{Z} e^{-E_i / kT} \qquad \text{(EQ. 2)}$$

While it is impractical to calculate the partition function Z, the Monte Carlo method sidesteps this problem by looking at the ratio of probabilities for two states. Considering two states S_1 and S_2, with assigned energy E_1 and E_2 respectively, the ratio of probabilities is given by EQ. 3.

$$\frac{e^{-E_1 / kT} / Z}{e^{-E_2 / kT} / Z} = e^{-(E_2 - E_1 / kT)} = e^{-(\Delta E_{21} / kT)} \qquad \text{(EQ. 3)}$$

Therefore, with any given state S_1 we can readily determine if an state S_2 is more likely to occur at equilibrium. If ΔE_{21} is negative (i.e., state S_2 has lower energy) the EQ. 3 will result in a value greater than one (i.e., S_2 is more likely to occur) and the movement from the state S_1 to the state S_2 is accepted. On the other hand, if the state S_2 has a higher energy than state S_1 (i.e., the move is uphill on the energy surface) the above term has a value between 0 and 1. Instead of plainly rejecting the movement based on the detrimental assigned energy value, a random number from a uniform distribution in the interval 0 to 1 [Aba94, Zve08] is used to determine if the movement is accepted or rejected. If the selected random number is less than the resulting value from EQ. 3, then the movement is accepted (even if detrimental), otherwise it is rejected.

This acceptance/rejection rule is called the Metropolis Criterion (or Metropolis–Hastings algorithm), a sampling method created by Wilfred Keith Hasting in 1970 [Has70], which is a generalization of another sampling method introduced by Nicholas Metropolis and colleagues in 1953 [Met53]. Figure 13 illustrates this algorithm.

The Metropolis–Hastings algorithm works in a way that, as more sample values are produced, the distribution of values more closely approximates the desired distribution. For the PSP problem that is a distribution of assigned energy converging to the global minimum. As the sample values are produced iteratively, the distribution of the next sample is dependent only on the current sample value. This makes possible to reconstruct the sequence of samples into a Markov chain, which becomes particularly useful when proving the correctness of the Monte Carlo method and variations.

By choosing movement via this criterion, the Monte Carlo method has the requirements, under suitable conditions, to locate the global energy minimum, which will be the state with the highest probability at equilibrium. The suitable

conditions required is that movements are of appropriate magnitude to allow efficient coverage of the available state (movements too far apart leaves gaps in the energy landscape while movements too near can result in intractable computational complexity). Another condition is that all the low energy states can be accessible from each other without crossing unduly high energy barriers [Zve08]. This latter condition can be diminished with the use of lower movement steps, but this greatly increases the computational complexity of the problem as already cited.

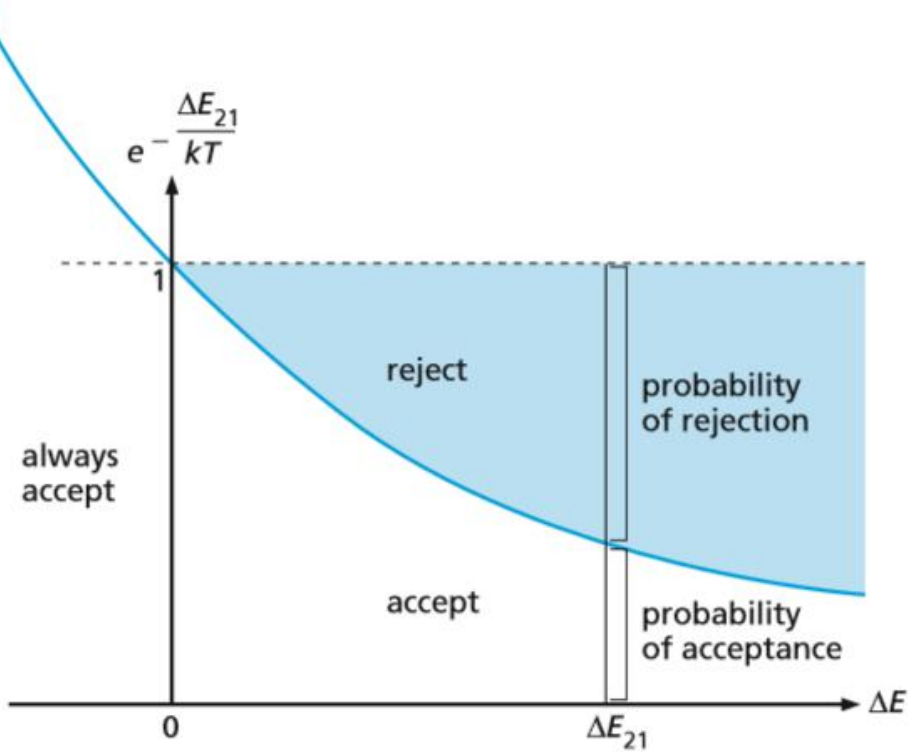

Figure 13 - Illustration of the Metropolis Criterion for the acceptance or rejection of movements in the Monte Carlo method. Figure extracted from [Zve08].

The Monte Carlo method proved to be an attractive solutions to the research community to lighten the computational costs of classical MD methods. Like conventional MD methods, the Monte Carlo method is also employed on several steps of the prediction, serving on multiple purposes such as refinement [Chi06, Ols14], predicting secondary structures [Hof14, Lin12, Lin09a], predicting side chain conformations [Nag12] or effectively attempting to predict the native structure of proteins [Lip17, Aba94, Car03, Cho04, Gib01, Har02, Jay06, Lip12, Lip14, Nar06, Ped97, Zha07].

2.7. Replica-Exchange Molecular Dynamics

The Replica-Exchange Molecular Dynamics (or REMD) method, also known as Multiple Markov Chain Method (or MMCM) or even parallel tempering, was created by Sugita and Okamoto in 1999 [Sug99] as a formulation for the MD algorithm using the replica exchange method, which

was originally proposed by Swendsen and Wang in 1986 [Swe86], extended by Geyer in 1991 [Gey91] and latter developed by Hukushima and Nemoto in 1995 [Huk95].

Since then, this method has been used in several fields of the bioinformatics area, including studies of structure-function relationship in proteins [Mic15], DNA [Mac14], RNA [Ber13, Roe14], protein stability [Hat14], folding dynamics [Eng13, Jan14, Xue15], and secondary structure prediction [Zha15].

The main objective of the REMD method is to overcome multiple-minima problem by exchanging non-interacting replicas of the system at several temperatures [Sug99]. It works by simulating the protein fold of several different copies of the target protein (also called replicas) in different temperatures in a parallel and independent way (i.e., the simulations do not influence each other). Some of these temperatures are way higher than those found in living beings, but necessary to reach wider, more energetic, protein structure conformations. A frequency rate stipulated by the user named EAF (or Exchange Attempt Frequency), such as 1ps, defines the period rate in which exchange attempts will be performed. At these events, adjacent replicas attempt to exchange temperatures with each other. This prompt replicas simulating in lower temperatures to acquire the energy necessary to break energy barriers and therefore overcome local-minimum states, and replicas simulating in higher temperatures to cool down and gradually converge into a single stable structure. The simulation continues until a certain number of steps (simulation time), or a certain degree of convergence, is reached. Figure 14 exemplifies this process.

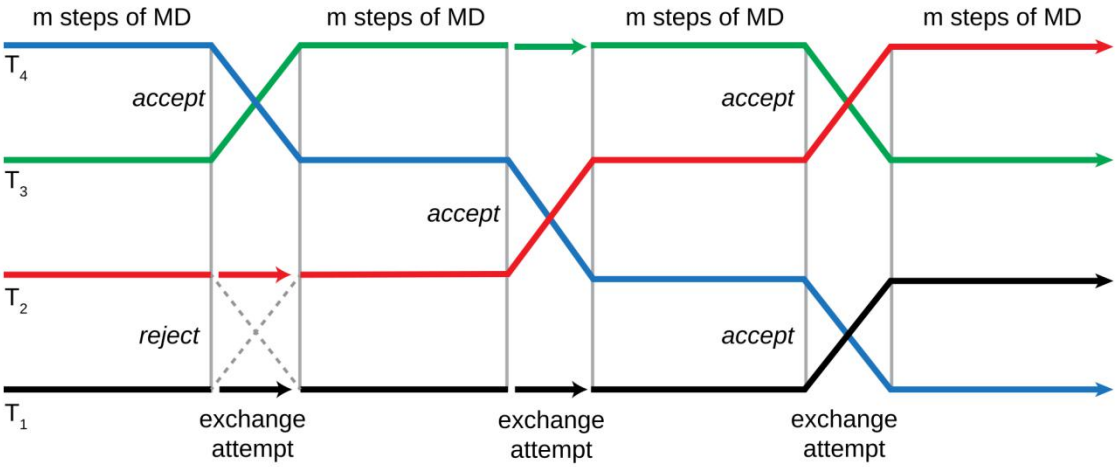

Figure 14 - Schematic of a REMD simulation. The colored lines represent the different replicas being simulated through an MD software. T_1, T_2, T_3 and T_4 represent different temperature levels. At fixed period rates, adjacent replicas attempt to exchange temperatures. Figure obtained from [Row01].

The theoretical basis for Monte Carlo simulations is the Markov chain theory and, as already cited, the sequence of samples can be reconstructed as a Markov chain. It is, therefore, desired that the configurations generated by the Markov chain sample the Boltzmann distribution, after an initial transient, “equilibration” period [Man99]. It is desired to sample the

Boltzmann distribution as it describes that states with lower energy will have a higher probability of occurring than states with higher energy [Atk09] (Figure 15).

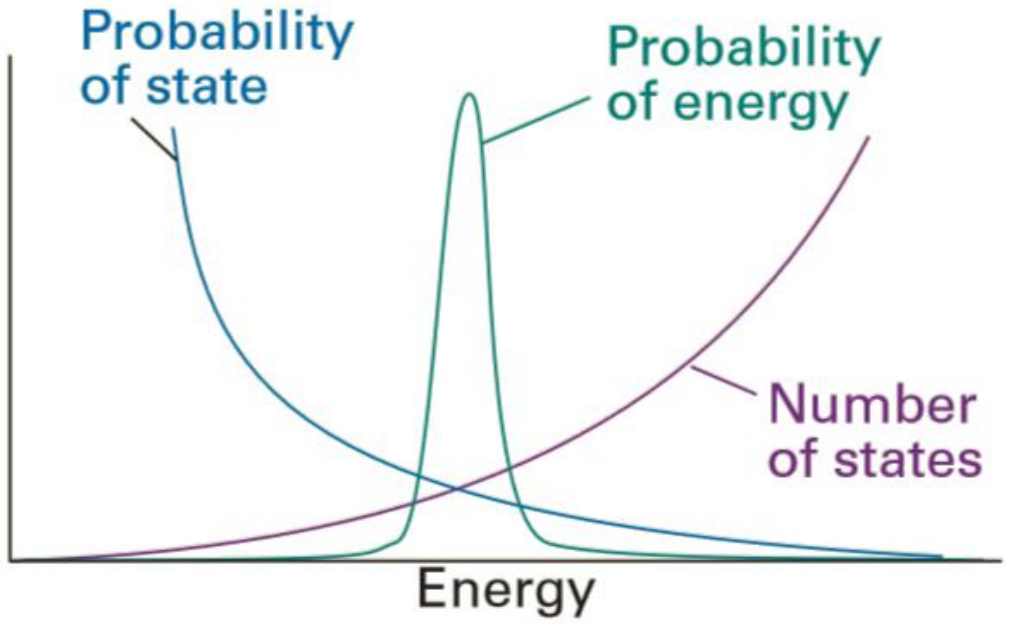

Figure 15 - Graphical representation of several aspects of the Boltzmann distribution. The x axis defines the free energy of states (i.e., adopted 3D structure). The purple line represent the number of possible states in that given energy value. The green line represent the probability of that given energy occurring, which is a Gaussian distribution. Finally, the blue line represents the probability of the state occurring in nature, where lower energy states will have a higher probability of occurring than higher energy states. Figure extracted from [Atk09].

Additionally, it is also desired that the limiting distribution of the Markov chain exists and is unique (i.e., the distribution converge into a single unique state) [Man99]. This result is assured if the Markov chain is regular and satisfies the detailed balance condition [Par88]. Manousiouthakis and Deem proved in 1999 [Man99] that a Monte Carlo simulation need only to abide by the (weaker) balance condition (or BC) to be considered correct. The BC, in turn, simply requires that a Boltzmann distribution is maintained at all times [Lip17, Man99]. This proves that most Monte Carlo methods and hybrid approaches are indeed capable of producing optimal results given enough time.

The REMD method, which uses the Metropolis Criterion, not only guarantee the balance condition, but also the detailed balance condition. Sugita and Okamoto described in their paper [Sug99] that the transition probability of an exchange attempt must be the same that the probability of accepting the inverse movement. This is show in EQ. 4, where $W_{REM}(X)$ is the product of all Boltzmann factors for each replica of the simulation for the state X, and $w(X \longrightarrow X')$ is the transition probability of state X to state X'. The states of the REMD simulation are defined as the current coordinates and

momenta for each replica being simulated. The product of all Boltzmann factors is explained in deep at the original paper [Sug99], but can be roughly translated as the weighted value of the probability of the state X occurring.

$$W_{REM}(X)w(X \longrightarrow X') = W_{REM}(X')w(X' \longrightarrow X) \qquad \text{(EQ. 4)}$$

The transition probability of state X to state X' is given by the Metropolis criterion in EQ. 5, where $x_m^{[i]}$ labels the replica i with temperature m.

$$w(X \longrightarrow X') \equiv w(x_m^{[i]} \mid x_{m+1}^{[j]}) = \begin{cases} 1, & \text{for } \Delta \leq 0, \\ \mathrm{e}^{-\Delta}, & \text{for } \Delta \rhd 0. \end{cases} \qquad \text{(EQ. 5)}$$

The delta of EQ. 5 is given by EQ. 6, where k is the Boltzmann constant, T is the temperature level m (in Kelvin), E is the energy assigned to the structure i or j with its respective atom coordinates q.

$$\Delta = \left[\frac{1}{kT_{m+1}} - \frac{1}{kT_m} \right] \left(E(q^{[i]}) - E(q^{[j]}) \right) \qquad \text{(EQ. 6)}$$

If the exchange attempt is successful, then the velocities of all the atoms in the replicas are rescaled uniformly by the square root of the ratio of the two temperatures described in EQ. 7, where p is momenta of atoms of their respective structures.

$$\begin{cases} p^{[i]'} \equiv \sqrt{\frac{T_n}{T_m}} p^{[i]}, \\ p^{[j]'} \equiv \sqrt{\frac{T_m}{T_n}} p^{[j]}. \end{cases} \qquad \text{(EQ. 7)}$$

The REMD method defined that only replicas on adjacent temperatures attempt to exchange temperatures since the fist term of the delta formula (defined by the bracket in EQ. 6) decrease exponentially the acceptance ratio of the exchange.

A variety of thermostats methods is also employed to control the energy present in the replicas, in order to maintain a balance of total energy. Popular thermostats used in REMD PSP simulations are the Anderson thermostat [And80], the Berendsen thermostat [Ber84], the Nosé-Hoover thermostat [Nos84, Hoo85], and the Langevin thermostat [Lon92].

The studies performed by Rosta, Buchete and Hummer [Ros09] have concluded that the Berednsen thermostat produce a balance shift regarding the folding states of predictions according to the temperature of the system, pointing that the folded state is overpopulated by about 10% at low

temperatures, and underpopulated at high temperatures. While this has an undesired impact on certain aspects of the results such as the enthalpy of folding [Ros09], this may contribute to better filtering the resulting ensemble. This will be discussed latter on this work.

As the REMD method simulate the protein fold of the target protein on several different temperature levels in an independent way, and also adding the temperatures exchange process, relying on the convergence of the results is virtually impossible. This is solved by taking “snapshots” of the simulating structures as they fold and unfold. The rate in which these snapshots are taken can be defined by the user, but the standard practice is taking a new snapshot every picosecond.

Although the REMD method was proposed about two decades ago, it is still one of the most used methods to simulate protein folding for its capability of breaking free from local-minimum states and for generating a wider array of structures compared to conventional simulation methods, making it a valuable tool for the scientific community. Examples of researches which used the REMD as the main form of predicting the 3D structure of a target protein include [Urb08, Ho06, Fuk02, You03]. Beyond predicting the structure of proteins, the REMD method is also employed on other areas such as crystallographic structure refinement, geometric parameters optimization and evaluation of the ligand-receptor interaction [Lip17].

The main drawback for this method, however, remains its high computational costs. Both for simulating the protein fold as well as analyzing its posterior results. While conventional MD methods can converge the fold simulation process and result in a single prediction, the REMD method do not share the same advantage. The solution of taking snapshots of the predictions result in a large ensemble of predictions with size equal to EQ. 8, where T_{Total} represents the total time of the simulation (in picoseconds), $P_{Snapshots}$ represents period time in which snapshots are taken (also in picoseconds) and $N_{replicas}$ represents the number of replicas used in the simulation.

$$S = T_{Total} \cdot P_{Snapshots} \cdot N_{replicas} \qquad \text{(EQ. 8)}$$

As an example, using only 10 replicas on a 50ns simulation (a short simulation time given some larger proteins are found to fold in a timeframe of seconds) and with a snapshot period of 1ps, the resulting ensemble of predictions would contain around half a million entries. Arguably impossible to be analyzed manually on a reasonable timeframe.

2.8. AMBER14

The AMBER [Cas05, Pea95, Amb17] is an example of molecular dynamics simulation software package that can run and analyze MD simulations for proteins, nucleic acids and carbohydrates [Lip17]. The package is composed

of two parts: (i) a set of molecular mechanical force fields for simulations of biomolecules and (ii) a package of simulations programs.

Generally, an AMBER simulation is composed of tree steps: (i) configuration of the system; (ii) simulation; and (iii) trajectory analysis. The AMBER also supports the MD simulation with explicit or implicit solvent [Nym08], the latter usually having a significant lower computational costs [Lip17]. The implementation of the implicit solvent models were developed using the Poisson-Boltzmann equation and the Generalized Born approximation model as basis [Onu02, Sti90], while the explicit solvent models are treated by the Particle-Mesh Ewald (PME) method [Dar98]. This work used the version 14.0 of the AMBER molecular dynamics package [Cas14].

2.9. CASP: Critical Assessment of Structure Prediction

Every 2 years since 1994, the international community of proteins structure prediction researchers assemble for the Critical Assessment of Structure Prediction (or CASP) [Uni17]. In this conference, several different methods of structure predictions are tested in a blind manner.

At the time that predictions are made, neither predictors or the organizers and assessors know the structures of the target proteins. The targets for predictions are either structures soon-to-be solved by X-ray crystallography or NMR spectroscopy, or structures that have just been solved (mainly by one of the structural genomics centers) and are kept on hold by the Protein Data Bank. The only information predictions have is the sequence of amino acids (or primary structure) of the target protein. Since its start, much has changed on the CASP proceeding methodology, be it the evaluating methods or in the participation categories. The latest instance of the CASP meeting happened in 2016, labeled CASP12, where the assessment categories were divided as follow:

1. **High Accuracy Modeling:** include domains where majority of submitted models are of sufficient accuracy for detailed analysis. Established numerical methods are used to evaluate main chain, side chains, atomic accuracy, and contacts, as well as hydrogen bonds and covalent geometry;

2. **Biological Relevance:** assess models on the basis of how well they provide answers to biological questions. Target providers are asked to say what questions prompted the determination of the experimental structure, and the ability of models to provide answers to those questions are compared with the extent to which the experimental structure can do so in addition to assessing aspects of accuracy that include sequence alignment, backbone accuracy, and side chain placement;

3. **Topology:** assess domains where all submitted models are of relatively low accuracy using the established CASP metrics together with assessor

judgment;

4. **Data Assisted:** assess how much the accuracy of models is improved by the addition of sparse data. Targets for which such data are available are re-released after initial data independent models have been collected, together with the available data. Data types need to include simulated and actual sparse NMR data, crosslinking data, and low angle X-ray scattering data;

5. **Contact Prediction:** assess the ability of methods to predict three dimensional contacts in targets structures;

6. **Refinement:** analyze success in refining models beyond the accuracy obtained in the initial submissions. Selected targets from among those released in the main modeling experiment are included. Assessors select one of the best models received during the prediction season, and reissue it as a starting structure for refinement;

7. **Assembly:** assess how well current methods can determine domain-domain, subunit-subunit, and protein-protein interactions;

8. **Accuracy Estimation:** assess the ability to provide useful accuracy estimates for models at the overall, residue, and atomic levels.

Some of these categories are also divided between (a) human and (b) server, as to minimize the impact of high end laboratories to less fortunate research groups. Most of the quality metrics used to assess the quality of predictions, which will be presented in Chapter 2.11, were extracted from this important conference.

2.10. Ramachandran Plot

The Sasisekharan-Ramakrishnan-Ramachandran map (also known only as Ramachandran map, diagram or plot), originally developed by Ramachandran, Ramakrishnan, and Sasisekharan in 1963 [Ram63], is a fundamental tool in the analysis of protein structures. It is a way to visualize energetically allowed regions for backbone dihedral angles ψ against φ of amino acid residues in protein structure based on the fundamental law that two atoms can't occupy the same space at the same time. This limit the conformational angles that peptide bonds can adopt without atoms colliding with each other into predictable bands.

There are four basic types of Ramachandran plots, depending on the stereo-chemistry of the amino acid: generic (which refers to the 18 non-glycine and non-proline amino acids), glycine, proline, and pre-proline (which refers to residues preceding a proline [Mac91]) [Ho05]. The ω angle at the peptide bond is normally 180°, since the partial-double-bond character keeps the peptide planar [Pau51]. For the ψ and φ angles, they can be observed in Figure 16, depicting a Ramachandran plot for the general case.

Because dihedral angle values are circular (i.e., 0° is the same as 360°), the edges of the Ramachandran plot "wrap" around. Therefore, the small strip of allowed values along the lower-left edge of the plot are a continuation of the large, extended-chain region at upper left.

The Ramachandran plot is used as a way to assess protein structure predictions and also as a component to absolute quality metrics which will be described in chapter 2.11.2.

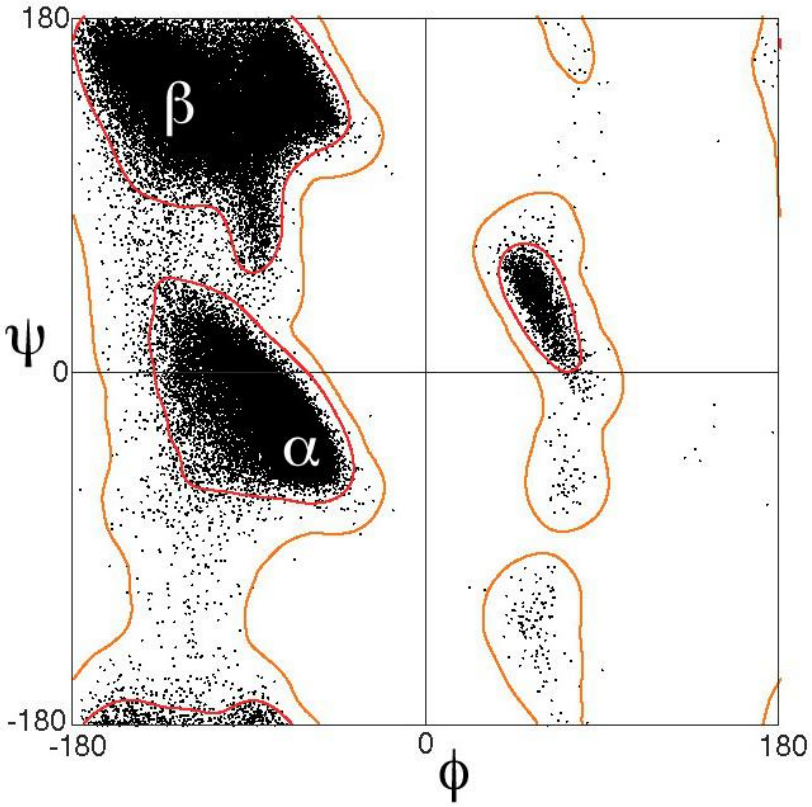

Figure 16 - A Ramachandran plot of the torsion angles that specify protein backbone conformation, with the regions of probable alpha-helix and beta-strand labeled as *α* and *β* respectively. The red lines denote favored regions. Brown lines denote allowed regions. White regions are prohibited angles conformations. Plot created based on the data-set of Lovell et al. 2003 [Lov03]. Figure obtained from [Dcr17].

2. 11. Quality metrics

Quality metrics are the main form of evaluating a predicted protein structure quality. They are divided into (a) Relative Quality Metrics and (b) Absolute Quality Metrics. Depending on the protein structure, its parameters and the availability of a model structure some metrics are best advised.

2. 11. 1. Relative quality metrics

Relative quality metrics refer to the quality metrics that use a model structure of the target protein, often determined experimentally by NMR

spectroscopy or X-ray crystallography, to calculate the quality of the given prediction. These metrics are precise and result in a reliable assessment, sometimes being used to test and compare novel absolute metrics. Although there are several different relative quality metrics in existence, only 3 were initially chosen to be analyzed in this work due to the limited time frame available. Table 1 presents an abridged version of their functionality.

Table 1 - Functionality Abridgment of Relative Quality Metrics

Relative Quality Metric	Functionality Abridgment
RMSD	Measures the distance between same atoms of two structures of the same protein.
GDT_TS	Calculates the largest set of amino acid residues' alpha carbon atoms in the model structure that falls within a defined distance cutoff of their position in the experimental structure.
QCS	Attempt to mimic visual inspection by human expert. Captures both global and local structural features, with emphasis on global topology.

2.11.1.1. RMSD

The Root-Mean-Square Deviation (or simply RMSD) is a measure originated from the statistic area and was latter adapted to bioinformatics. It is used to compare two structures of the same protein by measuring the distance between the same atoms of both structures.

The RMSD is the most commonly used relative quality metric. Possibly because the result is a measure of length units in Ångströms (The Ångström, Å, named after the physicist Anders J. Ångström, is equal to 0.1 nm. Although not an SI unit, it is used universally by structural biologists to describe atomic distances—it is approximately the length of a typical COH bond). Therefore it is easy to interpret its results, where a larger RMSD value means the predicted structure differs more from the model structure in terms of distance. On the other hand, the lower the output value is, the closer the atoms are to each other. Ultimately, the best possible output is 0, when both structures are effectively the same when superimposed.

Equation EQ. 9 shows how the RMSD is calculated, were δ is the distance between the atom i from the predicted structure and the same atom in the model structure or the mean position of the N equivalent atoms.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \qquad \text{(EQ. 9)}$$

Often the RMSD is is calculated using only the backbone heavy atoms (*C*, *N* and *O* atoms), or even using solely the C_α atoms. It is also common that, during the RMSD calculation, translations and rotations are performed on one

of the structures with the intent of obtained the best superposition, which minimizes the resulting RMSD value [Lip17]. Given two sets of n atoms v and w, the RMSD is defined by EQ. 10.

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2} \quad \text{(EQ. 10)}$$

EQ. 10 can be further expanded into a more visible equation showed in EQ. 11, where x, y and z denotes the position of the atom i in the Cartesian coordinate system.

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2\right)} \quad \text{(EQ. 11)}$$

2.11.1.2. GDT_TS

The Global Distance Test (or GDT_TS to represent “total score”, or simply GDT) is used to measure the similarity between two structures of the same protein, resulting a value between 0 and 1 (where 0 represents 0% similarity and 1 represents 100% similarity). This metric was developed with the intent of being more accurate than the common RSMD. As RSMD only measures the rough mean distance between atoms, a completely different structure from the native structure may have the same RSMD score than a very similar structure, but with very distant end points from the native structure. The GDT metric is more sensible to factors like these, recognizing outlier regions of individual loop regions in a structure that are otherwise significant accurate [Zem99, Zem03]. The GDT_TS is calculated using the formula show in EQ. 12, where GDT_{Pn} is an estimation of the percent of residues that can fit under distance cutoff <= n.0 Ångströms (1.0 Ångström for $P1$, 2.0 Ångströms for $P2$, and so on) [Pre17a].

$$GDT\_TS = \frac{\left(GDT_{P1} + GDT_{P2} + GDT_{P4} + GDT_{P8}\right)}{4.0} \quad \text{(EQ. 12)}$$

2.11.1.3. QCS

The Quality Control Score (or simply QCS) is a method developed to contribute to CASP in terms of automatic evaluation of predicted structures. This score is deemed particularly useful to compare poor predictions. According to its authors, the QCS metric reflects their manual evaluations experience and aims to capture global features of models defined by the mutual arrangement of secondary structure elements (or SEEs).

The QCS is calculated using a weighted sum of six different scores (EQ. 13), where w_n is a weight multiplier, S_P is a score attributed to the position of SEEs, S_L is a score attributed to the length of SEEs, S_H is a score attributed to the handedness of SSE triplets (i.e., handedness defines the position of a third SSE when two SSEs are being considered), S_A is a score attributed to the angle between SSE vectors, S_I is a score attributed to the interaction between SSE vectors and finally SC is a score attributed to the contact between all C_α atoms. More information about individual scores can be found in the original paper [Con11].

$$QCS = \frac{100}{\sum_{i=1}^{6} w_i}\left(w_1 S_P + w_2 S_L + w_3 S_H + w_4 S_A + w_5 S_I + w_6 S_C\right) \qquad \text{(EQ. 13)}$$

Overall, QCS is in agreement with manual inspection and correlates well with GDT_TS. However, QCS can reveal models with a better global topology that are missed by GDT_TS. This metric is not only suitable to select candidates for manual inspections in the CASP assessment, but also can be useful as an independent and objective method to assess the quality of structure prediction with emphasis on the global topology [Con11].

Although the QCS metric is used with frequency on the CASP conference, it was deemed to be have an unnecessary degree of precision for testing the SnapFi tool, where the simple RMSD and GDT_TS scores are both readily available in the AMBER package and provide an output value more easily interpreted by the vastness of the research community. While this metric was presented among the others in the Plan of Study and Research of this work in 2016, it was ultimately discarded.

2.11.2. Absolute quality metrics

Absolute quality metrics refer to quality metrics that do not rely on model structures. These metrics are especially important to evaluate the predictions of *ab initio* methods, where an homologous (or even remotely homologous) protein is not provided. Due to the lack of a 3D structure for comparison, measuring a predicted protein structure's quality becomes a complex task.

To perform such task, a standardized ranking method is then made necessary. Many different absolute metrics for ranking predictions were developed by several authors, each using different approaches and mathematical formulas. These metrics are capable of distinguishing and classifying different predicted structures through the assessment of different properties of the structure such as the torsion angles of the amino acid residues, a measurement of how "common" a stereochemical conformation is compared to already known structures, comparisons with the native states of other non-related proteins, etc. Because of that, different absolute quality metrics tend to perform a better assessment of certain proteins types and structures with certain conformations than others.

Most absolute quality metrics, however, are based on potential energy formulas. Most of them are measured in kcal/mol while others just denote energy values without a proper unit. By using the Anfinsen's thermodynamic hypothesis [Anf73], that means that the lowest the value resulted from these metrics, the better the predicted structure is.

In general, there are two types of potential energy functions for protein structure prediction: physics-based and knowledge-based. Physics-based potential functions are developed from *ab initio* quantum chemical calculations, whereas knowledge-based potential functions are developed from statistical analysis of known protein structures. The knowledge-based potentials may be further divided into two categories: all-atom potentials and coarse-grained (semi)-residue-based potentials [Lu08]. In many applications, knowledge-based potential functions outperform the physics-based potentials [Bra05, Sko06].

It is worth mentioning that absolute quality metrics are unable to classify a predicted structure as the native state of the protein if one is made, contrary to relative quality metrics. Although not having an optimal accuracy, they are a viable alternatives for when the most accurate methods are unavailable due to the lack of funds and/or time (since NMR and X-ray crystallography are both expensive and time-consuming options, often demanding several weeks or even months).

For this work, 8 different absolute metrics were chosen according to their popular use and quality when evaluating predicted protein structures, and also for their overall availability and facility of installation and use. Additionally, the energy minimization of the predicted structures was also proposed to be used as a way to assess the predicted structures during the course of this study. The amount of metrics chosen was defined to fit in the time frame available for this work. Table 2 presents an abridged version of their functionality. Except for the GFactor and Probscore absolute quality metrics, all scores are based on potential energy functions, which means that lower score values are better than higher score values.

Table 2 - Functionality Abridgment of Absolute Quality Metrics

Absolute Quality Metric	Functionality Abridgment
DFIRE	An all-atom potential energy function based on a distance-scaled, finite ideal gas reference state.
dDFIRE	A "dipolar" DFIRE potential energy function based on the orientation of angles involved in the dipole-dipole interactions.
DOPE	A potential energy function grounded entirely in the probability theory, using probability density functions (pdf).
GFactor	A logs-odds score based on how "normal" a given stereochemical property is.
GOAP	A generalization of previous approaches of orientation-dependant energy potentials that consider only representatives atoms or blocks of

	side-chains and polar atoms.
OPUS-PSP	An orientation-dependent statistical all-atom energy potential derived from side chain packing.
Probscore	A quality metric created based on three different scores generated by the MolProbity service, providing a single number that represents the central MolProbity protein quality statistics.
RWplus	A distance-dependent and orientation-dependent potential energy function that uses a random-walk ideal chain as the reference state.
Minimized Energy	The minimized energy of the structure calculated using a potential energy function with the sander module, both provided by AMBER 14.

2.11.2.1. DFIRE

The DFIRE is a distance-scaled, finite ideal-gas reference (DFIRE) state proposed by Zhou and Zhou in 2002 [Zho02]. The ideal gas state used as a basis to construct the DFIRE state is a theoretical gas state composed by several randomly moving point particles whose only interaction is a perfectly elastic collision. This theoretical state is useful due to it obeying the ideal gas law first proposed by Émile Clapeyron in 1834 and presented in EQ. 14, where P is the pressure of the gas, V is the volume of the gas, n is the amount (in moles) of the substance, R is the ideal gas constant (equal to the product of the Boltzmann constant and the Avogadro constant) and T is the absolute temperature of the gas (in Kelvin). This simplified equation of state is useful for facilitating simple analysis such as statistical mechanics, which in turn enable the development of knowledge based energy potential formulas.

$$PV = nRT \qquad \text{(EQ. 14)}$$

The DFIRE reference state is used to construct a residue-specific all-atom potential of mean force from a database of 1011 nonhomologous (less than 30% homology) protein structures with resolution less than 2 Å. This all-atom potential based on the proposed state is used as an absolute quality metric (referred only as DFIRE in this work), which, according to the authors, is able to recognize more native proteins than previously developed, residue-specific, all-atom knowledge-based potentials.

2.11.2.2. dDFIRE

This quality metric is a “dipolar” DFIRE (dDFIRE) energy function based on the orientation of angles involved in the dipole-dipole interactions proposed by Yang and Zhou in 2008 [Yan08]. Each polar atom is treated as a dipole and the orientation of the dipole is defined by the bond vectors that connect the polar atom with other heavy atoms.

The dDFIRE energy function is then extracted from protein structures based on the distance between two atoms and the three angles involved in dipole–dipole interactions. According to its authors "[...] it provides a consistent treatment for the possible orientation-dependent interaction between polar and nonpolar atom and between polar atoms and non-hydrogen bonded. Moreover, an integrated treatment of distance and angle dependence produces a parameter-free statistical energy function" [Yan08].

2.11.2.3. DOPE

The Discrete Optimized Protein Energy (DOPE) is an atomic distance dependent statistical potential created from a sample of native structures that does not depend on any adjustable parameters. It was proposed by Shen and Sali in 2006 [She06]. DOPE is based on an improved reference state that corresponds to non-interacting atoms in a homogeneous sphere with the radius dependent on a sample native structure; thus accounting for the finite and spherical shape of the native structures.

The DOPE potential was extracted from a nonredundant set of 1472 crystallographic structures. It is grounded entirely in the probability theory, using probability density functions (pdf). A series of these functions are created and then paired to form joint pdfs. These joint pdfs are used in the final DOPE formula that comprehends several mathematical steps, which are explained in deep in the original paper [She06].

2.11.2.4. GFactor

The G-factor (referred as GFactor in this work) absolute quality metric provides a measure of how "normal", or alternatively how "unusual", a given stereochemical property is. It was published by Laskowski and colleagues in 1996 [Las96]. The properties for which GFactor is computed are the combinations for each residue angles ϕ-ψ, the combination for each residue $\chi 1$ - $\chi 2$, and finally the residues' $\chi 1$ values. It is essentially a logs-odds score based on the observed distribution of these given properties in high-resolution X-ray crystal structures.

A low G-factor score represents that the property corresponds to a low-probability conformation while a high G-factor score represents a high probability conformation. More precisely, values below −0.5 represent unusual property where as values below −1.0 represent high unusualness [Gan12]. Positive values, on the other hand, denote increasingly "regular" conformations.

2.11.2.5. GOAP

The generalized orientation-dependent all-atom potential (GOAP) absolute quality metric is a potential energy function proposed by Zhou and

Skolnick in 2011 [Zho11]. It depends on the relative orientation of the planes associated with each heavy atom in interacting pairs.

This metric is a generalization of previous approaches of orientation-dependant energy potentials (such as DFIRE, DOPE, etc) that consider only representatives atoms or blocks of side-chains and polar atoms. It can be decomposed into a distance-dependent part, which is treated identically as in DFIRE, and an angle-dependent part (denoted GOAP AnGular, or GOAP_AG). More details about the GOAP_AG potential can be found in the original paper [Zho11].

According to the authors “GOAP naturally integrates orientation dependent polar atoms interactions, hydrogen-bonding, and side-chain interactions”.

2.11.2.6. OPUS-PSP

The OPUS-PSP absolute quality metric is an orientation-dependent statistical all-atom Potential derived from Side chain Packing (hence the PSP) proposed by Lu, Dousis and Ma in 2008 [Lu08]. The side-chain packing is an important determinants of protein structure, as sequence identities of all polypeptide chains are solely designated by side-chains.

The basis of the OPUS-PSP hinges solely on side-chain packing interactions described by a unique basis set of rigid-body building blocks. This basis set is formed by decomposing the chemical structures of 20 amino acid residues into 19 block types. An energy potential function is then generated from the orientation-specific packing statistics of pairs of those blocks in a non-redundant structural database.

The OPUS-PSP absolute quality metric was designed to bridge the gap between all-atom and residue-based potentials and overcome a series of drawbacks of both methods. Overall, this quality metric is a generally applicable potential for protein structure modeling, specially for handling side-chain conformations, one of the most difficult steps in high-accuracy protein structure prediction and refinement.

2.11.2.7. Probscore

MolProbity is a general purpose web service created by several authors [Che10, Dav07]. It is available at no costs at [Mol17]. One of its functions is to provide broad-spectrum solidly based evaluation of structure quality at both the global and local level for both proteins and nucleic acids. It relies heavily on the power and sensitivity provided by optimized hydrogen placement and all-atom contact analysis, complemented by updated versions of covalent-geometry and torsion-angle criteria.

The aggregated MolProbity score (or MolProbity score), referred as Probscore in this work, is an absolute quality metric created based on three different scores generated by the MolProbity service, providing a single number that represents the central MolProbity protein quality statistics. It is a log-weighted combination of the clashscore, percentage Ramachandran not

favored and percentage bad side-chain rotamers, giving one number that reflects the crystallographic resolution at which those values would be expected. Therefore, a structure with a numerically lower MolProbity score than its actual crystallographic resolution is, quality-wise, better than the average structure at that resolution. The formula to calculate the MolProbity score along a brief description of its composing terms can be found at [Pre17b].

2.11.2.8. RWplus

The RW potential is a distance-dependent atomic potential proposed by Zhang and Zhang in 2010 [Zha10] that uses a random-walk ideal chain as the reference state. An ideal chain (or freely-jointed chain) is the simplest model to describe polymers, where it can be considered as the segments of an ideal polymer and is defined by a random walk (i.e., random movement) in three dimension space in which any kind of interactions among monomers is neglected. The orientation-dependent all-atom potential energy function can also capture the feature of side-chain packing such as the OPUS-PSP absolute quality metric.

The RWplus is a hybrid energy potential composed of a distance dependent energy term from the original RW and an implemented orientation dependent term. 20 vector pairs were defined to describe the side-chain orientation of 20 amino acids. The orientation term was then generated from the orientation specific packing statistics of those vector pairs in a nonredundant high-resolution structural database and used to built the RWplus energy potential function.

Because the ideal chain has no amino acid-specific interactions between the subunits but keeps the sequence continuity, it mimics the generic entropic elasticity and connectivity of polymer protein molecules, which could not be described by other reference states such as ideal gas systems used in DFIRE and DOPE absolute quality metrics. As a result, the RW potential has a steeper energy at short distances than these analytical energy potentials functions, which helps the RW potential to capture strong signals at short-range interactions. The hybrid potential RWplus was found to indeed improve the ability of regular RW in recognizing the native-like structural features.

2.11.2.9. Minimized Energy

The energy minimization of the structures was also proposed as a way to assess the predictions of the simulations. In theory, the structure prediction with lowest minimized energy would more closely match the native structure of the protein considering the Anfinsen's thermodynamic hypothesis [Anf73], which states that the native state of the protein correspond to the global minimum of the energy landscape (Figure 11).

The energy minimization is a form of geometry optimization in contrast to molecular dynamics simulations. While the latter simulates the motion of

molecules regarding several aspects like time, temperature, chemical forces, velocities, the Newton's laws of motion, among several others physical properties, thus producing a trajectory of the folding pathway, geometry optimization, on the other hand, aims to reach the global minimum disregarding basic physical boundaries. While both could ultimately achieve the same results, they arrive at it via different approaches.

The tleap module of AMBER 14 was used to generate the required input files for the energy minimization and the sander module, also provide in the AMBER, was used to calculate the minimized energy. The standard configuration of the sander module use the gradient descent optimization algorithm (Figure 17) to calculate energy minimization. This algorithm finds the local minimum by utilizing an iterative method where each step is taken at the direction of the negative function's gradient (or at the approximate gradient), which correspond to the direction closest to the minimum.

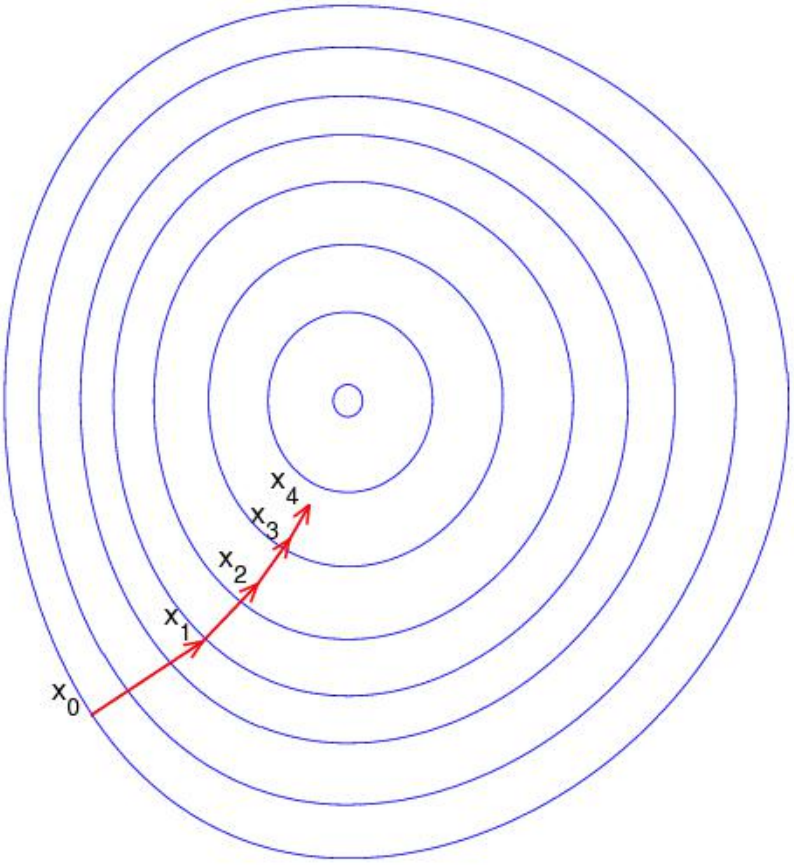

Figure 17 - Illustration of gradient descent on a series of level sets, represented by the blue lines. The central point corresponds to the global minimum of the function. The red arrows represent the 4 iterative steps taken by the algorithm. Figure obtained from [Ale17].

The force field used to calculate the potential energy was the ff12SB force field provided by AMBER 14. Most of the parameters needed for this process were defined as the default values used in several tutorials of energy minimization provided by AMBER and were the following:

1. The **maximum cycles of minimization (maxcyc)** was defined as 1000 cycles. Although discovered that this value could be significantly lowered after consultation with orienting professor Dr. Osmar Norberto de Souza, the parameter was left intentionally high as a mean to evaluate the performance difference between more or less energy minimization cycles. Due to the results obtained, which will be discussed in chapter 6.6.2, running the energy minimization with less steps became unnecessary.

2. The **non-bonded cutoff (cut)**, that defines the radius (in Ångströms) in which atomic interactions are computed, was defined as 9999 (effectively infinite cutoff).

3. The **maximum radius for generalized born (rgbmax)** defines the effective maximum distance for considering pairs of atoms to contribute to the calculation of the effective Born radii. In short, it defines the threshold in which the solvation interactions will affect the atoms of the system. The default value of 9999 (effectively infinite cutoff) was used.

4. The **implicit generalized born (igb)** parameter defines which type (if any) of generalized Born solvation model will be used. The value of 1 was used, which correspond to the pairwise generalized Born solvation model proposed by Hawkins, Cramer and Truhlar in 1995 [Haw95, Haw96].

5. The **number to print progress (ntpr)** value defines the number of steps that must occur before the minimization progress is printed on the output file. This value was set to 100 cycles.

3. MOTIVATION AND OBJECTIVES

This chapter addresses the motivation for performing this work along with the broad and specific objectives of the research.

3.1. Motivation

The PSP problem emerged in the 1960 decade [Anf73] and, until today, no definitive solution has been found. Given the biological importance of proteins and its NP-Complete complexity, the PSP problem is one of the big challenges of modern science [Cre09, Dil12]. In a review published in 2012, Dill and MacCallum [Dil12] emphasized the advances attained by the scientific community across the globe regarding this area and the lingering importance of newer and more accurate methods to predict the structure of proteins based on their amino acid sequences.

Achieving such thing in a fast and cheap manner would enable research groups to unveil still obscure processes of life, such as aging and memory creation, as well as helping treat enduring diseases such as cancer or drug resistant bacteria and viruses. Not to mention possible applications of great impact for industries such as the biopharmaceutical and synthetic materials industry. The increase of participants in the CASP conference every two years is a strong indicator of the increased number of interested researches in solving this problem.

While there are ways to experimentally determine the structure of proteins, such as Nuclear Magnetic Resonance (NMR) and X-ray crystallography, these methods are time-consuming, require specialized personal and are also very costly. Beyond that, experiments in vitro or in vivo can pose a hazardous risk to health if poorly performed. Faster and cheaper ways to achieve the same, or at least similar results, are necessary. The in silico experimentation is an attractive solution to these problems. The main problem with the *in silico* approaches currently available is that, beyond not being 100% precise, they also either fail to predict novel conformational folds, fail to search the energy landscape of possible structure conformations in a acceptable time frame or yield a massive amount of structure predictions which must be analyzed posteriorly.

The REMD method is one of the best DM methods for simulating a protein fold. Being able to predict novel protein folds as well as provenly being able to locate the global energy minimum given enough simulation time. Its main drawback fits the last category of the drawbacks of *in silico* methods: it yields a massive amount of structure predictions which must be analyzed posteriorly. If the time to analyze these structures can be shortened, it could reduce both the computational and personnel costs for future projects significantly.

The main motivator for this work is, therefore, finding a way to alleviate the analytical workload produced by such methods, preferentially without significantly impacting its resulting quality.

3.2. Broad Objectives

The broad objectives of this proposal are the creation of a tool that filters out unsatisfactory protein structures predictions on REMD PSP simulations, targeting to reduce the overall volume of data that need to be analyzed for future studies based on this method.

3.3. Specific Objectives

The specific objectives of this proposal are the following:

1. Devise a method based on the cited quality metrics to limit the amount of data that needs to be analyzed on REMD PSP simulations;

2. Devise a novel metric based on these results capable of extracting the best structure, or an ensemble of best structures, from REMD PSP simulations;

3. Test the devised methods on REMD PSP simulations;

4. Analyze the results compared to the existing literature.

4. RELATED WORKS

There are many described works that aim to improve the time efficiency of REMD PSP simulations. A pilot search were performed on the PubMed database [Pub17], ACM Digital Library [Ass17], Wiley Online Library [Joh17], IEEE database [IEE17] and the Google Academics search engine [Goo17] using only the search string "REMD". Results were filtered manually by the author according to their relevance to this study. A total of 20 works were gathered in this step and can be found in Table 3.

Table 3 - Pilot Search Results

Title	Year	Reference
Accelerating the replica exchange method through an efficient all-pairs exchange	2007	[Bre07]
An improved replica-exchange sampling method: Temperature intervals with global energy reassignment	2007	[Li07a]
Optimization of replica exchange molecular dynamics by fast mimicking	2007	[Hri07]
Hamiltonian replica exchange molecular dynamics using soft-core interactions	2008	[Hri08]
Optimized Explicit-Solvent Replica Exchange Molecular Dynamics from Scratch	2008	[Nad08]
Asynchronous Replica Exchange for Molecular Simulations	2008	[Gal08]
TIGER2: An improved algorithm for temperature intervals with global exchange of replicas	2009	[Li09]
Replica exchange simulation method using temperature and solvent viscosity	2010	[Ngu10]
How hot? Systematic convergence of the replica exchange method using multiple reservoirs.	2010	[Rus10]
Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2)	2011	[Wan11]
Replica Exchange Statistical Temperature Molecular Dynamics Algorithm	2012	[Kim12]
Superposition-Enhanced Estimation of Optimal Temperature Spacings for Parallel Tempering Simulations	2014	[Bal14]
Increasing the sampling efficiency of protein conformational transition using velocity-scaling optimized hybrid explicit/implicit solvent REMD simulation	2015	[Yu15]
TIGER2 with solvent energy averaging (TIGER2A): An accelerated sampling method for large molecular	2015	[Li15]

systems with explicit representation of solvent		
Greedy replica exchange algorithm for heterogeneous computing grids	2015	[Loc15]
Improving the Replica-Exchange Molecular-Dynamics Method for Efficient Sampling in the Temperature Space	2015	[Che15]
Enhanced Conformational Sampling Using Replica Exchange with Concurrent Solute Scaling and Hamiltonian Biasing Realized in One Dimension	2015	[Yan15]
Accelerating molecular simulations of proteins using Bayesian inference on weak information	2015	[Per15]
GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations	2015	[Jun15]
Large Scale Asynchronous and Distributed Multi-Dimensional Replica Exchange Molecular Simulations and Efficiency Analysis	2015	[Xia15]

From these 20 cited works, the common terms present in the abstract, title and keywords fields, and also pertaining this described study, were extracted. They were the following:

- Molecular Dynamics
- MD
- Simulation
- REMD
- Replica Exchange
- Sampling
- Optimization
- Filter
- Filtering

Based on these common terms, a refined search string using a Boolean logic was built aiming to better filter the large amount of studies using the REMD method:

(REMD OR replica exchange) AND (sampling OR optimization OR filter OR filtering)

While regular Molecular Dynamics can also contribute to optimize the REMD method, the usage of this term in the final search string massively increased the amount of studies found to the point of becoming intractable to analyze them in the available time frame for this work and it was, therefore, discarded.

Using the presented final search string, an extensive search was performed on the PubMed database [Pub17] as well as the ACM Digital Library [Ass17], Wiley Online Library [Joh17], IEEE Xplore Digital Library

[IEE17], Scopus database [Els17] and the ACS Publications database [Ame17]. These virtual databases are widely known among the scientific community and largely used.

Searching for “Replica Exchange”, due to the Steeming process adopted by the Wiley Online Library and the ACS Publications database which also use variants of the word, would result in incorrect searches like “Replicated” and the final ensemble would contain a massive amount of undesired works. The term was then excluded from the search on these two databases.

A total of 2032 studies were found. Table 4 show the contribution of each database to this total.

Table 4 - Number of Studies Found on Each Database

Database	Number of Studies Found	% of Total
ACM Digital Library	9	≈ 0%
ACS Publications Library	651	≈ 32%
IEEE Xplore Digital Library	21	≈ 1%
PubMED Database	484	≈ 24%
Scopus Database	674	≈ 33%
Wiley Online Library	193	≈ 10%

From this large portion of articles, book chapters and revisions, only a few were chosen as related work to the described study. Regarding the acceptance and rejection of the studies, the following criteria were used:

1. Only studies written in the English language were considered.

2. The title of the studies were read. Those considered completely out of the scope of this work were rejected.

3. The abstracts were then read. Those considered out of the scope of this work were rejected.

4. The full text of the remaining studies were then read and accepted or rejected based on the following criteria:

    a) Studies of MD directed to specific proteins or protein types were rejected.

    b) Studies which increased the quality output of the REMD method, but not its computational cost were rejected.

    c) The remaining studies which successfully decreased the computational costs of the REMD or regular MD methods were accepted.

After these criteria were applied, the remaining ensemble of studies was significantly filtered and totaled 77 entries, in which only 56 were unique.

Table 5 show the exact contribution of each database to this total. Adding the studies found in the pilot search to this ensemble, a total of 63 unique related works were found.

Table 5 - Related Works Found for Each Database (Containing Duplicates)

Database	Number of Studies Found	% of Total
ACM Digital Library	2	≈ 3%
ACS Publications Library	11	≈ 14%
IEEE Xplore Digital Library	5	≈ 7%
PubMED Database	27	≈ 35%
Scopus Database	32	≈ 42%
Wiley Online Library	0	= 0%

Although the idea for creating a tool capable of filtering an ensemble of structure predictions based on existing absolute quality metrics is rather simple, no such tool was found in the literature review performed. Several different approaches to optimize the REMD method using both hardware and software solutions were found however.

In order to better analyze the studies found, a simple classification form was created based on the optimization approach of the REMD method. The vast majority of the different correlated works found can be fitted into 3 distinct categories created by the author. Table 6 depicts the unique studies found and their assigned optimization category.

Table 6 - Related Works with Assigned Optimization Category

Title	Year	Optimization Category	Reference
Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation	2003	Simulation Convergence Efficiency	[Rhe03]
A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling	2006	Simulation Convergence Efficiency	[Aff06]
Improved Efficiency of Replica Exchange Simulations Through Use of a Hybrid Explicit/Implicit Solvation Model	2006	Calculation or Hardware Efficiency	[Oku06]
Accelerating the Replica Exchange Method Through an Efficient All-Pairs Exchange	2007	Simulation Convergence Efficiency	[Bre07]
An Extremal Optimization Search Method for the Protein Folding Problem: The Go-Model Example	2007	Simulation Convergence Efficiency	[Shm07]
An Improved Replica-Exchange Sampling Method: Temperature Intervals with Global Energy	2007	Simulation Convergence Efficiency	[Li07a]

Reassignment			
Dynamics and Optimal Number of Replicas in Parallel Tempering Simulations	2007	Number of Replicas Efficiency	[Nad07a]
Folding Simulations with Novel Conformational Search Method	2007	Number of Replicas Efficiency	[Son07]
Grid-Based Asynchronous Replica Exchange	2007	Calculation or Hardware Efficiency	[Li07b]
Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir	2007	Simulation Convergence Efficiency	[Oku07]
Molecular Dynamics Simulations Using Temperature-Enhanced Essential Dynamics Replica Exchange	2007	Simulation Convergence Efficiency	[Kub07]
Optimization of Replica Exchange Molecular Dynamics by Fast Mimicking	2007	Special	[Hri07]
Optimizing Replica Exchange Moves For Molecular Dynamics	2007	Simulation Convergence Efficiency	[Nad07b]
Serial Replica Exchange	2007	Calculation or Hardware Efficiency	[Haf07]
Smart Resolution Replica Exchange: An Efficient Algorithm for Exploring Complex Energy Landscapes	2007	Simulation Convergence Efficiency	[Liu07]
A Global Optimization Scheme: Kernel Replica Exchange Simulation Method for Protein Folding	2008	Number of Replicas Efficiency	[Mu08]
Asynchronous Replica Exchange for Molecular Simulations	2008	Calculation or Hardware Efficiency	[Gal08]
Fragment Replica-Exchange Method for Efficient Protein Conformation Sampling	2008	Number of Replicas Efficiency	[Suz08]
Hamiltonian Replica Exchange Molecular Dynamics Using Soft-Core Interactions	2008	Number of Replicas Efficiency	[Hri08]
Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration	2008	Simulation Convergence Efficiency	[Faj08]
Optimized Explicit-Solvent Replica Exchange Molecular Dynamics from Scratch	2008	Number of Replicas Efficiency	[Nad08]
Enhanced Conformational Sampling Of Nucleic Acids By a New Hamiltonian	2009	Calculation or Hardware	[Cur09]

Replica Exchange Molecular Dynamics Approach		Efficiency	
Optimal Replica Exchange Method Combined with Tsallis Weight Sampling	2009	Simulation Convergence Efficiency	[Kim09]
TIGER2: An Improved Algorithm for Temperature Intervals with Global Exchange of Replicas	2009	Simulation Convergence Efficiency	[Li09]
How Hot? Systematic Convergence of the Replica Exchange Method Using Multiple Reservoirs.	2010	Simulation Convergence Efficiency	[Rus10]
Replica Exchange Simulation Method Using Temperature and Solvent Viscosity	2010	Number of Replicas Efficiency	[Ngu10]
Massively Parallelized Replica-Exchange Simulations of Polymers on GPUs	2011	Calculation or Hardware Efficiency	[Gro11]
Optimization of Monte Carlo Trial Moves for Protein Simulations	2011	Calculation or Hardware Efficiency	[Bet11]
Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2)	2011	Number of Replicas Efficiency	[Wan11]
Coulomb Replica-Exchange Method: Handling Electrostatic Attractive and Repulsive Forces for Biomolecules	2012	Number of Replicas Efficiency	[Ito12]
Efficient Conformational Sampling in Explicit Solvent Using a Hybrid Replica Exchange Molecular Dynamics Method	2012	Number of Replicas Efficiency	[Cha12]
Free Energy Guided Sampling	2012	Simulation Convergence Efficiency	[Zho12]
pH-Replica Exchange Molecular Dynamics in Proteins Using a Discrete Protonation Method	2012	Simulation Convergence Efficiency	[Das12]
Replica Exchange Statistical Temperature Molecular Dynamics Algorithm	2012	Number of Replicas Efficiency	[Kim12]
A Convective Replica-Exchange Method for Sampling New Energy Basins	2013	Simulation Convergence Efficiency	[Spi13]
A Framework for Flexible and Scalable Replica-Exchange on Production Distributed CI	2013	Calculation or Hardware Efficiency	[Rad13]
A Hadoop Approach to Advanced Sampling Algorithms in Molecular	2013	Calculation or Hardware	[Niu13]

Dynamics Simulation on Cloud Computing		Efficiency	
Implementing Replica Exchange Molecular Dynamics Using Work Queue	2013	Calculation or Hardware Efficiency	[Smi13]
K MapReduce: A Scalable Tool for Data-Processing and Search/Ensemble Applications on Large-Scale Supercomputers	2013	Calculation or Hardware Efficiency	[Mat13]
MuSTAR MD: Multi-scale Sampling Using Temperature Accelerated and Replica Exchange Molecular Dynamics	2013	Simulation Convergence Efficiency	[Yam13]
Optimization of Umbrella Sampling Replica Exchange Molecular Dynamics by Replica Positioning	2013	Simulation Convergence Efficiency	[Das13]
Accelerate Sampling in Atomistic Energy Landscapes Using Topology-Based Coarse-Grained Models	2014	Simulation Convergence Efficiency	[Zha14]
Scalable replica-exchange framework for Wang-Landau sampling	2014	Calculation or Hardware Efficiency	[Vog14]
Superposition-Enhanced Estimation of Optimal Temperature Spacings for Parallel Tempering Simulations	2014	Simulation Convergence Efficiency	[Bal14]
Theory of Adaptive Optimization for Umbrella Sampling	2014	Simulation Convergence Efficiency	[Par14]
A Generic Implementation of Replica Exchange with Solute Tempering (REST2) Algorithm in NAMD for Complex Biophysical Simulations	2015	Calculation or Hardware Efficiency	[Jo15]
A LAMMPS Implementation of Volume–Temperature Replica Exchange Molecular Dynamics	2015	Simulation Convergence Efficiency	[Liu15]
Accelerating Molecular Simulations of Proteins Using Bayesian Inference on Weak Information	2015	Simulation Convergence Efficiency	[Per15]
Asynchronous Replica Exchange Software for Grid and Heterogeneous Computing	2015	Calculation or Hardware Efficiency	[Gal15]
Conformational Sampling Enhancement of Replica Exchange Molecular Dynamics Simulations Using Swarm Particle Intelligence	2015	Simulation Convergence Efficiency	[Kam15]
Enhanced Conformational Sampling Using Replica Exchange with Concurrent Solute Scaling and	2015	Number of Replicas Efficiency	[Yan15]

Hamiltonian Biasing Realized on One Dimension			
GENESIS: a Hybrid-Parallel and Multi-Scale Molecular Dynamics Simulator with Enhanced Sampling Algorithms for Biomolecular and Cellular Simulations	2015	Calculation or Hardware Efficiency	[Jun15]
Greedy Replica Exchange Algorithm For Heterogeneous Computing Grids	2015	Calculation or Hardware Efficiency	[Loc15]
Increasing The Sampling Efficiency of Protein Conformational Transition Using Velocity-Scaling Optimized Hybrid Explicit/Implicit Solvent Remd Simulation	2015	Calculation or Hardware Efficiency	[Yu15]
Improving the Replica-Exchange Molecular-Dynamics Method for Efficient Sampling in the Temperature Space	2015	Calculation or Hardware Efficiency	[Che15]
Large Scale Asynchronous and Distributed Multi-Dimensional Replica Exchange Molecular Simulations and Efficiency Analysis	2015	Calculation or Hardware Efficiency	[Xia15]
TIGER2 With Solvent Energy Averaging (TIGER2A): An Accelerated Sampling Method for Large Molecular Systems with Explicit Representation of Solvent	2015	Simulation Convergence Efficiency	[Li15]
A Population-Based Conformational Optimal Algorithm Using Replica-Exchange in Ab-Initio Protein Structure Prediction	2016	Simulation Convergence Efficiency	[Zha16]
Coarse kMC-based replica exchange algorithms for the accelerated simulation of protein folding in explicit solvent	2016	Number of Replicas Efficiency	[Pet16]
Hadoop-Based Replica Exchange Over Heterogeneous Distributed Cyberinfrastructures	2016	Calculation or Hardware Efficiency	[Pla17]
Multiscale Implementation of Infinite-Swap Replica Exchange Molecular Dynamics	2016	Simulation Convergence Efficiency	[Yu16]
Walking Freely In The Energy And Temperature Space By The Modified Replica Exchange Molecular Dynamics Method	2016	Simulation Convergence Efficiency	[Chen16]
Efficient Conformational Search Based on Structural Dissimilarity Sampling: Applications for Reproducing	2017	Simulation Convergence Efficiency	[Har17]

Structural Transitions of Proteins			

Many solutions that improve the efficiency of REMD simulations are based on parallelizing the process (often by temperatures) on a computational efficient grid or reducing the effort needed to compute the factors on each step of the simulation. There are also methods that divide the workload of each step by assigned a certain number of atoms and/or residues into several different processors or even different machines. Both kinds of approaches are usually fitted to run on large and powerful computational grids, often designed specifically for this purpose, such as the Anton massively parallel supercomputer [Sha07a]. There are also methods that rely on reducing the computational burden of MD simulations, often grouping atoms into residues or removing atoms deemed unnecessary for the final quality of the predictions. While these cited methods achieve faster run simulations, they do not significantly alter the REMD process itself in terms of replicas and total simulation time. These optimization approaches were then labeled as "Calculation or Hardware Efficiency".

Methods within the "Number of Replicas Efficiency" optimization category are the second largest group of the related works found. These approaches (exclusive to the REMD method) aim to optimize the simulation by reducing the amount of replicas that need to be run simultaneously without significantly affecting the sampling efficiency, that is, the capacity of the replicas to reach a wide array of potential 3D structures. By reducing the number of replicas of a REMD PSP simulation, even if its only one, the computational cost is significantly reduced (see EQ. 8). Disregarding the cost of posterior analysis, these optimization approaches have greater impact on computer with lower performance however, as a powerful enough computer (or cluster) capable of simulating all replicas in parallel in a totally independent way is not really affected by this optimization.

The number of existing works that use this approach to optimize the REMD method is a strong indicator that the simulation process is somewhat faulty at this point or at least lack refinement, that is, the number of replicas that need to be simulated in a conventional REMD PSP simulation is greater than the ideal number, or at least does not contribute significantly in terms of satisfactory predictions. This idea will be brought back posteriorly in chapter 6.

Lastly, the vast majority of related works, reaching almost half the studies cited, were classified within the "Simulation Convergence Efficiency" category. This optimization approach is defined by slight modifications of the REMD or regular MD methods aiming to increase the convergence of the system. The computational gain in these approaches rely solely on reducing the total steps (simulation time) of the system without significantly affecting the quality of the resulting ensembles. This real computational gains of these approaches, however, ultimately dependent on protein being simulated, as a protein fold that does not converge easily will still have a high computational cost. On the other side, running the simulation more rapidly, even its only a nanosecond, result in a significant computational gain, especially in the REMD method (see EQ. 8). These approaches, therefore, have the most significant impact of the 3

categories. Figure 18 shows the exact distribution of the assigned categories to the related works found.

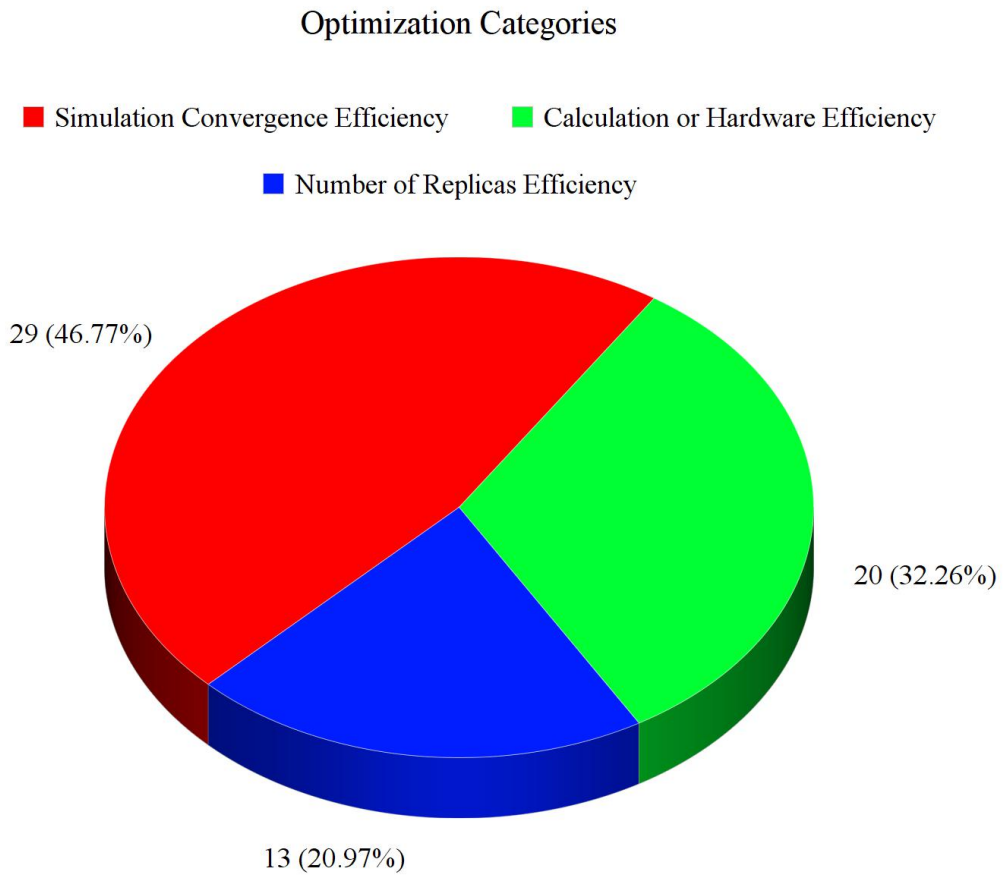

Figure 18 - Pie chart of the optimization categories assigned to related works. Plot created using the online Chart Tool [Zyg17].

Apart from these 3 categories, only a single study was found, marked as “Special” in Optimization Category field of Table 6. This work uses the approach of mimicking the REMD PSP simulation process in order to find optimal configurations, as well as other useful information, for later real simulations.

It is worth noticing that the cited works, except for 3 studies, were published 10 years ago, and more than half of them just 5 years ago. That means the problem of REMD simulations optimization is still a topic of great interest in the research community and also a strong indicator that the method still has room for more improvements.

5. METHODOLOGY

This chapter will describe the methodology applied in this work regarding the REMD PSP simulations test datasets used, software resources employed and the how the relative quality metrics were calculated.

5.1. REMD PSP Simulations Test Dataset

In order to analyze possible ways to filters out unsatisfactory protein structures predictions on REMD PSP simulations, a dataset of such simulations was needed. Colleagues Lipinksi-Paes and Norberto de Souza kindly granted part of the dataset used in their latest work [Lip17] (denoted as F protocols) through personal communication, which contained REMD PSP simulations of 9 different proteins.

The dataset of REMD PSP simulations contained proteins of α, β and αβ conformations, all simulated using the pairwise generalized Born solvation model proposed by Hawkins, Cramer and Truhlar in 1995 [Haw95, Haw96]. This solvation model is an implicit solvation models that significantly reduces the computational costs of the simulations. The MD software used to simulate the test dataset was the AMBER 14 [Cas14] with the PMEMD module, which is an extensively-modified program provided by the AMBER package. This module enables the use of Graphics Processing Units (or GPUs) to massively accelerate MD simulations that use the generalized Born solvation model or other specific explicit solvent models.

All the simulations of the test dataset started with the extended proteins' structures constructed with the tleap module of AMBER, a basic preparation programs for AMBER simulations. An initial step of energy minimization, necessary to avoid errors during the protein fold, was also executed. The simulations were then performed under 50 nanoseconds each, with a snapshot frequency of 1 picosecond. The cutoff radius for calculating the potential energy of the structures was defined as the fixed value of 8.0 Å and the exchange attempt frequency of the REMD simulations were defined as 0,020 ps^{-1}. Table 7 summarizes the information about the proteins simulated in the test dataset.

Table 7 - Proteins Simulated in the REMD Dataset

PDB ID	Amino Acids Sequence	Number of Residues	Class	Reference
1L2Y	NLYIQWLKDGGPSSGRPPPS	20	α	[Nei02]
1RIJ	ALQELLGQWLKDGGPSSGRPPPS	23	α	[Liu04]
1VII	MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF	36	α	[Mck97]
2WXC	GSQNNDALSPAIRRLLAEWNLDASAIKGTGVGGRLTREDVEKH	47	α	[Neu09]

	LAKA			
1UAO	GYDPETGTWG	10	β	[Hon04]
1LE1	SWTWENGKWTWKX	13	β	[Coc01]
1E0L	GATAVSEWTEYKTADGKTYYY NNRTLESTWEKPQELK	37 (26)	β	[Mac00]
1FME	EQYTAKYKGRTFRNEKELRDFI EKFKGR	28	αβ	[Sar01]
1PSV	KPYTARIKGRTFSNEKELRDFLE TFTGR	28	αβ	[Dah97]

The protein 1E0L is composed of 37 amino acid residues, however a great part of those residues only constitute random coils. Because of that, this protein was simulated using only 26 residues that compose the secondary structures of the protein (three β sheets), ranging from the 6th residue to the 31th residue.

Each REMD PSP simulations of the test dataset used up to 16 different temperatures (or replicas) ranging from 269.50 Kelvins to 537.54 Kelvins. Although 537.54 Kelvins is a temperature notably higher than those found on living beings, it is necessary to break local minima as described in chapter 2.7. At such high temperatures, however, unwanted rotations around the peptide bond might occur leading to non-physical chiralities [Roe01]. Chirality restraint on the backbone were then applied to the simulation through the use of the "makeCHIR_RST" script provided with the AMBER 14 package.

The exact number of temperatures (or replicas) chosen for each REMD PSP simulation can be observed in Table 8 and were defined according to the number of atoms and the degrees of freedom of the system. The temperatures were obtained using the temperature predictor web server created by Patriksson and Spoel in 2008 [Pat08], which yields an array of temperatures capable of achieving the desired exchange probability for any given system.

Table 8 - Number of Replicas for Each Protein

PDB ID	Number of Residues	Number of Replicas	Total Number of Structures Predicted
1UAO	10	8	400,000
1LE1	13	10	500,000
1L2Y	20	12	600,000
1RIJ	23	12	600,000
1FME	28	14	699,650*
1PSV	28	14	700,000
1VII	36	14	700,000
1E0L	37 (26)	14	700,000
2WXC	47	16	800,000

* It is still unclear as what caused this particular simulation to result in less predictions, but nevertheless the amount of missing structures represent only 0.05% of the desired total amount of 700.000 and therefore shouldn't significantly impact the subsequent analysis.

The SHAKE algorithm [Ryc77] was also applied to restraint bonds involving hydrogen atoms [Liu07] in the simulations. Typically these bonds can be neglected to improve the computational efficiency of the simulation without significantly impacting its final quality. Finally, the force field used in the simulations was the ff12SB force field present in the AMBER 14 package, recommended by the developers of AMBER at the time for simulating proteins and nucleic acids. This force field is a continuation of the previous ff99SB force field [Hor06].

5.2. Case Study of a REMD PSP Simulation: Protein 1UNC

Aiming to test a series of initial hypothesis and answer preliminary questions, the 3D structure of the Human Villin C-Terminal Headpiece Subdomain, PDB ID 1UNC [Ver04], was chosen for a case of study. This small protein is composed of only 35 amino acid residues that form 3 different alpha helices joined together by a tightly packed hydrophobic nucleus [Lip17]. It was chosen for the case study for two main reason: (i) it was the first protein simulated from the test dataset and thus was readily available at the time and (ii) the protein is one of the smallest proteins to posses such large amount of secondary structure elements. While its small size contribute for quickly calculating absolute quality metrics, the amount of secondary structures elements present in its structure is a valuable asset to evaluate their efficacy in assessing the predicted structures obtained from the REMD PSP simulations.

This protein was simulated under the same conditions as the other 9 described in chapter 5.1, with the only difference that the simulation was performed in triplicates with different random seeds. A total of 14 different temperatures were used in these simulations. This generated a total of 700.000 structure predictions.

Per request of colleague Thiago Lipinski Paes, this protein was also used to test a few hypothesis outside the proposed work spectrum of this dissertation. Although only marginally related to the work here described, the results found are believed to be sufficiently valuable and worth reporting to future colleagues interested in further studying it. This part of the work was labeled as additional hypothesis for clearer distinction.

In order to test these additional hypothesis, an auxiliary test dataset was also granted by colleagues Lipinksi-Paes and Norberto de Souza. In this auxiliary test dataset, the protein 1UNC was simulated using 8 different simulations protocols, also performed in triplicates. Among the methods used, the CuT-REMD proposed by Lipinski-Paes and Noberto de Souza in 2017 [Lip17] was employed along Cutoff Molecular Dynamics, conventional REMD and conventional MD. The Cutoff Molecular Dynamics (denoted Cu-MD in this study) is a modification of conventional MD methods that use increasingly cutoff radius for energy calculations, until a maximum value is reached. The CuT-REMD has the same strategy, but is a modification of REMD simulations instead of conventional MD simulations. Both methods use a permanence time that denotes the simulation time spent on a determined cutoff radius. After such time passes, the cutoff radius is increased. The standard value of 1

Å was used. Table 9 shows a summary of the protocols used, along with a identification label to facilitate posterior analysis.

Table 9 - Simulations Protocols Used for Testing the Additional Hypothesis

Label	Simulation Method	Initial Cutoff Radius	Max. Cutoff Radius	Permanence Time	Exchange Attempt Frequency
A	CuT-REMD	4.0 Å	8.0 Å	1 ns	1.000 ps^{-1}
B	CuT-REMD	4.0 Å	8.0 Å	1 ns	0.020 ps^{-1}
C	CuT-REMD	4.0 Å	8.0 Å	2 ns	1.000 ps^{-1}
D	CuT-REMD	4.0 Å	8.0 Å	2 ns	0.025 ps^{-1}
E	Cu-MD	4.0 Å	8.0 Å	1 ns	-
F	Cu-MD	4.0 Å	8.0 Å	2 ns	-
G	REMD	8.0 Å	8.0 Å	-	1.000 ps^{-1}
H	MD	8.0 Å	8.0 Å	-	-

The results were clustered using common methods of the literature [Lip17]. Only the structures obtained in the first 4 temperatures of the REMD PSP simulation were used for this clustering. The clusters were computed using the cpptraj [Pea95] module of AMBER and more specifically the average-linkage algorithm [Sha07b]. This algorithm works the following way:

1. All the predicted structures given as input to the algorithm compose a pool of non-clustered structures.

2. For each structure in the pool of non-clustered structures, the RMSD of all other structures in the pool is calculated using the selected one as model.

3. Structures with RMSD equal or lower than ε Ångströms are considered neighbors. The value of epsilon (ε) is defined by the user. The standard value of 2.0 Å was used in this work [Lin11, Dau99].

4. The structure with the highest number of neighbors is established as the centroid of a cluster containing all its neighbors. The structures of this cluster are eliminated from the pool of non-clustered structures.

5. This process is repeated from step 2 until the pool of non-clustered structures is empty.

Using this algorithm, a series of non-overlapping clusters of structures are obtained [Dau99]. The cluster-to-cluster distance is defined as the average of all distances between individual points of the two clusters. The clustering algorithm was also configured to only use the Cα carbons of residues present in the secondary structures of the experimental structure. This procedure is commonly used when clustering structures predicted by MD simulations and is used to avoid situations of loops and termini disrupting the clusters [Per15].

5.3. Relative Metrics Calculation

Aiming to better assess the quality of the predicted structures from the REMD PSP simulations, the RMSD and GDT_TS relative quality metrics were chosen as the main forms of evaluation. While the GDT_TS is capable of restricting the influence of random loops (or coils) in its formula, the RMSD is highly sensitive to these irregular conformations that, for the most part, can be safely ignored from the quality assessment for not contributing much to the structure function.

For calculating the GDT_TS relative quality metric, all residues of the proteins were considered. For the RMSD calculation, on the other hand, only an intervals of the residues were considered. In summary, the first and last few residues were discarded wherever possible, that is, when a secondary structure was not present. These first and last residues generally configure random coils. Extended regions of random coils were also discarded, which happened only in the largest 2WXC protein of the test dataset. The interval of residues used to calculate the RMSD of each protein can be seen in Table 10.

Table 10 - RMSD Calculation Intervals for Each Protein

PDB ID	Total Number of Residues	Residue Interval Used to Calculate RMSD
1UAO	10	1-10
1LE1	13	1-12
1L2Y	20	3-18
1RIJ	23	2-22
1FME	28	2-28
1PSV	28	2-27
1VII	36	3-32
1E0L	37 (26)	1-26
2WXC	47	10-28 & 36-47

For most part, the GDT_TS relative quality metric was given preference over the RMSD score for 2 reasons: (i) the GDT_TS provides a higher quality assessment than the simple RMSD, and (ii) while the RMSD score ranges between 0 and infinite, the GDT_TS results in a fixed interval ranging between 0 and 1, which facilitates further analysis and the creation of smaller and clearer graphics.

6. DISCUSSION AND RESULTS

In this chapter the SnapFi tool will be presented along the proposed filtering methodology. Firstly, the analytical data filtering process being proposed in this study will be presented. The SnapFi tool and its development process will then be discussed. After that, the case study of the protein 1UNC will be analyzed, followed by the study of the entire test dataset. Lastly a proposed filtering method that can be easily applied with the SnapFi tool will be presented.

6.1. Analytical Data Filtering Methodology

Based on the optimization categories presented in Chapter 4, the current workflow of optimizing and running a REMD PSP simulation can be seen in Figure 19.

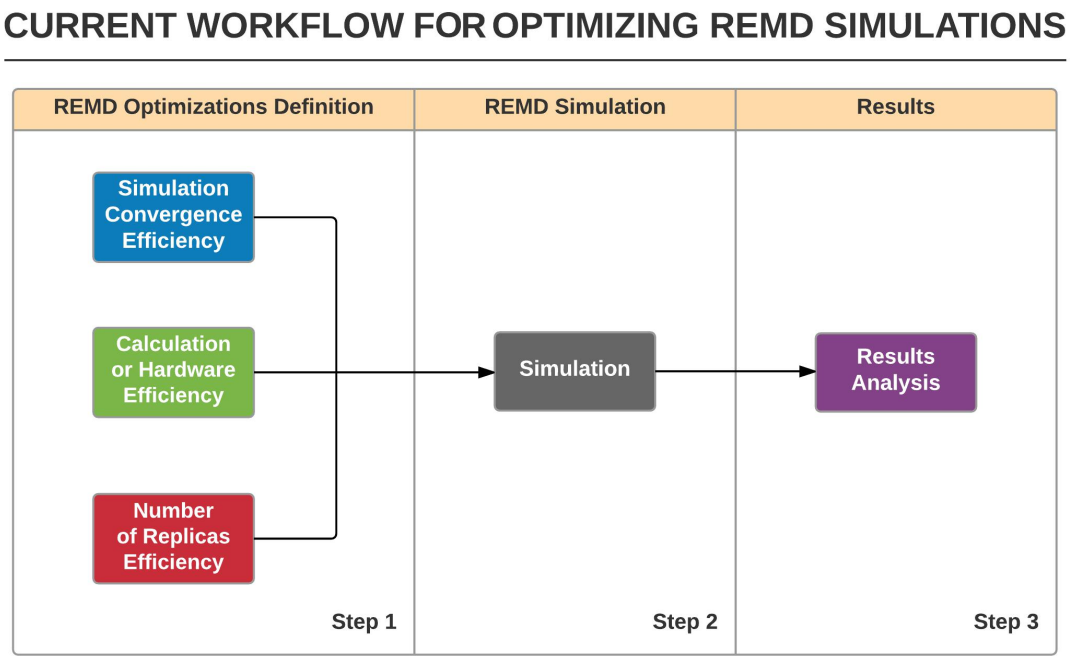

Figure 19 - Current workflow diagram of optimizing and running a REMD PSP simulation. Diagram created using the online Lucidchart tool [Luc17].

Out of the 3 different optimizations approaches, only the simulation convergence efficiency and number of replicas efficiency categories may affect the final number of structure predicted of the REMD PSP simulation. Since the proposed optimization approach of using absolute quality metrics to filter the structure prediction ensemble generated at the end of the REMD PSP simulation does not change the simulation process itself, it can, therefore, be used in parallel with the other cited optimization methods, further optimizing REMD PSP simulations. This process was labeled as Analytical Data Filtering.

This optimization approach proposed has the potential to improve even further the efficiency of REMD PSP simulations by targeting an area still untouched at large by optimization advancements. It would, therefore, introduce a fourth optimization approach, and an intermediary third step in the workflow, in additional to those already presented. This can be better seen in the workflow diagram of Figure 20.

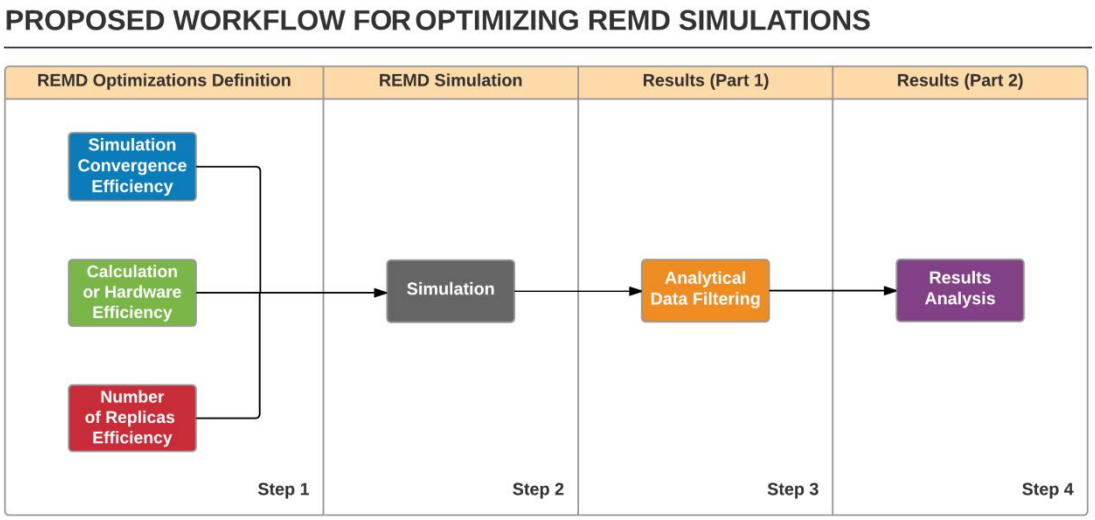

Figure 20 - Proposed workflow diagram of optimizing and running a REMD PSP simulation. Diagram created using the online Lucidchart tool [Luc17].

The proposed SnapFi tool aims to fill this gap by creating an algorithm that is able to filter part of the resulting data, eliminating unsatisfactory protein structure predictions from the predicted structures ensemble, and thus optimizing the posterior analysis of the simulation results. Furthermore, per helpful advice of professor Dr. Duncan Dubugras Ruiz during the progress review seminar, the tool was designed in a modular fashion so that it could not only be used in parallel with the cited optimization approaches, but also accommodate novel absolute quality metrics yet to be developed that may become part of new filtering approaches devised by other research colleagues. Such studies are greatly encouraged by the authors, as the present investigation couldn't test all the possible filtering conformations. This will be further discussed in chapter 6.5.

6.2. Introducing the SnapFi Tool

The SnapFi tool is composed by 11 quality metric extraction scripts, 8 "module" scripts that include utilities and filtering algorithms, and a main iteration script that controls the flow of the program. They were all programmed using version 3.6.1 of the Python programming language [Pyt17]. The Python is a powerful programming language easy to code and is supported by various operational systems.

The 11 metric extraction scripts extract the absolute and relative quality metrics described in chapter 2.11. They take as input a text file containing a

list of structures' IDs. These IDs are assigned to each structure inside an ensemble of predicted structures (resulted of a REMD PSP simulation) by another utility script, which loads such ensembles and extract each predicted structure as a separate *pdb* file (necessary for extracting the quality metrics). The quality metric extraction scripts therefore load these separate *pdb* files and assign them a respective quality score. The final output generated is a text file containing a mapping of the structures' IDs and their assigned scores. Some quality metrics also have additional auxiliary files necessary for their execution that are not worth mentioning in this dissertation, but are explained in depth on their respective softwares. In summary, the quality metric extraction scripts are:

- *dDFIRE.py script,*
- *DFIRE.py script,*
- *DOPE.py script,*
- *GFactor.py script,*
- *GOAP script,*
- *OPUS_PSP.py script,*
- *Probscore.py script,*
- *RW_Plus.py script,*
- *Energy.py script,*
- *GDT.py script and,*
- *RMSD.py script.*

It is worth mentioning that some of the quality metric extraction scripts provided in the SnapFi tool will require slight adjustments by the user (generally just an environment setting line) due to different installation locations of the tools involved in calculating the quality metrics.

Regarding the modules scripts, they have different functions such as extracting the individual predicted structures from ensembles of predictions and filtering unsatisfactory predictions based on a given algorithm. They are:

- *Filter_Number.py*

  Receives a quality metric mapping text file generated by a quality metric extraction script, sorts the list according to the quality metric scores and returns the first *x* number of entries, where *x* is a number received as input and specified by the user.

- *Filter_Percent.py*

  Same as the Filter_Number.py script, except that instead of returning a fixed *x* number of entries, it returns a percentage of first entries received as input and specified by the user.

- *Load_PDB.py*

  Retrieve all predicted structures from one or multiple ensembles of predictions. Each predicted structure is saved into a separate *pdb* file with an assigning ID. This script iteratively call the Retrieve_Model.py

script, which effectively retrieve a single given structure from an ensemble of structures in a efficient way.

- *Retrieve_Metric.py*

Filters a quality metric mapping text file based on another file containing a list of IDs. Useful to eliminate certain structures in the mapping when running filtering scripts that read all the IDs in the mapping file to filter unsatisfactory predictions.

- *Retrieve_Model.py*

Retrieve a single given predicted structure from an ensemble of predicted structures into a single separate *pdb* file in a efficient way using the *sed* command of Linux. This command is able to read a text file at any given line without having to read the preceding lines. In order to further optimize this process, based on the structure of the ensemble file, an index file is created containing the starting line where each structure is saved. Thus when retrieving a structure, the index file is used to determine exactly where the *sed* command must read.

- *Threshold_Fixed.py*

Filters a given quality metric mapping text file based on a threshold value provided by the user. The user must also specify which type of comparison is used (i.e., greater or equal than, greater than, lower than or lower or equal than).

- *Threshold_Dynamic.py*

This script generates a threshold based on the quality metric scores contained in the quality metric mapping text file provided as input and filter the structures contained in the file using it. Firstly the mapping file is sorted according to the scores of the quality metric. A position given by the user as input is then used to extract the base value of the threshold. A threshold margin value, also provided by the user, is then applied to this base value. As an example, if using the 1º ranked score, that is equal to 10, as the base position and applying a threshold margin of -0.2 (negative 20%), the final threshold obtained is 8. The user must also specify which type of comparison is used (i.e., greater or equal than, greater than, lower than or lower or equal than). In the cited example, using the greater than comparison type, all structures with its assigned quality metric score bellow 8 would be filtered. The user can also specify if the mapping file is sorted ascending or descending.

- *Voting.py*

This script receives two or more text files containing a list of IDs. It can be a quality metric mapping text file for instance. The script then assign a number of “votes” for each structure ID contained on a input file. In summary, the final number of “votes” for each structure is defined by the amount of times that structure appears in one of the given input files. Based on the number of “votes” required, specified by the user as input,

the structures with enough “votes” are written in the final output file. The output file, therefore, contains a list of structure IDs.

Finally, the main iteration script that control the flow of the program is labeled as SnapFi.py. This script simply set environment variables and execute modules based on the filtering configuration file received as input. The filtering configuration file was designed to work similarly as a programming language. The “#” symbol denotes a commentary, which exclude posterior text from execution. Each execution step is denoted by the structure: STEP *Name* = *Module(Parameters)*, where *Name* is the name assigned by the user to the step, *Module* is the path to the module to be executed and *Parameters* are the parameters required by the module. An example of a filtering configuration file can be seen in Figure 21.

```
#FILTERING SCRIPT

STEP Data = Modules/Load_PDB.py(Ensemble.pdb)

#Retrieve Metrics

STEP dDFIRE = Metrics/dDFIRE/dDFIRE.py(Data)

STEP GFactor = Metrics/GFactor/GFactor.py(Data)

#Filter Structures According to Best 25% Values

STEP F_dDFIRE = Modules/Filter_Percent.py(0.25, dDFIRE, REVERSED=TRUE)

STEP F_GFactor = Modules/Filter_Percent.py(0.25, GFactor)

#Retrieve the Common Structures of the 2 Ensembles

STEP Final = Modules/Voting.py(2, F_dDFIRE, F_GFactor)
```

Figure 21 - Example of a filtering configuration file. In this example, the absolute quality metrics dDFIRE and GFactor are extracted from an ensemble, the best 25% scores for each metric are extracted and the union of their results (acquired by the use of the Voting.py module) is performed, generating a list of filtered structures IDs.

The modules scripts were also created in a way to facilitate writing the filtering configuration file, where the quality metric mapping text file or text files containing a list of structures IDs can be passed by parameter by simply specifying their respective STEP name. This can also be seen in all steps subsequent to the first one in the example filtering configuration file showed in Figure 21. For a clear example, the “dDFIRE” step use the list of structures IDs generated by the “Data” step as input.

As already cited, the tool was modeled with the specific intent of working together with other optimization techniques to the REMD method, proposed by several different authors. More importantly, its modular structures not only support a myriad of different filtering configurations but also easily support the development of new modules and the integration of new quality metrics.

The overall SnapFi filtration process can be seen in the workflow diagram in Figure 22.

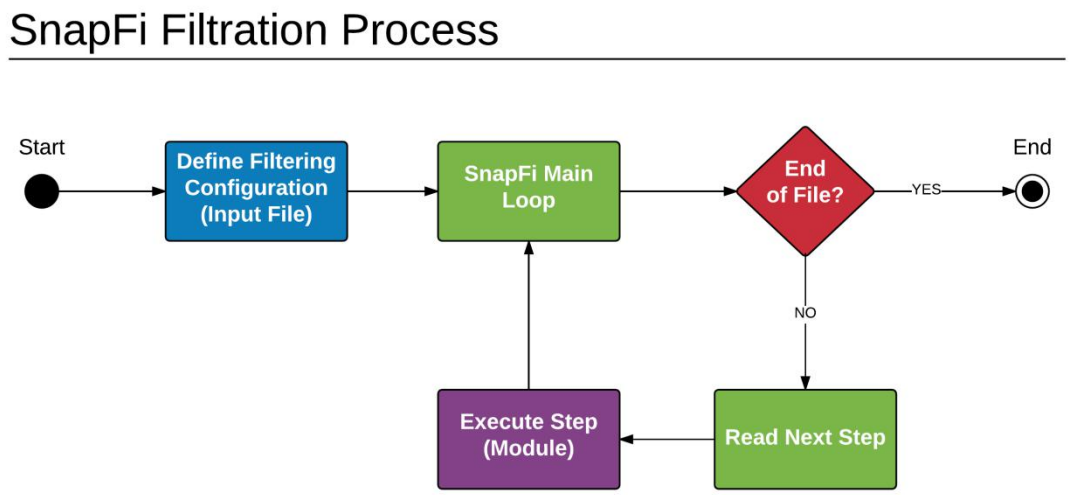

Figure 22 - Workflow diagram of the SnapFi filtration process. The definition of filtering configuration file must be made manually by the user, while the rest of the process is executed automatically by the SnapFi.py script.

Currently, the SnapFi tool only supports the Linux operational system. Although making the tool available for other operational system would not be significantly demanding in terms of time and difficulty, most programs related to the quality metrics and MD softwares are built exclusively to the Linux operational system. Thus making the developed SnapFi as a multiplatform tool was deemed unnecessary for the time being.

The SnapFi suite (and all related scripts) is distributed under the version 3 of the GNU General Public License, published by the Free Software Foundation [Fre17]. For more details, the GNU General Public License must be consulted (https://www.gnu.org/licenses/). It is possible to redistribute and modify all files included in the SnapFi tool, provided they respect the license terms. The SnapFi tool suite is freely available online in: https://github.com/Racaoma/SnapFi to all user in the hope that it will be useful.

The quality metric calculation softwares couldn't be provided along with the SnapFi scripts due to copyright terms, but for each quality metric a Readme text file was included containing instruction and links to download the required programs.

6.3. Discussion & Results: Case Study of the Protein 1UNC

This section will present the case study of the protein 1UNC regarding several formulated hypothesis about REMD PSP simulations and the efficiency of the quality metrics chosen.

6. 3. 1. First Formulated Hypothesis

The first formulated hypothesis to be tested was:

The high temperatures of REMD PSP simulations are important for simulation aspects, but can be safely discarded for post-simulation analysis.

This hypothesis was made based on several observations of common practices on different studies found in the literature and also analyzing the physical aspects of the REMD process itself. These observations are:

- During a REMD PSP simulation, replicas can break local minimums in high temperatures, but can't easily converge on them. Given enough simulation time, they will eventually exchange temperatures until reaching colder temperatures with easier convergence. While a high temperature may capture a single or few snapshots of the native-like structure, colder temperatures have a much higher chance of converging such structure and captures many snapshots. High temperatures in such cases are at least redundant for post-simulation analysis.

- Studies in the literature, such as [Zhe11, Roe14, Dau99], which analyzed REMD PSP simulations through the use of clustering algorithms, employed the method of retrieving only a few replicas (all with low temperatures) to reduced the amount of data to be analyzed.

- Personal reports of colleagues in the LABIO group who used the REMD PSP simulation method stated that native-state like structures were generally found at lower temperatures and seldom found at the elevated temperature ranges.

In order to test this, using each of the absolute quality metrics presented (except for the energy minimization, which were not included in this study so far), for each REMD PSP simulation of the triplicates, the 700,000 predicted structures were ranked. From this, the top 1% ranked structures were selected and compared to its respective GDT_TS, using the experimentally obtained structure as model. Additionally, it was also observed in which temperature the structures were extracted. Figure 23 depicts this analysis. Structures extracted from the same temperature were clustered together to facilitate visualization.

It is possible to observe that the vast majority of the top 1% scores structures are found within the lowest temperatures of the REMD PSP simulation. This pattern is also observable on all triplicates simulations of the protein 1UNC. In order to verify the impact of the high temperatures, the same analysis was performed, but instead using the 1% worst scored structures. This can be seen in Figure 24.

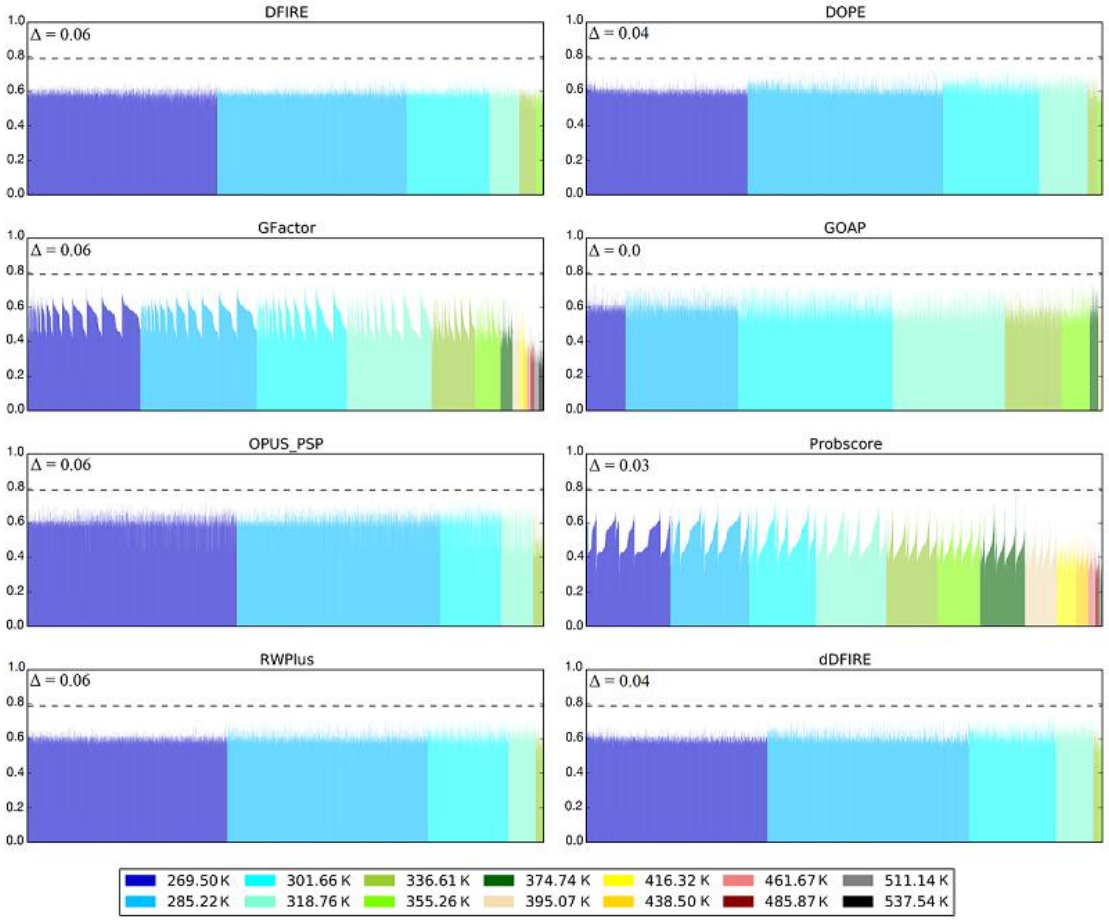

Figure 23 - Top 1% scored structures clustered by temperature (x axis) according to each absolute quality metric versus GDT_TS (y axis). Data relative to the first REMD PSP simulation of the protein 1UNC. The dotted line represent the real best GDT_TS found within the entire REMD PSP simulation, while the delta (Δ) depicts the difference between this value and the best GDT_TS found within each quality metric ensemble of top scored structures. A value of 0.0 indicates that the quality metric was capable of extracting the best predicted structure using only 1% top scored structures.

By comparing both analysis from Figure 23 and Figure 24, it becomes clear that, while the top scored structures lies within the lowest temperatures, the worst scored structures are found in the highest temperatures. Although it may be possible that this pattern is related to the absolute quality metrics chosen, this hypothesis was discarded by the author due to the pattern appearing in all chosen quality metrics and their highly different method of scoring the structures.

This pattern found supports the hypothesis that the highest temperatures of a REMD PSP simulation can be safely ignored from posterior analysis without significantly affecting the quality of the results. It is also worth noticing that the Probscore metric showed little variance regarding the Δ GDT_TS between the top 1% best predicted structures and the top 1% worst predicted structures. Posterior studies, which will be presented further on this dissertation, proved that this metric (along a few others) were unsuitable to filter good predicted structures from unsatisfactory predicted structures. This

same analysis was performed using 2% and 3% of the top predicted structures without significant changes.

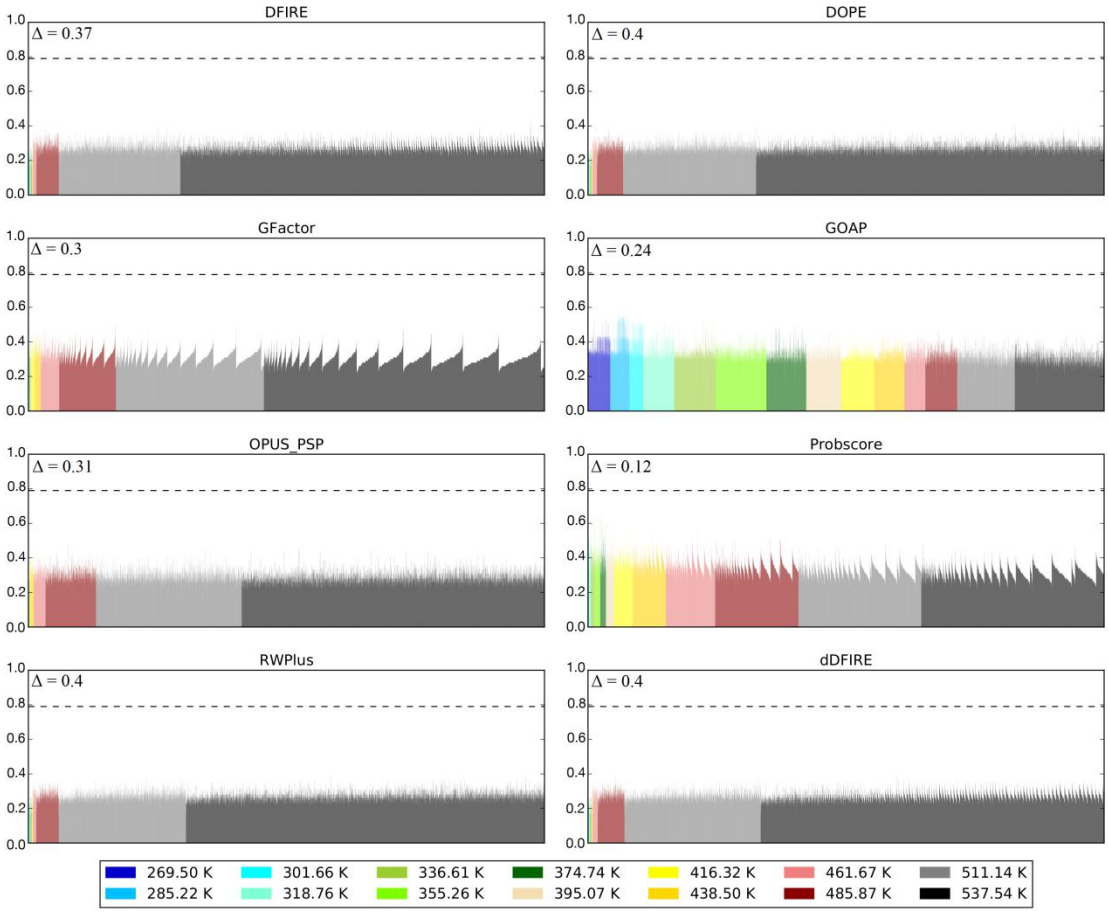

Figure 24 - Worst 1% scored structures clustered by temperature (x axis) according to each absolute quality metric versus GDT_TS (y axis). Data relative to the first REMD PSP simulation of the protein 1UNC. The dotted line represent the real best GDT_TS found within the entire REMD PSP simulation, while the delta (Δ) depicts the difference between this value and the best GDT_TS found within each quality metric ensemble of top scored structures. A greater delta value indicate a higher disparity in the retrieved ensemble to the best structure predicted in the entire REMD PSP simulation.

6. 3. 2. Second Formulated Hypothesis

The second formulated hypothesis to be tested was:

Analyzing only the first few temperatures of REMD PSP simulations is enough for capturing the best predicted structures.

This hypothesis is similar to the first formulated hypothesis, but its focus is on the number of temperatures that must be included in post-simulation analysis. In order to test this, a cumulative distribution analysis was performed that evaluated the distribution of the predicted structures according to their GDT_TS scores and the temperature in which they were extracted. The

GDT_TS scores were clustered together in bands to facilitate the analysis. Each band contained scores ranging from its base value (e.g. 0.5) until its base value plus 0.099 (e.g. 0.599). Only values above 0.5 were considered for this analysis for clearer visualization. Figure 25, 26 and 27 depict this analysis for the each triplicate simulation of the protein 1UNC.

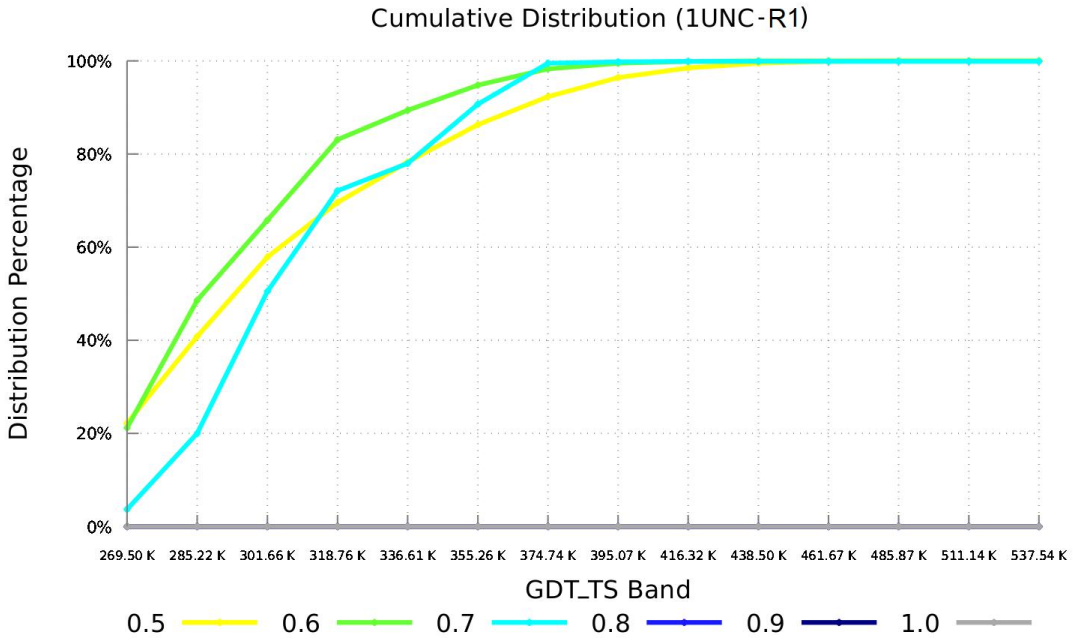

Figure 25 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the 1st REMD PSP simulation of the protein 1UNC (labeled R1).

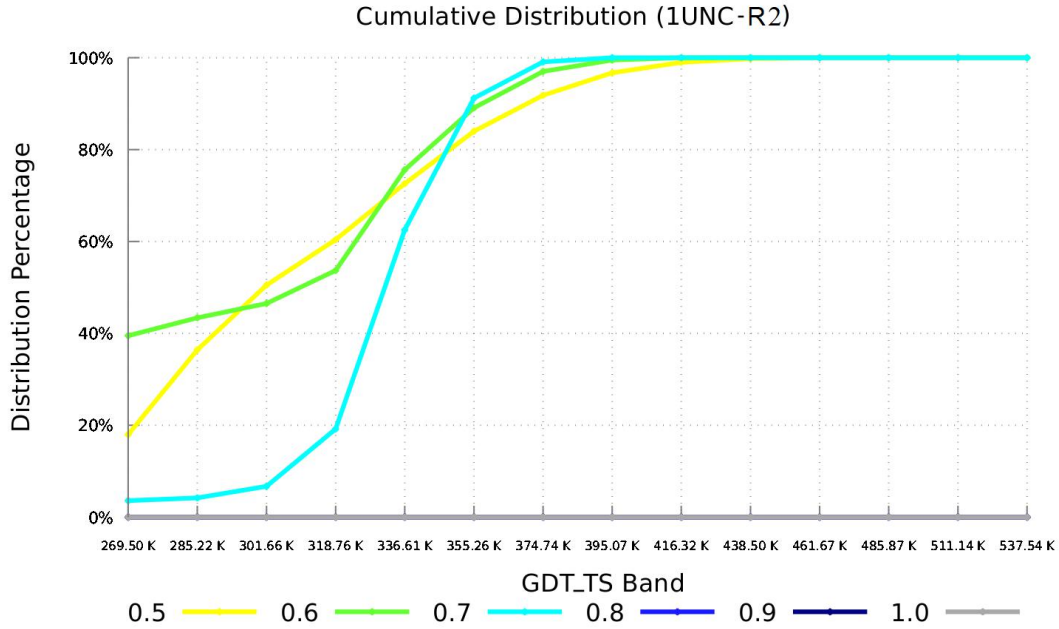

Figure 26 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted.

Data relative to the 2nd REMD PSP simulation of the protein 1UNC (labeled R2).

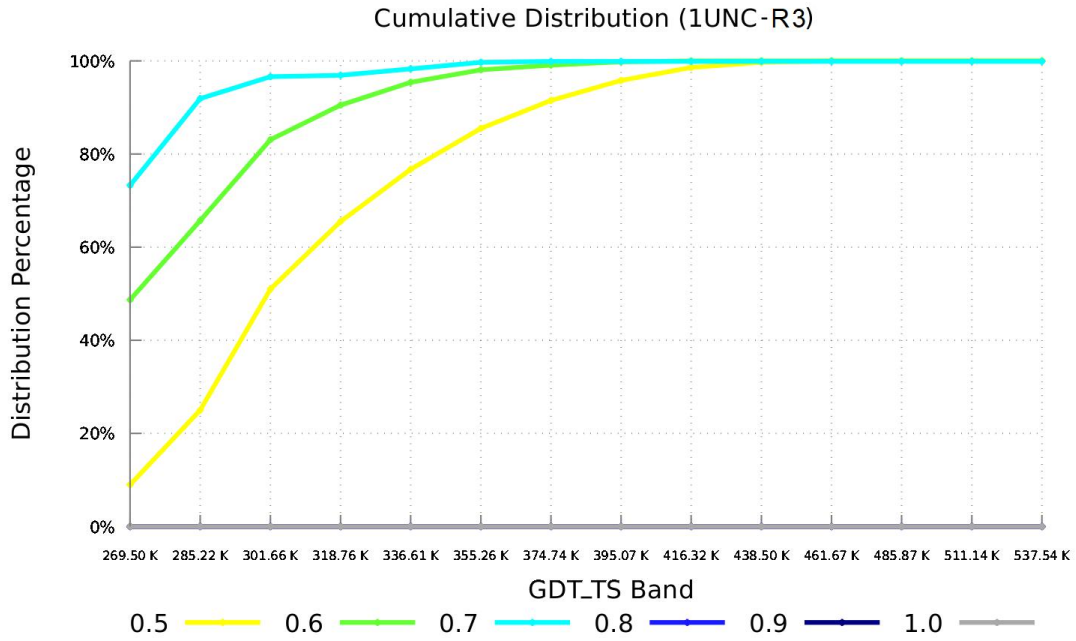

Figure 27 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the 3rd REMD PSP simulation of the protein 1UNC (labeled R3).

It is important to observe that each temperature excluded from post-simulation analysis has the potential to significantly reduce the amount of data that need to be analyzed, but may also discard native-like structure predictions. Achieving a balance between efficiency and quality is therefore vital. In order to achieve this balance, the number of temperatures in which 80% of the best GDT_TS band converge was established as the desired amount of temperatures to be analyzed. This values is believed to contain the vast majority of predictions while still reducing significantly the amount of data that needs to be analyzed posteriorly.

In Figures 25, 26 and 27, it is possible to see that the convergence of results heavily depend on the simulation itself, whereas the last triplicate simulation converged 80% of the best GDT_TS band merely on the first temperature and the second simulation converged only at the sixth temperature. Further data was then deemed necessary to safely established what number of temperatures was required to achieve the desired balance between quality and efficiency.

6.3.3. Third Formulated Hypothesis

The third formulated hypothesis to be tested was:

In terms of the quality of the results produced, some absolute quality metrics are able to outperform others, potentially enabling the exclusion of some of them for further analysis.

To test this hypothesis, using each of the absolute quality metrics presented (except for the energy minimization, which were not included in this study so far), for each of the REMD PSP simulation of the triplicates, and for each temperature of these simulations, the predicted structures were ranked. From this, the top 1% ranked structures were selected and compared to its respective GDT_TS, using the experimentally obtained structure as model. For each temperature, a different quartile distribution plot was created. This analysis can be observed in Figures 28, 29 and 30.

Although some quality metrics did appear to yield better results than some more often, this pattern was at least irregular. Further analysis on the entire test dataset, which will be presented in chapter 6.6.2, confirmed this. This same analysis was performed using 2% and 3% of the top predicted structures without significant changes. A valuable information that could be extracted from this analysis, however, was that the best predicted structure from REMD PSP simulations were outliers to the distribution. This observation will be very helpful when analyzing the additional hypothesis in chapter 6.4.

It is also worth noticing that the presented results also reinforce the first formulated hypothesis as the quartiles tend to have a much higher GDT_TS score at lower temperatures.

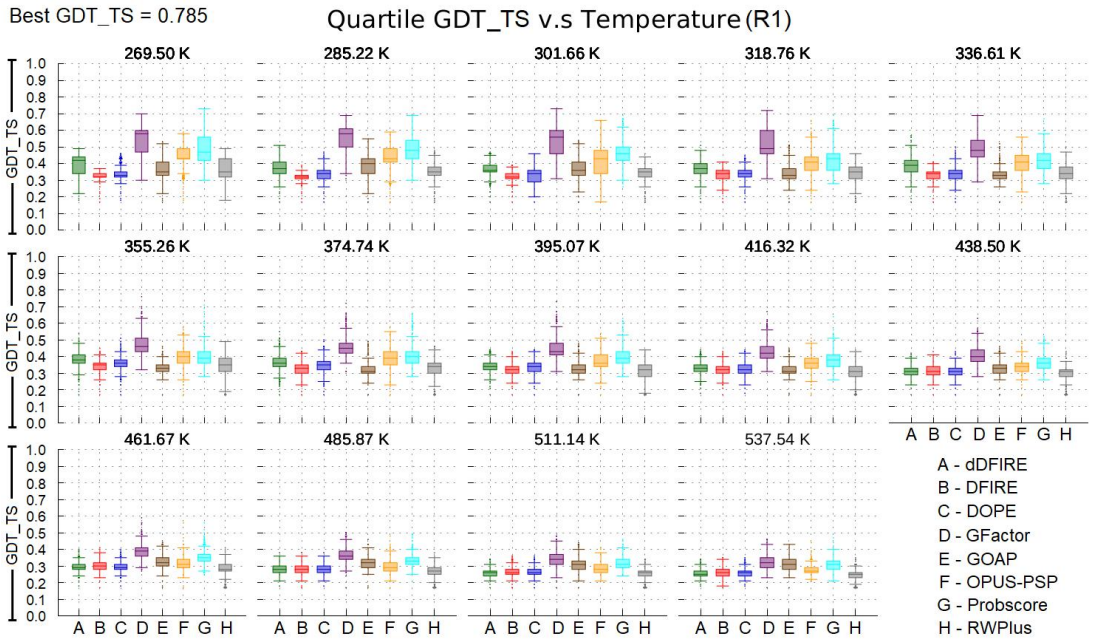

Figure 28 - Quartile distribution of GDT_TS (y axis) plot for the top 1% best scored structures according to each absolute quality metrics (x axis) and for each temperature of the simulation. Data relative to the 1st REMD PSP simulation of the protein 1UNC (labeled R1).

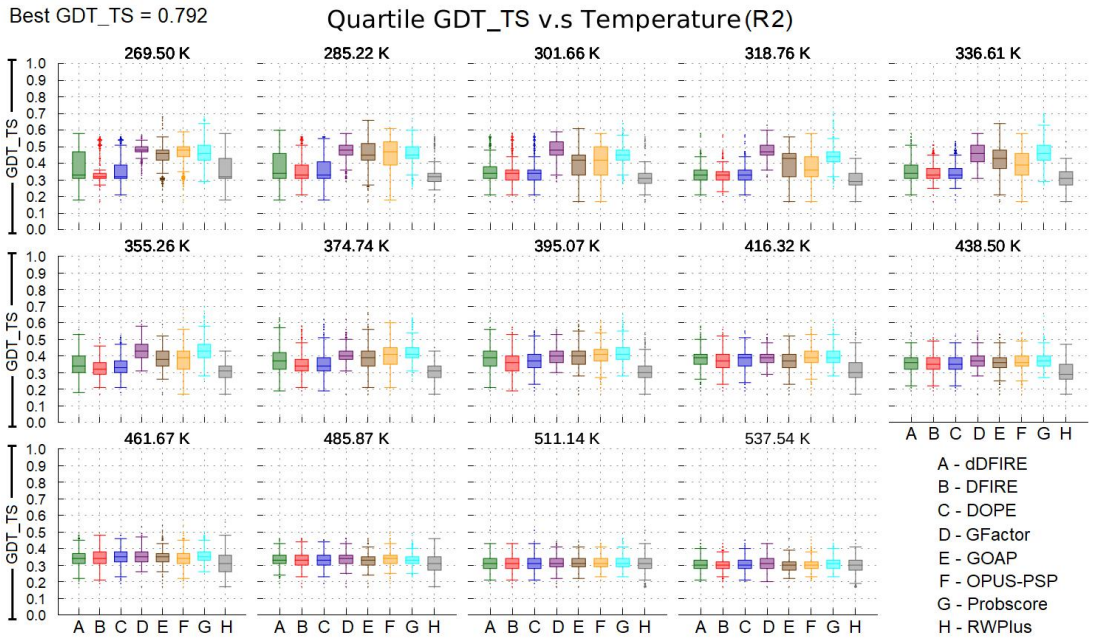

Figure 29 - Quartile distribution of GDT_TS (y axis) plot for the top 1% best scored structures according to each absolute quality metrics (x axis) and for each temperature of the simulation. Data relative to the 2nd REMD PSP simulation of the protein 1UNC (labeled R2).

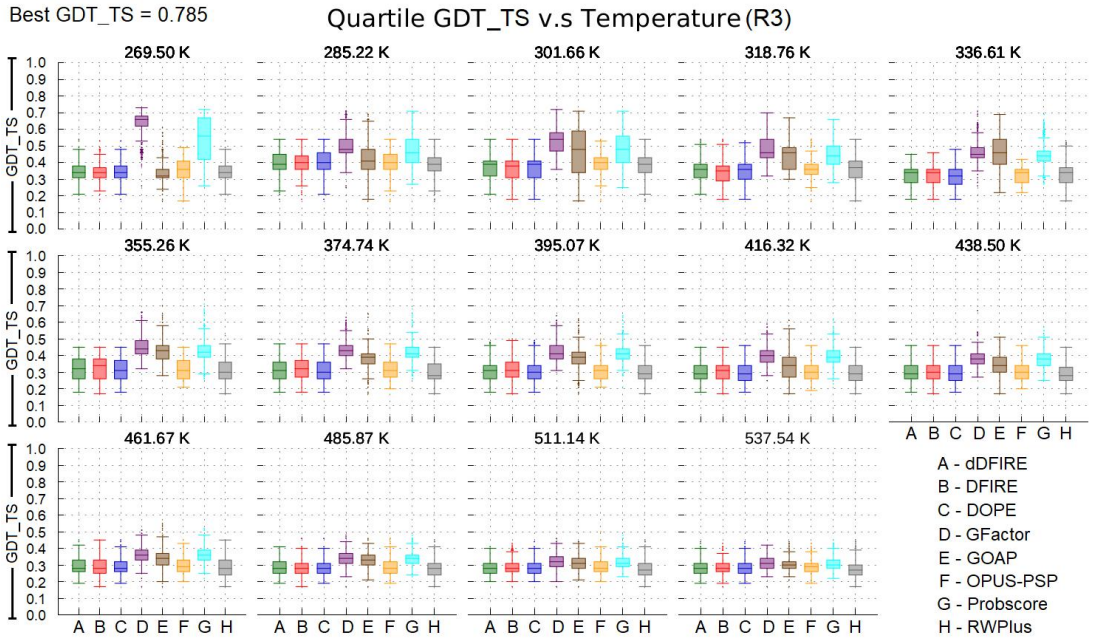

Figure 30 - Quartile distribution of GDT_TS (y axis) plot for the top 1% best scored structures according to each absolute quality metrics (x axis) and for each temperature of the simulation. Data relative to the 3rd REMD PSP simulation of the protein 1UNC (labeled R3).

6.4. Testing Additional Hypothesis Using the Protein 1UNC

This section will present the additional hypothesis formulated to test and question the clustering methods used in the literature per request of colleague Thiago Lipinski Paes. It is worth mentioning again that these hypothesis are only marginally related to the work here described, yet the results found are believed to be sufficiently valuable and worth reporting to future colleagues interested in further studying it.

6.4.1. First Additional Formulated Hypothesis

The first additional formulated hypothesis to be tested was:

Using the 5 most populated clusters resulted from clustering algorithms performed on protein structure prediction simulations is an acceptable strategy to reduce the amount of data that need to be analyzed posteriorly.

The strategy of using the most populated clusters, or the single most populated one, was adopted on a few studies found in the literature [Lip17, Kan11, Dau99, Chu13]. It is based on the assumption that the most populated cluster will contain the most diversity of structures, and therefore will have a higher chance of containing the best or a set of best predictions of the simulation. In a study performed in 2009 by researchers Lin and Shell [Lin09b], it was also found that spikes in population of the most populated cluster may be a reasonable signal of convergence in a REMD PSP simulation.

In order to verify the formulated hypothesis, the auxiliary test dataset was analyzed according to the RMSD values obtained from the centroid of the clusters, which represents the structures contained in them. This analysis can be seen in Table 11.

Table 11 - Analysis of Clusters' Centroids According to RMSD & Population

Simulation Protocol & Triplicate Number	Population & RMSD of the 5 Most Populated Clusters					Top Scored Cluster by RMSD & its Population
	1st	2nd	3rd	4th	5th	
A1 - Pop.	14,239	12,622	12,386	11,626	10,157	462
A1 - RMSD	8.00	8.57	5.47	7.62	4.52	1.58
A2 - Pop.	31,123	14,044	10,740	10,679	8,660	64
A2 - RMSD	7.16	7.47	7.16	6.06	6.64	3.23
A3 - Pop.	36,089	27,331	20,582	15,199	6,976	80
A3 - RMSD	4.17	6.66	6.21	6.50	6.44	3.08
B1 - Pop.	26,866	20,707	20,098	11,713	10,845	11,713
B1 - RMSD	6.17	6.75	3.57	2.77	3.58	2.77
B2 - Pop.	30,969	20,700	17,053	11,816	10,208	20,700
B2 - RMSD	5.95	2.77	6.79	6.13	6.02	2.77

B3 - Pop.	44,365	31,982	18,406	6,713	6,085	7
B3 - RMSD	7.35	6.5	7.07	7.20	6.54	4.16
C1 - Pop.	22,471	22,418	17,139	17,080	8,706	397
C1 - RMSD	7.86	7.09	7.11	6.24	6.95	2.03
C2 - Pop.	35,919	35,276	16,055	13,442	12,306	3,873
C2 - RMSD	7.40	4.24	7.28	5.82	6.00	2.40
C3 - Pop.	40,353	37,907	18,102	15,108	12,925	80
C3 - RMSD	4.06	6.22	6.26	7.20	6.07	3.12
D1 - Pop.	27,443	19,315	17,327	10,400	9,254	6,131
D1 - RMSD	2.43	6.07	5.49	4.62	5.89	2.07
D2 - Pop.	21,382	20,570	15,037	12,593	12,504	58
D2 - RMSD	7.84	6.38	5.12	6.48	8.22	2.22
D3 - Pop.	22,784	16,550	12,413	11,630	11,036	1,540
D3 - RMSD	3.70	7.22	6.40	4.30	3.73	3.00
E1 - Pop.	18,529	17,662	16,365	14,934	10,466	7
E1 - RMSD	8.75	6.54	9.06	6.15	6.18	3.09
E2 - Pop.	36,047	26,727	26,187	15,471	7,794	35
E2 - RMSD	6.81	6.79	7.27	8.23	8.45	4.96
E3 - Pop.	23,244	19,376	16,144	8,522	3,871	175
E3 - RMSD	6.30	6.34	5.88	6.63	7.18	4.45
F1 - Pop.	22,795	18,093	16,359	16,338	15,579	11
F1 - RMSD	7.24	7.16	7.73	7.48	6.98	5.67
F2 - Pop.	34,501	30,775	14,570	14,160	9,498	12
F2 - RMSD	6.94	8.89	5.70	7.18	6.9	3.90
F3 - Pop.	29,002	28,599	24,803	9,166	7,970	16
F3 - RMSD	8.32	8.70	3.43	7.31	8.11	2.82
G1 - Pop.	23,888	20,841	17,793	15,675	7,574	41
G1 - RMSD	3.40	7.24	7.60	4.97	6.71	3.08
G2 - Pop.	28,582	26,099	21,119	18,271	17,448	9
G2 - RMSD	6.97	7.04	6.53	7.53	6.00	2.72
G3 - Pop.	34,980	20,716	20,628	14,507	9,889	4
G3 - RMSD	3.45	5.77	6.81	4.51	7.11	2.97
H1 - Pop.	45,988	43,895	23,844	21,039	17,054	84
H1 - RMSD	7.52	7.63	4.51	7.21	7.40	3.51
H2 - Pop.	29,940	29,058	15,093	8,666	7,562	663
H2 - RMSD	5.33	6.42	5.39	7.60	7.38	4.94
H3 - Pop.	36,558	19,126	18,348	14,151	8,531	377
H3 - RMSD	7.52	7.18	7.46	7.66	8.79	4.56

It is possible to observe in Table 11 that rarely the most populated clusters were, indeed, the best scored ones in terms of RMSD. This patterns was also observed using the GDT_TS relative score. In most cases, the cluster with few structures were the best scored ones. It is worth noticing that they weren't, however, within the least populated clusters.

An hypothesis to describe this phenomenon is that, by observing the results obtained from the analysis of the third formulated hypothesis described in chapter 6.3.2, the best predicted structures were outliers to their

distribution. Thus attempting to cluster such structures with others is not a good strategy. As for the fact that they weren't within the least populated clusters, it is hypothesized that due to the rough energy landscape conformation of proteins and the folding funnel theory described in chapter 2.4, there is a strong tendency for the simulations to converge at local or global minimums. Thus even outlier structures, but close to the native structure of the target protein, may have similar ones to be clustered with.

Unfortunately only 1 protein could be tested in this form (the protein 1UNC), so it would be hasty to completely reject the formulated hypothesis. An extended study regarding this subject would, however, be of great value to the scientific community. Due to the limited time frame for this work, such study couldn't be done by the author.

6. 4. 2. Second Additional Formulated Hypothesis

The second additional formulated hypothesis to be tested was:

By using absolute quality metrics and the population of the clusters, a way to filter unsatisfactory clusters can be found.

To test this hypothesis, several different filtering configurations were tested. Overall, the efficiency of the quality metrics vary considerably between different protein structure prediction simulations. This can be seen in Table 12. The population of the cluster also varied significantly. As an example, for the first and second execution of the *B* protocol, which can be seen in Table 11, the most populated clusters indeed captured the best scored structures considering only the centroid of the clusters. For the rest of the simulations, the opposite happened where only low populated cluster were the top scored ones. Due to this variance and the time dedicated to test the additional hypothesis, a filter couldn't be found.

Table 12 - RMSD Scores (Rounded to Float-Point Precision of 2 Digits) According to Top Scored Structure of Each Absolute Quality Metric

Simulation ID	Real Best RMSD	RMSD Extracted from Top Scored Structure According to Each Absolute Quality Metric								
		dDFIRE	DFIRE	DOPE	GFactor	GOAP	OPUS-PSP	Probscore	RWPlus	Minimized Energy
A1	1,58	2.08	3.32	3.32	17.63	8.07	4.06	8.98	3.32	7.62
A2	3,23	6.78	6.78	6.78	12.98	4.14	5.43	7.89	6.78	3.30
A3	3,08	6.21	6.21	6.21	9.67	4.32	6.50	15.98	6.21	6.21
B1	2,77	2.77	2.77	2.77	18.35	2.77	7.30	8.18	2.77	6.17
B2	2,77	7.57	7.57	7.57	9.22	7.57	6.70	4.79	7.57	2.77
B3	4,16	6.72	6.72	6.72	7.25	6.67	5.76	14.41	6.72	6.5
C1	2,03	7.23	7.23	7.23	10.12	8.02	7.95	8.98	7.23	7.09
C2	2,40	6.76	5.81	5.81	10.49	7.00	5.82	6.17	5.81	7.40
C3	3,12	6.22	4.06	6.22	7.20	4.41	6.22	6.91	6.22	6.22
D1	2,07	5.98	6.78	6.78	11.09	3.65	5.98	6.12	2.07	5.90
D2	2,22	5.87	6.10	2.41	10.95	3.20	7.01	9.64	2.41	6.38
D3	3,00	3.89	3.47	3.47	11.26	7.20	3.73	11.06	3.47	4.16
E1	3,09	4.96	5.87	5.87	16.48	3.09	5.31	6.04	5.87	6.15
E2	4,96	6.81	6.65	6.79	14.41	7.07	5.81	5.59	6.17	6.92
E3	4,45	6.23	6.37	5.68	7.83	5.68	5.15	5.27	6.38	6.30
F1	5,67	7.75	7.16	7.16	7.50	7.51	7.77	6.74	7.16	7.51
F2	3,90	6.55	8.53	8.53	11.03	8.25	7.59	7.32	6.55	6.26
F3	2,82	7.81	7.34	7.34	10.38	4.88	7.81	9.57	7.34	3.43
G1	3,08	6.71	6.71	6.71	8.03	6.07	6.79	6.49	6.71	4.97
G2	2,72	3.86	3.86	3.86	6.67	3.86	3.86	7.52	3.86	3.42
G3	2,97	3.45	3.45	3.45	10.84	3.99	7.19	8.87	3.45	3.45
H1	3,51	6.90	4.32	3.82	5.42	3.67	4.23	6.90	3.82	4.51
H2	4,94	5.62	5.62	5.62	7.70	5.62	7.31	8.47	5.62	7.20
H3	4,56	4.56	4.82	4.82	6.21	6.53	6.13	6.56	6.13	7.18

6.5. Preliminary Filtering Configuration Proposed

Based on the results obtained in the case study of the protein 1UNC, a preliminary filtering configuration was proposed. Based on the assumption that the third simulation of the protein 1UNC represents a worst-case scenario and that the first 6 temperatures are needed to efficiently retrieve the best predictions without significantly impacting the final quality of the ensemble, this represents a cutoff of at least 40% of the original data volume (the greater the number of temperatures in the REMD PSP simulation, the greater this value becomes).

It is a fact that a single absolute quality metric isn't capable of retrieving the best predicted structure among an ensemble of predictions a 100% of the time. Otherwise a whole section of the CASP would be unnecessary as well as this, and other similar, studies. Utilizing more than just a single absolute quality metric is desirable to guarantee the quality of the structures. If, however, instead of applying all quality metrics on the entire resulting ensemble, a pipeline of different quality metrics is established where each extract the top 10% predicted structures from the previous resulting ensemble (where the first ensemble is the original predicted structures ensemble), the volume of data can be significantly reduced.

As an example, the computational cost of analyzing the entire predicted structures ensembles of a REMD PSP simulation with all chosen quality metrics would be equal to EQ. 15, where n is the size of the ensemble and x is the number of quality metrics chosen. The Big-Theta notation is commonly in computer science to determine the computational costs of algorithms. As algorithms usually have different execution times according to the input received, the Big-Theta notation asymptotically bound the growth of the running time of a given algorithm according to the function inside its parenthesis. In short, given EQ. 15, it means that the running time of the algorithm is (asymptotically) at least $k_1 \cdot n \cdot x$ and at most $k_2 \cdot n \cdot x$, for any constants k_1 and k_2.

$$\Theta(nx) \qquad \text{(EQ. 15)}$$

This cost can be reduced by eliminating the high temperatures as proposed. Considering a scenario where only the first 6 temperatures out of the original 10 are selected for running the predefined quality metrics, the analysis would then have a computational cost equal to EQ. 16.

$$\Theta(0.6nx) \qquad \text{(EQ. 16)}$$

Applying the proposed pipeline over this filtered ensemble would, therefore, result in the computational cost of EQ. 17.

$$\Theta\left(\sum_{i=1}^{x}\frac{0.6n}{10^{i-1}}\right) \quad \text{(EQ. 17)}$$

Such formula will result in a recurring decimal as more metrics are used, which can be seen in EQ. 18.

$$\Theta\left(0.\overline{6}n\right) \quad \text{(EQ. 18)}$$

It is possible to see a significant complexity reduction when comparing the initial cost in EQ. 15 to the final cost presented in EQ. 18. This proposed process thus represents a cutoff of at least 33% computational cost when compared to the standard approach.

An extensive search for such a filtering configuration was then performed. Unfortunately, the amount of possible filtering configuration is massive, as this filtering approach configures a permutation without repetition. The number of possible filtering configurations is then given by EQ. 19, where n denotes the number of possible absolute quality metrics that can be used in the filter and r denotes the final number of quality metrics that indeed will be incorporated in the filter.

$$\frac{n!}{(n-r)!} \quad \text{(EQ. 19)}$$

Using only the cited 8 absolute quality metrics plus the energy minimization value, this would result in 362,880 unique filtering configurations. A value unduly large to be tested. Several heuristics were then used to test the different filtering conformations. Unfortunately, none was able to reach the desirable quality. Whereas they failed to filter a substantial amount of structures, failed to uphold a minimum quality level or had inferior quality over applying a single quality metric to filter the structures. A larger test dataset was then needed to detect possible patterns in the quality metrics. The test dataset described in chapter 5.1 was the used to continue this study.

6. 6. Discussion & Results: REMD PSP Simulations Test Dataset

The first goal when analyzing the test dataset was to verify the pending hypothesis presented in chapter 6.3 that couldn't be concluded due to the lack of data.

6.6.1. *Verification of the Second Formulated Hypothesis*

In order to test the number of temperatures that must be analyzed posterior to a REMD PSP simulation without impacting the quality of the results, a cumulative distribution analysis of the GDT_TS bands was again performed on each protein REMD PSP simulation. Two analysis were performed: one for discovering in which temperature (first temperature, second temperature, and so on) 50% of convergence of the top GDT_TS bands is attained (Table 13), and the other attaining 80% of convergence (Table 14). The position of the temperature was used instead of the temperature itself because the temperatures employed in REMD PSP simulations vary according to each target protein. The tables are a summary of the results found in the cumulative distribution plots, which can be seen at Appendix A.

Table 13 - Temperatures in which 50% Convergence of the Top GDT_TS Bands are Attained.

	Temperature in Which GDT_TS Bands Attain 50% Convergence					
Protein	GDT_TS 0.5	GDT_TS 0.6	GDT_TS 0.7	GDT_TS 0.8	GDT_TS 0.9	GDT_TS 1.0
1UAO	6th	6th	1st	1st	1st	2nd
1LE1	8th	3rd	4th	8th	-	-
1L2Y	5th	6th	4th	3rd	4th	4th
1RIJ	5th	1st	4th	4th	4th	-
1FME	5th	9th	-	-	-	-
1PSV	4th	7th	-	-	-	-
1VII	4th	6th	-	-	-	-
1E0L	9th	-	-	-	-	-
2WXC	9th	-	-	-	-	-

Table 14 - Temperatures in which 80% Convergence of the Top GDT_TS Bands are Attained.

	Temperature in Which GDT_TS Bands Attain 80% Convergence					
Protein	GDT_TS 0.5	GDT_TS 0.6	GDT_TS 0.7	GDT_TS 0.8	GDT_TS 0.9	GDT_TS 1.0
1UAO	8th	7th	5th	4th	3rd	4th
1LE1	9th	7th	7th	9th	-	-
1L2Y	9th	8th	6th	5th	5th	4th
1RIJ	8th	5th	6th	5th	5th	-
1FME	8th	10th	-	-	-	-
1PSV	7th	10th	-	-	-	-
1VII	6th	7th	-	-	-	-
1E0L	10th	-	-	-	-	-
2WXC	10th	-	-	-	-	-

It was observed in this analysis that two REMD PSP simulations could successfully predict the native state of their target protein according to the GDT_TS relative quality metric. The proteins that were successfully predicted were the 1L2Y and 1UAO. Figure 31 and 32, respectively, depicts the cartoon drawings of the predicted structure versus their experimentally determined structure.

Figure 31 - Overlapping cartoon drawings of the main chain of the top predicted structure of protein 1L2Y (green) versus the experimentally determined structure (red) [Nei02]. Figure obtained using the PyMOL program [Sch17].

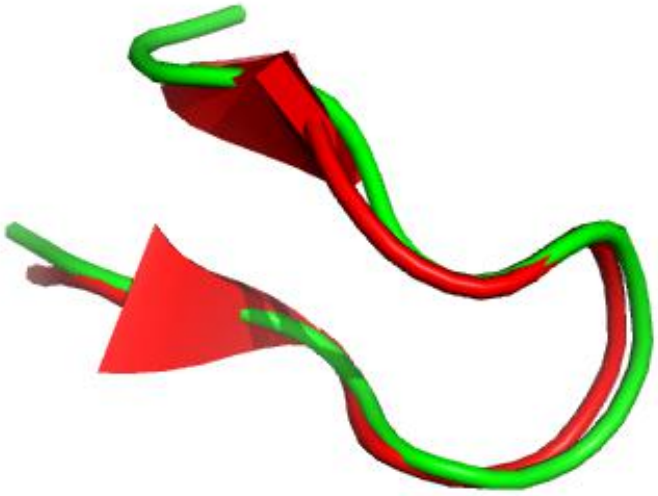

Figure 32 - Overlapping cartoon drawings of the main chain on the top predicted structure of protein 1UAO (green) versus the experimentally determined structure (red) [Hon04]. Figure obtained using the PyMOL program [Sch17].

It is worth noticing that the top predicted structure of the protein 1UAO couldn't successfully predict the β sheets of the native state structure. Both the RMSD and GDT_TS relative quality metrics, however, do not use the secondary structures in their scoring function, as described in section 2.11.1, but rather use the distance between the same atoms of the model structure and the predicted structure. Regarding atoms positioning, the predicted structure of protein 1UAO is (almost) the same as the experimentally determined structure.

Regarding the results presented in tables 13 and 14, it was observed that the better the quality of the REMD PSP simulation, the more rapidly the top GDT_TS bands converge in terms of temperature. It is possible to see that the highest GDT_TS band always converge at low temperatures, once again reinforcing the hypothesis that high temperatures can be safely discarded from post-simulation analysis.

It is hypothesized that this effect may be due low simulation time in which the proteins were simulated (only 50ns). When a simulation successfully achieve predicting the native state or a native-like state of the target protein, it will converge at low temperatures as high temperatures provide too much energy for the conformation to stabilize properly. These predicted structures will then remain stable on the low temperatures, while escaping to other local minimums on the energy landscape when exchanging to higher ones, thus reducing its GDT_TS score. This effect cause the low temperatures to contain the best GDT_TS scores and higher temperatures to contain less ideal structure conformations, but still with reasonable quality.

On the other hand, REMD PSP simulations that couldn't reach the highest GDT_TS scores are in the process of converging those high quality predictions, and thus will converge their best prediction only at higher temperature ranges (in which they are being created). Figure 33 depicts the cumulative distribution analysis created for the protein 1UAO, where the top 4 GDT_TS bands can be seen converging clearly faster than the others.

This hypothesis, however, result in both good and bad factors. Although the observation that the top predicted structures converge at low temperatures in excellent for filtering the unsatisfactory predicted structures from the ensemble generated at the end of a REMD PSP simulation, the observation that middle to low GDT_TS bands take longer to converge may imply in the volatility of this filtering technique. That is, without beforehand the resulted quality of the simulation, filtering more or less temperatures may have a significant impact on the quality or not. The obtained results, therefore, imply that the exact optimal number of temperatures which can be filtered depend on the final quality of REMD PSP simulation performed.

6.6.2. *Verification of the Third Formulated Hypothesis*

In order to test the overall efficiency of the cited absolute quality metrics, an histogram of the top GDT_TS bands was created for each protein REMD PSP simulation. The first observation made was that, while the 1UNC REMD PSP simulation could extract the top predicted structures using only the top 1% scored structures from the absolute quality metrics with reasonable

success, most of the others REMD PSP simulations couldn't achieve the same feat. After a series of test, it was observed that retrieving 20% of the top scored structured according to the absolute quality metrics retrieved an acceptable balance between quality and efficiency.

While most absolute quality metrics performed well in the REMD PSP simulations that produced the best results, some did not achieve the desired quality on the REMD PSP simulations that resulted in poorer predictions. Among those metrics is the Probscore, GFactor and OPUS-PSP. The minimized energy of the predicted structures also did not perform well in such cases, contrary to what was expected. Table 15, 16 and 17 shows the number of structures within the achieved GDT_TS bands of the 1PSV, 1VII and 1LE1 protein structure prediction simulations respectively, extracted using the top 20% scored structures by each absolute quality metric. It is worth noticing that while the 1PSV protein has an αβ conformation and 1LE1 has a β conformation (β sheets are more difficult to be predicted), the 1VII protein has an α conformation (generally easier to predict).

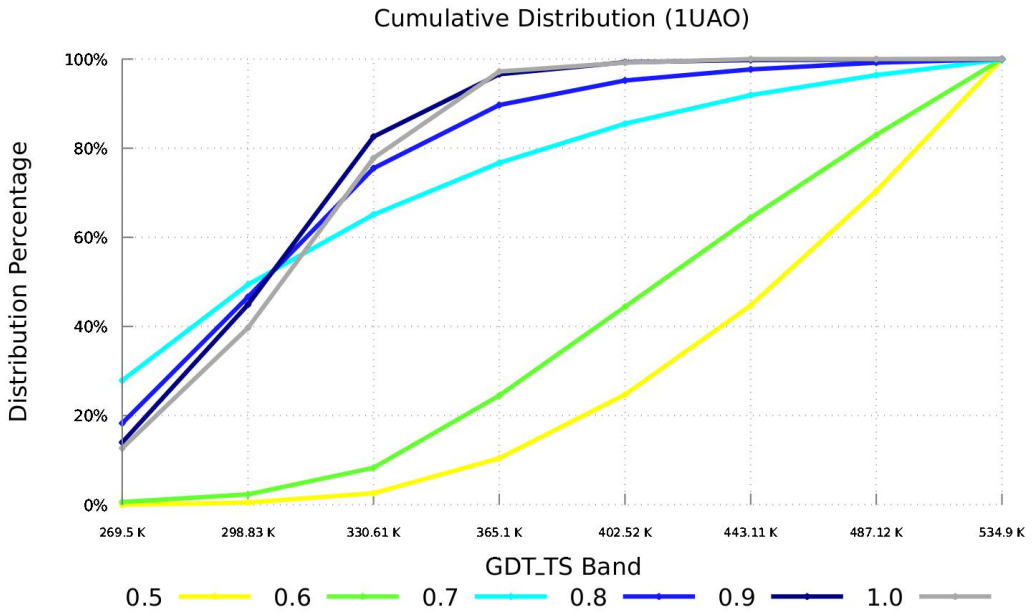

Figure 33 - Cumulative distribution plot of the top GDT_TS bands for the protein 1UAO REMD PSP simulation.

Table 15 - Number of Predicted Structures Within the Achieved GDT_TS Bands of the 1PSV PSP Simulation, Filtered Using the Top 20% Scored Structures According to Each Absolute Quality Metrics.

Quality Metric	GDT_TS Bands Extracted from the 20% Scored Structures						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
dDFIRE	0	0	167	90,418	49,334	81	0
DFIRE	0	0	852	102,702	36,363	83	0
DOPE	0	0	427	100,252	39,235	86	0
GFactor	0	104	11,341	97,473	31,069	13	0
GOAP	0	0	2,005	106,742	31,148	105	0
M. Energy	0	0	6,839	105,387	27,773	1	0

OPUS-PSP	0	0	4,356	106,446	29,139	59	0
Probscore	0	4093	74,984	54,939	5,971	13	0
RWPlus	0	0	305	100,083	39,530	82	0
Simulation Total	0	10,308	258,427	362,373	68,779	113	0

Table 16 - Number of Predicted Structures Within the Achieved GDT_TS Bands of the 1VII PSP Simulation, Filtered Using the Top 20% Scored Structures According to Each Absolute Quality Metrics.

Quality Metric	GDT_TS Bands Extracted from the 20% Scored Structures						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
dDFIRE	0	0	2,703	93,054	43,139	1,104	0
DFIRE	0	0	6,307	95,889	37,093	711	0
DOPE	0	0	5,480	94,644	38,935	941	0
GFactor	87	282	5,944	97,292	36,179	219	0
GOAP	0	12	10,633	88,831	39,209	1,315	0
M. Energy	0	0	3,947	99,960	36,020	73	0
OPUS-PSP	0	12	8,411	108,242	23,015	320	0
Probscore	2	26,358	79,798	28,000	5,638	204	0
RWPlus	0	0	5,051	97,033	37,162	754	0
Simulation Total	102	59,908	270,537	289,788	77,860	1,805	0

Table 17 - Number of Predicted Structures Within the Achieved GDT_TS Bands of the 1LE1 PSP Simulation, Filtered Using the Top 20% Scored Structures According to Each Absolute Quality Metrics.

Quality Metric	GDT_TS Bands Extracted from the 20% Scored Structures							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
dDFIRE	0	0	126	26,442	66,340	6,817	275	0
DFIRE	0	0	105	22,403	66,023	11,156	312	1
DOPE	0	0	106	23,435	66,445	9,773	240	1
GFactor	0	0	65	17,251	59,376	22,595	713	0
GOAP	0	0	123	24,319	70,066	5,451	41	0
M. Energy	0	0	76	23,105	50,436	25,945	438	0
OPUS-PSP	0	0	64	20,857	62,237	16,456	386	0
Probscore	0	0	91	18,359	68,281	12,931	335	3
RWPlus	0	0	107	24,545	65,605	9,421	321	1
Simulation Total	0	0	352	81,785	344,467	71,134	2,249	13

Although the Probscore absolute quality metric performed well in the 1LE1 simulation, it did not achieve the same quality results on virtually all others PSP simulations. As discussed in chapter 6.3.1, this metric tends to be volatile. The GFactor, OPUS-PSP and the minimized energy also show the same volatility, which can be seen in the presented tables, where GFactor retained low GDT_TS bands on the 1PSV and 1VII PSP simulation after

filtering, the OPUS-PSP retained mostly of the middle bands (0.4 or 0.5) instead of retaining the higher ones, and finally the minimized energy was unable to retain the higher GDT_TS bands on the all of them. This lack of reliability is not something desired on a filtering method. Based on these results, the 3 absolute quality metrics and the energy minimization were then excluded for posterior analysis.

As the energy minimization achieved undesired results running with 1000 cycles of energy minimization, as described in section 2.11.2.9, the proposal of testing the quality of the results with less minimization cycles (also described in that section) became irrelevant. Due to the larger size of the protein 2WXC (thus implying in a much higher computational cost) and the observed inefficiency of the energy minimization to assess the quality of predicted structures, this scoring method was not tested on that protein.

It is still worth noticing that the efficiency (in term of quality) of the others absolute quality metrics also varied significantly between different proteins, where the GOAP absolute quality metric presented the very best results in proteins 1PSV and 1VII, and yet the very worst results in protein 1LE1. Moreover, not only between REMD PSP simulations of different proteins, but also different REMD PSP simulations of the same protein presented a considerably variance between the quality of the absolute quality metrics. This can be seen in Table 12 presented in section 6.4.2. Although varying considerably, the remaining absolute quality metrics, except for a few isolate cases such as the GOAP case of protein 1LE1, showed a reasonable stable quality in filtering the unsatisfactory predictions. This can be seen in Table 18.

Table 18 - Top RMSD Score From the Entire REMD PSP Simulation of Each Protein Versus the RMSD Score Extracted by the Top Scored Structure According to Each Absolute Quality Metric

Absolute Quality Metric	RMSD Scores (Rounded to Float-Point Precision of 2 Digits)								
	1E0L	1FME	1L2Y	1LE1	1PSV	1RIJ	1UAO	1VII	2WXC
dDFIRE	6.40	5.29	1.00	6.95	6.20	4.16	1.34	6.83	9.07
DFIRE	6.57	5.23	1.22	5.58	6.21	3.56	1.93	6.86	9.31
DOPE	6.67	5.46	1.34	5.86	6.19	2.32	1.93	6.39	9.45
GFactor	8.53	6.01	4.55	4.72	6.42	3.08	3.50	4.52	12.92
GOAP	6.68	5.47	3.29	5.98	6.29	4.22	1.26	6.86	8.84
M. Energy	6.42	5.89	5.58	5.52	6.34	3.69	2.79	6.26	-
OPUS-PSP	6.38	5.98	3.12	4.44	7.15	5.36	3.77	4.80	9.47
Probscore	6.75	8.03	5.25	7.37	7.93	6.81	5.17	7.27	8.25
RWPlus	6.52	5.51	0.97	5.75	6.21	3.14	1.93	6.86	9.31
Real Best RMSD	4.78	2.85	0.34	1.40	3.09	0.82	0.36	2.55	3.96

It is important to observe in Table 18 that the efficiency of the absolute quality metrics vary considerably between different REMD PSP simulations and a “generally best” absolute quality metric cannot be pinpointed, nor any discernable pattern could be found regarding a possible relation between the

quality metrics' efficiency and the protein conformation (α, β or αβ), size, amino acid composition, etc.

Moreover, the percentage of top scored structures (20%) can also be decreased substantially if the simulation produced good results (close to the native state of the target protein) and could need a slight increment if the simulation performed poorly. It was observed that in the 1L2Y and 1UAO PSP simulations (which produced the best results), the same quality in its filtering could be achieved, that is, successfully maintaining a high proportion of the higher GDT_TS bands while effectively filtering the lower GDT_TS bands, by using only the top 10% or fewer of the predicted structures according to the absolute quality metrics. This effect can be seen in Figures 34 and 35.

The 1E0L PSP simulation, which produced the worst results on the other hand, could achieve significantly better results, in terms of the distribution of GDT_TS bands, if 25% of the top scored structures are used for the filtering. This effect can be seen in Figures 36 and 37.

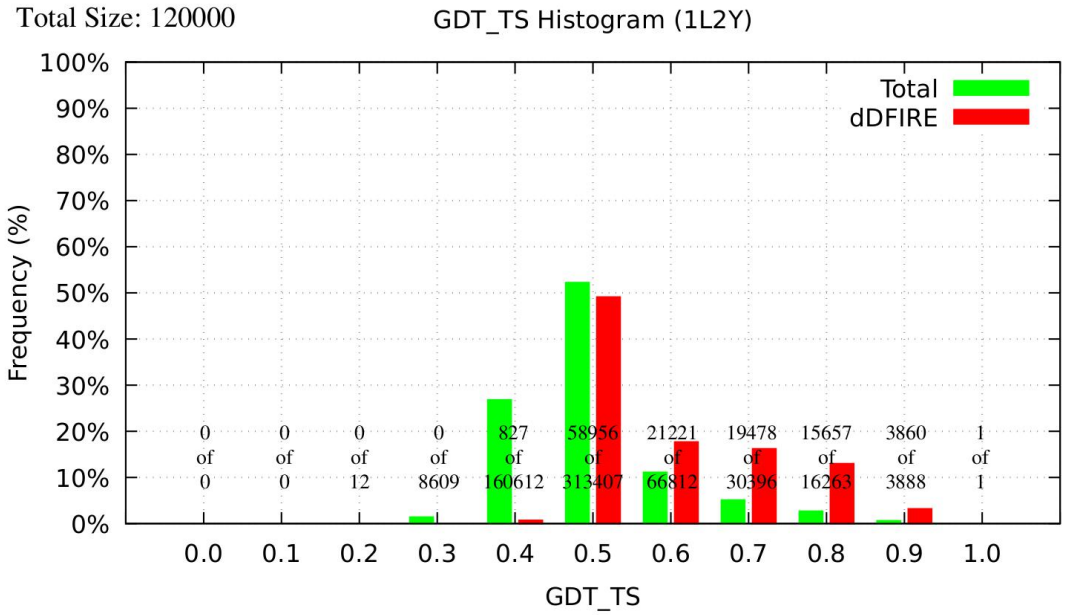

Figure 34 - GDT_TS histograms of the REMD PSP simulation of protein 1L2Y. The red bars are the GDT_TS histogram attained using the top 20% scored structures according to the dDFIRE absolute quality metric. The green bars are the overall GDT_TS histogram of the entire REMD simulation.

6.7. Final Filtering Configuration Proposed

Based on the results obtained from the case study of the protein 1UNC and the test dataset containing 9 different proteins, a series of filters were tested. Unfortunately, not all conformation could be tested, as described in chapter 6.5. Based on the variance observed in the absolute quality metrics, a priority was given to filters which could maintain a certain level of quality between different REMD PSP simulations.

The best filtering configuration found used the following steps:

1. Extract the structures predicted at the first 6 temperatures from the ensemble of predicted structures resulted from the target REMD PSP simulation.

2. Extract the dDFIRE, DFIRE, DOPE, GOAP and RWPlus absolute quality metrics from the resulting ensemble of step 1.

3. Apply the Dynamic Threshold script (described in section 6.2) present in the SnapFi tool to each of these quality metrics. The threshold base was defined as the top scored structure (position 0), the threshold margin is defined as negative 20% (-0.20) and the comparison type used was set as less or equal (LE). In summary, this will generate a threshold value relative to the top scored structure with a discount of 20%. All the structures that are above this value are filtered.

4. Apply the Voting script (described in section 6.2) present in the SnapFi tool to all the the results of step 3. The number of votes is defined as 2. In summary, this step will ensure that each structure in the filtered ensembles generated by step 3 must be considered as “satisfactory” (i.e., not filtered) by at least 2 different absolute quality metrics from the possible total of 5 absolute quality metrics.

This filtering method can be easily performed using the filtering configuration input file presented in Figure 38. The results obtained using the proposed filter, for each protein in the test dataset of REMD PSP simulations, can be observed in Figures 39 to 47.

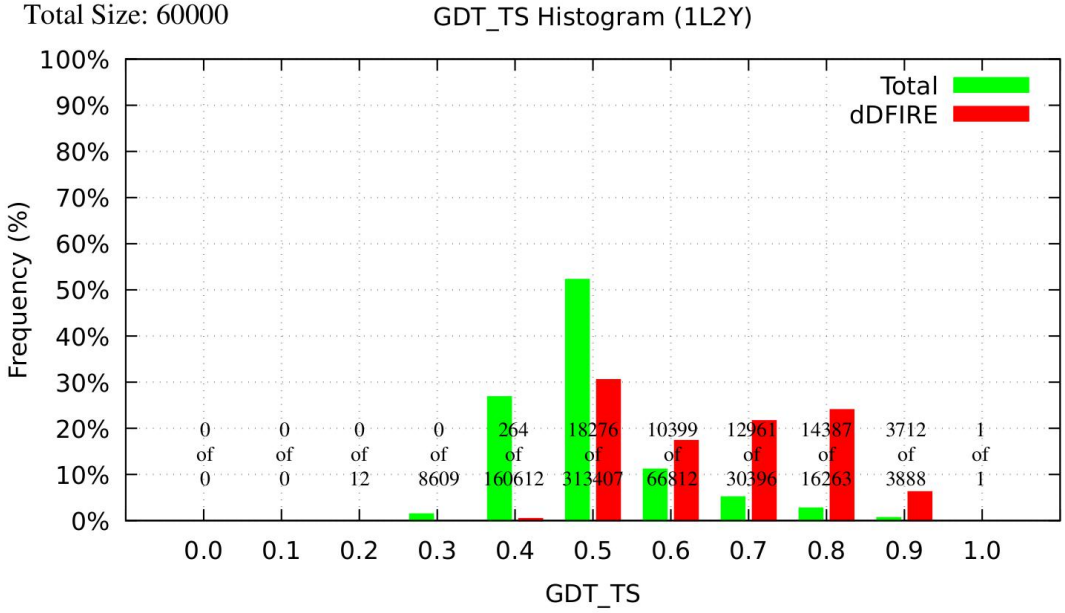

Figure 35 - GDT_TS histograms of the REMD PSP simulation of protein 1L2Y. The red bars are the GDT_TS histogram attained using the top 10% scored structures according to the dDFIRE absolute quality metric. The green bars are the overall GDT_TS histogram of the entire REMD simulation.

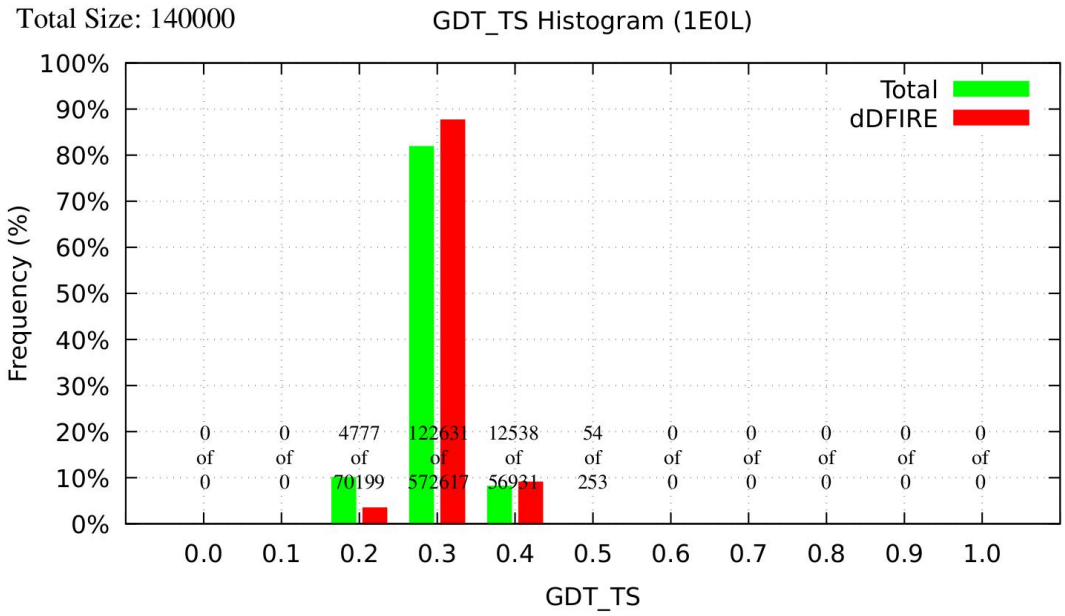

Figure 36 - GDT_TS histograms of the REMD PSP simulation of protein 1E0L. The red bars are the GDT_TS histogram attained using the top 20% scored structures according to the dDFIRE absolute quality metric. The green bars are the overall GDT_TS histogram of the entire REMD simulation.

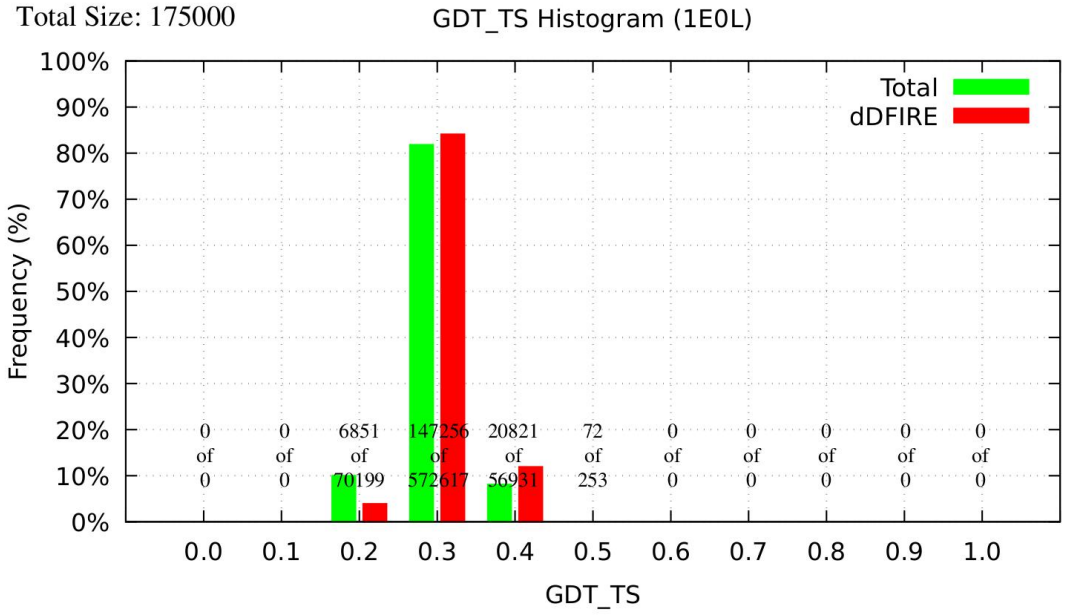

Figure 37 - GDT_TS histograms of the REMD PSP simulation of protein 1E0L. The red bars are the GDT_TS histogram attained using the top 25% scored structures according to the dDFIRE absolute quality metric. The green bars are the overall GDT_TS histogram of the entire REMD simulation.

```
#Load the PDB

STEP DATA = Modules/Load_PDB.py(Ensemble.pdb)

#Extract Quality Metrics

STEP dDFIRE = Metrics/dDFIRE/dDFIRE.py(DATA)

STEP DFIRE = Metrics/dDFIRE/DFIRE.py(DATA)

STEP DOPE = Metrics/DOPE/DOPE.py(DATA)

STEP GOAP = Metrics/GOAP/GOAP.py(DATA)

STEP RWPlus = Metrics/RW_Plus/RW_Plus.py(DATA)

#Apply the Dynamic Threshold Script

STEP F_dDFIRE = Modules/Threshold_Dynamic.py(0, -0.20, dDFIRE, LE, REVERSED=FALSE)

STEP F_DFIRE = Modules/Threshold_Dynamic.py(0, -0.20, DFIRE, LE, REVERSED=FALSE)

STEP F_DOPE = Modules/Threshold_Dynamic.py(0, -0.20, DOPE, LE, REVERSED=FALSE)

STEP F_GOAP = Modules/Threshold_Dynamic.py(0, -0.20, GOAP, LE, REVERSED=FALSE)

STEP F_RWPlus = Modules/Threshold_Dynamic.py(0, -0.20, RWPlus, LE, REVERSED=FALSE)

#Apply the Voting Script

STEP FILTER_TEMP = Modules/Voting.py(2, F_dDFIRE, F_DFIRE, F_DOPE, F_GOAP, F_RWPlus)
```

Figure 38 - Filtering configuration input file for the SnapFi tool which execute the filtering method described in section 6.7. The exclusion of the higher temperatures (step 1) can be easily done by either the cpptraj module of AMBER 14 or by ignoring the respective temperature ensembles when using the Load_PDB.py script.

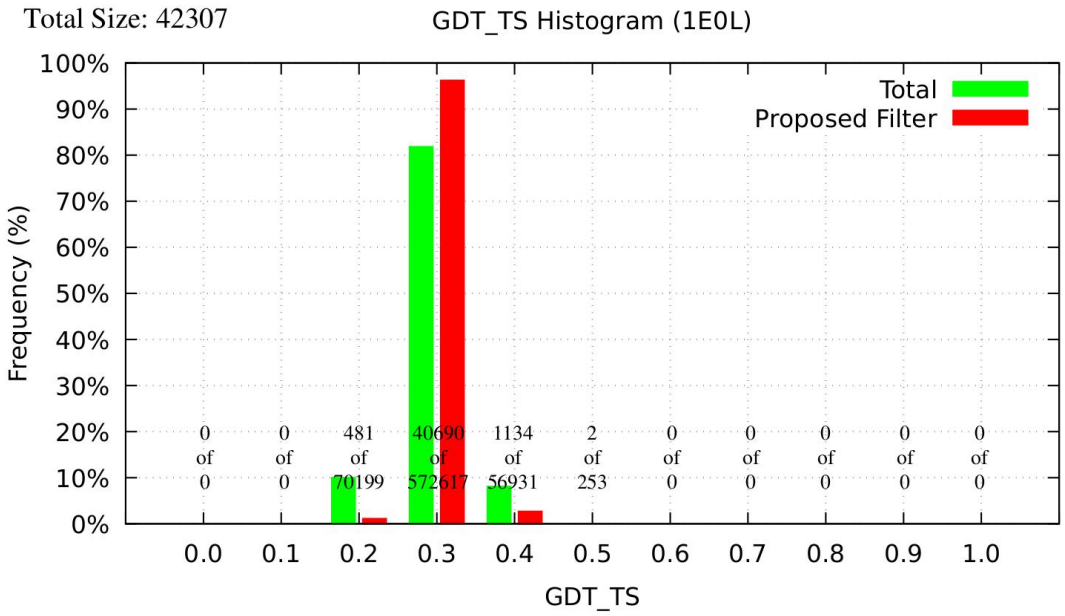

Figure 39 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1E0L (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

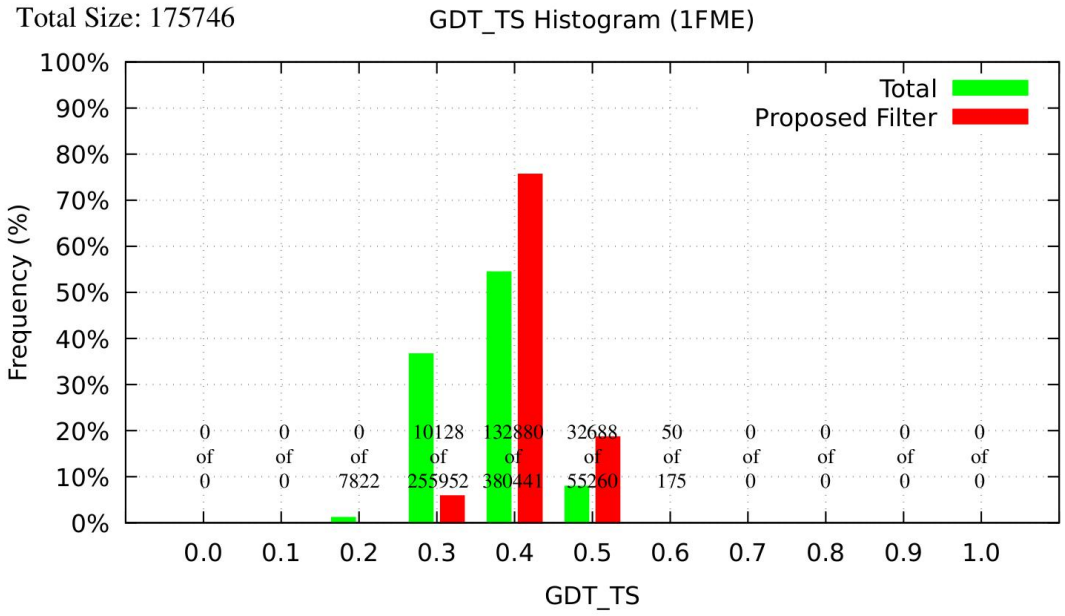

Figure 40 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1FME (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

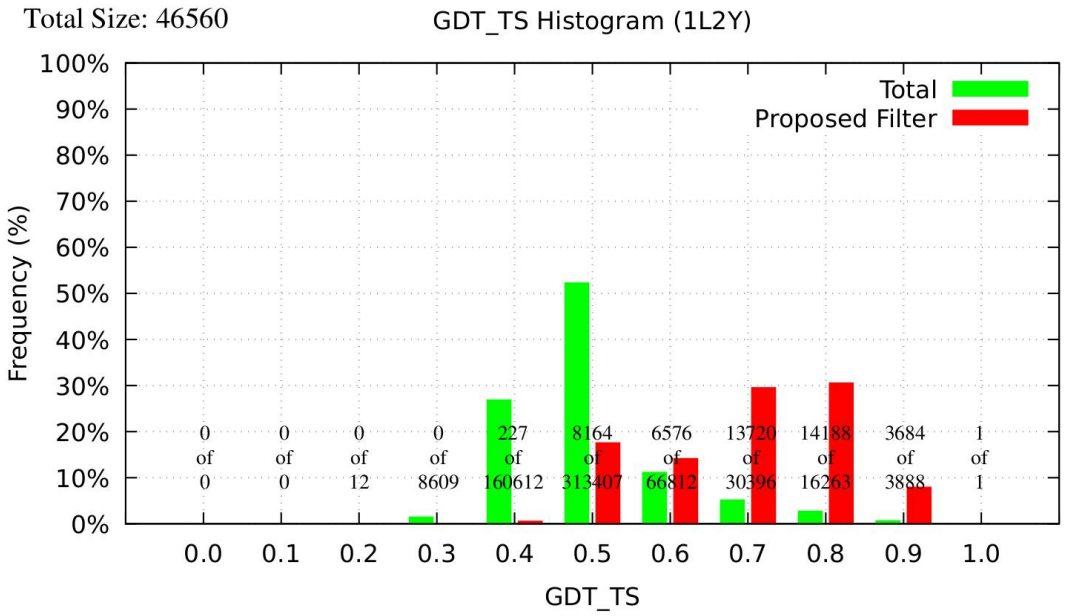

Figure 41 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1L2Y (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

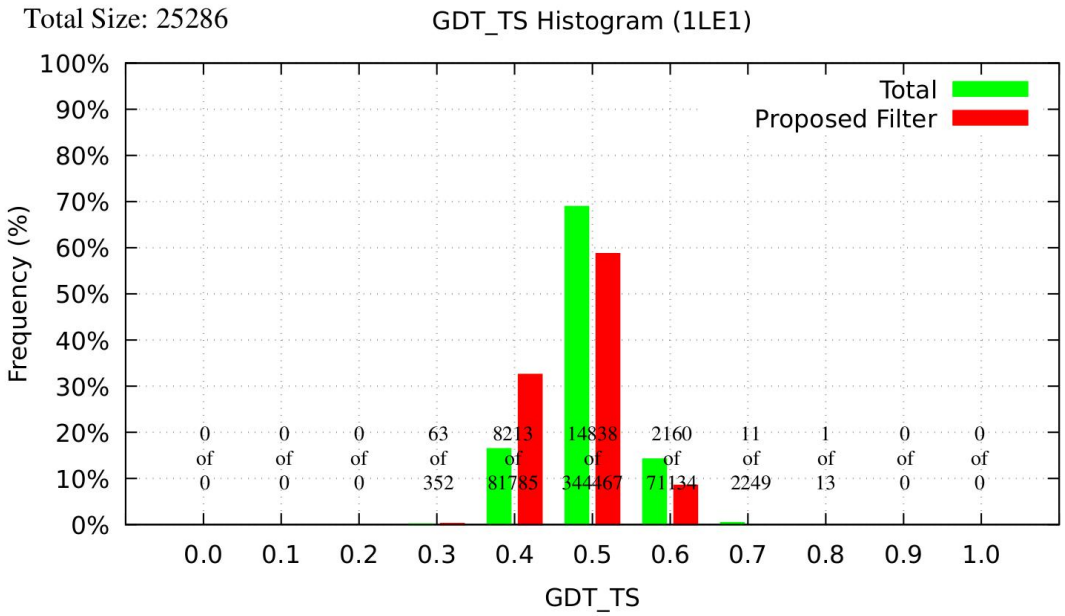

Figure 42 - TGDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1LE1 (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

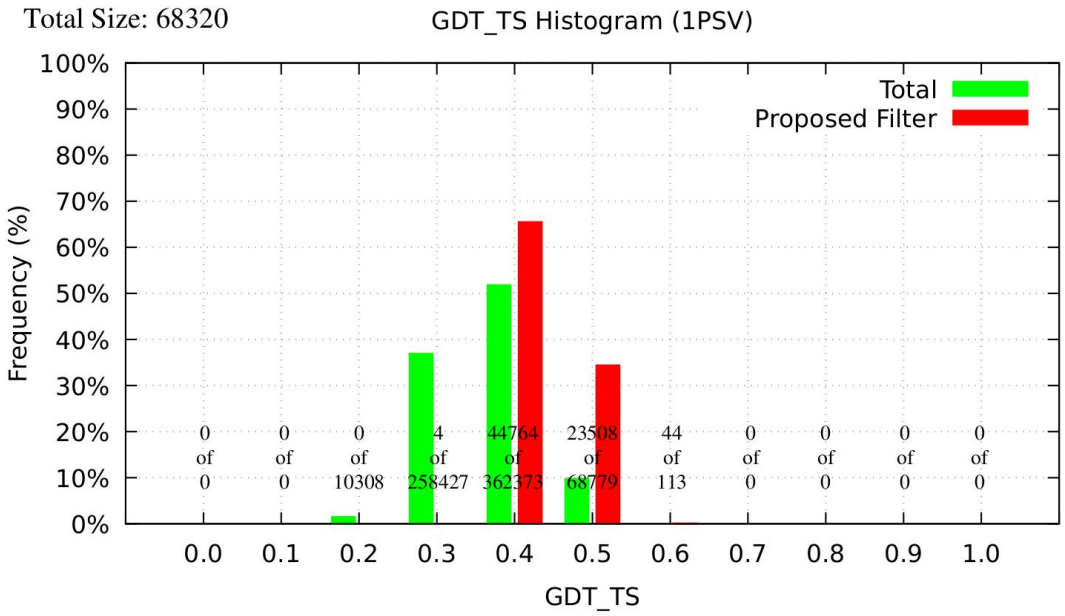

Figure 43 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1PSV (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

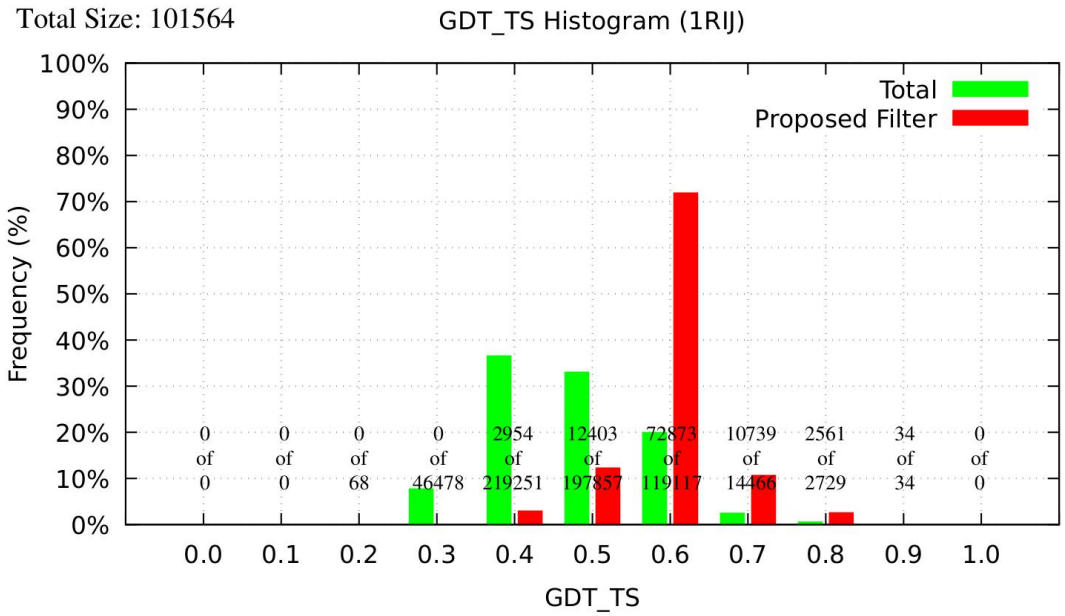

Figure 44 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1RIJ (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

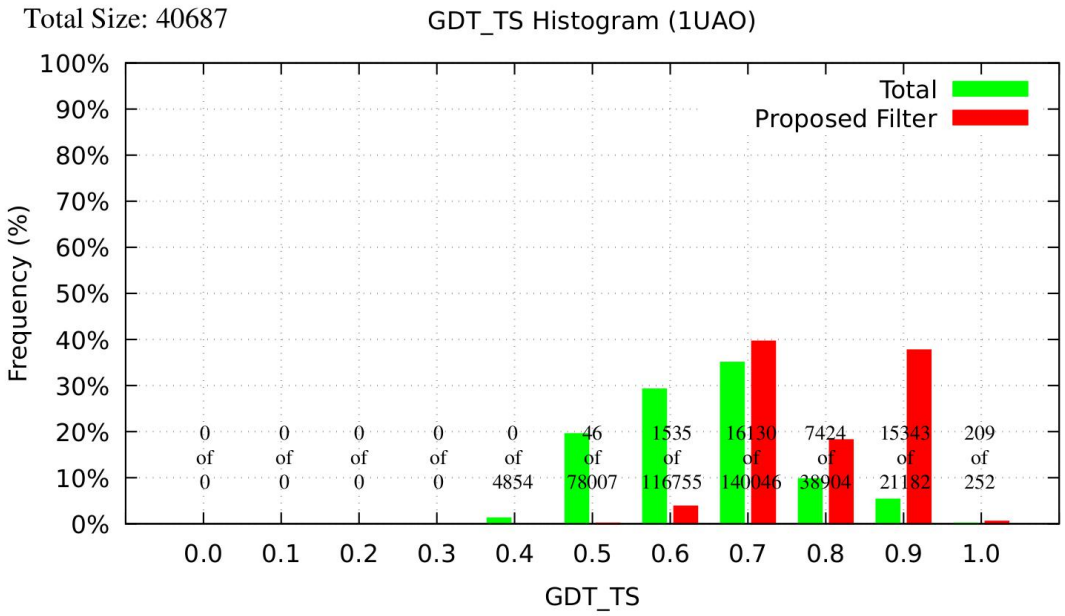

Figure 45 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1UAO (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

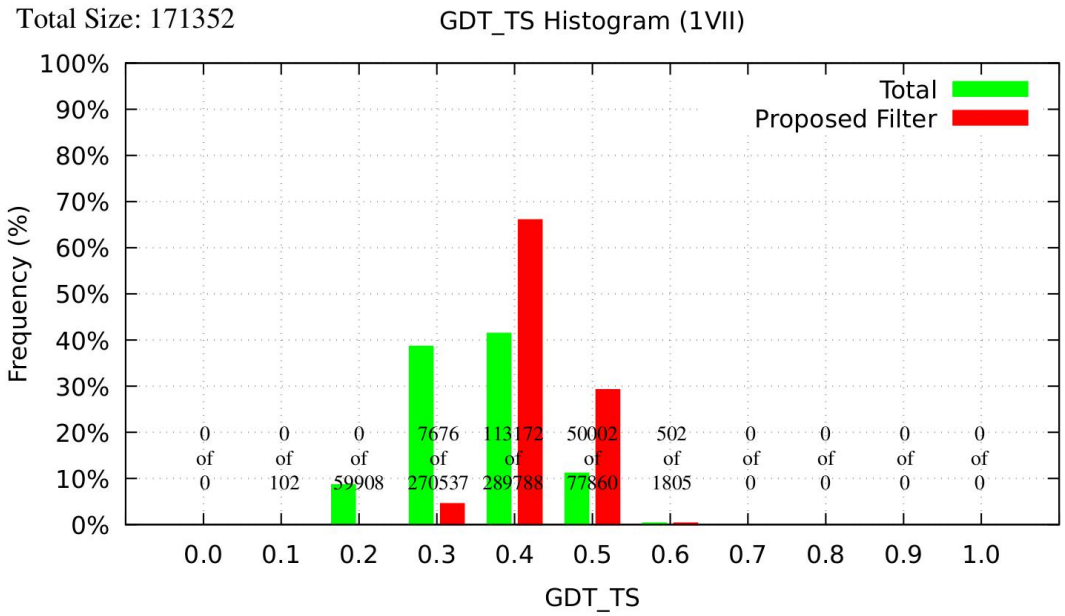

Figure 46 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 1VII (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

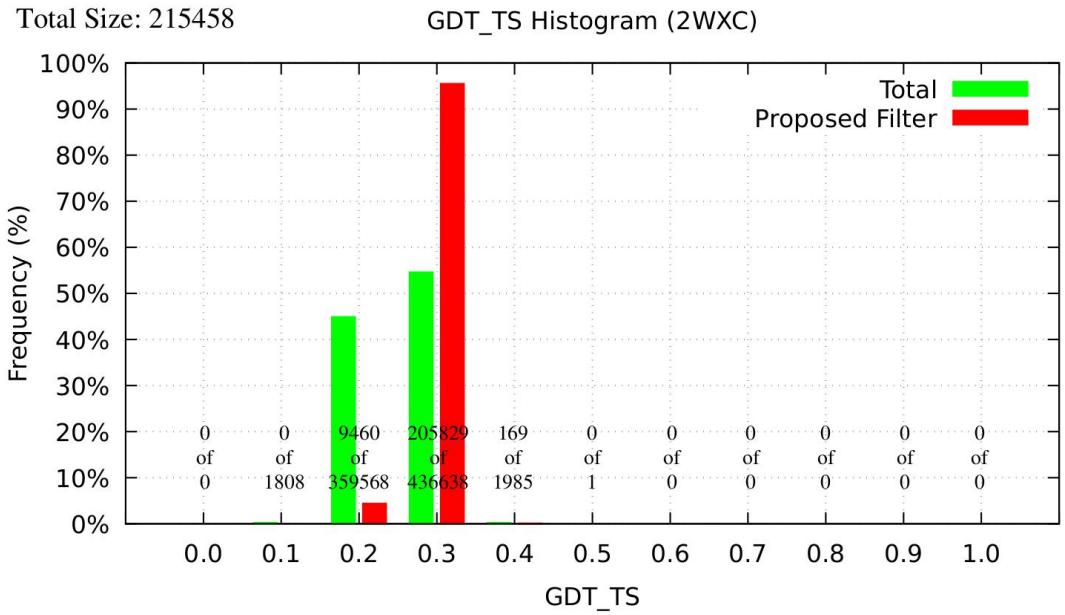

Figure 47 - GDT_TS histograms of all predicted structures generated by the REMD PSP simulation of the protein 2WXC (green bars) and the remaining structures after the proposed filtering method is applied (red bars).

Unfortunately, the proposed filter was unable to achieve positive results on all REMD PSP simulations. Namely, the PSP simulation of the proteins 1LE1 and 1E0L achieves a inferior distribution of the GDT_TS compared to the original ensemble of predicted structures. Moreover, the number of structures filtered also varied significantly, which can be seen in Table 19.

Table 19 - Number of Structures Contained in the Initial and Filtered Ensembles of Predicted Structures for Each Simulation

Protein ID	Initial Ensemble Size	Filtered Ensemble Size	Percentage of Structures Filtered
1E0L	700,000	42,307	~94%
1FME	699,650	175,746	~75%
1L2Y	600,000	46,560	~92%
1LE1	500,000	25,286	~95%
1PSV	700,000	63,320	~91%
1RIJ	600,000	101,564	~83%
1UAO	400,000	40,687	~90%
1VII	700,000	171,352	~76%
2WXC	800,000	215,458	~73%

Nonetheless, the proposed filtering method successfully filtered more than two thirds the initial volume of predicted structures on all simulations with significant quality gains on most cases. It is also important to notice that the filtering methodology can be adjusted to achieve a better quality distribution at the expense of reducing its filtering capability by increasing the threshold margin applied to the dynamic threshold script or increasing the

amount of temperatures used. As previously cited, the modular architecture of the SnapFi tool is capable of handling adjustments in its filtering configurations in an easy manner.

Unfortunately, there is no clear indicator whereas the proposed filter must be further adjusted or not by the user when using it on a new simulation. The proposed filtering methodology was found to be a good middle term between safe filtering, i.e. not filtering out the best predictions, and total volume of structures filtered. If the user desires so, the threshold values can be increased, or the number of temperatures limited even further, to filter out more structures at the expense of risking filtering good predictions. As already cited, if the simulation retrieved several structures close to the native structure of the target protein, this increase shouldn't significantly impact the final quality of the results. It is worth noticing that the largest computational cost of the proposed filtering methodology resides in calculating the quality metrics (step 2 of the filter) and once this process is done, the threshold values can be adjusted and rerun in a short time.

Regarding the efficiency of the proposed filtering method, the use of the 5 different absolute quality metrics on the entire ensemble generated at the end of a REMD PSP simulation diminish its capability of reducing the computational cost of analyzing such simulations. Unfortunately, no other filtering configuration tested that used less absolute quality metrics could be found. However, the proposed filtering configuration, although not as efficient as desired, is capable of generating more reliable results than just applying a single absolute quality metrics to filter unsatisfactory structures. This is due to the unreliable efficiency, in term of quality, of the absolute quality metrics explained in section 6.6.2.

7. CONCLUSIONS

In this chapter the conclusions extracted from the performed study are described. Firstly, the conclusions regarding the developed tool are presented, followed by the conclusions extracted from the tests performed to clarify the formulated hypothesis of this work. The conclusions of the proposed filtering methodology are then described along with its limitations and advantages. Finally, the final considerations are presented.

7.1. SnapFi Tool

In this study a new optimization area still untouched at large by the scientific community was proposed, that is the Analytical Data Filtering optimization. This approach rely on filtering the resulted predicted protein structure at the end of a MD simulation (with heavy emphasis in the REMD method). Due to that, it can be coupled with other optimization approaches to further optimize such simulations. As no other such method was found in the literature that aims to reduce the amount of data posterior to a REMD PSP simulation, a comparison between results of the different methods, which was a specific objective of this study, became unnecessary.

The SnapFi tool was then presented, which aims to cover this proposed optimization area. The tool, although rather simple, is capable of executing its designed objective remarkably well. Moreover, its modular architecture enables the SnapFi to be easily upgraded with new features and support for novel quality metrics. Despite being designed to filter the ensemble of predictions generated at the end of a REMD PSP simulation, it can also be used on conventional MD methods and variants, given it also generates such .pdb ensembles files.

7.2. Formulated Hypothesis

The greatest find in this study lies in the confirmation of the hypothesis that high temperatures of REMD PSP simulation can be safely discarded from posterior analysis as all the performed tests confirmed this. The exact number of temperatures that can be discarded may vary from simulation to simulation however. Simulations that produced good results (i.e. many predictions close to the native state of the protein) can be filtered more and simulations that produced inferior results may take advantage of more temperatures (i.e., a weaker filter). The data obtained from the performed tests point that in worst case scenarios, only the lower 10 temperatures can be used for further analysis without significantly impacting the quality of the results. It is believed that in REMD PSP simulations that must use more temperatures, the simulation itself didn't achieved a minimum quality level that justifies the effort to further analyze the results, that is, the highest GDT_TS values found were lower than 0.5. In other words, the predicted structures are so far from the

native state of the target protein that further analysis might be deemed unnecessary.

It was also found that the cited absolute quality metrics show a heavy variance between different REMD PSP simulation, even if the simulation targets the same protein. This variance made the discovery of an universal filter method (i.e. without any form of adjustment) impossible, as well as the formulation of a novel quality metric based on other quality metrics, which was one of the specific objective of this study. This finding, nonetheless, is an important remark, proving that using just a single absolute quality metric as a mean of retrieving the best predicted structures is a flawed process prone to errors.

It was also discovered that using the top populated clusters, presented in the additional hypothesis of this study, is also a flawed process and thus should be avoided. A deeper study may be required in this part, but observed results point that the top predicted structures of a REMD PSP simulation are generally outliers to the distribution and thus often do not form big clusters.

Regarding the quality of the absolute quality metrics, contrary to what was expected, the minimized energy proved to be an inefficient way of filtering predicted structures. This goes against the Anfinsen's thermodynamic hypothesis [Anf73], considering that predicted structures with higher minimized energy were closer to the native state of the target protein than predicted structures with lower minimized energy. It is hypothesized two reasons for this finding: (i) the way relative quality metrics are calculated, based solely on atom positioning, does not correctly represent the energy landscape of the protein folding, thus a structure with less potential energy may generate a worst score than a structure with their atoms closer to the native state of the protein, and (ii) the force field used in this work (the ff12SB) produce incorrect results. As force fields are being updated on a regular basis to correct flaws on their evaluations, the second hypothesis is more likely the reason behind this.

7.3. Proposed Filtering Methodology

The filtering method presented in chapter 6.7 was able to significantly filter the amount of structures in the ensemble of predicted structures of a REMD PSP simulation. The best results managed to bring the total number of structures predicted down to less than 10% of its original number, while also increasing the proportion of structures with high GDT_TS. Unfortunately, the proposed filter still carries on the variance of the absolute quality metrics which it is composed of. On a few occasions it produces significantly inferior results, in terms of quality, than the total ensemble of predicted structures.

Notably, the filter achieves the best results when applied to a simulation that could successfully predict the native state or native-like state of the target protein. In such occasions, the amount of structures that can be filter is also vastly superior. The proposed filter, therefore, vary its efficiency (in terms of computational time) and quality (in terms of filtering out unsatisfactory predictions) depending on the quality of the simulation.

Nonetheless, it is believed that the proposed filter managed to maintain a reasonable distribution of the top GDT_TS bands even if producing an inferior one compared to the total ensemble. The reduced effort of analyzing the filtered ensemble may, in some cases, compensate its loss of quality. Based on the study being performed, small adjustments in the proposed filter can be made to better fit the balance between efficiency and quality.

In summary, the limitations of the proposed filtering methodology are as following:

1. The efficiency of the proposed filter vary its efficiency according to the simulation it is applied;

2. Due to needing to compute several absolute quality metrics, the proposed filter might still have a high computational cost;

3. There is no clear indicator whereas the proposed filter must be further adjusted or not by the user when using it on a new simulation.

Whereas the advantages of the proposed filter are as following:

1. As the high temperatures are always excluded, the larger the simulation is (regarding number of replicas), the more predicted structures are able to be filtered out at the very first step;

2. When compared to using a single absolute quality metric, the proposed filter, due to using several of them, does not vary its efficiency so much;

3. The proposed filter is able to filter out more than two thirds the initial volume of predicted structures, even when producing a worse proportion of high GDT_TS bands compared to the initial ensemble. This reduction on the volume of data alone might be enough to justify its loss of overall quality.

7.4. Final Considerations

Due to the limited time frame for this study, some valuable tests couldn't be performed. The impact of the different thermostats in REMD PSP simulations regarding the amount of temperatures that can be filtered is an important example. Such further studies could improve our understanding of the impact of the temperatures in a REMD PSP simulation or even limit the results found on this study to simulations that use specific thermostats.

Moreover, it is also believed that the proposed tool is capable of working together with several other REMD optimization techniques and also with conventional MD methods, given that it is properly configurated first. Due to the amount of optimization approaches available in the literature this could not be tested in full.

Overall, even though there are some limitations to the presented method, the amount of unsatisfactory structures filtered and the final quality of the

filtered ensemble is very encouraging. It is important to notice that novel filtering methodologies can be formulated with ease using the SnapFi tool, some of which might even solve some of the limitations found. Finally, the conclusions extracted from the formulated hypothesis is also of great significance to the scientific community and might be the target of new correlated studies.

REFERENCES

[Aba94] Abagyan, R.; Totrov, M. “Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins”. *Journal of Molecular Biology*, 235-3, Jan 1994, pp. 983–1002.

[Aff06] Affentranger, R.; Tavernelli, I.; Iorio, E. E. D. “A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling”. *Journal of Chemical Theory and Computation*, vol. 2-2, Jan 2006, pp. 217-228.

[Ald59] Alder, B. J.; Wainwright, T. E. "Studies in Molecular Dynamics. I. General Method". *The Journal of Chemical Physics*, vol. 31-2, Aug 1959, pp. 459.

[Ale17] Alexandrov, O. “An illustration of the gradient descent method”. Captured on: https://commons.wikimedia.org/wiki/File:Gradient_descent.png, May 2017.

[Ale95] Alexandrov, N. N.; Nussinov, R.; Zimmer, R. M. “Fast protein fold recognition via sequence to structure alignment and contact capacity potentials”. In: Pacific Symposium on Biocomputing, 1995, pp. 53–72.

[Alt90] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J.; “Basic local alignment search tool”. *Journal of Molecular Biology*, vol 215-3, Oct 1990, pp. 403-410.

[Amb17] Amber MD. “Amber Home Page”. Captured on: http://ambermd.org/, May 2017.

[Ame17] American Chemical Society. “ACS Publications Home Page”. Captured on: http://pubs.acs.org/, Jun 2017.

[And80] Andersen, H. C. “Molecular dynamics simulations at constant pressure and/or temperature”. *The Journal of Chemical Physics*, vol. 72-4, Feb 1980, pp. 2384-2393.

[Anf73] Anfinsen, C. B. “Principles that govern the folding of protein chains”. *Science*, vol. 181-96, Jul 1973, pp. 223–230.

[Ass17] Association for Computing Machinery. “ACM Digital Library”. Captured on: http://dl.acm.org/, Jun 2017.

[Ast19] Aston, F.W. “LXXIV. A positive ray spectrograph”. *Philosophical Magazine Series 6*, vol. 38-228, 1919, pp. 707-714.

[Ast31] Astbury W. T.; Street A. "X-ray studies of the structures of hair, wool and related fibres. I. General". *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 230-(681-693), Jan 1932, pp. 75-101.

[Ast33] Astbury W. T. "Some Problems in the X-ray Analysis of the Structure of Animal Hairs and Other Protein Fibers". *Transactions of the Faraday Society*, vol. 29-140, 1933, pp. 193-211.

[Ast34] Astbury W. T.; Woods, H. J. "X-ray studies of the structures of hair, wool and related fibres. II. The molecular structure and elastic properties of hair keratin". *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 114-788, Jan 1934, pp. 314-316.

[Ast35] Astbury, W. T.; Sisson, W. A. "X-ray studies of the structures of hair, wool and related fibres. III. The configuration of the keratin molecule and its orientation in the biological cell". *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 150-871, Jul 1935, pp. 533-551.

[Atk09] Atkins, P.; Paula, J. d.; Friedman, Ron. “Quanta, Matter and Change: A Molecular Approach to Physical Chemistry”. W. H. Freeman, 2009, 782p.

[Bal14] Ballard, A. J.; Wales, D. J. “Superposition-Enhanced Estimation of Optimal Temperature Spacings for Parallel Tempering Simulations”. *Journal of Chemical Theory and Computation*, vol. 10-12, Nov 2014, pp. 5599-5605.

[Bec07] Beck, D. A. C.; White, G. W. N.; Daggett, V. “Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations”. *Journal of Structural Biology*, vol. 157-3, Mar 2007, pp. 514–523.

[Ben17] Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. "GenBank". *Nucleic Acids Research*, vol. 45-Database Issue, Jan 2017, pp. D37-D42.

[Ber00] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. “The Protein Data Bank”. *Nucleic Acids Research*, vol. 28-1, Jan 2000, pp. 235-242.

[Ber13] Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Swails, J. M.; Roitberg, A. E.; Cheatham, T. E. “Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide”. *Journal of Chemical Theory and Computation*, vol. 10-1, Nov 2013, pp. 492–499.

[Ber84] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. “Molecular dynamics with coupling to an external bath”. *The Journal of Chemical Physics*, vol. 81-8, Oct 1984, pp. 3684-3690.

[Bet11] Betancourt, M. R. “Optimization of Monte Carlo Trial Moves for Protein Simulations”. *The Journal of Chemical Physics*, vol. 134-1, Jan 2011, pp. 014104.

[Bia14] Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L.; e Schwede, T. “SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information”. *Nucleic Acids Research*, vol.42-Web Server Issue, Jul 2014, pp. 252–258.

[Bla13] Blaszczyk, M.; Jamroz, M.; Kmiecik, S.; Kolinski, A. “CABS-fold: Server for the de novo and consensus-based prediction of protein structure”. *Nucleic Acids Research*, vol. 41-Web Server Issue, Jul 2013, pp. 406–411.

[Bow91] Bowie, J. U.; Luthy, R.; Eisenberg, D. “A method to identify protein sequences that fold into a known three-dimensional structure”. *Science*, vol. 253-5016, Jul 1991, pp. 164–170.

[Bow09] Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. “Progress and challenges in the automated construction of Markov state models for full protein systems.” *The Journal of Chemical Physics*, vol. 131-12, Sept 2009, pp. 124101.

[Bra05] Bradley, P.; Malmström, L.; Qian, B.; Schonbrun, J.; Chivian, D.; Kim, D. E.; Meiler, J.; Misura, K. M.S.; Baker, D. “Free modeling with Rosetta in CASP6”. *Proteins*, vol. 61-S7, pp. 128–134.

[Bra12] Bramucci, E.; Paiardini, A.; Bossa, F.; Pascarella, S. “PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL”. *BMC Bioinformatics*, vol. 13-4, Mar 2012, pp. 1-6.

[Bre07] Brenner, P.; Sweet, C. R.; VonHandorf, D.. Izaguirre, J. A. “Accelerating the replica exchange method through an efficient all-pairs exchange”. *The Journal of Chemical Physics*, vol. 126-7, Feb 2007, pp. 074103.

[Bro83] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. “CHARMM: A program for macromolecular energy, minimization, and dynamics calculations”. *Journal of Computational Chemistry*, vol. 4-2, Jun 1983, pp. 187–217.

[Car03] Carnevali, P.; Tóth, G.; Toubassi, G.; Meshkat, S. N. “Fast protein structure prediction using monte carlo simulations with modal moves”. *Journal of the American Chemical Society*, vol. 125-47, Oct 2003, pp. 14244–14245.

[Cas05] Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. “The Amber biomolecular simulation programs”. *Journal of Computational Chemistry*, vol. 26-16, Dec 2005, pp. 1668–1688.

[Cas14] Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. “Amber 14 Reference Manual”. Captured on: http://ambermd.org/doc12/Amber14.pdf, Jun 2017.

[Cha12] Chaudhury, S.; Olson, M. A.; Tawa, G.; Wallqvist, A.; Lee, M. S. “Efficient Conformational Sampling in Explicit Solvent Using a Hybrid Replica Exchange Molecular Dynamics Method”. *Journal of Chemical Theory and Computation*, vol. 8-2, Feb 2012, pp. 677-687.

[Che05] Cheng,J.; Randall, A. Z.; Sweredoski, M. J.; Baldi, P. “SCRATCH: a protein structure and structural feature prediction server“. *Nucleic Acids Research*, vol. 1-33(Web Server issue), Jul 2005, pp. 72–76.

[Che10] Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. “MolProbity: all-atom structure validation for macromolecular crystallography”. *Acta Crystallographica Section D: Biological Crystallography*, vol. 66-Pt1, Jan 2010, pp. 12–21.

[Che15] Chen, C.; Xiao, Y.; Huang, Y. “Improving the Replica-Exchange Molecular-Dynamics Method for Efficient Sampling in the Temperature Space”. *Physical Review E*, vol. 91-5, May 2015, pp. 052708.

[Che16] Chen, C.; Huang, Y. “Walking Freely In The Energy And Temperature Space By The Modified Replica Exchange Molecular Dynamics Method”. *Journal of Computational Chemistry*, vol. 37-17, Apr 2016, pp. 1565–1575.

[Chi03] Chikenji, G.; Fujitsuka, Y.; Takada, S. “A reversible fragment assembly method for de novo protein structure prediction”. *The Journal of Chemical Physics*, vol. 199-13, Sept 2003, pp. 6895.

[Chi06] Chinchio, M.; Czaplewski, C.; Ołdziej, S.; Scheraga, H. A. “A hierarchical multiscale approach to protein structure prediction: Production of lowresolution packing arrangements of helices and refinement of the best models with a united-residue force field”. *Multiscale Modeling and Simulation*, vol. 5-4, Dec 2006, pp. 1175–1195.

[Cho04] Chou, C. I.; Han, R. S.; Li, S. P.; Lee, T. K. “Guided simulated annealing method for optimization problems”. *Physical Review E*, vol. 67-(6 Pt 2), Jun 2003, pp. 066704.

[Chr05] Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Van Gunsteren, W. F. “The GROMOS software for biomolecular simulation: GROMOS05”. *Journal of Computational Chemistry*, vol. 26-16, Dec 2005, pp. 1719–1751.

[Chu13] Chu, W.; Zhang, J.; Zheng, Q.; Chen, L.; Zhang, H. “Insights into the Folding and Unfolding Processes of Wild-Type and Mutated SH3 Domain by Molecular Dynamics and Replica Exchange Molecular Dynamics Simulations”. *PLOS ONE*, vol. 8-5, May 2013, pp. 1-9.

[Coc01] Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. “Tryptophan Zippers: Stable, Monomeric β-Hairpins”. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98-10, May 2001, pp. 5578–5583.

[Con11] Cong, Q.; Kinch, L. N.; Pei, J.; Shi, S.; Grishin, V. N.; Li, W.; Grishin, N. V. “An automatic method for CASP9 free modeling structure prediction assessment”. *Bioinformatics*, vol. 27-24, Dec 2011, pp. 3371–3378.

[Cor95] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. “A second generation force field for the simulation of proteins, nucleic acids, and organic molecules”. *Journal of the American Chemical Society*, vol. 117-19, May 1995, pp. 5179–5197.

[Cre09] Crescenzi, P.; Goldman, D.; Papadimitriou, C.; Piccolboni, A.; Yannakakis, M. “On the complexity of protein folding”. *Journal of Computational Biology*, vol. 5-3, Mar 2009, pp. 423-465.

[Cur09] Curuksu, J.; Zacharias, M. “Enhanced Conformational Sampling Of Nucleic Acids By a New Hamiltonian Replica Exchange Molecular Dynamics Approach”. *The Journal of Chemical Physics*, vol. 130-10, Mar 2009, pp. 104110.

[Dah97] Dahiyat, B. I.; Sarisky, C. A.; Mayo, S. L. “Protein Design: Towards Fully Automated Sequence Selection”. *Journal of Molecular Biology*, vol. 273-4, Nov 1997, pp. 789-796.

[Dal12] Dall’Agno, K. C. d. M. “Um estudo sobre a predição da estrutura 3D aproximada de proteínas utilizando o método CReF com refinamento”. Tese de Doutorado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2012, 132p.

[Dar98] Darden, T.; York, D.; Pedersen, L. “Particle mesh Ewald: An N· log(N) method for Ewald sums in large systems”. *The Journal of Chemical Physics*, vol. 98-12, Aug 1998, pp. 10089–10092.

[Das12] Dashti, S. D.; Meng, Y.; Roitberg, A. E. “pH-Replica Exchange Molecular Dynamics in Proteins Using a Discrete Protonation Method”. *The Journal of Physical Chemistry B*, vol. 116-30, Jun 2012, pp. 8805-8811.

[Das13] Dashti, S. D.; Roitberg, A. E. “Optimization of Umbrella Sampling Replica Exchange Molecular Dynamics by Replica Positioning”. *Journal of Chemical Theory and Computation*, vol. 9-11, Oct 2013, pp. 4692-4699.

[Dau99] Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Gunsteren, W. F.; Mark, A. E. “Peptide Folding: When Simulation Meets Experiment”. *Angewandte Chemie International Edition*, vol. 38-(1-2), Jan 199, pp. 236–240.

[Dav07] Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B.; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. “MolProbity: all-atom contacts and structure validation for proteins and nucleic acids”. *Nucleic Acids Research*, vol.35-(Web Server Issue), Jul 2007, pp. 375–383.

[Dcr17] Dcrjsr. “Ramachandran_plot_general_100K”. Captured on: https://commons.wikimedia.org/wiki/File:Ramachandran_plot_general_100K.jpg, Jun 2017.

[Dil12] Dill, K. A.; MacCallum, J. L. “The protein-folding problem, 50 years on”, *Science*, vol. 338-6110, Nov 2012, pp. 1042–1046.

[Dor10a] Dorn, M.; Norberto de Souza, O. “A3N: An artificial neural network n-gram-based method to approximate 3-d polypeptides structure prediction”. *Expert Systems with Applications*, vol. 37-12, Dec 2010, pp. 7497–7508.

[Dor10b] Dorn, M.; Norberto de Souza, O. “Mining the protein data bank with CReF to predict approximate 3-D structures of polypeptides”. International Journal of Data Mining and Bioinformatics, vol. 4-3, Jun 2010, pp. 281–299.

[Dor13] Dorn, M.; Buriol, L. S.; Lamb, L. C. “A molecular dynamics and knowledge-based computational strategy to predict native-like structures of polypeptides”. *Expert Systems with Applications*, vol.40-2, Feb 2013, pp. 698–706.

[Dua98] Duan, Y.; Kollman, P. A. “Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution”. *Science*, vol. 282-5389, Oct 1998, pp. 740–744.

[Edm50] Edman, P. “Method for Determination of the Amino Acid Sequence in Peptides”. *Acta Chemica Scandinavica*, vol. 4-1, 1950, pp. 283–293.

[Els17] Elsevier B.V. “Scopus”. Captured on: https://www.scopus.com/home.uri, Jun 2017.

[Eng13] English, C. A.; García, A. E. “Folding and unfolding thermodynamics of the TC10b Trp-cage miniprotein”. *Physical Chemistry Chemical Physics*, vol. 16-7, Dec 2013, pp. 2748–2757.

[Faj08] Fajer, M.; Hamelberg, D.; McCammon, J. A. “Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration”. *Journal of Chemical Theory and Computation*, vol. 4-10, Sept 2008, pp. 1565-1569.

[Fin97] Finkelstein, A. V.; Badretdinov, Y. A. “Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold”. *Folding and Design*, vol. 2-2, Apr 1997, pp. 115-121.

[Flo06] Floudas, C. A.; Fung, H. K.; McAllister, S. R.; Mönnigmann, M.; Rajgaria, R. “Advances in protein structure prediction and de novo protein design: A review”. *Chemical Engineering Science*, vol. 61-3, Feb 2006, pp. 966–988.

[Flo07] Floudas, C. A. “Computational methods in protein structure prediction". *Biotechnology and Bioengineering*, vol. 97-2, Jun 2007, pp. 207–213.

[Fra91] Frauenfelder, H.; Sligar, S.; Wolynes, P. “The energy landscapes and motions of proteins”. *Science*, vol. 254-5038, Dec 1991, pp. 1598-1603.

[Fre17] Free Software Foundation. “Free Software Foundation Home Page”. Captured on: http://www.fsf.org/, Jun 2017.

[Fuk02] Fukunishi, H.; Watanabe, O.; Takada, S. “On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction”. *Journal of Chemical Physics*, vol. 116-20, May 2002, pp. 9058–9067.

[Gal00] Galzitskaya, O. V.; Skoogarev, A. V.; Ivankov, D. N.; Finkelstein, A. V. “Folding nuclei in 3D protein structures”. In: Pac. Symp. Biocomput., 2000, pp. 131-142.

[Gal08] Gallicchio, E.; Levy, R. M.; Parashar, M. “Asynchronous Replica Exchange for Molecular Simulations”. *Journal of Computational Chemistry*, vol. 29-5, Apr 2008, pp. 788–794.

[Gal15] Gallicchio, E.; Xia, J.; Flynn, W. F.; Zhang, B.; Samlalsingh, S.; Mentes, A.; Levy, R. M. “Asynchronous replica exchange software for grid and heterogeneous computing”. *Computer Physics Communications*, vol. 196-1, Nov 2015, pp. 236-246.

[Gan12] Ganguly, B.; Prasad, S. “Homology Modeling and Functional Annotation of Bubaline Pregnancy Associated Glycoprotein 2”. *Journal of Animal Science and Biotechnology*, vol. 3-1, May 2012, pp. 13.

[Gey91] Geyer, C. J. “Markov chain Monte Carlo maximum likelihood”. *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, 1991, pp. 156–163.

[Gib01] Gibbs, N.; Clarke, A. R.; Sessions, R. B. “Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model”. *Proteins*, vol. 43-2, May 2001, pp. 186–202.

[Gin03] Ginalski, K.; Pas, J.; Wyrwicz, L.S.; vonGrotthuss, M.; Bujnicki, J.M.; Rychlewski, L. “ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure”. *Nucleic Acids Research*, vol. 31-13, Jul 2003, pp. 3804–3807.

[Gni14] Gniewek, P.; Kolinski, A.; Kloczkowski, A.; Gront, D. “BioShell-Threading: versatile monte carlo package for protein 3d threading”. *BMC Bioinformatics*, vol. 15-1, Jan 2014 , pp. 22.

[Goo17] Google. “Google Academics”. Captured on: https://scholar.google.com, Jun 2017.

[Gro11] Gross, J.; Janke, W.; Bachmann, M. “Massively Parallelized Replica-Exchange Simulations of Polymers on GPUs”. *Computer Physics Communications*, vol. 182-8, Aug 2011, pp. 1638-1644.

[Haf07] Hagen, M.; Kim, B.; Liu, P.; Friesner, R. A.; Berne, B. J. “Serial Replica Exchange”. *The Journal of Physical Chemistry B*, vol. 111-6, Jan 2007, 1416-1423.

[Har02] Hardin, C.; Pogorelov, T. V.; Luthey-Schulten, Z. “Ab initio protein structure prediction”. *Current Opinion in Structural Biology*, vol. 12-2, Apr 2002, pp. 176–181.

[Han97] Hansmann, U. H. E. “Parallel tempering algorithm for conformational studies of biological molecules”. *Chemical Physics Letters*, vol. 281-(1-3), Dec 1997, pp. 140-150.

[Har17] Har, R.; Shigeta, Y. “Efficient Conformational Search Based on Structural Dissimilarity Sampling: Applications for Reproducing Structural Transitions of Proteins”. *Journal of Chemical Theory and Computation*, vol. 13-3, Feb 2017, pp. 1411-1423.

[Has70] Hastings, W. K. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. *Biometrika*, vol. 75-1, Apr 1970, pp. 97-109.

[Hat14] Hatch, H. W.; Stillinger, F. H.; Debenedetti, P. G. “Computational study of the stability of the miniprotein trp-cage, the GB1 β-hairpin, and the AK16 peptide, under negative pressure”. *The Journal of Physical Chemistry B*, vol. 118-28, Feb 2014, pp. 7761–7769.

[Haw95] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. “Pairwise Solute Descreening of Solute Charges from a Dielectric Medium”. *Chemical Physics Letters*, vol. 246-1, Nov 1995, pp. 122-129.

[Haw96] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. “Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium”. *The Journal of Physical Chemistry*, vol. 100-51, Dec 1996, pp. 19824-19839.

[Hel08] Helles, G. “A comparative study of the reported performance of ab initio protein structure prediction algorithms". *Journal of the Royal Society Interface*, vol. 5-21, Apr 2008, pp. 387–396.

[Ho05] Ho, B. K.; Brasseur, R. “The Ramachandran Plots of Glycine and Pre-proline”. *BMC Structural Biology*, vol. 5-1, Aug 2005, pp. 14;

[Ho06] Ho, B. K.; Dill, K. A. “Folding very short peptides using molecular dynamics”. *PLoS computational biology*, vol. 2-4, Apr 2006, pp. 27.

[Hof14] Hoffmann, F.; Vancea, I.; Kamat, S. G.; Strodel, B. “Protein structure prediction: Assembly of secondary structure elements by basin-hopping”. *ChemPhysChem*, vol. 15-15, Jul 2014, pp. 3378–3390.

[Hon04] Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. “10 Residue Folded Peptide Designed by Segment Statistics”, *Structure*, vol. 12-8, Aug 2004, pp. 1507-1518.

[Hor06] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. “Comparison of Multiple AMBER Force Fields and Development of Improved Protein Backbone Parameters”. *Proteins*, vol. 65-3, Mar 2006, pp. 712–725.

[Hoo85] Hoover, W. G. “Canonical dynamics: Equilibrium phase-space distributions”. *Physical Review A*, vol. 31-3, Mar 1985, pp. 1695-1697.

[Hri07] Hritz, J.; Oostenbrink, C. “Optimization of replica exchange molecular dynamics by fast mimicking”. *The Journal of Chemical Physics*, vol. 127-20, Nov 2007, pp. 204104.

[Hri08] Hritz, J.; Oostenbrink, C. “Hamiltonian replica exchange molecular dynamics using soft-core interactions”. *The Journal of Chemical Physics*, vol. 128-14, Apr 2008, pp. 144121.

[Huk95] Hukushima, K.; Nemoto, K. “Exchange monte carlo method and application to spin glass simulations”. *Journal of the Physical Society of Japan*, vol. 65-6, Dec 1995, pp. 1604-1608.

[IEE17] IEEE. “IEEE Xplore Digital Library”. Captured on: http://ieeexplore.ieee.org/Xplore/home.jsp, Jun 2017.

[Ito12] Itoh, S. G.; Okumura, H. “Coulomb Replica-Exchange Method: Handling Electrostatic Attractive and Repulsive Forces for Biomolecules”. *Journal of Computational Chemistry*, vol. 34-8, Nov 2012, pp. 622–639.

[Jag08] Jagielska, A.; Wroblewska, L.; Skolnick, J. “Protein model refinement using an optimized physics-based all-atom force field”. *Proceedings of the National Academy of Sciences*, vol. 105-24, Jun 2008, pp. 8268–8273.

[Jan14] Jani, V.; Sonavane, U. B.; Joshi, R. “REMD and umbrella sampling simulations to probe the energy barrier of the folding pathways of engrailed homeodomain”. *Journal of Molecular Modeling*, vol. 20-6, Jun 2014, pp. 2283.

[Jar05] Jaroszewski, L.; Rychlewski, L.; Li, Z.; Li, W.; Godzik, A. “FFAS03: A server for profile-profile sequence alignments”. *Nucleic Acids Research*, vol. 33-(Web Server Issue), Jul 2005, pp. 284–288.

[Jay06] Jayaram, B.; Bhushan, K.; Shenoy, S. R.; Narang, P.; Bose, S.; Agrawal, P.; Sahu, D.; Pandey, V. “Bhageerath: An energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins“. *Nucleic Acids Research*, vol. 34-21, Nov 2006, pp. 6195–6204.

[Jo15] Jo, S.; Jiang, W. “A generic implementation of replica exchange with solute tempering (REST2) algorithm in NAMD for complex biophysical simulations”, *Computer Physics Communications*, vol. 197-1, Dec 2015, pp. 304-311.

[Joh17] John Wiley & Sons. “Wiley Online Library”. Captured on: http://onlinelibrary.wiley.com/, Jun 2017.

[Joh18] Johndoct. “Amino Acid Structure”. Captured on: https://commons.wikimedia.org/wiki/File:Amino-acid-structure.jpg, May 2018.

[Jon01] Jones, D. T. “Predicting novel protein folds by using FRAGFOLD”. *Proteins*, vol. 45-S5, Jan 2001, pp. 127–132.

[Jon92] Jones, D. T.; Taylor, W. R.; Thornton, J. M. “A new approach to protein fold recognition”. *Nature*, vol. 358-6381, Jul 1992, pp. 86–89.

[Jon99] Jones, D. T. “Protein secondary structure prediction based on position specific scoring matrices”. *Journal of Molecular Biology*, vol. 292-2, Sept 1999 pp. 195–202.

[Jor96] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. “Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids”. *Journal of the American Chemical Society*, vol. 118-45, Nov 1996, pp. 11225-11236.

[Jun15] Jung, J.; Mori, T.; Kobayashi, C.; Matsunaga, Y.; Yoda, T.; Feig, M. Sugita, Y. “GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations”. *WIREs Computational Molecular Science,* vol. 5-4, May 2015, pp. 310–323.

[Käl12] Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. “Template-based protein structure modeling using the RaptorX web server”. *Nature Protocols*, vol. 7-8, Jul 2012, pp. 1511–1522.

[Kam15] Kamberaj, H. “Conformational Sampling Enhancement of Replica Exchange Molecular Dynamics Simulations Using Swarm Particle Intelligence”. *The Journal of Chemical Physics*, vol. 143-12, Sept 2015, pp. 124105.

[Kan11] Kannan, S; Zacharias, M; “Role of the closing base pair for d(GCA) hairpin stability: free energy analysis and folding simulations”. *Nucleic Acids Research*, vol. 39-19, Oct 2011, pp. 8271-8280.

[Kes10] Kessel, A.; Ben-Tal, N. “Introduction to Proteins: Structure, Function, and Motion”. CRC Press, 2010, 654p.

[Kim09] Kim, J.; Straub, J. E. “Optimal Replica Exchange Method Combined with Tsallis Weight Sampling”. *The Journal of Chemical Physics*, vol. 130-14, Apr 2009, pp. 144114.

[Kim12] Kim, J.; Straub, J. E.; Keyes, T. “Replica Exchange Statistical Temperature Molecular Dynamics Algorithm”. *The Journal of Physical Chemistry B*, vol. 116-29, Apr 2012, pp. 8646-8653.

[Kou10] Koulgi, S.;Sonavane, U.; Joshi, R. “Insights into the folding pathway of the Engrailed Homeodomain protein using replica exchange molecular dynamics simulations”. *Journal of Molecular Graphics & Modelling*, vol. 29-3, Nov 2010, pp. 481–491.

[Kri04] Krieger, E.; Darden, T.; Nabuurs, S. B.; Finkelstein, A.; Vriend, G. “Making optimal use of empirical energy functions:force-field parameterization in crystal space”. *Proteins*, vol. 57-4, Dec 2004, pp. 678–683.

[Kub07] Kubitzki, M. B.; Groot, B. L. “Molecular Dynamics Simulations Using Temperature-Enhanced Essential Dynamics Replica Exchange”. *Biophysical Journal*, vol. 92-12, Jun 2007, pp. 4262-4270.

[Lam16] Lamiable, A.; Thevenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tuffery, P. “PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex”. *Nucleic Acids Research*, vol. 44-W1, Jul 2016, pp. 449–454.

[Las93] Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. “PROCHECK: a program to check the stereochemical quality of protein structures”. *Journal of Applied Crystallography*, vol. 26-2, Apr 1993, pp. 283-291.

[Las96] Laskowski, R. A.; Rullmann, J. A. C.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. "AQUA and PROCHECK-NMR: Programs for Checking the Quality of Protein Structures Solved by NMR". *Journal of Biomolecular NMR*, vol. 8-4, Dec 1996, pp. 477-486.

[Lee02] Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. “Molecular dynamics in the endgame of protein structure prediction”. *Journal of Molecular Biology*, vol. 313-2, May 2002, pp. 417–430.

[Lee04] Lee, J.; Kim, S. Y.; Joo, K.; Kim, I. “Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing”. *Proteins*, vol. 56-4, Sept 2004, pp. 704–714.

[Leh12] Lehninger, A.; Nelson, D. L.; Cox, M. M. “Lehninger principles of biochemistry”. W.H. Freeman, 2012, 1340p.

[Leh93] Lehninger, A.; Nelson, D. L.; Cox, M. M. “Principles of biochemistry”. Worth Publishers, 1993, 1090p.

[Lei07] Lei, H.; Wu, C.; Liu, H.; Duan, Y. “Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations”. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104-12, Mar 2007, pp. 4925–4930.

[Lei08] Lei, H.; Wu, C.; Wang, Z. X.; Zhou, Y.; Duan, Y. “Folding processes of the B domain of protein A to the native state observed in all-atom ab initio folding simulations”. *The Journal of Chemical Physics*, vol. 128-23, Jun 2008, pp. 235105.

[Lei09] Lei, H.; Wang, Z. X.; Wu, C.; Duan, Y. “Dual folding pathways of an α/β protein from all-atom ab initio folding simulations”. *The Journal of Chemical Physics*, vol. 131-16, Oct 2009, pp. 165105.

[Leo92] Leopold, P. E.; Montal, M.; Onuchic, J. N.; “Protein folding funnels: A kinetic approach to the sequence-structure relationship”. *Proceedings of the National Academy of Sciences*, vol. 89-18, Sept 1992, pp. 8721-8725.

[Les05] Lesk, A. “Introduction to bioinformatics”. Oxford University Press, 2005, 360p.

[Lev68] Levinthal, C. “Are there pathways for protein folding?”. *Journal of Medical Physics*, vol. 65-1, 1968, pp. 44-45.

[Lev76] Levitt, M.; Chothia, C. “Structural patterns in globular proteins”. *Nature*, vol. 261-5561, Jun 1976, pp. 552–558.

[Li07a] Li, X.; O’Brien, C. P.; Collier, G.; Vellore, N. A.; Wang, F.; Latour, R. A.; Bruce, D. A.; Stuart, S. J. “An improved replica-exchange sampling method: Temperature intervals with global energy reassignment”. *The Journal of Chemical Physics*, vol. 127-16, Oct 2007, pp. 164116.

[Li07b] Li, Z.; Parashar, M. "Grid-Based Asynchronous Replica Exchange" In: 2007 8th IEEE/ACM International Conference on Grid Computing, 2007, pp. 201-208.

[Li09] Li, X.; Latour, R. A.; Stuart, S. J. “TIGER2: An improved algorithm for temperature intervals with global exchange of replicas”. *The Journal of Chemical Physics*, vol. 130-17, May 2009, pp. 174106.

[Li15] Li, X.; Snyder, J. A.; Stuart, S. J.; Latour, R. A. “TIGER2 with solvent energy averaging (TIGER2A): An accelerated sampling method for large molecular systems with explicit representation of solvent”. *The Journal of Chemical Physics*, vol. 143-14, Oct 2015, pp. 144105.

[Lin09a] Lindert, S.; Staritzbichler, R.; Wötzel, N.; Karakas¸, M.; Stewart, P. L.; Meiler, J. “EM-fold: De novo folding of α/-helical proteins guided by intermediate resolution electron microscopy density maps”. *Structure*, vol. 17-7, Jul 2009, pp. 990–1003.

[Lin09b] Lin, E.; Shell, M. S. “Convergence and Heterogeneity in Peptide Folding with Replica Exchange Molecular Dynamics”. *Journal of Chemical Theory and Computation*, vol. 5-8, Jul 2009, pp. 2062-2073.

[Lin11] Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. “How Fast-Folding Proteins Fold”. *Science*, vol. 334-6055, Oct 2011, pp. 517-520.

[Lin12] Lindert, S.; Alexander, N.; Wötzel, N.; Karakas, M.; Stewart, P. L.; Meiler, J. “Ab initio protein modeling into cryoEM density maps using EM-Fold”. *Biopolymers*, vol. 97-9, Feb 2012, pp. 669–677.

[Lip12] Lipinski-Paes, T.; Norberto de Souza, O. “Cooperative multi-agent system for protein structure prediction”. In: 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, pg. 117.

[Lip14] Lipinski-Paes, T. e Norberto de Souza, O. (2014). “Masters: A general sequence-based multiagent system for protein tertiary structure prediction”. *Electronic Notes in Theoretical Computer Science,* vol. 306-1, Jul 2014, pp. 45–59.

[Lip17] Lipinski-Paes, T. “Cut-REMD: Um Novo Método Para Predição De Estruturas Terciárias De Proteínas Baseado Em Raio De Corte Incremental”. Tese de Doutorado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2017, 201p.

[Liu04] Liu, Y.; Liu, Z.; Androphy, E.; Chen, J.; Baleja, J. D. “Design and Characterization of Helical Peptides that Inhibit the E6 Protein of Papillomavirus”. *Biochemistry*, vol. 43-23, May 2004, pp. 7421-7431.

[Liu07] Liu, P.; Voth, G. A. “Smart Resolution Replica Exchange: An Efficient Algorithm for Exploring Complex Energy Landscapes”. *The Journal of Chemical Physics*, vol. 126-4, Jan 2007, pp. 045106.

[Liu15] Liu, L.; Kuo, J. “A LAMMPS implementation of volume–temperature replica exchange molecular dynamics”. *Computer Physics Communications*, vol. 189-1, Apr 2015, pp. 119-127.

[Loc01] Locker, C. R.; Hernandez, R.; “A minimalist model protein with multiple folding funnels”. *Proceedings of the National Academy of Sciences*, vol. 98-16, Jul 2001, pp. 9074–9079.

[Loc15] Lockhart, C.; O'Connor, J.; Armentrout, S.; Klimov, D. K. “Greedy replica exchange algorithm for heterogeneous computing grids”. *Journal of Molecular Modeling*, vol. 21-9, Sept 2015, pp. 243.

[Lon92] Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. “Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N’-methylamide”. *Biopolymers*, vol. 32-5, May 1992, pp. 523-535.

[Lov03] Lovell, S. C.; Davis, I. W.; Arendall, W. B.; Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. “Structure Validation by Cα Geometry: ϕ, ψ and Cβ deviation”. *Proteins*, vol. 50-3, Jan 2003, pp. 437–450.

[Lu08] Lu, M.; Dousis, A. D.; Ma, J. “OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing”. *Journal of Molecular Biology*, vol. 376-1, Feb 2008, pp. 288–301.

[Luc17] Lucid Software Inc. “LucidChart”. Captured on: https://www.lucidchart.com, Jun 2017.

[Lus01] Luscombe, N. M.; Greenbaum, D.; Gerstein, M. "What is bioinformatics? A proposed definition and overview of the field". *Methods of Information in Medicine*, vol. 40-4, 2001, pp. 346–358.

[Lyr14] Lyras, D. P.; Metzler, D. “ReformAlign: improved multiple sequence alignments using a profile-based meta-alignment approach”. *BMC Bioinformatics*, vol. 15-1, Aug 2014, pp. 265.

[Mac00] Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. “Structural Analysis of WW Domains and Design of a WW Prototype”. *Nature Structural & Molecular Biology*, vol. 7-5, May 2000, pp. 375-379.

[Mac13] Machado, V. S. “Método CReF Para Predição Da Estrutura 3D Aproximada: Revisão De Literatura Sobre Estruturas De Proteínas”. Monografia, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2013, 50p.

[Mac14] Maciejczyk, M.; Spasic, A.; Liwo, A.; Scheraga, H. A. “DNA duplex formation with a coarse-grained model”. *Journal of Chemical Theory and Computation*, vol. 10-11, Sept 2014, pp. 5020–5035.

[Mac91] MacArthur, M. W.; Thornton, J. M. “Influence of Proline Residues on Protein Conformation”, *Journal of Molecular Biology*, vol. 218-2, Mar 1991, pp. 397-412.

[Man99] Manousiouthakis, V. I.; Deem, M. W. “Strict detailed balance is unnecessary in Monte Carlo simulation”. *The Journal of Chemical Physics*, vol. 110-6, Jan 1999, pp. 2753–2756.

[Mar00] Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; e Sali, A. “Comparative protein structure modeling of genes and genomes”. *Annual Review of Biophysics and Biomolecular Structure*, vol. 29-1, Jun 2000, pp. 291–325.

[Mar12] Marks, D. S.; Hopf, T. A.; Sander, C. “Protein structure prediction from sequence variation”. *Nature Biotechnology*, vol. 30-11, Nov 2012, pp. 1072–1080.

[Mat13] Matsuda, M.; Maruyama, N.; Takizawa, S. "K MapReduce: A Scalable Tool for Data-Processing and Search/Ensemble Applications on Large-Scale Supercomputers". In: 2013 IEEE International Conference on Cluster Computing (CLUSTER), 2013, pp. 1-8.

[Mcc77] McCammon, J. A.; Gelin, B. R.; Karplus, M. “Dynamics of folded proteins”. *Nature*, vol. 267-5612, Jun 1977, pp. 585–590.

[Mck97] McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. “NMR Structure of the 35-Residue Villin Headpiece Subdomain”. *Nature Structural Biology*, vol. 4-3, Mar 1997, pp. 180-184.

[Mel12] Melo, M. C. R.; Bernardi, R. C.; Fernandes, T. V. A.; Pascutti, P. G. “GSAFold: A new application of GSA to protein structure prediction”. *Proteins*, vol. 80-9, Aug 2012, pp. 2305–2310.

[Met53] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *The Journal of Chemical Physics*, vol. 21-6, Jun 1953, pp. 1087-1092.

[Mic07] Seringhaus, M.; Gerstein, M. “Chemistry Nobel Rich in Structure”. *Science*, vol. 315-5808, Jan 2007, pp. 40-41.

[Mic15] Michino, M.; Shi, L. “Computational Approaches in the Structure–Function Studies of Dopamine Receptors”, Dopamine Receptor Technologies, 2015, pp. 31-42.

[Mir14] Mirjalili, V.; Noyes, K.; Feig, M. “Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging”. *Proteins*, 82-S2, Feb 2014, pp. 196–207.

[Mol17] Duke University School of Medicine. “MolProbity Main Page”. Captured on: http://molprobity.biochem.duke.edu/, Jun 2017.

[Mu08] Mu, Y.; Yang, Y.; Xu, W. “A Global Optimization Scheme: Kernel Replica Exchange Simulation Method For Protein Folding”. *Journal of Theoretical and Computational Chemistry*, vol. 7-2, Apr 2008, pp. 177-187.

[Nad07a] Nadler, W.; Hansmann, U. H. E. “Dynamics and Optimal Number of Replicas in Parallel Tempering Simulations”. *Physical Review E*, vol. 76-6, Dec 2007, pp. 065701.

[Nad07b] Nadler, W.; Hansmann, U. H. E. “Optimizing Replica Exchange Moves For Molecular Dynamics”. *Physical Review E*, vol. 76-5(Pt 2), Dec 2007, pp. 057102

[Nad08] Nadler, W.; Hansmann, U. H. E. “Optimized Explicit-Solvent Replica Exchange Molecular Dynamics from Scratch”. *The Journal of Physical Chemistry B*, vol. 112-34, Jul 2008, pp. 10386-10387.

[Nag12] Nagata, K.; Randall, A.; Baldi, P. “SIDEpro: A novel machine learning approach for the fast and accurate prediction of side-chain conformations”. *Proteins*, vol. 80-1, Jan 2012, pp. 142–153.

[Nar06] Narang, P.; Bhushan, K.; Bose, S.; Jayaram, B. “Protein structure evaluation using an all-atom energy based empirical scoring function”. *Journal of Biomolecular Structure and Dynamics*, vol. 23-4, Feb 2006, pp. 385–406.

[Nei02] Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. “Designing a 20-Residue Protein”. *Nature Structural Biology*, vol. 9-6, Apr 2002, pp. 425-430.

[Neu09] Neuweiler, H.; Sharpe, T. D; Rutherford, T. J.; Johnson, C. M.; Allen, M. D.; Ferguson, N.; Fersht, A. R. “The Folding Mechanism of BBL: Plasticity of Transition-State Structure Observed within an Ultrafast Folding Protein Family”. *Journal of Molecular Biology*, vol. 390-5, Jul 2009, pp. 1060-1073.

[Ngu10] Nguyen, P. H. “Replica exchange simulation method using temperature and solvent viscosity”. *The Journal of Chemical Physics*, vol. 132-14, Apr 2010, pp. 144109.

[Niu13] Niu, J.; Bai, S.; Khosravi, E.; Park, S. J. "A Hadoop approach to advanced sampling algorithms in molecular dynamics simulation on cloud computing" In: 2013 IEEE International Conference on Bioinformatics and Biomedicine, 2013, pp. 452-455.

[Nos84] Nosé, S. “A unified formulation of the constant temperature molecular dynamics methods”. *The Journal of Chemical Physics*, vol. 81-1, Jul 1984, pp. 511-519.

[Nym08] Nymeyer, H. “How efficient is replica exchange molecular dynamics? An analytic approach”. *Journal of Chemical Theory and Computation*, vol. 4-4, Mar 2008, pp. 626– 636.

[Nym98] Nymeyer, H.; Garcia, A. E.; Onuchic, J. N.; “Folding funnels and frustration in off-lattice minimalist protein landscapes”. *Proceedings of the National Academy of Sciences*, vol. 95-11, May 1998, pp. 5921–5928.

[Oku06] Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. “Improved Efficiency of Replica Exchange Simulations Through Use of a Hybrid Explicit/Implicit Solvation Model”. *Journal of Chemical Theory and Computation*, vol. 2-2, Feb 2006, pp. 420-433.

[Oku07] Okur, A.; Roe, D. R.; Cui, G.; Hornak, V.; Simmerling, C. “Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir”. *Journal of Chemical Theory and Computation*, vol. 3-2, Jan 2007, pp. 557-568.

[Ols14] Olson, B.; Shehu, A. “Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction”. In: Proceedings of the 6th International Conference on Bioinformatics and Computational Biology, 2014, pp. 143–148.

[Onu02] Onufriev, A.; Case, D. A.; Bashford, D. “Effective born radii in the generalized born approximation: The importance of being perfect”. *Journal of Computational Chemistry*, vol. 23-14, Nov 2002, pp. 1297–1304.

[Onu97] Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. “Theory of protein folding: The energy landscape perspective”. *Annual Review of Physical Chemistry*, vol. 48-1, Oct 1997, pp. 545-600.

[Osg00] Osguthorpe, D. J. "Ab initio protein folding”. *Current Opinion in Structural Biology*, vol. 10-2, Apr 2000, pp. 146–152.

[Par12] Park, I. H.; Gangupomu, V.; Wagner, J.; Jain, A.; Vaidehi, N. “Structure refinement of protein low resolution models using the GNEIMO constrained dynamics method”. *Journal of Physical Chemistry B*, vol. 116-8, Mar 2012, pp. 2365–2375.

[Par14] Park, S.; Im, W. “Theory of Adaptive Optimization for Umbrella Sampling”. *Journal of Chemical Theory and Computation*, vol. 10-7, Jun 2014, pp. 2719-2728.

[Par88] Parisi, G. Statistical Field Theory, Volume 66 of Frontiers in Physics. Addison-Wesley Pub, 1988, 343p.

[Pat08] Patriksson, A.; Spoel, D. “A Temperature Predictor for Parallel Tempering Simulations”. *Physical Chemistry Chemical Physics*, vol. 10-15, Feb 2008, pp. 2073-2077.

[Pau51] Pauling, L.; Corey, R.B.; Branson, H. R. “The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain”. *Proceedings of the National Academy of Sciences*, vol. 37-4, Apr 1951, pp. 205-211.

[Pav11] Pavlopoulou, A.; Michalopoulos, I. “State-of-the-art bioinformatics protein structure prediction tools (Review)”. *International Journal of Molecular Medicine*, vol. 28-3, Sept 2011, pp. 295-310.

[Pea95] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel,G.; Kollman, P. “AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules”. *Computer Physics Communications*, vol. 91-(1-3), Sept 1995, pp. 1-41.

[Ped97] Pedersen, J. T.; Moult, J. “Protein folding simulations with genetic algorithms and a detailed molecular description”. *Journal of Molecular Biology*, vol. 269-2, Jun 1997, pp. 240–259.

[Per15] Perez, A.; MacCallum, J. L.; Dill, K. A. “Accelerating molecular simulations of proteins using Bayesian inference on weak information”. *Proceedings of the National Academy of Sciences*, vol. 112-38, Aug 2015, pp. 11846-11851.

[Pet16] Peter, E. K.; Shea, J.; Pivkin, I. V. “Coarse kMC-Based Replica Exchange Algorithms for the Accelerated Simulation of Protein Folding in Explicit Solvent”. *Physical Chemistry Chemical Physics*, vol. 18-18, Apr 2016, pp. 13052-13065.

[Pla17] Platania, R.; Shams, S.; Chiu, C..; Kim, N.; Kim, J.; Park, S. “Hadoop-Based Replica Exchange Over Heterogeneous Distributed Cyberinfrastructures”. *Concurrency and Computation: Practice and Experience*, vol. 29-4, Jun 2016, pp. 3878.

[Pre17a] Prediction Center. “LGA”. Captured on: http://predictioncenter.org/local/lga/lga_description.html, Jun 2017.

[Pre17b] Prediction Center. “Results Table HELP”. Captured on: http://www.predictioncenter.org/casp11/doc/help.html#molprb_Score, Jun 2017.

[Pub17] PubMed. “Home - PubMed - NCBI”. Captured on: https://www.ncbi.nlm.nih.gov/pubmed/, Jun 2017.

[Pyt17] Python Software Foundation. “Python”. Captured on: https://www.python.org/, Jun 2017.

[Rad13] Radak, B.K.; Lee, T.; He, P; Romanus, M.; Weidner, O.; Dai, W.; Gallicchio, E.; Deng, N.; York, D. M.; Levy, R. M.; Jha, S. “A framework for flexible and scalable replica-exchange on production distributed CI”. In: XSEDE, 2013, pp. 26:1-26:8.

[Rah64] Rahman, A. “Correlations in the Motion of Atoms in Liquid Argon”. *Physical Review*, vol. 136-2A, Oct 1964, pp. 405-411.

[Ram63] Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. “Stereochemistry of polypeptide chain configurations”. *Journal of Molecular Biology*, vol. 7-1, Jul 1963, pp. 95-99.

[Rhe03] Rhee, Y. M.; Pande, V. S. “Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation”. *Biophysical Journal*, vol. 84-2, Feb 2003, pp. 775-786.

[Roe01] Roe, D.; Okur, A.; Simmerling, C. “Replica Exchange Simulations with AMBER 10”. Captured on: http://ambermd.org/tutorials/advanced/tutorial7/, Jun 2017.

[Roe14] Roe, D. R.; Bergonzo, C.; Cheatham, T. E. “Evaluation of Enhanced Sampling Provided by Accelerated Molecular Dynamics with Hamiltonian Replica Exchange Methods”. *The Journal of Physical Chemistry B*, vol. 118-13, Apr 2014, pp. 3543–3552.

[Roh04] Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. “Protein Structure Prediction Using Rosetta”. *Methods in Enzymology*, vol. 383-1, 2004, pp. 66–93.

[Ros09] Rosta, E.; Buchete, N.-V.; Hummer, G. “Thermostat artifacts in replica exchange molecular dynamics simulations”. *Journal of Chemical Theory and Computation*, vol. 5-5, Apr 2009 ; pp. 1393–1399.

[Roy10] Roy, A.; Kucukural, A.; Zhang, Y. “I-TASSER: a unified platform for automated protein structure and function prediction”. *Nature Protocols*, vol. 5-4, Apr 2010, pp. 725–738.

[Row01] Rowley, C. “Schematic of a replica exchange molecular dynamics simulation”. Captured on: https://commons.wikimedia.org/wiki/File:Schematic_of_a_replica_exchange_molecular_dynamics_simulation.svg, May 2017.

[Rus10] Ruscio, J. Z.; Fawzi, N. L.; Head-Gordon, T. “How hot? Systematic convergence of the replica exchange method using multiple reservoirs.” *Journal of Computational Chemistry*, vol. 31-3, Feb 2010, pp. 620–627.

[Ryc77] Ryckaert, J.; Ciccotti, G.; Berendsen, H. J. C. “Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes”. *Journal of Computational Physics*, vol. 23-3, Mar 1977, pp. 327-341.

[Sar01] Sarisky, C. A.; Mayo, S. L. “The βββα Fold: Explorations in Sequence Space”. *Journal of Molecular Biology*, vol. 307-5, Apr 2001, pp. 1411-1418.

[Sch17] Schrödinger. “PyMOL Home Page”. Captured on: https://www.pymol.org/, Jun 2017.

[Sei05] Seibert, M. M.; Patriksson, A.; Hess, B.; van der Spoel, D. “Reproducible polypeptide folding and structure prediction using molecular dynamics simulations”. *Journal of Molecular Biology*, vol. 354-1, Nov 2005, pp. 173–183.

[Sha07a] Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. “Anton, a Special-purpose Machine for Molecular Dynamics Simulation”. *SIGARCH Comput. Archit. News*, vol 35-2, Jun 2007, pp. 1-12.

[Sha07b] Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. “Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms”. *Journal of Chemical Theory and Computation*, vol. 3-6, Oct 2007, pp. 2312-2334.

[She06] Shen, M.; Sali, A. “Statistical potential for assessment and prediction of protein structures”. *Protein Science : A Publication of the Protein Society*, vol. 15-11, Nov 2006, pp. 2507–2524.

[Shm07] Shmygelska, A. “An extremal optimization search method for the protein folding problem: the go-model example”. In: Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation, 2007, pp. 2572-2579.

[Sim99] Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. “Ab initio protein structure prediction of CASP III targets using ROSETTA”. *Proteins*, vol. 37-S3, Nov 1999, pp. 171–176.

[Sko06] Skolnick, J. “In quest of an empirical potential for protein structure prediction”. *Current Opinion in Structural Biology*, vol. 16-2, Apr 2006, pp. 166-171.

[Sky15] Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; Mourik, T.; Mitchell, J. B. O. “A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution”. *Physical Chemistry Chemical Physics*, vol. 17-9, Jan 2015, pp. 6174-6191.

[Smi13] Smith-Romero, M.; Niu, J.; Allen, A.; Khosravi, A.; Bai, S. "Implementing Replica Exchange Molecular Dynamics Using Work Queue" In: 2013 IEEE International Conference on Bioinformatics and Biomedicine, 2013, pp. 63-64.

[Soc98] Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. “Protein folding mechanisms and the multidimensional folding funnel”. *Proteins*, vol. 32-2, Aug 1998, pp. 136-158.

[Söd05] Söding, J. “Protein homology detection by HMM-HMM comparison”. *Bioinformatics*, vol. 21-5, May 2005, pp. 951–960.

[Son07] Son, W.; Jang, S.; Pak, Y.; Shin, S. “Folding simulations with novel conformational search method”. *The Journal of Chemical Physics*, vol. 126-10, Mar 2007, pp. 104906.

[Spi13] Spill, Y. G.; Bouvier, G.; Nilges, M. “A convective replica-exchange method for sampling new energy basins”. *Journal of Computational Chemistry,* vol. 34-2, Jan 2013, pp. 132–140.

[Spl17] Splettstoesser, T. “Folding Funnel Schematic”. Captured on: https://commons.wikimedia.org/w/index.php?curid=28353539, May 2017.

[Sri95] Srinivasan, R.; Rose, G. D. “LINUS: A hierarchic procedure to predict the fold of a protein”. *Proteins*, vol. 22-2, Jun 1995, pp. 81– 99.

[Sti90] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. “Semianalytical treatment of solvation for molecular mechanics and dynamics”. *Journal of the American Chemical Society*, vol. 112-16, Aug 1990, pp. 6127–6129.

[Sto17] Stote, R.; Dejaegere, A.; Kuznetsov, D.; Falquet, L. “Potential Energy Functions”. Captured on: http://www.ch.embnet.org/MD_tutorial/pages/MD.Part2.html, May 2017.

[Sue03] Suenaga, A. “Replica-exchange molecular dynamics simulations for a small-sized protein folding with implicit solvent”. Journal of Molecular Structure: THEOCHEM, vol. 634-(1-3), Sept 2003, pp. 235–241.

[Sug99] Sugita, Y.; Okamoto, Y. “Replica-exchange molecular dynamics method for protein folding”. *Chemical Physics Letters*, vol. 314-(1–2), Nov 1999, pp. 141–151.

[Suz08] Suzuki, M.; Okuda, H. “Fragment Replica-Exchange Method for Efficient Protein Conformation Sampling”. *Molecular Simulation*, vol. 34-3, May 2008, pp. 267-275.

[Swe86] Swendsen, R. H.; Wang, J.-S.; “Replica Monte Carlo Simulation of Spin-Glasses”. *Physical Review Letters*, vol. 57-21, Nov 1986, pp. 2607-2609.

[Teo04] Teodorescu, O.; Galor, T.; Pillardy, J.; Elber, R. “Enriching the sequence substitution matrix by structural information”. *Proteins*, vol. 51-1, Oct 2004, pp. 41–48.

[Tie07] Tie, Y.; Kovalevsky, A. Y.; Boross, P.; Wang, Y.; Ghosh, A. K.; Tözsér, J.; Harrison, R. W.; Weber, I. T. “Atomic resolution crystal structures of HIV-1 protease and mutants V82A and I84V with saquinavir”. *Proteins Structure Function and Bioinformatics*, vol. 67-1, Apr 2007, pp. 232-242.

[Tra04] Tramontano, A. “Integral and differential form of the protein folding problem“. *Physics of Life Reviews*, vol. 1-2, Jul 2004, pp. 103-127.

[Tra07] Tramontano, A. “Protein structure prediction. concepts and applications”. John Wiley & Sons, 2006, 228p.

[Tuc99] Tuckerman, M. E.; Martyna, G. J. “Understanding modern molecular dynamics: Techniques and applications”. *The Journal of Physical Chemistry B*, vol. 104-2, Dec 1999, pp. 159–178.

[Uni17] University of California. “Protein Structure Prediction Center”. Captured on: http://predictioncenter.org/, May 2017.

[Urb08] Urbic, T.; Urbic, T.; Avbelj, F.; Dill, K. A. “Molecular simulations find stable structures in fragments of protein G”. *Acta Chimica Slovenica*, vol. 2008-55, Jan 2008, pp. 385–395.

[Ver04] Vermeulen, W.; Vanhaesebrouck, P.; Van Troys, M.; Verschueren, M.; Fant, F.; Goethals, M.; Ampe, C.; Martins, J. C.; Borremans, F. A. M. “Solution Structures of the C-Terminal Headpiece Subdomains of Human Villin and Advillin, Evaluation of Headpiece F-Actin-Binding Requirements”. *Protein Science*, vol. 13, May 2004, pp. 1276–1287

[Vog14] Vogel, T.; Li, Y. W.; Wüst, T.; Landau, D. P. “Scalable replica-exchange framework for Wang-Landau sampling”. *Physical Review E*, vol. 90-2, Aug 2014, pp. 023302.

[Wan11] Wang, L.; Friesner, R. A.; Berne, B. J. “Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2)”. *The Journal of Physical Chemistry B*, vol. 115-30, Jun 2011, pp. 9431-9438.

[Was17a] Washington University - Biology Department. “Peptide Bonds and Protein Structure”. Captured on: http://www.nslc.wustl.edu/courses/bio2960/labs/02Protein_Structure/PS2011.htm, May 2017.

[Was17b] Washington University - Biology Department. “The Molecular Biology of Sickle Cell Anemia”. Captured on: http://www.nslc.wustl.edu/sicklecell/part2/molecular.html, May 2017.

[Wol97] Wolynes, P. G. “Folding funnels and energy landscapes of larger proteins within the capillarity approximation”. *Proceedings of the National Academy of Sciences*, vol. 94-12, Jun 1997, pp. 6170-6175.

[Xia15] Xia, J.; Flynn, W. F.; Gallicchio, E.; Zhang, B. W.; He, P.; Tan, Z., Levy, R. M. “Large Scale Asynchronous and Distributed Multi-Dimensional Replica Exchange Molecular Simulations and Efficiency Analysis”. *Journal of Computational Chemistry*, vol. 36-23, Sept 2015, pp. 1772–1785.

[Xu00] Xu, Y.; Xu, D. “Protein threading using PROSPECT: design and evaluation”. *Proteins*, vol. 40-3, Aug 2000, pp. 343–354.

[Xu08] Xu, W.; Mu, Y. “Ab initio folding simulation of Trpcage by replica exchange with hybrid Hamiltonian”. *Biophysical Chemistry*, vol. 137-(2-3), Oct 2008, pp. 116–125.

[Xu12] Xu, D.; Zhang, Y. “Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field”. *Proteins*, vol. 80-7, Jul 2012, pp. 1715–1735.

[Xue15] Xue, X.; Yongjun, W.; Zhihong, L. “Folding of SAM-II riboswitch explored by replica-exchange molecular dynamics simulation”. *Journal of Theoretical Biology*, vol. 365-1, Jan 2014, pp. 265–269.

[Yam13] Yamamori, Y.; Kitao, A. “MuSTAR MD: Multi-scale Sampling Using Temperature Accelerated and Replica Exchange Molecular Dynamics”. *The Journal of Chemical Physics*, vol. 139-14, Oct 2013, pp. 145105.

[Yan08] Yang, Y.; Zhou, Y. “Specific interactions for ab initio folding of protein terminal regions with secondary structures”. *Proteins*, vol. 72-2, Feb 2008, pp. 793–803.

[Yan11] Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y. “Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates”. *Bioinformatics*, vol. 27-15, Aug 2011, pp. 2076–2082.

[Yan15] Yang, M.; Huang, J.; MacKerell, A. D. “Enhanced Conformational Sampling Using Replica Exchange with Concurrent Solute Scaling and Hamiltonian Biasing Realized in One Dimension”. *Journal of Chemical Theory and Computation*, vol. 11-6, May 2015, pp. 2855-2867.

[You03] Young, M. R.; Pande, V. S. “Multiplexed-replica exchange molecular dynamics method for protein folding simulation”. *Biophysical Journal*, vol. 84-21, Feb 2003, pp. 775–786.

[Yu15] Yu, Y.; Wang, J.; Shao, Q.; Shi, J.; Zhu, W. “Increasing the sampling efficiency of protein conformational transition using velocity-scaling optimized hybrid explicit/implicit solvent REMD simulation”. *The Journal of Chemical Physics*, vol. 142-12, Mar 2015, pp. 125105.

[Yu16] Yu, T.; Lu, J.; Abrams, C. F.; Vanden-Eijnden, E. “Multiscale Implementation of Infinite-Swap Replica Exchange Molecular Dynamics”. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113-42, Oct 2016, pp. 11744–11749.

[Zag02] Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. “Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide distributed computing”. *Journal of Molecular Biology*, vol. 323-5, Nov 2002, pp. 927–937.

[Zem03] Zemla A. "LGA - a Method for Finding 3D Similarities in Protein Structures", *Nucleic Acids Research*, vol.31-13, Jul 2003, pp. 3370-3374.

[Zem99] Zemla, A.; Venclovas, Č.; Moult, J.; Fidelis, K. “Processing and analysis of CASP3 protein structure predictions”. *Proteins*, vol. 37-S3, Nov 1999, pp. 22–29.

[Zha04] Zhang, Y.; Skolnick, J. “Automated structure prediction of weakly homologous proteins on a genomic scale”. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101-20, May 2004, pp. 7594–7599.

[Zha05] Zhang, Y.; Arakaki, A. K.; Skolnick, J. R. “TASSER: An automated method for the prediction of protein tertiary structures in CASP6”. *Proteins*, vol. 61-S7, Sept 2005, pp. 91–98.

[Zha07] Zhang, J.; Lin, M.; Chen, R.; Liang, J.; Liu, J. S. “Monte carlo sampling of near-native structures of proteins with applications”. *Proteins*, vol. 66-1, Jan 2007, pp. 61–68

[Zha10] Zhang, J.; Zhang, Y. “A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction”. *PLOS ONE*, vol. 5-10, Oct 2010, pp. 15386.

[Zha14] Zhang, W. Chen, J. “Accelerate Sampling in Atomistic Energy Landscapes Using Topology-Based Coarse-Grained Models”. *Journal of Chemical Theory and Computation*, vol. 10-3, Jan 2014, pp. 918-923.

[Zha15] Zhang, Y.; Sagui, C. “Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI”. *Journal of Molecular Graphics and Modelling*, 55-1, Feb 2015, pp. 72–84.

[Zha16] Zhang, G.; Yu, X.; Zhou, X.; Hao, X. "A population-based conformational optimal algorithm using replica-exchange in ab-initio protein structure prediction" In: 2016 Chinese Control and Decision Conference (CCDC), 2016, pp. 701-706.

[Zhe11] Zheng, W.; Gallicchio, E.; Deng, N.; Andrec, M.; Levy, R. M. “Kinetic Network Study of the Diversity and Temperature Dependence of Trp-Cage Folding Pathways: Combining Transition Path Theory with Stochastic Simulations”. *The Journal of Physical Chemistry B*, vol. 115-6, Feb 2011, pp. 1512–1523.

[Zho02] Zhou, H.; Zhou, Y. “Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction”. *Protein Science : A Publication of the Protein Society*, vol. 11-11, Nov 2002, pp. 2714–2726.

[Zho11] Zhou, H.; Skolnick, J. “GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction”. *Biophysical Journal*, vol. 101-8, Oct 2011, pp. 2043–2052.

[Zho12] Zhou, T.; Caflisch, A. “Free Energy Guided Sampling”. *Journal of Chemical Theory and Computation*, vol. 8-6, Apr 2012, pp. 2134-2140.

[Zim17] Zimmerman, M.; Snow, B. “Defining Proteins”. Captured on: http://catalog.flatworldknowledge.com/bookhub/reader/3728?e=zimmerman_1.0-ch06_s01, May 2017.

[Zve08] Zvelebil, M. J.; Baum, J. O. “Understanding Bioinformatics”. Garland Science, 2008, 772p.

[Zyg17] Zygomatic. “Chart Tool”. Captured on: https://www.onlinecharttool.com/, Jun 2017.

APPENDIX A - CUMULATIVE DISTRIBUTION PLOTS OF THE TOP GDT_TS BANDS FOR EACH REMD SIMULATION OF TEST DATASET

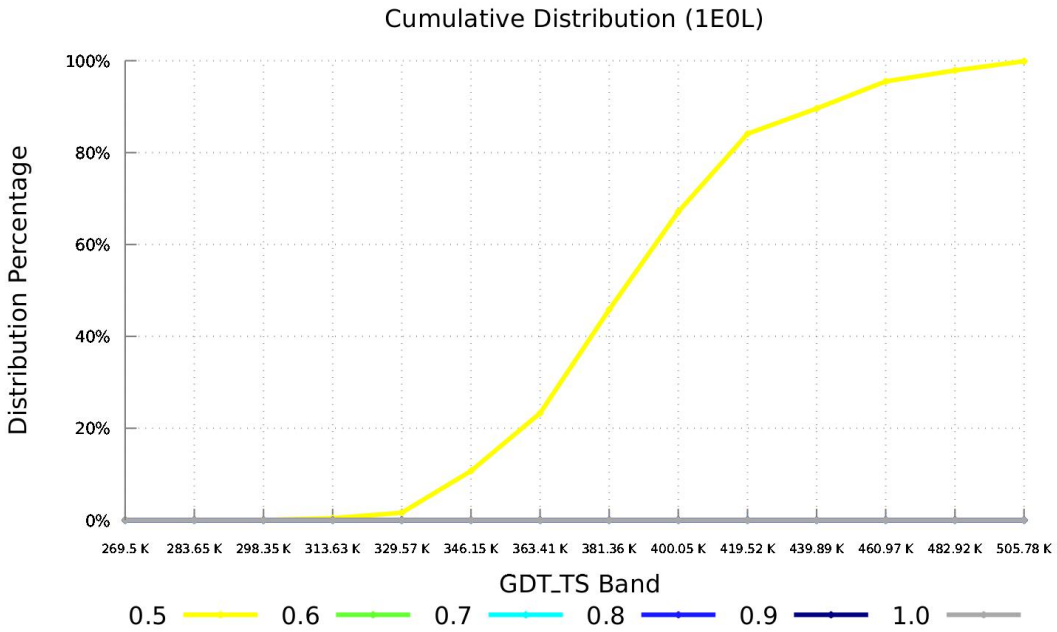

Figure A.1 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1E0L.

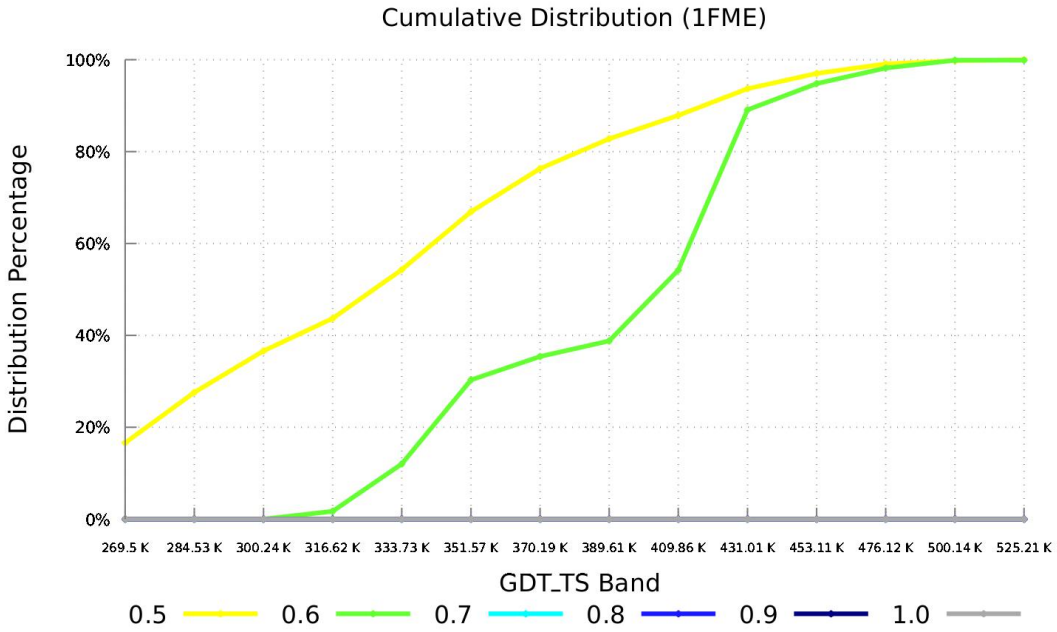

Figure A.2 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1FME.

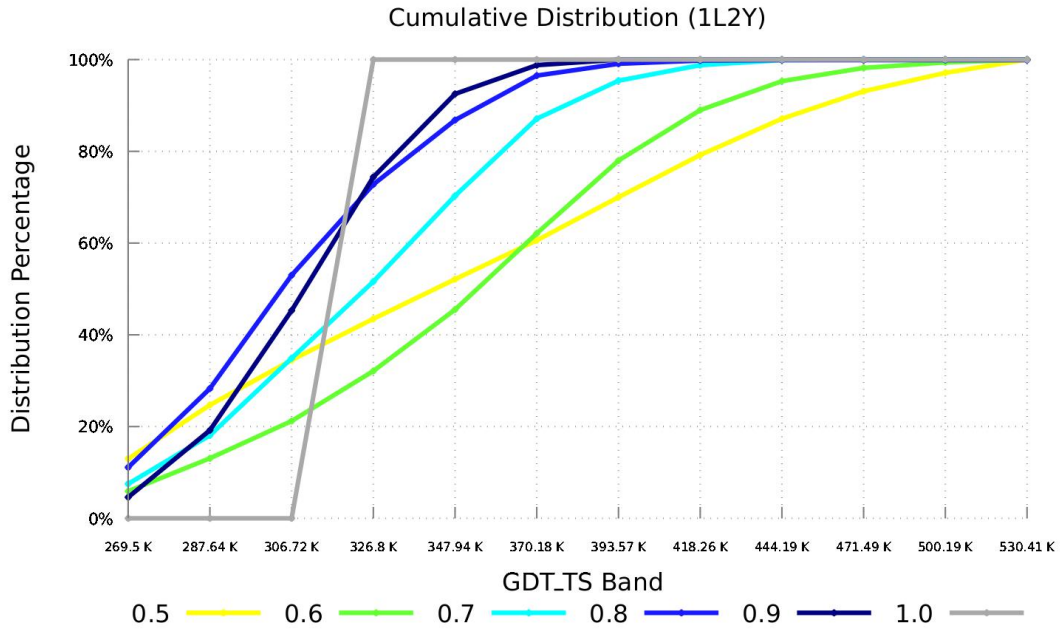

Figure A.3 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1L2Y.

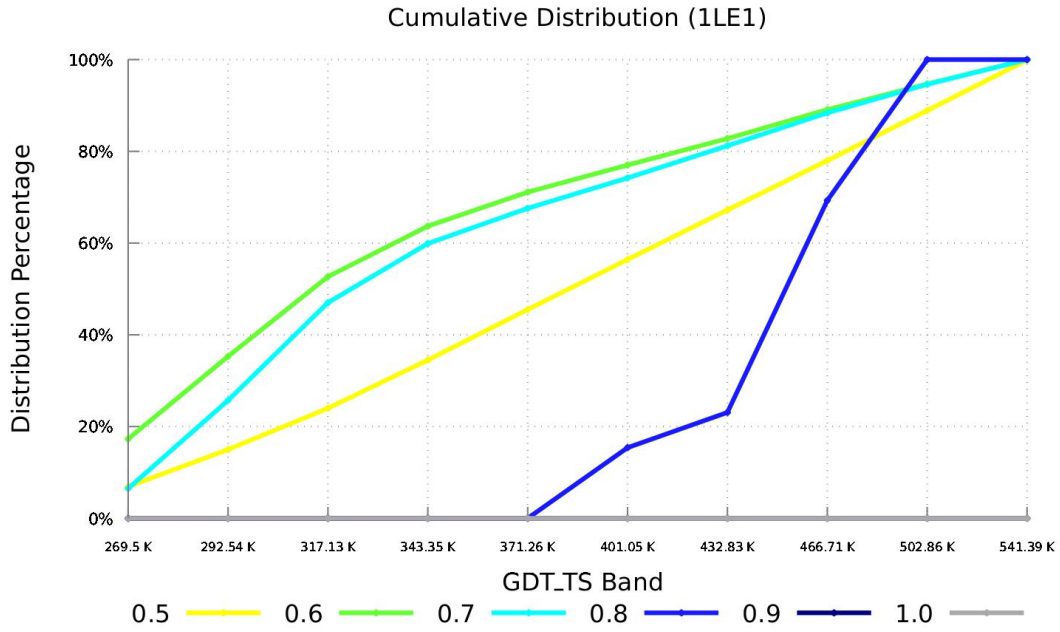

Figure A.4 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1LE1.

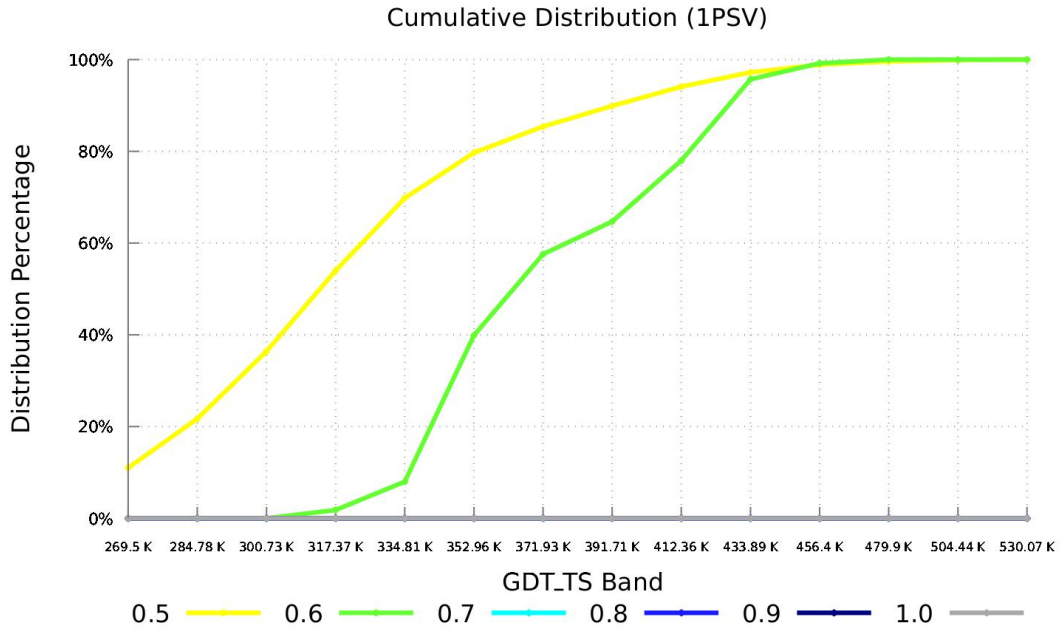

Figure A.5 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1PSV.

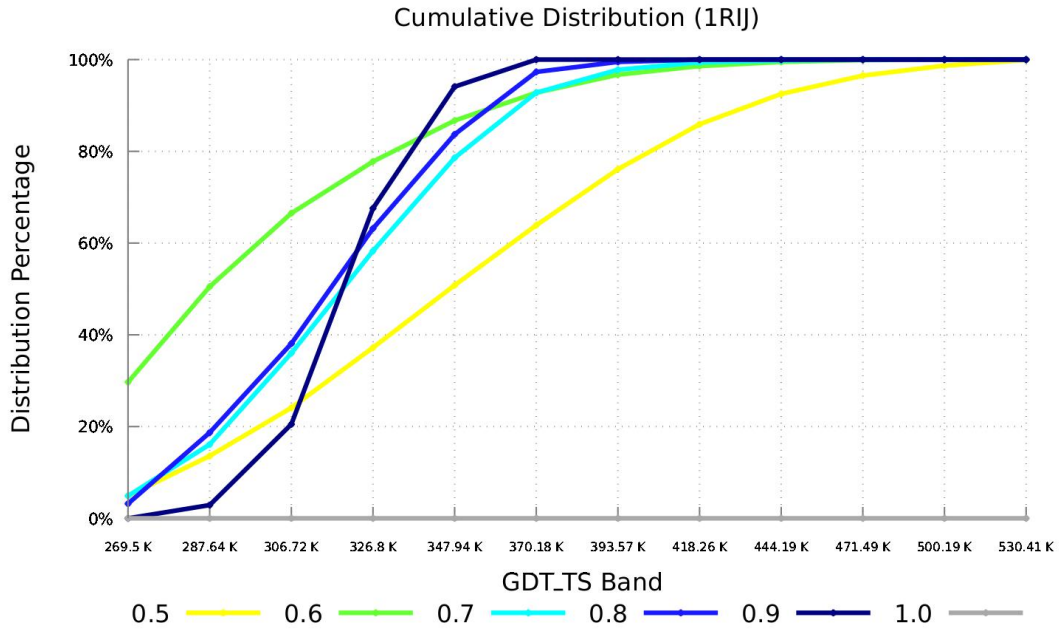

Figure A.6 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1RIJ.

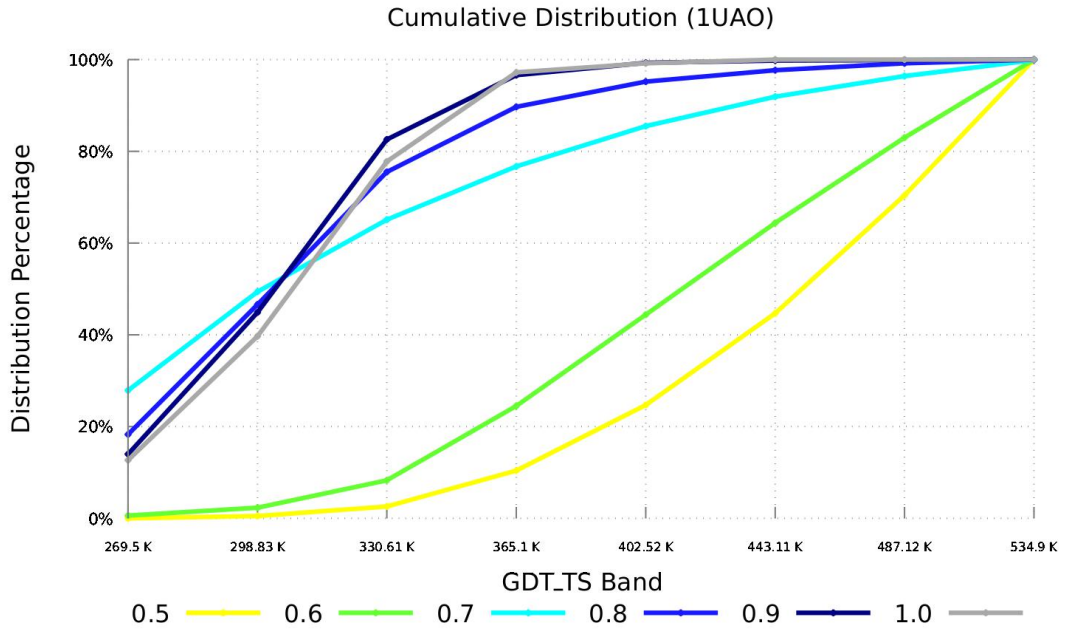

Figure A.7 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1UAO.

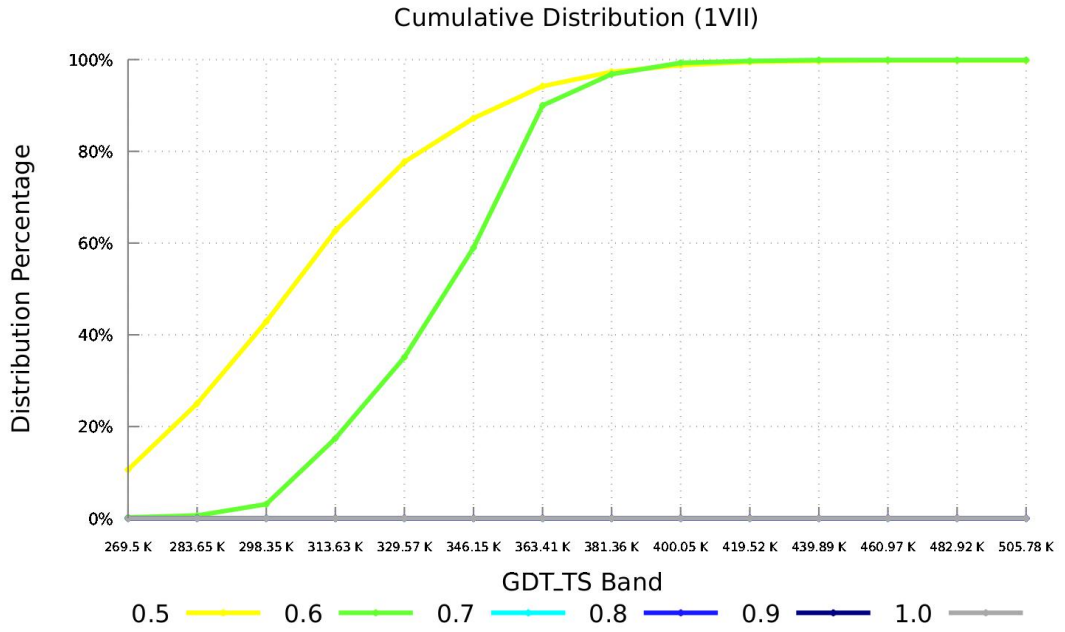

Figure A.8 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 1VII.

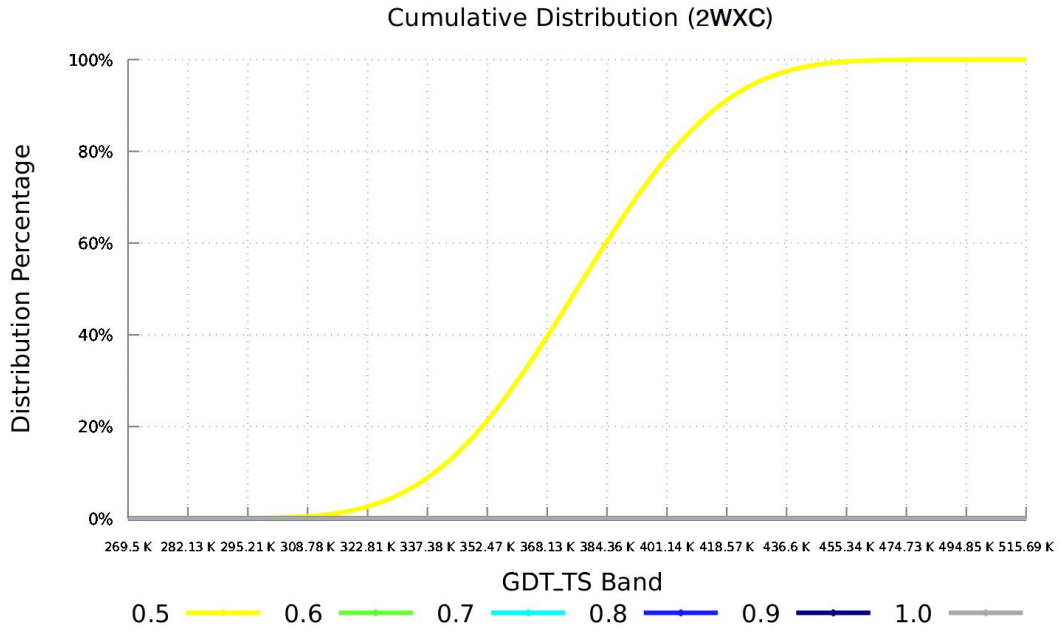

Figure A.9 - Cumulative distribution of the top scored structures predicted according to GDT_TS versus the temperature in which they were extracted. Data relative to the REMD PSP simulation of the protein 2WXC.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria Acadêmica
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: proacad@pucrs.br
Site: www.pucrs.br/proacad