

O problema de predição de estrutura de proteínas é atualmente um dos problemas em aberto mais importantes da Ciência da Computação e Bioinformática. Este trabalho apresenta uma abordagem que combina diversos recursos científicos e tecnológicos direcionados a solução do problema. Dentre eles, pode-se destacar o uso de inserção de fragmentos como um método de incorporar informação do domínio do problema, a utilização da rotina de Hooke-Jeeves para efetuar buscas locais e clusterização de conformações guiadas por RMSD. O método proposto foi avaliado com as 10 proteínas mais utilizadas na literatura e os resultados obtidos foram competitivos com o estado da arte.

Orientador: Prof. Dr. Rafael Stubs Parpinelli

JOINVILLE, 2019

ANO 2019 RENAN SAMUEL DA SILVA

A SELF ADAPTIVE GREEDY EVOLUTIONARY ALGORITHM USING MONTE CARLO FRAGMENT INSERTION AND CONFORMATION CLUSTERING



UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
APLICADA – PPGCA

DISSERTAÇÃO DE MESTRADO

**A SELF ADAPTIVE GREEDY EVOLUTIONARY
ALGORITHM USING MONTE CARLO FRAGMENT
INSERTION AND CONFORMATION CLUSTERING**

RENAN SAMUEL DA SILVA

JOINVILLE, 2019

**SANTA CATARINA STATE UNIVERSITY - UDESC
CENTER OF TECHNOLOGICAL SCIENCES - CCT
GRADUATE PROGRAM IN APPLIED COMPUTING - PPGCA**

RENAN SAMUEL DA SILVA

**A SELF ADAPTIVE GREEDY EVOLUTIONARY ALGORITHM USING
MONTE CARLO FRAGMENT INSERTION AND CONFORMATION
CLUSTERING**

JOINVILLE

2019

RENAN SAMUEL DA SILVA

**A SELF ADAPTIVE GREEDY EVOLUTIONARY ALGORITHM USING
MONTE CARLO FRAGMENT INSERTION AND CONFORMATION
CLUSTERING**

Master thesis submitted to the Computer Science Department at the College of Technological Science of Santa Catarina State University in fulfillment of the partial requirement for the Master's degree in Applied Computing.

Advisor: Dr. Rafael Stubs Parpinelli

JOINVILLE

2019

**Ficha catalográfica elaborada pelo programa de geração automática da
Biblioteca Setorial do CCT/UDESC,
com os dados fornecidos pelo(a) autor(a)**

Silva, Renan Samuel da
A SELF ADAPTIVE GREEDY EVOLUTIONARY
ALGORITHM USING MONTE CARLO FRAGMENT
INSERTION AND CONFORMATION CLUSTERING / Renan
Samuel da Silva. -- 2019.
158 p.

Orientador: Rafael Stubs Parpinelli
Dissertação (mestrado) -- Universidade do Estado de
Santa Catarina, Centro de Ciências Tecnológicas, Programa
de Pós-Graduação em Computação Aplicada, Joinville, 2019.

1. Monte Carlo. 2. Inserção de fragmento. 3. Problema de
Predição de Estrutura de Proteínas. 4. Controle de
Parâmetros. 5. Clusterização. I. Parpinelli, Rafael Stubs. II.
Universidade do Estado de Santa Catarina, Centro de
Ciências Tecnológicas, Programa de Pós-Graduação em
Computação Aplicada. III. Título.

**A SELF ADAPTIVE GREEDY EVOLUTIONARY ALGORITHM USING MONTE
CARLO FRAGMENT INSERTION AND CONFORMATION CLUSTERING**

por

Renan Samuel da Silva

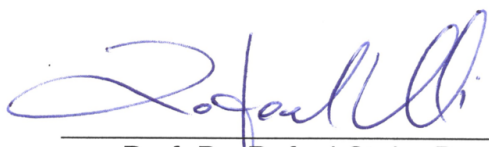
Esta dissertação foi julgada adequada para obtenção do título de

Mestre em Computação Aplicada

Área de concentração em “Ciência da Computação”,
e aprovada em sua forma final pelo

CURSO DE MESTRADO ACADÊMICO EM COMPUTAÇÃO APLICADA
DO CENTRO DE CIÊNCIAS TECNOLÓGICAS DA
UNIVERSIDADE DO ESTADO DE SANTA CATARINA.

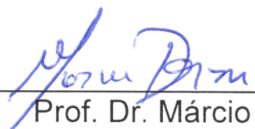
Banca Examinadora:



Prof. Dr. Rafael Stubs Parpinelli
CCT/UDESC (Orientador/Presidente)



Prof. Dr. Chidambaram Chidambaram
CCT/UDESC



Prof. Dr. Márcio Dorn
UFRGS

Joinville, SC, 30 de agosto de 2019.

To my Father that is no longer with us and
to my Mother that always stood by me.

ACKNOWLEDGEMENTS

Firstly I would like to thank my mother, Tatiana. Without her continuous and never ending support I would not been here. Not only did she financially and emotionally support me, she also did encourage me since I was very young to study and go after my dreams. This shaped me into who I am.

I also would like to deeply thank Rafael Parpinelli for supporting me over this journey and helping me push the limits of my knowledge. I also thank all my friends that were with me since my first day at the university. In special Andressa Umetsu, Rafael Castro, Vinicius Zuchi, Tiago Heinrich, Aurelio Grott, Mateus Boiani, William Pereira, Karll Henning, Dalmo Neto, Lucas Zanatta and many others. They all made my days more bearable and fun. To all the teachers that served both as friends and a role models, in special professor Omir. I had the pleasure of working for more than two years with him. The experience that I gained while working with him helped define who I am today as an academic. To my team leads at JobScore, Thiago and Wesklei, who can understand what is like to be working and doing a master degree at the same time and helped along the path as well.

Finally I would like to thank Andressa Umetsu for becoming one of the closest friends that I have. Her friendship made me a happier person.

“It is not because things are difficult that we do not dare, but because we do not dare, things are difficult.”

Seneca, Moral letters to Lucilius/Letter 104,
verse 26

ABSTRACT

The Protein Structure Prediction Problem (PSPP) is currently one of the most important and challenging open problems in both computer science and in structural bioinformatics. Being able to accurately predict protein conformations would have a significant impact on several fields. Proteinopathies would be better understood, intelligent protein based drugs could be designed more easily and the overall understanding of the protein functions in organisms would be further increased. Even with the problem being several decades old, no viable general solution was found. Currently, only small protein are able to be predicted with satisfactory levels of accuracy. As such, this work has as its main goal to attempt in improving the prediction power of ab initio methods by utilizing an self adaptive evolution algorithm using Monte Carlo based fragment insertion and conformational clustering. With this, a proven meta heuristic is utilized as the core of the conformation sampling process with fragment insertion feeding domain specific information into the process. The online parameter control allows the method to adapt to proteins of different features and also to different stages of the optimization process. The main contribution of this work is a novel combination of online parameter control, fragment insertion, conformational clustering and metaheuristics. In order to assess the performance of the proposed methods they were compared against each other and two methods from the literature using a series of statistical methods. The proposed methods were then compared against several methods in the literature, utilizing only the RMSD. Finally, it was concluded that the proposed methods were able to outperform Rosetta, one of the state-of-art methods. Furthermore, the RMSD of the protein targets were either the best one so far, or the second best in the literature, over more than 20 works reviewed.

Keywords: Monte Carlo, Fragment Insertion, Protein Structure Prediction Problem,

Parameter Control, Clustering.

RESUMO

O Problema de Predição de Estrutura de Proteína é atualmente um dos mais importantes e desafiadores problemas em aberto na ciência da computação e na bioinformática estrutural. Ser capaz de prever a estrutura de proteínas com acurácia teria um impacto significativo em múltiplas áreas do conhecimento. Proteinopatias seriam melhor entendidas, drogas inteligentes baseadas em proteínas poderiam ser projetadas mais facilmente e o conhecimento em geral de proteínas nos seres vivos seria expandido. Apesar do problema de predição ter várias décadas, até hoje nenhuma solução generalizada viável foi encontrada. Atualmente, apenas pequenas proteínas podem ter sua estrutura predita com um nível desejável de acurácia. Deste modo, este trabalho tem como principal objetivo tentar melhorar o poder preditivo de métodos *ab initio* através da utilização de um algoritmo evolucionário auto adaptativo usando inserção de fragmentos baseada em Monte Carlo e clusterização de conformações. Com isto, uma poderosa meta heurística será utilizada como núcleo para o processo de amostragem de conformações com a inserção de fragmentos agregando informação do domínio do problema. O controle online de parâmetros permitirá que o método se adapte a proteínas de diferentes características e também a diferentes etapas do processo de otimização. A principal contribuição deste trabalho está na inovativa proposta de combinar controle online de parâmetros, inserção de fragmentos, clusterização de conformações e meta heurísticas. Com o objetivo de aferir a performance dos métodos propostos, estes foram comparados contra dois outros métodos da literatura, utilizando uma série de métodos estatísticos. Os métodos propostos foram então comparados contra diversos outros na literatura, utilizando apenas o RMSD. Foi concluído que os métodos propostos foram capazes de obter performance superior a Rosetta, um dos métodos estado da arte. Finalmente, o RMSD das proteínas alvo foram ou a melhor ou a segunda melhor encontradas na literatura, dentre mais de 20 trabalhos analisados.

Palavras-chave: Monte Carlo, Inserção de fragmento, Problema de Predição de Es-

trutura de Proteínas, Controle de Parâmetros, Clusterização.

LIST OF FIGURES

Figure 1 – Amino Acid Structure (Source: (GARRETT; GRISHAM, 1999))	32
Figure 2 – The Peptide Bond (Source: (GARRETT; GRISHAM, 1999))	34
Figure 3 – Dihedral angles of an amino acid (Source: Author)	34
Figure 4 – An α -helix (Source: (LODISH et al., 2008))	36
Figure 5 – A parallel β -sheet (Source: (GARRETT; GRISHAM, 1999))	37
Figure 6 – An anti-parallel β -sheet (Source: (GARRETT; GRISHAM, 1999)) . .	37
Figure 7 – An anti-parallel pleated-sheet (Source: (GARRETT; GRISHAM, 1999))	38
Figure 8 – The native conformation of protein 1CTF (Source: RCSB PDB) . . .	39
Figure 9 – The quaternary structure of the hemoglobin. Each color denotes one protein chain. The green point in the middle marks the point of radial symmetry. (Source: RCSB PDB)	39
Figure 10 – 19 aligned samples of a protein determined by Nuclear Magnetic Resonance (NMR) (SOURCE (ADAPTED): (WUTHRICH, 1989)) . .	47
Figure 11 – The Protein Computational Model (Source: Author)	64
Figure 12 – The Proposed Search Procedure (Source: Author)	68
Figure 13 – Boxplot presenting the RMSD for the protein predictions with the competing methods. For the two proposed methods, the best-by-rmsd data was used	76
Figure 14 – Boxplot presenting the <code>scorefxn</code> for the protein predictions with the competing methods. For the two proposed methods, the best-by-energy data was used	76
Figure 15 – Pareto front of the methods using RMSD and Energy	81
Figure 16 – The best energy and average energy for proteins 1enh and 2mr9. The continuous line represents the best energy. The dashed line represents the average of the population. The X Axis represents the function evaluations and the Y Axis indicates the value of the <code>score3</code> energy function being optimized.	82
Figure 17 – The average pairwise RMSD for proteins 1rop and 1utg. The X Axis represents the function evaluations and the Y Axis indicates the value of the <code>score3</code> energy function being optimized.	83
Figure 18 – RMSD comparison of PPF-MC and PPF-REMC.	85
Figure 19 – Energy comparison of PPF-MC and PPF-REMC.	86
Figure 20 – The predicted conformations (in green/light gray) compared to the native conformation (in red/darker gray). The RMSD between the predicted and native conformation is show between parenthesis. . .	95

Figure 23 – Predicted Conformation of 1acw (green, light) over the Native Conformation	151
Figure 24 – Predicted Conformation of 1ail (green, light) over the Native Conformation	152
Figure 25 – Predicted Conformation of 1crn (green, light) over the Native Conformation	153
Figure 26 – Predicted Conformation of 1enh (green, light) over the Native Conformation	154
Figure 27 – Predicted Conformation of 1l2y (green, light) over the Native Conformation	155
Figure 28 – Predicted Conformation of 1rop (green, light) over the Native Conformation	155
Figure 29 – Predicted Conformation of 1utg (green, light) over the Native Conformation	156
Figure 30 – Predicted Conformation of 1wqc (green, light) over the Native Conformation	157
Figure 31 – Predicted Conformation of 1zdd (green, light) over the Native Conformation	158
Figure 32 – Predicted Conformation of 2mr9 (green, light) over the Native Conformation	158

LIST OF TABLES

Table 1 – The 20 Naturally Occurring Amino Acids	32
Table 2 – Side Chain Angles	35
Table 3 – Comparison of the works in the literature. FI means Fragment In- sertion. OPC means Parameter Control. LS means local search. CC means Conformation Clustering. KD means Knowledge Domain. A dash indicates that the relevant criteria in the column was not met. The last line categorizes the methods proposed in this work	61
Table 4 – The Machine Setup	72
Table 5 – Target proteins and their features	73
Table 6 – Parameters utilized in the proposed methods	74
Table 7 – Kruskal-Wallis for best-by-rmsd using RMSD	78
Table 8 – Kruskal-Wallis for best-by-energy using scorefxn	78
Table 9 – Summary of Mann-Whitney best-by-rmsd using RMSD compared against the previous method	78
Table 10 – Summary of Mann-Whitney best-by-energy using scorefxn compared against the previous method	78
Table 11 – Summary of Mann-Whitney best-by-rmsd using RMSD compared against rosetta	79
Table 12 – Summary of Mann-Whitney best-by-energy using scorefxn compared against rosetta	79
Table 13 – Summary of the method in the Pareto front	80
Table 14 – RMSD change caused by repacking	86
Table 15 – Summary of Mann-Whitney using score3, using the energy before repacking. The two proposed methods are compared against rosetta.	87
Table 16 – Processing times, in seconds, for PPF-MC and PPF-REMC	89
Table 17 – Function evaluations spent on Hooke-Jeeves, for PPF-MC and PPF- REMC	89
Table 18 – A comparison of the RMSD from the best prediction	91
Table 19 – GDT-TS for PPF-MC and PPF-REMC	93
Table 20 – TM Score for PPF-MC and PPF-REMC	93

Table 21 – Summary of rosetta, the two proposed methods and sade-remc results. The RMSD and scorefxn is considered. Only the first five proteins are presented. First column shows the protein name. Second column shows the method name. Third and fourth column presents RMSD and energy. The first value corresponds to the best (lowest) value found in 50 runs. The values between parenthesis are the mean and the standard deviation.	121
Table 22 – Summary of rosetta, the two proposed methods and sade-remc results. The RMSD and scorefxn is considered. Only the last five proteins are presented. First column shows the protein name. Second column shows the method name. Third and fourth column presents RMSD and energy. The first value corresponds to the best (lowest) value found in 50 runs. The values between parenthesis are the mean and the standard deviation.	122
Table 23 – Shapiro wilk for best-by-rmsd using scorefxn	123
Table 24 – Shapiro wilk for best-by-rmsd using RMSD	123
Table 25 – Shapiro wilk for best-by-energy using scorefxn	124
Table 26 – Shapiro wilk for best-by-energy using RMSD	124
Table 27 – Mann-Whitney for 1acw best-by-rmsd using RMSD	125
Table 28 – Mann-Whitney for 1ail best-by-rmsd using RMSD	125
Table 29 – Mann-Whitney for 1crn best-by-rmsd using RMSD	126
Table 30 – Mann-Whitney for 1enh best-by-rmsd using RMSD	126
Table 31 – Mann-Whitney for 1l2y best-by-rmsd using RMSD	126
Table 32 – Mann-Whitney for 1rop best-by-rmsd using RMSD	126
Table 33 – Mann-Whitney for 1utg best-by-rmsd using RMSD	127
Table 34 – Mann-Whitney for 1wqc best-by-rmsd using RMSD	127
Table 35 – Mann-Whitney for 1zdd best-by-rmsd using RMSD	127
Table 36 – Mann-Whitney for 2mr9 best-by-rmsd using RMSD	127
Table 37 – Mann-Whitney for 1acw best-by-energy using scorefxn	129
Table 38 – Mann-Whitney for 1ail best-by-energy using scorefxn	129
Table 39 – Mann-Whitney for 1crn best-by-energy using scorefxn	130
Table 40 – Mann-Whitney for 1enh best-by-energy using scorefxn	130
Table 41 – Mann-Whitney for 1l2y best-by-energy using scorefxn	130
Table 42 – Mann-Whitney for 1rop best-by-energy using scorefxn	130
Table 43 – Mann-Whitney for 1utg best-by-energy using scorefxn	131
Table 44 – Mann-Whitney for 1wqc best-by-energy using scorefxn	131
Table 45 – Mann-Whitney for 1zdd best-by-energy using scorefxn	131
Table 46 – Mann-Whitney for 2mr9 best-by-energy using scorefxn	131

Table 47 – The 50 individual GDT-TS measurements for PPF-MC over all ten proteins.	144
Table 48 – The 50 individual GDT-TS measurements for PPF-REMC over all ten proteins.	145
Table 49 – The 50 individual TM-Score measurements for PPF-MC over all ten proteins.	148
Table 50 – The 50 individual TM-Score measurements for PPF-REMC over all ten proteins.	149

LIST OF ABBREVIATIONS AND ACRONYMS

AMBER	Assisted Model Building with Energy Refinement
ABC	Artificial Bee Colony
ACO	Ant Colony Optimization
APL	Angle Probability List
CASP	Critical Assessment of Structure Prediction
CC	Conformation Clustering
CHARMM	Chemistry at HARvard Macromolecular Mechanics
DE	Differential Evolution
DNA	Deoxyribonucleic Acid
EDA	Estimation of Distribution Algorithm
EA	Evolutionary Algorithms
FA	Firefly Algorithm
FCC	Face Centered Cubic
FFI	Forced Fragment Insertion
FI	Fragment Insertion
GA	Genetic Algorithms
GDT-TS	Global Distance Test Total Score
GRASP	Greedy Randomized Adaptive Search Procedure
GROMOS	GRoningen MOlecular Simulation
HM	Homology Modeling
HP	Hydrophobic-Polar
ILS	Iterative Local Search
LSOP	Large Scale Optimization Problem
MH	Meta Heuristics
MC	Monte Carlo
MCMC	Monte Carlo Markov Chain

NMR	Nuclear Magnetic Resonance
PDB	Protein Databank
PSO	Particle Swarm Optimization
PSPP	Protein Structure Prediction Problem
PSP	Protein Structure Prediction
PCA	Principal Component Analysis
PPF	Protein Prediction Framework
RNA	Ribonucleic Acid
REU	Rosetta Energy Unit
REMC	Replica Exchange Monte Carlo
RMSD	Root-Mean-Square Deviation
SaDE	Self Adaptive Differential Evolution
SaEA	Self Adaptive Evolutionary Algorithm
SASA	Solvent-accessible surface area
SVD	Single Variable Decomposition
SA	Simulated Annealing
SI	Swarm Intelligence
TM	Thread Modeling
TM-Score	Thread Modeling Score
TSP	Traveling Salesman Problem

LIST OF SYMBOLS

NP	The population size used in populational algorithms
1ZDD	Disulfide-Stabilized mini protein A domain
1CRN	Crambin
1ENH	DNA binding protein of Drosophila
1AIL	Influenza A RNA binding protein
1A3N	Deoxy Human Hemoglobin
1CTF	Ribosomal Protein from E. Coli
1PLW	Methionine-Enkephalin
4ZRY	Human Keratin
D	Problem dimensionality
F	DE's scaling factor
Cr	DE's crossover ratio
$\vec{w}$	The new solution vector
$\vec{x}_r$	Random solution vector
$\vec{x}_{best}$	Best solution vector
C_α	Alpha Carbon of the protein backbone
χ_i	The i -th dihedral angle of the protein side chain
ϕ	The dihedral angle in the $N - C_\alpha$ atom pair
ψ	The dihedral angle in the $C_\alpha - C$ atom pair
ω	The dihedral angle in the $C - N$ atom pair

CONTENTS

1	INTRODUCTION	25
1.1	MOTIVATION	29
1.2	OBJECTIVES	29
1.3	LIST OF PUBLICATIONS	29
1.4	DOCUMENT STRUCTURE	30
2	LITERATURE REVIEW	31
2.1	PROTEINS	31
2.1.1	Amino Acids	31
2.1.2	Protein Structures	35
2.1.3	Protein Secondary Structure Prediction	39
2.2	PROTEIN STRUCTURE PREDICTION	40
2.2.1	Ab Initio Methods	42
2.2.1.1	<i>The Rosetta Suite</i>	44
2.2.2	Knowledge Based Methods	45
2.2.3	Evaluating Protein Errors	46
2.3	OPTIMIZATION	48
2.3.1	Parameter Control in EAs	50
2.3.2	Self-Adaptive Differential Evolution	52
2.3.3	Monte Carlo Methods	54
2.3.4	Hooke-Jeeves Local Search	55
2.4	CLUSTERING	56
2.5	RELATED WORK	57
3	A HYBRID METHOD BASED ON MONTE CARLO FRAGMENT SAM- PLING FOR PROTEIN STRUCTURE PREDICTION	63
3.1	COMPUTATIONAL REPRESENTATION	63
3.2	ENERGY FUNCTIONS EMPLOYED	64
3.3	CONFORMATION SAMPLING	65
4	EXPERIMENTS, RESULTS AND ANALYSIS	71
4.1	DESIGN OF EXPERIMENTS	71
4.2	ENERGY AND RMSD ANALYSIS	74
4.2.1	Pareto Front Analysis	80
4.3	CONVERGENCE AND DIVERSITY ANALYSIS	82

4.4	MONTE CARLO VS REPLICA EXCHANGE MONTE CARLO	84
4.5	REPACKING IMPACT	85
4.6	PROCESSING TIME AND FUNCTIONS EVALUATIONS	88
4.7	COMPARISON WITH COMPETING METHODS	90
4.8	GDT-TS AND TM-SCORE METRICS	92
4.9	VISUAL REPRESENTATION OF THE PREDICTIONS	94
5	FINAL CONSIDERATIONS AND FUTURE WORK	97
5.1	FUTURE WORKS	99
	BIBLIOGRAPHY	101
	APPENDIX A – PROTEIN PRE PROCESSING	113
A.1	FRAGMENT GENERATION CODE	113
A.2	FRAGMENT PICKER	118
A.3	FRAGMENT PICKER CONFIGURATION	119
	APPENDIX B – SUMMARY OF RMSD AND ENERGY RESULTS	121
	APPENDIX C – SHAPIRO-WILK TEST RESULTS	123
	APPENDIX D – MANN-WHITNEY TEST RESULTS FOR RMSD	125
	APPENDIX E – MANN-WHITNEY TEST RESULTS FOR SCOREFXN129	125
	APPENDIX F – CONVERGENCE GRAPHS FOR ALL PROTEINS	133
	APPENDIX G – GDT-TS RAW DATA	143
	APPENDIX H – TM-SCORE RAW DATA	147
	APPENDIX I – GRAPHICAL REPRESENTATION OF THE PREDICTED CONFORMATIONS	151