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

UDESC

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 predizer a estrutura de proteínas com acurácia teria um impacto significativo em multiplas áread do conhecimento. Proteinopatias seriam melhor entendidas, drogas inteligentes baseadas em proteínas poderia ser projetadas mais facilmente e o conhecimento em geral de proteínas nos seres vivos seria expandido. Apesar do problema de predição ter várias decadas, até hoje nenhuma solução generalizada viável foi encontrada. Atualmente, apenas pequenas proteínas podem ter sua estrutura predita com um nível desejável de acurária. 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 inserçao 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 parametros permitirá que o método se adapte a proteínas de diferentes caracteristicas e também a diferentes etapas do processo de otimização. A principal contribuição deste trabalho está na inovativa proposta de combinar controle online de parametros, 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 literuatura, utilizando apenas o RMSD. Foi concluído que os métodos propostos foram capazes the obter performance superior a Rosetta, um os 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 Estrutura de Proteínas, Controle de Parâmetros, Clusterização.

LIST OF FIGURES

LIST OF TABLES

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 Assessement 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 Inteligence
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