Pontifícia Universidade Católica
do Rio de Janeiro

Edwin Germán Maldonado Távara

Algoritmo Genético Multiobjetivo na Predição de Estruturas Proteicas no Modelo Hidrofóbico - Polar

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Marley María Bernárdez Rebuzzi Vellasco
Co-orientador: André Vargas Abs da Cruz

Rio de Janeiro

Abril de 2012

Edwin Germán Maldonado Távara

Algoritmo Genético Multiobjetivo na Predição de Estruturas Proteicas no Modelo Hidrofóbico - Polar

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Marley María Bernárdez Rebuzzi Vellasco
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Dr. André Vargas Abs da Cruz
Co-Orientador
ICA/DEE/PUC-Rio

Profa. Karla Tereza Figueiredo Leite
UESO

Prof. André Carlos Ponce de Leon Ferreira de Carvalho
USP

Prof. Fabio Lima Custódio
LNCC

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de Janeiro, 16 de Abril de 2012

Edwin Germán Maldonado Távara

Graduou-se em Engenharia de Computação pela Universidad Nacional de Trujillo - Perú 2001.

Ficha Catalográfica

Maldonado Távara, Edwin Germán

Algoritmo genético multiobjetivo na predição de estruturas proteicas no modelo hidrofóbico polar / Edwin Germán Maldonado Távara; orientadores: Marley M. B. R. Vellasco, André Vargas Abs da Cruz. – 2012.
110 f. : il. (color.) ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2012.
Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Algoritmos genéticos multiobjetivo. 3. Modelo hidrofóbico-polar. 4. Predição de estruturas de proteínas. 5. Compactação da estrutura de proteínas. I. Vellasco, Marley M. B. R. II. Cruz, André Vargas Abs da. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

A minha mãe Luz.

Eli ri do medo, nada o assusta, ele não recua
diante da espada, o som da trompete não o deixa no lugar.
Filme Secretariat

Agradecimentos

O Deus por ter me guiado nesta jornada.

À CAPES e á PUC-Rio pelo apoio financeiro, sem os quais este trabalho não poderia ter sido realizado.

À Profa. Dra. Marley Maria B. R. Vellasco e ao Dr. André Vargas Abs da Cruz pelo apoio, incentivo e confiança depositada em mim.

Aos meus pais Luz e Germán por estarem sempre presentes ao meu lado, me apoiando-me ao longo da minha vida.

Ao meu irmão William por ser meu exemplo de vida e por ter me apoiado em toda minha educação.

Ao meu irmão David pelas orientações e a ajuda incondicional.

À minha irmã Nancy pela ajuda e bons conselhos.

À minha Ana Cecília pelo amor, carinho, apoio e, principalmente, pela paciência e compreensão ao longo da elaboração desta tese.

Ao meu amigo Gustavo pelo apoio e conversas sobre algoritmos genéticos e bioinformática que contribuiu ao desenvolvimento deste trabalho.

Ao meu amigo Harold, pela ajuda proporcionada na revisão da gramática desde trabalho.

Aos meus amigos da sala 604, pelo apoio e colaboração contínua.

Resumo

Távara, Edwin Germán Maldonado; Vellasco, Marley María Bernárdez Rebuzzi. **Algoritmo genético multiobjetivo na predição de estruturas proteicas no modelo Hidrofóbico - Polar.** Rio de Janeiro, 2012. 110 p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O problema da predição das estruturas de proteínas (*Protein Structure Prediction (PSP)*) é um dos desafios mais importantes na biologia molecular. Pelo fato deste problema ser muito difícil, têm sido propostos diferentes modelos simplificados para resolvê-lo. Um dos mais estudados é o modelo, Hidrofóbico-Polar (HP), o modelo HP fornece uma estimativa da energia da proteína com base na soma de interações entre pares de aminoácidos hidrofóbicos (contatos H-H). Entretanto, apesar das simplificações feitas no modelo HP, o problema permanece complexo, pertencendo à classe NP-Difícil. Muitas técnicas têm sido propostas para resolver este problema entre elas, técnicas baseadas em algoritmos genéticos. Em muitos casos, as técnicas baseadas em AG foram usadas com sucesso, mas, no entanto, abordagens utilizando AG muitas vezes não tratam adequadamente as soluções geradas, prejudicando o desempenho da busca. Além disso, mesmo que eles, em alguns casos, consigam atingir o mínimo de energia conhecido para uma conformação, estes modelos não levam em conta a forma da proteína um fator muito importante na hora de obter proteínas mais compactas. Foi desenvolvido um algoritmo genético multiobjetivo para PSP no modelo HP, de modo de avaliar de forma mais eficiente, as conformações produzidas. O modelo utiliza como avaliação uma combinação baseada no número de colisões, número de contatos hidrofóbicos, compactação dos aminoácidos hidrofóbicos e hidrofílicos, obtendo, desta forma estruturas mais naturais e de mínima energia. Os resultados obtidos demonstram a eficiência desse algoritmo na obtenção de estruturas proteicas compactas providenciando indicadores da compactação dos aminoácidos hidrofóbicos e hidrofílicos da proteína.

Palavras-chave

Algoritmos Genéticos Multiobjetivo; Modelo Hidrofóbico-Polar; Predição de Estruturas de Proteínas; Compactação da Estrutura de Proteínas.

Abstract

Távara, Edwin Germán Maldonado; Vellasco, Marley María Bernárdez Rebuzzi (Advisor). **Multiobjetive Genetic Algorithm for Predicting Protein Structures in Hydrophobic – Polar Model.** Rio de Janeiro, 2012. 110 p. Msc. Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The problem of protein structured prediction (PSP) is one of the most important challenges in molecular biology. Because this problem is very difficult, different simplified models have been proposed to solve it. One of the most studied is the Hydrophobic-Polar model HP this model provides an estimate of the protein energy based on the sum of hydrophobic contacts. However, despite the simplifications made in the HP model, the problem remains complex, belonging to the class of NP-Hard problems. Many techniques have been proposed to solve this problem as genetic algorithms. In many cases the GA techniques have been used successfully, but, however, with GA approaches often do not adequately address the generated solutions, impairing the performance of the search. Furthermore, in some cases would attain the minimum energy for a known conformation, these models do not take care the protein shape, a very important factor to obtain more compact proteins. This work developed a multiobjective genetic algorithm to PSP in HP model evaluating more efficiently, the conformations produced. This model is a combination of assessment based on the collisions numbers, hydrophobic contacts, hydrophobic and hydrophilic core compression, obtaining thus more natural structures with minimum energy. The results demonstrate the efficiency of this algorithm to obtain protein structures indicators providing compact compression of the hydrophobic and hydrophilic core protein.

Keywords

Multiobjective Genetic Algorithms; Hydrophobic-Polar Model; Prediction of Protein Structure; Compactation of Protein Structure.

Sumário

1 Introdução 16

1.1. Motivação 16

1.2. Objetivo Geral 18

1.3. Descrição do Trabalho e Contribuições 19

1.4. Organização da Dissertação 20

2 Fundamentos 21

2.1. Conceitos de Otimização Multiobjetivo 21

2.1.1. Otimização Mono-Objetivo 21

2.1.2. Otimização Multiobjetivo 22

2.1.3. Conceitos de dominância e de otimalidade de Pareto 24

2.2. Algoritmos Genéticos Multiobjetivo 27

2.3. Enovelamento de Proteínas. 28

2.3.1. Síntese de proteínas 29

2.3.2. Aminoácidos 31

2.4. Ligação Peptídica 36

2.5. Cadeia Principal 37

2.6. Gráfico de Ramachandran 39

2.7. Níveis estruturais nas Proteínas 40

2.7.1. Estrutura primária 40

40

2.7.2. Estrutura Secundária 40

2.7.3. Estrutura terciária 41

2.7.4. Estrutura quaternária 42

2.8. Predição de estruturas de proteínas 42

2.8.1. Predição por modelagem comparativa. 42

2.8.2. Predição por reconhecimento de enovelamento – Threading 44

2.8.3. Predição ab initio (primeiros princípios) 44

2.8.4. Predição de novo 46

2.9. Modelo Hidrofóbico-Polar 48
2.10. Raio de giro 50

3 Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas no Modelo Hidrofóbico Polar (AGMO-HP) 51
3.1. O modelo do Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas no Modelo Hidrofóbico Polar (AGMO-HP) 52
3.1.1. Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas Naturais (AGMO-HP) 52
3.1.2. Representação dos indivíduos 54
3.1.3. Função de Aptidão 54
3.1.4. População Inicial 55
3.1.5. Operadores genéticos 56

4 Experimento e Resultados 75
4.1. Conjunto de Teste 75
4.2. Configuração do AGMO-HP 77
4.3. Resultados para sequências de 27 monômeros 78
4.4. Resultados para sequências de 48 monômeros 85
4.5. Resultados para sequências de proteínas reais. 91

5 Conclusões e trabalhos futuros 95

Referências Bibliográficas 97
Apêndice 104

Lista de figuras

Figura 2.1: a) Modelo em metal elaborado por Watson e Crick. (b) Esquema do modelo de DNA publicado no texto de 1953 (WATSON e CRICK, 1953) 30
Figura 2.2: Tripletos de formação dos aminoácidos. Fonte: http://djalmasantos.wordpress.com/2010/11/15/codigo-genetico/ 31
Figura 2.3: Representação gráfica da estrutura química de um Aminoácido Fonte: Elaboração própria 31
Figura 2.4: Relação dos 20 aminoácidos com seus códigos de três e de uma letra 32
Figura 2.5: Representação química dos aminoácidos apolares e alifáticos Fonte: (LEHNINGER, NELSON e COX, 2002) 33
Figura 2.6: Representação química dos aminoácidos aromáticos. Fonte: (LEHNINGER, NELSON e COX, 2002) 34
Figura 2.7: Representação química dos aminoácidos não carregados e polares. Fonte: (LEHNINGER, NELSON e COX, 2002) 35
Figura 2.8: Representação química dos aminoácidos básicos. Fonte: (LEHNINGER, NELSON e COX, 2002) 35
Figura 2.9: Representação química dos aminoácidos carregados negativamente ou ácidos. Fonte: (LEHNINGER, NELSON e COX, 2002) 36
Figura 2.10: Ligação Peptídica. 37
Fonte: http://www.infoescola.com/bioquimica/ligacao-peptidica/ 37
Figura 2.12: Ângulos de torção (diedras) da cadeia principal da proteína. 38
Figura 2.13: Ângulos de torção diedras da cadeia lateral do aminoácido lisina. 38
Figura 2.14: Mapa de Ramachandran: a região mais favorável é apresentada em vermelho, região permitida é apresentada em amarelo, região ainda aceitável é apresentada em amarelo claro e a região não permitida em branco. A região em vermelho no canto superior esquerdo representa a região de folhas β paralelas e antiparalelas. A região em vermelho no centro esquerdo, e no centro direito representa a região de hélices α à direita e a esquerda respectivamente.
Fonte : (LASKOWSKI, MACARTHUR, *et al.*, 1993). 39

Figura 2.15: Estrutura primaria de uma Proteína. 40
Fonte: http://www.profpc.com.br/Química_das_células.htm 40
Figura 2.16: Estrutura Secundária: A) α-Hélice, B) Folha-β, C) Vol ta Fonte: http://mrclay13bio.wikispaces.com/protein+structure 41
Figura 2.17: Representação Ribbon da estrutura terciária da Lisozima. 41
Figura 2.18: Estrutura quaternária numa representação tipo 42
ribbon da Hemoglobina. 42
Fonte: http://portaldoprofessor.mec.gov.br/fichaTecnicaAula.html?aula=1599 42
Figura 2.19: Conformações malha quadrática (2.19a) e malha cubica (2.19b). Fonte: (FIDANOVA e LIRKOV) 49
Figura 3.1: Algoritmo Genético Multiobjetivo proposto para predição de Estruturas Proteicas. 53
Figura 3.2: Exemplo de representação dos cromossomas no modelo HP. É amostrada a sequência absoluta de um cromossomo e a sua dinâmica para a obtenção da malha em 3D. Na malha, os aminoácidos hidrofóbicos em vermelho e os hidrofílicos em preto. 54
Figura 3.3: Algoritmo para geração dos indivíduos da população inicial do algoritmo genético multiobjetivo proposto 56
Figura 3.4: Algoritmo para calculo dos créditos quando é executado o crossover. 57
Figura 3.5 : Cruzamento de um ponto em dois cromossomas para sequência de comprimento dez. 58
Figura 3.6: Cruzamento de dois pontos para em uma sequência de comprimento dez. 59
Figura 3.7: Cruzamento multiponto em sequências de comprimento dez 60
Figura 3.8: Operador de Crossover Hidrofóbico – Multiponto para sequências de comprimento doze. 61
Figura 3.9: Operador de cruzamento hidrofílico–multiponto para sequências de comprimento doze. 62
Figura 3.10: Algoritmo para cálculo dos créditos quando é executado o mutação. 63
Figura 3.11: Operador de Mutação Simples 64
Figura 3.12: Operador de Mutação com busca exaustiva Acumulada 65

Figura 3.13: Operador de Mutação Troca de Segmento sem Colisões 67
Figura 3.14: Operador de mutação simples acumulada 68
Figura 3.15: Mecanismo de geração dos cromossomos por troca
de pares para no Operador de mutação 2-op com memoria. 69
Figura 3.16: Dinâmica da roleta baseada em créditos 70
Figura 3.17: Método do torneio para seleção parental,
para duas conformações C1, C2. 73
Figura 3.18: Método da substituição parental. Esta figura apresenta um
esquema de como um novo indivíduo é inserido na população parental
através do critério de semelhança baseado na compactação da proteína. 74
Figura 4.1: Estruturas tridimensionais preditas para as
sequências 27.1 e 27.2, para os modelos AGMOS-HP e AGMO-HP. 80
: Estruturas tridimensionais preditas para as sequências 27.3 e 27.4,
para os modelos AGMS-HP e AGMO-HP. 80
Figura 4.2: Estruturas tridimensionais preditas para as
sequências 27.3 e 27.4, para os modelos AGMOS-HP e AGMO-HP. 81
Figura 4.3: Estruturas tridimensionais preditas para as sequências
27.5 e 27.6, para os modelos AGMOS -HP e AGMO-HP. 82
Figura 4.4: Estruturas tridimensionais preditas para as
sequências 27.7 e 27.8, para os modelos AGMOS -HP e AGMO-HP. 83
Figura 4.5: Estruturas tridimensionais preditas para as
sequências 27.9 e 27.10, para os modelos AGMOS -HP e AGMO-HP. 84
Figura 4.6: Estruturas tridimensionais preditas para as
sequências 48.1 e 48.2, para os modelos AGMOS-HP e AGMO-HP. 86
Figura 4.7: Estruturas tridimensionais preditas para as
sequências 48.3 e 48.4, para os modelos AGMOS -HP e AGMO-HP. 87
Figura 4.8: Estruturas tridimensionais preditas para as
sequências 48.5 e 48.6, para os modelos AGMOS -HP e AGMO-HP. 88
Figura 4.9: Estruturas tridimensionais preditas para as
sequências 48.7 e 48.8, para os modelos AGMOS -HP e AGMO-HP. 89
Figura 4.10: Estruturas tridimensionais preditas para as
sequências 48.9 e 48.10, para os modelos AGMOS -HP e AGMO-HP. 90
Figura 4.11: Estruturas tridimensionais preditas para as

sequências 46 e 58, para os modelos AGMOS-HP e AGMO-HP. 92
Figura 4.12: Estruturas tridimensional predita para a sequência de 103 monómeros, para os modelos AGMOS -HP e AGMO-HP. 93
Figura 4.13: Estruturas tridimensional predita para a sequência de 136 monómeros, para os modelos AGMOS -HP e AGMO-HP. 94

Lista de tabelas

Tabela 4.1: Conjunto de teste com sequências de 27 monômeros 76
Tabela 4.2: Conjunto de teste com sequências de 48 monômeros 76
Tabela 4.3: Conjunto de teste com sequências de proteínas reais. 77
Tabela 4.4: Configuração para sequências curtas(SC) do AGMO-HP 77
Tabela 4.5: Configuração para sequências longas(SL) do AGMO-HP 78
Tabela 4.6: Resultados obtidos com sequências de 27 monômeros em comparação com outros métodos. São amostrados o número de contatos hidrofóbicos media e desvio padrão. Também são apresentados os resultados obtidos da compactação dos aminoácidos hidrofóbicos (Comp. NúcleoH) e hidrofílicos (Comp. NúcleoP). 79
Tabela 4.7: Resultados obtidos com sequências de 48 monômeros em comparação com outros métodos, são amostrados o número de contatos hidrofóbicos media e desvio padrão. Também são apresentados os resultados obtidos da compactação dos seus aminoácidos hidrofóbicos (Comp. Núcleo) e hidrofílicos (Comp. Núcleo). 85
Tabela 4.8: Resultados obtidos com sequências reais de 46, 58, 103 e 136 monómeros em comparação com outros métodos, são amostrados o número de contatos hidrofóbicos media e desvio padrão. Também são apresentados os resultados obtidos da compactação dos seus aminoácidos hidrofóbicos (Comp. NucleoH) e hidrofílicos (Comp. NucleoP). 91

1
Introdução

1.1.
Motivação

As proteínas são muito importantes nos organismos biológicos. Por exemplo, a queratina é uma proteína estrutural essencial para formação de cabelo e unhas; outras proteínas, tais como a actina e miosina, tornam possível o movimento muscular. As enzimas são catalisadoras biológicas que participam, por exemplo, da digestão dos alimentos, os anticorpos são proteínas de reconhecimento que formam parte de nosso sistema imunológico e outras proteínas ajudam no controle dos sinais do cérebro e copiam genes durante a divisão celular (BUI e SUNDARRAJ, 2005). Por outro lado proteínas anormais, com algum erro na sua estrutura, podem ser a causa de doenças tais com, fibrose cística, Alzheimer, encefalopatia espongiforme bovina (EEB), vulgarmente conhecida como a "doença da vaca louca", entre outras (COHEN e KELLY, 2003).

As proteínas são formadas por aminoácidos, sendo que todas são constituídas partindo-se de um conjunto de 20 (vinte) aminoácidos, combinados sequencialmente através de ligações. A estrutura de uma proteína é subdividida em quatro níveis de organização, crescentes em complexidade: estrutura primária (sequência de aminoácidos), secundária (padrões regulares de dobramento), terciária (arranjo tridimensional de todos os aminoácidos) e quaternária (arranjo de duas ou mais cadeias proteicas no espaço tridimensional).

A função das proteínas depende principalmente da sua estrutura terciária e portanto, o conhecimento da estrutura terciária de uma proteína possibilita a identificação de tratamentos mais eficazes para as doenças. Este conhecimento também pode ser utilizado no desenvolvimento de novos medicamentos. Por exemplo, a replicação do vírus do AIDS depende da protease HIV.

Se pudéssemos identificar uma molécula que permanecesse ligada ao sítio ativo da enzima HIV protease, a função normal desta enzima poderia ser anulada (FOGEL e CORNE, 2003).

Um dos problemas atuais mais importantes da biologia molecular é o do enovelamento das proteínas (*protein folding problem*) que consiste em encontrar a estrutura tridimensional de uma proteína em seu estado natural a partir de sua estrutura primária (YUE e DILL, 1993). Embora já existam métodos experimentais para a determinação da estrutura terciária de proteínas já formadas, tais como a cristalografia de raios X e a espectrografia de ressonância magnética nuclear, esses métodos são muito trabalhosos e demandam muito tempo (OLIVARES e GARCIA, 2004). Apesar de já terem sido determinadas as sequências de mais de 250.000 proteínas, atualmente somente são conhecidas 35.000 estruturas terciárias, as quais estão disponíveis em um banco de dados de proteínas na Internet (Protein Data Bank).

Paralelamente aos métodos experimentais, técnicas computacionais vêm sendo desenvolvidas para prever estruturas de proteínas, entre elas, os modelos baseados em rede. Tratam-se de modelos simplificados que podem ser utilizados para extrair os princípios básicos do enovelamento de proteínas, fazer previsões e unificar o conhecimento de muitas propriedades diferentes das proteínas (YUE e DILL, 1995). Esses modelos frequentemente fornecem dados que são utilizados para desenvolver modelos mais complexos, como o de átomos explícitos (HART e NEWMAN, 2005).

Um dos modelos mais estudados é o modelo Hidrofóbico-Polar (HP) (DILL., 1985) (LAU e DILL., 1989), o modelo HP fornece uma estimativa da energia da conformação com base na soma de interações entre pares de aminoácidos hidrofóbicos. No entanto, mesmo tratando de um modelo simplificado, sua solução é um problema NP-Difícil (BERGER e LEIGHTON, 1998).

Muita pesquisa tem sido realizada com a aplicação de algoritmos genéticos para maximizar o número de contatos HH no modelo HP. (UNGER e MOULT, 1993), (ARNOLD L. PATTON, 1995), (MEHUL M. e PETER V. , 1997), (KRASNOGOR, 2002), (TETTAMANZI), (KRASNOGOR, W E. , *et al.*, 1999), (BUI e SUNDARRAJ, 2005), (CUSTÓDIO, 2008). Apesar de estes trabalhos serem utilizados com sucesso na predição de estruturas proteicas baseadas no

modelo HP, estes trabalhos apresentam, em algumas ocasiões, características que podem prejudicar o seu desempenho quando, por exemplo, se utiliza mecanismos de correção de conformações inválidas. Neste mecanismo existe a necessidade de avaliar diversas vezes as soluções até atingir uma conformação válida. Além disso, o uso de mecanismos de reparação pode ter um efeito negativo desfazendo os contatos HH em passos mais avançados da evolução (CUSTÓDIO, 2008).

Por outro lado, estes trabalhos não fazem tratamento algum da forma da compactação da proteína, chegando a encontrar conformações de energia mínima que, em muitos dos casos, dificilmente poderiam existir na natureza. Nesse sentido, surge à necessidade de criar um modelo que possa gerar conformações válidas de mínima energia cujas formas sejam mais compactas, e com um esforço computacional apropriado. Neste contexto, percebe-se que muitos objetivos precisam ser satisfeitos: tratamento das conformações inválidas que não signifique um esforço computacional muito alto, obtenção de estruturas proteicas com a mínima energia possível e estruturas proteicas compactas e globulares. Deste modo, de forma a considerar todos estes objetivos, neste trabalho é desenvolvido um modelo baseado em Algoritmos Genéticos Multiobjetivos.

1.2. Objetivo Geral

Fundamentado na discussão anterior, o objetivo desta dissertação é propor um algoritmo genético multiobjetivo para a predição de estruturas terciárias de proteínas, baseado no Modelo Hidrofóbico Polar, para geração de estruturas proteicas compactas e globulares. Como objetivos específicos pode-se relacionar:

- Validar o modelo desenvolvido com as sequências mais estudadas na literatura;
- Garantir que o modelo tenha a capacidade de predizer estruturas proteicas naturais e compactas;
- Desenvolver um programa para a predição de estruturas proteicas baseado no Modelo Hidrofóbico Polar.

1.3. Descrição do Trabalho e Contribuições

A pesquisa foi estruturada na seguinte ordem: i) estudo dos diferentes modelos de predição de proteínas e identificação das principais limitações de cada um deles; ii) criação de um modelo baseado num algoritmo genético multiobjetivo, denominado de Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas Naturais (AGMO-HP); iii) implementação de uma ferramenta computacional baseada nesse modelo;

iv) e, finalmente, comparação dos resultados desse modelo com os estudos de caso propostos por (UNGER e MOULT, 1993).

Em resumo as principais contribuições deste trabalho são:

- Criação de um Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas Naturais (AGMO-HP). Para este modelo foram definidos os objetivos do algoritmo genético, a representação dos indivíduos, os operadores genéticos e as estratégias de substituição parental.
- Desenvolvimento de uma metodologia pseudoadaptativa das percentagens de aplicação dos operadores genéticos inspirada nos trabalhos de (DAVIS, 1996) (DE JONG, 1975) (JULSTROM, 1995).
- Criação de uma ferramenta computacional para o desenvolvimento de testes em linguagem C# 2.0 e Matlab 2010 nos ambientes Windows 7 e Linux Ubuntu, permitindo, desta forma, a realização de experimentos com o objetivo de demonstrar a potencialidade e aplicabilidade do modelo AGMO-HP.

1.4. Organização da Dissertação

Este trabalho esta dividido da seguinte forma:

- No capítulo 2 é apresentado um resumo dos fundamentos teóricos essenciais para a compreensão deste trabalho.
- No capítulo 3, o modelo proposto de Algoritmo Genético Multiobjetivo é apresentado em detalhes.
- O capítulo 4 mostra os resultados obtidos em diversos estudos de caso.
- Por fim, o capítulo 5 apresenta as conclusões sobre modelo, sugerindo ideias para a continuação deste trabalho.

2
Fundamentos

2.1.
Conceitos de Otimização Multiobjetivo

2.1.1.
Otimização Mono-Objetivo

Um problema de otimização mono-objetivo é composto por uma função objetivo (a minimizar ou a maximizar) e, normalmente, por várias restrições que todas as soluções admissíveis têm de satisfazer. De uma maneira geral, um problema de otimização mono-objetivo pode ser formulado da seguinte forma:

Maximizar	$\phi(\xi)$		(1)
Sujeito a	$\gamma_{\varphi}(\xi)\geq 0,$	$\varphi=1,2,3...\vartheta$	(2)
	$\eta_{\kappa}(\xi)=0,$	$\kappa=1,2,3...K$	(3)
	$\xi_{\iota}^{inf}\leq\xi_{\iota}\leq\xi_{\iota}^{sup}$	$\iota=1,2,3...n$	(4)

O problema tem associado ϑ+K restrições funcionais, das quais ϑ são de desigualdade, $\gamma_{\varphi}(\xi)$, e K de igualdade, $\eta_{\kappa}(\xi)$. As restrições de desigualdade podem ser do tipo "$\geq$" ou "$\leq$", apesar da forma geral apresentar apenas restrições do tipo "$\geq$", o que não implica perda de generalidade.

Uma solução ξ é um vector de n variáveis de decisão: $\xi = (\xi_1,..., \xi_n)$. As restrições do último conjunto denominam-se por limites das variáveis de decisão, as quais obrigam a que cada variável de decisão ξ_l assuma um valor entre um limite inferior, ξ_{ι}^{inf}, e um limite superior, ξ_{ι}^{sup}. Estes limites constituem o espaço das decisões ou espaço das soluções.

Uma solução ξ que satisfaça todas as ϑ+K restrições funcionais e todos os $2n$ limites das variáveis de decisão denomina-se por solução admissível. Ao conjunto de todas as soluções admissíveis dá-se o nome de região admissível, Ξ.

Se a função objetivo e todas as restrições de um problema são funções lineares em relação à ξ, então este problema denomina-se de linear. Pelo contrário, se a função objetivo ou alguma das restrições são não lineares em relação à ξ, então este problema denomina-se de não linear.

Na resolução de um problema de otimização mono-objetivo, pretende-se determinar a solução ótima, ou seja, a solução admissível que aperfeiçoe a função objetivo, cujo valor é único, mesmo que existam soluções ótimas alternativas (CHORRO S. B, 2007).

2.1.2. Otimização Multiobjetivo

Um problema de otimização é do tipo Multiobjetivo se existem duas ou mais funções objetivo associadas ao problema. Neste tipo de problemas, o conceito de solução ótima, característico de problemas de otimização Mono-Objetivo, não se pode aplicar, uma vez que uma solução admissível que otimize um dos objetivos, não otimiza, em geral, os objetivos restantes, quando estes estão em conflito.

De uma maneira geral, um problema de otimização Multi-Objetivo pode ser formulado da seguinte forma:

Minimizar	$\phi_m(\xi)$	m= 1, 2,..., M
Sujeito a	$\xi \in \Xi$	((1)–(3))

Cada uma das M funções objetivo, $\phi(\xi) = (\phi_1(\xi), \phi_2(\xi), \ldots, \phi_m(\xi))^T$, com $M \geq 2$, pode ser a minimizar ou a maximizar. Como foi referida anteriormente, esta formulação não implica perda de generalidade.

Segundo (CHORRO S. B, 2007), uma das diferenças fundamentais entre otimização Mono-Objetivo e otimização Multiobjetivo é que, ao espaço das variáveis de decisão há que acrescentar outro espaço multidimensional gerado pelas funções objetivo, denominado de espaço das funções objetivo, ou simplesmente, espaço dos objetivos. Para cada solução $\xi = (\xi_1,...,\xi_n)$ no espaço das variáveis de decisão, existe um ponto que lhe corresponde no espaço dos objetivos, denotado por $\phi(\xi) = \zeta = (\zeta_1, \ldots, \zeta_M)^T$, com $\zeta_M = \phi_m(\xi)$ e m = 1,...,M

A região admissível no espaço das funções objetivo (o conjunto de todas as imagens dos pontos em Ξ) pode então ser definida da seguinte forma:

$$Z = \{\zeta \in \Re^M : \zeta = (\phi_1(\xi), \phi_2(\xi), \ldots, \phi_M(\xi)), \xi \in \Xi\}$$

Desta forma, a noção de solução ótima usada em problemas de otimização mono-objetivo é substituída pela noção de solução não dominada (também designada por ótima de Pareto, eficiente ou não inferior).

Uma solução não dominada é uma solução admissível para a qual não é possível melhorar simultaneamente todas as funções objetivo; isto é, a melhoria numa função objetivo apenas pode ser alcançada por degradação de pelo menos uma das outras. Segundo (CLÍMACO, ANTUNES e ALVES, 2003) uma solução admissível é dominada por outra, se ao passar-se da primeira para a segunda existir melhoria de pelo menos uma função objetivo, permanecendo inalteradas as restantes.

No conjunto das soluções não dominadas, também denominado por frente ótima de Pareto ou apenas frentes de Pareto têm de ser satisfeitas as duas condições seguintes:

- Quaisquer duas soluções deste conjunto têm de ser não dominadas entre si.
- Qualquer solução que não pertença a este conjunto é dominada por, pelo menos, uma solução deste conjunto.

Quando as funções objetivo de um problema de otimização multiobjetivo estão em conflito, normalmente o conjunto de soluções não dominadas é enorme. Desta forma, é difícil escolher uma solução deste conjunto, sem dispor de informação adicional relativa ao problema a resolver. No entanto, na ausência de tal informação, as soluções não dominadas não são comparáveis entre si, o que motiva as abordagens a determinar o número máximo possível de soluções não dominadas do problema. Assim, pode-se caracterizar o problema de otimização multiobjetivo como a determinação de um conjunto de soluções não dominadas com as seguintes características (CHORRO S. B, 2007):

- Esteja o mais próximo possível da frente ótima de Pareto real.
- Seja o mais diversificado possível.

A primeira característica é inerente a qualquer tarefa de otimização; também em otimização mono-objetivo se pretende determinar uma solução admissível que

garanta o valor ótimo para o modelo matemático. A segunda característica é específica da otimização multiobjetivo. Apenas com um conjunto de soluções dispersas é possível garantir a existência de um bom conjunto de soluções de compromisso entre os objetivos. Uma vez que a otimização multiobjetivo atua em dois espaços, das variáveis de decisão e dos objetivos, a diversidade das soluções pode ser definida nestes dois espaços. Apesar de, na maioria dos problemas, a diversidade num dos espaços significar, normalmente, a diversidade no outro espaço, isto pode não acontecer em alguns problemas.

2.1.3. Conceitos de dominância e de otimalidade de Pareto

A maioria dos algoritmos de otimização multiobjetivo usa o conceito de dominância, em particular quando há necessidade de comparar duas soluções, para averiguar se há dominância de uma sobre a outra.

Uma solução $\xi^1 \in \Xi$ domina outra solução $\xi^2 \in \Xi$, se e só se ξ^1 não é pior do que ξ^2 para todos os objetivos e ξ^1 é estritamente melhor do que ξ^2 para pelo menos um dos M objetivos.

Apesar de serem diferentes os conceitos de solução eficiente e de solução não dominada, quando utilizados de forma genérica nesta dissertação não é feita qualquer distinção entre eles. Enquanto que o conceito de solução eficiência se refere, geralmente, a pontos do espaço de decisão, o conceito de solução não dominada é utilizado para pontos do espaço dos objetivos; isto é, uma solução não dominada é a imagem de uma solução eficiente.

Matematicamente, e considerando todas as funções objetivo a minimizar, tem-se (CLÍMACO, ANTUNES e ALVES, 2003):

- Uma solução $\xi^1 \in \Xi$ é eficiente se e só se não existe outra solução $\xi^2 \in \Xi$ tal que $\zeta_m(\xi^2) \leq \zeta_m(\xi^1)$ para todo o m (m = 1, 2, ..., M) e $\zeta_m(\xi^2) < \zeta_m(\xi^1)$ para pelo menos um m.
- Um ponto do espaço dos objetivos $\zeta_1 \in Z$, com $\zeta_1 = (\zeta_1(\xi^1),..., \zeta_m(\xi^1))$ e $\xi^1 \in \Xi$, diz-se não dominado se e só se ξ^1 é uma solução eficiente.

Uma solução de compromisso satisfatória para o problema multiobjetivo deverá ser não dominada, cujos valores das funções objetivo são satisfatórios para o Agente de Decisão (AD) e de tal modo que seja aceitável como solução final.

Desta forma, é apenas sobre o conjunto das soluções não dominadas que deve recair a atenção do analista e do AD.

Entre quaisquer duas soluções não dominadas a uma melhoria em pelo menos um dos objetivos encontra-se sempre associado um sacrifício em pelo menos um dos outros objetivos.

Isto é, verifica-se sempre uma compensação entre objetivos no conjunto das soluções não dominadas.

A relação de dominância entre duas soluções, tal como foi definida atrás, é muitas vezes referida como uma relação de dominância fraca: ξ^1 domina fracamente uma solução ξ^2, se ξ^1 não é pior do que ξ^2 em todos os M objetivos e é estritamente melhor do que ξ^2 em pelo menos um dos M objetivos. A partir da definição de dominância fraca, pode-se obter a definição de dominância estrita (forte), da forma que se segue.

Uma solução ξ^1 domina estritamente (fortemente) uma solução ξ^2, se a solução ξ^1 é estritamente melhor do que a solução ξ^2 em todos os M objetivos. Desta forma, se uma solução ξ^2 domina estritamente (fortemente) uma solução ξ^2, a solução ξ^1 também domina fracamente a solução ξ^2, mas o recíproco não é verdadeiro.

Tal como existem soluções ótimas globais e locais em otimização mono-objetivo, também poderão existir frentes ótimas de Pareto globais e locais em otimização multiobjetivo.

Ao conjunto de soluções não dominadas de toda a região admissível dá-se o nome de frente ótima de Pareto global. À frente ótima de Pareto global é referida simplesmente como frente ótima de Pareto ou frente de Pareto. Uma vez que as soluções deste conjunto são não dominadas em relação a qualquer solução da região admissível, elas são as melhores soluções do problema de otimização multiobjetivo.

Define-se, também, frente ótima de Pareto local da seguinte forma ((DEB, 1999) e (MIETTINEN, 1999)): se para cada solução x de um conjunto P" não existe nenhuma solução y (na vizinhança de x, tal que $\|y-x\|_\infty \leq \varepsilon$, em que ε é um número positivo muito pequeno), que domine qualquer solução de P", então as soluções que pertencem ao conjunto P" constituem um conjunto ótimo de Pareto local.

De acordo com esta definição, uma frente ótima de Pareto global é também uma frente ótima de Pareto local.

A definição de dominância forte pode ser usada para definir frente fracamente não dominada. Dado um conjunto de soluções, P, à frente fracamente não dominada, P', é formada pelas soluções que não são fortemente dominadas relativamente a qualquer outra solução do conjunto P (DEB, 2001).

As definições de frente fracamente não dominadas globais e locais podem também ser feitas de forma semelhante, a partir da definição de frente fracamente não dominada.

2.2. Algoritmos Genéticos Multiobjetivo

Os métodos de programação matemática multiobjetivo calculam, em geral, em cada iteração uma única solução não dominada, através da otimização de funções escalares substitutas. Estas funções agregam temporariamente em uma única dimensão as múltiplas funções objetivo, de tal modo que a solução ótima de uma função escalar substituta é uma solução não dominada do problema multiobjetivo. Assim, se for necessário caracterizar com alguma exaustividade o conjunto das soluções não dominadas, estas abordagens podem exigir um esforço computacional apreciável. As características das soluções calculadas, em particular a respectiva diversidade, estão bastante dependentes dos parâmetros de informação de preferências incluídas nas funções escalares substitutas (CLÍMACO, ANTUNES e ALVES, 2003).

Ao trabalharem com populações de soluções, os AGs têm a potencialidade de determinar várias soluções não dominadas numa só execução do algoritmo. O processo de otimização multiobjetivo deve ter em conta os três aspectos seguintes:

- Minimizar a distância da frente de Pareto obtida à frente ótima de Pareto real.
- Maximizar a extensão da frente de Pareto obtida, isto é, as soluções não dominadas obtidas devem abranger a maior gama de valores possíveis para cada objetivo.

Atendendo a estes três aspectos, quando se implementa um AG em otimização multiobjetivo, devem ter-se sempre em atenção as duas questões seguintes:

a) Como aperfeiçoar a função de aptidão e a seleção, de forma a conduzir a pesquisa em direção à frente ótima de Pareto real.
b) Como manter uma população diversificada de soluções, de forma a prevenir a convergência prematura e alcançar uma frente de Pareto extensa e bem distribuída.

2.3.
Enovelamento de Proteínas.

As proteínas (polipeptídeos) são macromoléculas envolvidas na maior parte das transformações em uma célula viva. Estes biopolímeros são formados por um alfabeto de 20 tipos diferentes de aminoácidos. A sequência de aminoácidos ou estrutura primária, forma, através de um processo de condensação, a cadeia polipeptídica da proteína. A ligação CO-NH resultante, uma ligação amida, é conhecida como ligação peptídica.

Esta cadeia quando em condições fisiológicas (ambiente nativo), adota uma única estrutura tridimensional (3D) ou conformação nativa. Isto é, quando a proteína é sintetizada ela se dobra para que parte da cadeia principal e da cadeia lateral, fundamentais para desempenhar a sua função, sejam postas em posição geométrica precisa. Esta dobra nativa, adotada pela cadeia polipeptídica, não sofre variação (GIBAS e JAMBECK, 2001), sendo única para uma dada sequência de aminoácidos.

A estrutura nativa determina a função bioquímica especifica da proteína na célula (BAXEVANIS, e OUELLETTE, 2005) (BRANDEN e TOOZ, 1998), a qual pode ser de catálise, de ligação, de transporte, entre outras (BRANDEN e TOOZ, 1998). Conhecer a estrutura 3D da proteína implica conhecer a sua função. A partir deste conhecimento é possível influenciar, através do desenvolvimento de compostos químicos, fármacos ou drogas, a ação que a proteína exerce no organismo.

Experimentalmente, a estrutura 3D de uma proteína pode ser obtida através de técnicas de cristalografia por difração de raios-X ou de ressonância magnética nuclear (NMR) (BAXEVANIS, e OUELLETTE, 2005) (TRAMONTANO e LESK, 2006). No entanto, o elevado custo e o alto grau de complexidade associados a estas técnicas fazem com que as mesmas sejam difíceis de realizar. Isto motivou a realização de novos estudos para desenvolvimento de técnicas que pudessem predizer de forma eficaz e eficiente a estrutura 3D de uma proteína.

O processo pelo qual uma sequência de aminoácidos atinge sua conformação em estado nativo é chamado de enovelamento ou dobramento (BRANDEN e TOOZ, 1998).

A sequência de aminoácidos e o ambiente em que estes estão inseridos são os fatores que, durante o processo de enovelamento, fazem com que a proteína assuma determinada conformação.

Experimentos realizados por Anfinsen (ANFINSEN, E, *et al.*, 1961) demostraram que a molécula de uma proteína quando desnaturada (estado desenovelado) por rompimento das condições em seu ambiente pode enovelar-se novamente em sua estrutura nativa quando as condições fisiológicas são restauradas.

A partir desta constatação, assume-se que as sequências de aminoácidos contêm toda a informação necessária para determinar a estrutura nativa da proteína. Sendo assim, muitas pesquisas foram desenvolvidas no sentido de se encontrar a estrutura nativa da proteína, dada a sequência de aminoácidos através de métodos computacionais.

O problema da predição da estrutura 3D de uma proteína, somente a partir da sequência de aminoácidos, tornou-se então um dos principais e maiores problemas ainda não resolvidos da biologia molecular estrutural (BRANDEN e TOOZ, 1998) e da Bioinformática.

2.3.1. Síntese de proteínas

O acido desoxirribonucleico (DNA), presente em todas as células vivas, consiste numa longa hélice dupla formada por um esqueleto de fosfato e açúcar e por pares de moléculas denominadas base. As duas metades da hélice são complementares, pois cada um dos quatro tipos de bases apenas pode se emparelhar com a base do tipo complementar figura 2.1.

Embora a sequência de bases do DNA seja contínua, diferentes segmentos constituem unidades funcionais independentes, denominados genes.

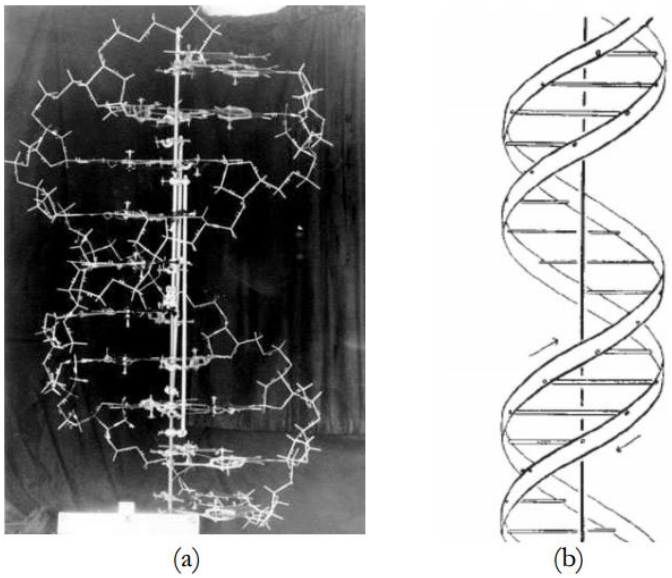

Figura 2.1: a) Modelo em metal elaborado por Watson e Crick. (b) Esquema do modelo de DNA publicado no texto de 1953 (WATSON e CRICK, 1953)

São estes que contêm a informação necessária à síntese das proteínas, macromoléculas essenciais para o metabolismo dos seres vivos, de que são exemplos às enzimas, os anticorpos e vários hormônios.

Quando uma célula recebe o sinal para produzir uma proteína, uma das cadeias da hélice dupla de DNA serve de molde para a síntese de uma sequência de bases complementar, denominada acido ribonucleico mensageiro (mRNA), num processo denominado transcrição. Cada tripleto ordenado de bases do mRNA, designado por codão, codifica uma de 20 moléculas, os aminoácidos, de que são feitas as proteínas, ou terminação da proteína, (figura 2.2). A informação contida no mRNA é traduzida numa sequência de aminoácidos, que se vão ligando uns aos outros numa cadeia linear, denominada cadeia polipeptídica. Cada proteína é formada por uma ou mais cadeias polipeptídicas.

Primeira base	Segunda base U	Segunda base C	Segunda base A	Segunda base G	Terceira base
U	UUU Fen	UCU Ser	UAU Tir	UGU Cis	U
	UUC Fen	UCC Ser	UAC Tir	UGC Cis	C
	UUA Leu	UCA Ser	UAA Fim	UGA Fim	A
	UUG Leu	UCG Ser	UAG Fim	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Tre	AAU Ans	AGU Ser	U
	AUC Ile	ACC Tre	AAC Ans	AGC Ser	C
	AUA Ile	ACA Tre	AAA Lis	AGA Arg	A
	AUG Met	ACG Tre	AAG Lis	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gli	U
	GUC Val	GCC Ala	GAC Asp	GGC Gli	C
	GUA Val	GCA Ala	GAA Glu	GGA Gli	A
	GUG Val	GCG Ala	GAG Glu	GGG Gli	G

Figura 2.2: Tripletos de formação dos aminoácidos. Fonte: http://djalmasantos.wordpress.com/2010/11/15/codigo-genetico/

2.3.2. Aminoácidos

Os aminoácidos são as unidades básicas que formam as proteínas. Um aminoácido é quimicamente composto por um átomo de carbono denominado C^{α}, o qual possui quatro diferentes ligantes: um grupo amino (-NH2), um grupo carboxílico (-COOH), um átomo de hidrogênio (H) e um grupo orgânico R também chamado de cadeia lateral ou radical. A figura 2.3 apresenta a estrutura química padrão de um aminoácido:

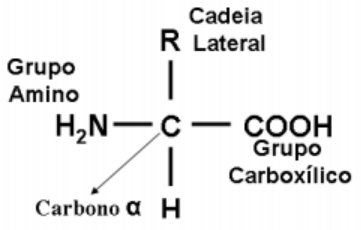

Figura 2.3: Representação gráfica da estrutura química de um Aminoácido Fonte: Elaboração própria

Existem 20 tipos diferentes de aminoácidos que se diferenciam por suas cadeias laterais. Estas podem possuir alguns átomos ou anéis aromáticos complexos. O grupo R de cada aminoácido caracteriza as suas propriedades físico-químicas (TRAMONTANO e LESK, 2006).

Os aminoácidos, por convenção internacional são identificados por abreviações de três letras (derivadas do seu nome em inglês) ou por um símbolo de uma letra (LEHNINGER, NELSON e COX, 2002). A figura 2.4 relaciona os 20 aminoácidos existentes e seus respectivos códigos abreviados de três letras e o símbolo de uma letra.

Em uma cadeia polipeptídica, a região N-terminal é aquela que possuí um grupo amino livre e a região C-terminal é aquela que possuí na cadeia polipeptídica um grupo carboxílico livre.

Os aminoácidos podem ser classificados pela natureza química de seus grupos R. Segundo (LEHNINGER, NELSON e COX, 2002), os aminoácidos podem ser divididos em cinco classes: grupos R apolares e alifáticos (1), grupos R aromáticos (2), grupos R não carregados e polares (3), grupos R carregados positivamente ou básicos (4) e grupos R carregados negativamente ou ácidos (5).

Conhecer as propriedades físico-químicas de cada aminoácido é fundamental para entender a bioquímica das proteínas. Estas propriedades são importantes, pois são elas que contribuem para que uma proteína encontre a sua estabilidade físico-química representando o seu estado nativo.

Aminoácido	Abreviação	Símbolo	Propriedades
Alanina	Ala	A	Não-polar, hidrofóbico
Cisteína	Cys	C	Polar, hidrofílico
Ácido Aspártico	Asp	D	Polar, hidrofílico
Ácido Glutâmico	Glu	E	Polar, hidrofílico
Fenilalanina	Phe	F	Não-polar, hidrofóbico
Glicina	Gly	G	Polar, hidrofílico
Histidina	His	H	Polar, hidrofílico
Isoleucina	Ile	I	Não-polar, hidrofóbico
Lisina	Lys	K	Polar, hidrofílico
Leucina	Leu	L	Não-polar, hidrofóbico
Metionina	Met	M	Não-polar, hidrofóbico
Asparagina	Asn	N	Polar, hidrofílico
Prolina	Pro	P	Não-polar, hidrofóbico
Glutamina	Gln	Q	Polar, hidrofílico
Arginina	Arg	R	Polar, hidrofílico
Serina	Ser	S	Polar, hidrofílico
Treonina	Thr	T	Polar, hidrofílico
Valina	Val	V	Não-polar, hidrofóbico
Triptofano	Trp	W	Não-polar, hidrofóbico
Tirosina	Tyr	Y	Polar, hidrofílico

Figura 2.4: Relação dos 20 aminoácidos com seus códigos de três e de uma letra

2.3.2.1.
Grupos R apolares e alifáticos

A primeira classe compreende os aminoácidos com cadeia lateral estritamente hidrofóbica e apolar, isto é, que não se dissolve em água. Esta classe engloba os aminoácidos: alanina (ala), valina (val), leucina (leu), isoleucina (ile), glicina (gly) e prolina (pro). As cadeias laterais destes aminoácidos contribuíram para a estabilização da estrutura da proteína pela promoção de interações hidrofóbicas em seu interior (LEHNINGER, NELSON e COX, 2002). A glicina (gly), apesar de ser um aminoácido apolar, não contribuiu efetivamente para a existência de interações hidrofóbicas.

O grupo amino secundário dos resíduos da prolina é mantido em uma conformação rígida que leva à redução da flexibilidade estrutural de regiões polipeptídicas em que este aminoácido ocorre (LEHNINGER, NELSON e COX, 2002). A Figura 2.5 apresenta a estrutura química dos aminoácidos classificados como apolares e alifáticos.

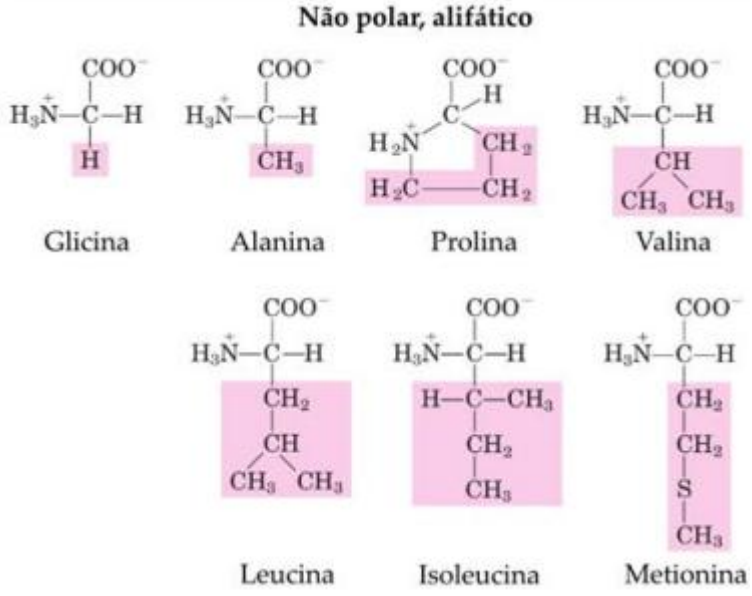

Figura 2.5: Representação química dos aminoácidos apolares e alifáticos
Fonte: (LEHNINGER, NELSON e COX, 2002)

2.3.2.2. Grupos R aromáticos

A segunda classe de aminoácidos compreende os aminoácidos que podem participar de interações hidrofóbicas. São eles: a fenilamina (phe), a tirosina (tyr) e o triptofano (trp). Estes aminoácidos com suas cadeias laterais aromáticas são relativamente apolares ou hidrofóbicos (LEHNINGER, NELSON e COX, 2002). A Figura 2.6 apresenta a estrutura química dos aminoácidos classificados como aromáticos.

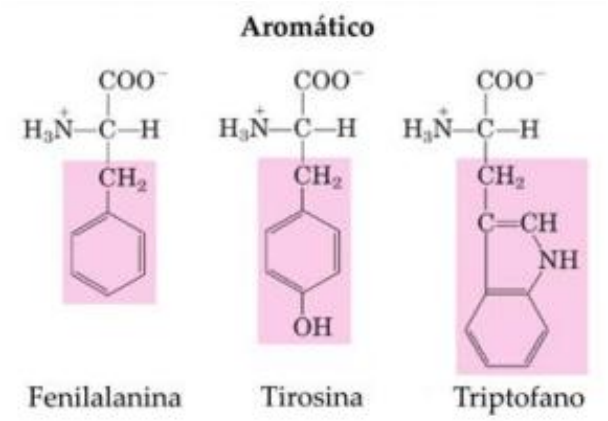

Figura 2.6: Representação química dos aminoácidos aromáticos. Fonte: (LEHNINGER, NELSON e COX, 2002)

2.3.2.3. Grupos R não carregados e polares

A terceira classe de aminoácidos abrange os aminoácidos não carregados, mas polares. Fazem parte destes grupos a serina (ser), a treonina (thr), a cisteína (cys), a asparagina (asn) e a glutamina (gln). Estes aminoácidos são mais solúveis em água do que os aminoácidos não polares, isto porque contêm grupos funcionais que formam ligações de hidrogênio com a água (LEHNINGER, NELSON e COX, 2002). A Figura 2.7 apresenta a estrutura química dos aminoácidos com grupamento R não carregado e polar.

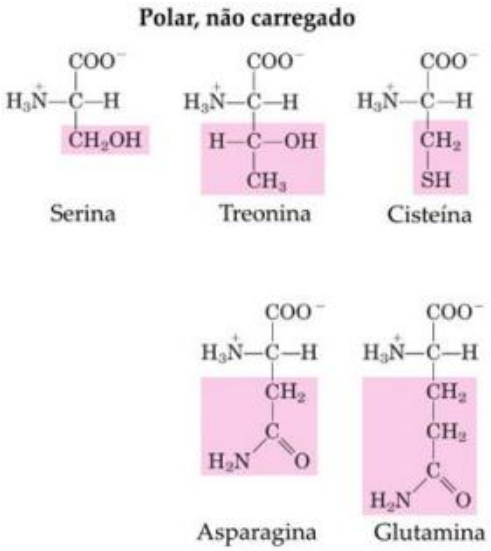

Figura 2.7: Representação química dos aminoácidos não carregados e polares. Fonte: (LEHNINGER, NELSON e COX, 2002)

2.3.2.4. Grupos R carregados positivamente ou básicos

Os grupamentos R mais hidrofílicos são aqueles que são positivamente ou negativamente carregados. Os aminoácidos pertencentes a quarta classe são aqueles nos quais os grupos R têm uma carga positiva líquida em pH 7 (LEHNINGER, NELSON e COX, 2002). Esta classe abrange os aminoácidos lisina (lis), arginina (arg) e histidina (his). A Figura 2.8 apresenta a estrutura química dos aminoácidos com grupamento R carregado positivamente.

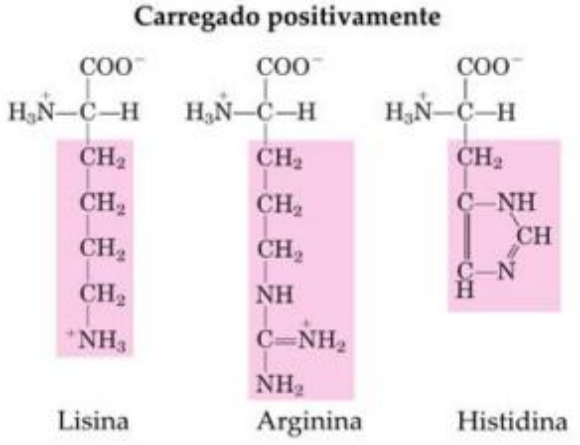

Figura 2.8: Representação química dos aminoácidos básicos. Fonte: (LEHNINGER, NELSON e COX, 2002)

2.3.2.5.Grupos R carregados negativamente ou ácidos

Os aminoácidos que possuem o grupamento R com uma carga negativa em pH 7 pertencem à quinta classe. Fazem parte desta classe os aminoácidos conhecidos como ácido aspártico e ácido glutâmico. A Figura 2.9 apresenta a estrutura química dos aminoácidos com grupamento R carregado negativamente

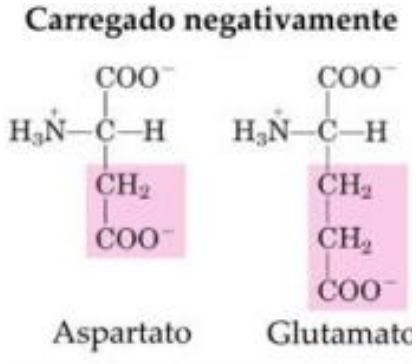

Figura 2.9: Representação química dos aminoácidos carregados negativamente ou ácidos. Fonte: (LEHNINGER, NELSON e COX, 2002)

2.4. Ligação Peptídica

Os aminoácidos, durante a síntese das proteínas, se ligam covalentemente de forma sequencial, formando um polímero ou cadeias polipeptídicas. Esta ligação recebe o nome de ligação peptídica e se forma entre o átomo de carbono (C) do grupo carboxílico de um aminoácido e o átomo de nitrogênio (N) do grupo amina de outro aminoácido. Os elementos que compõem a água são removidos como um coproduto da reação. A água (H_2O) se forma a partir do –OH do grupo carboxila de um dos aminoácidos e de um átomo de H do grupo -NH2 do outro aminoácido. A figura 2.10 esquematiza a formação de uma ligação peptídica entre dois resíduos de aminoácidos.

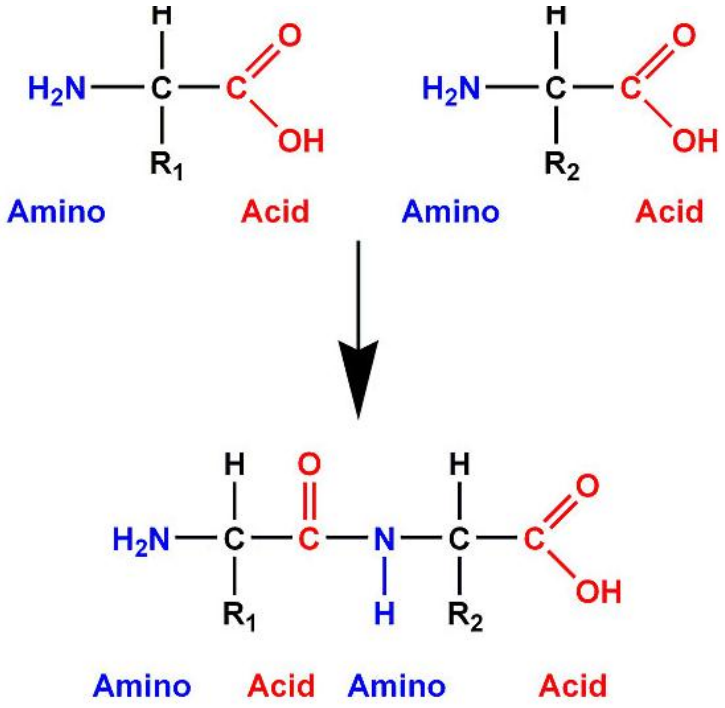

Figura 2.10: Ligação Peptídica.

Fonte: http://www.infoescola.com/bioquimica/ligacao-peptidica/

2.5. Cadeia Principal

Também chamada de cadeia polipeptídica, a repetição do conjunto –N-C_α-C- em uma proteína é chamado de cadeia principal. A direção da cadeia polipeptídica é determinada do grupo amino terminal (grupo N-terminal) até o grupo carboxila terminal (grupo C-terminal) em um polipeptídio.

É fácil ver que podemos calcular o número possível de peptídeos ou proteínas para uma cadeia de n aminoácidos simplesmente elevando 20 a enésima potencia (20^n). Tomando uma proteína típica de 60 resíduos de aminoácidos, o número de proteínas que pode ser feito a partir de 20 aminoácidos é $20^{60} = 10^{78}$ (BETTELHEIM e BROWN, 2012).

O ângulo de torção da ligação entre o carbono α e a carboxila -C_α-C- e chamado de psi (ψ). O ângulo de rotação da ligação entre o carbono α e a amina (-C_α-N-) é chamado de phi (ϕ). O ângulo de torção da ligação peptídica é

denominado ômega (ω) que, para manter o arranjo planar, assume valores próximos de 0^0 ou 180^0, figura 2.11.

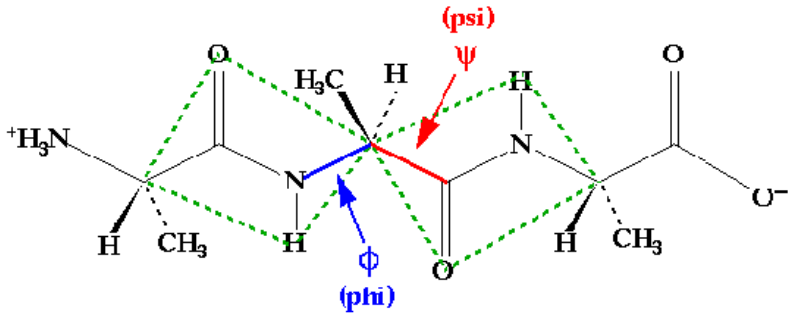

Figura 2.11: Ângulos de torção (diedras) da cadeia principal da proteína.

"*A rigidez da ligação peptídica permite que as proteínas tenham formas tridimensionais bem definidas e a liberdade de rotação em ambos os lados da unidade peptídica e igualmente importante porque permite que as proteínas se enovelem de muitos modos diferentes. De fato, os ângulos phi e psi determinam o caminho que a cadeia polipeptídica adota no espaço as cadeias laterais também possuem ângulos de torção que definem a sua conformação*" (CUSTÓDIO, 2008). São denotados χ_1, χ_2, χ_3 e χ_4 como, por exemplo, na cadeia lateral da lisina figura 2.12.

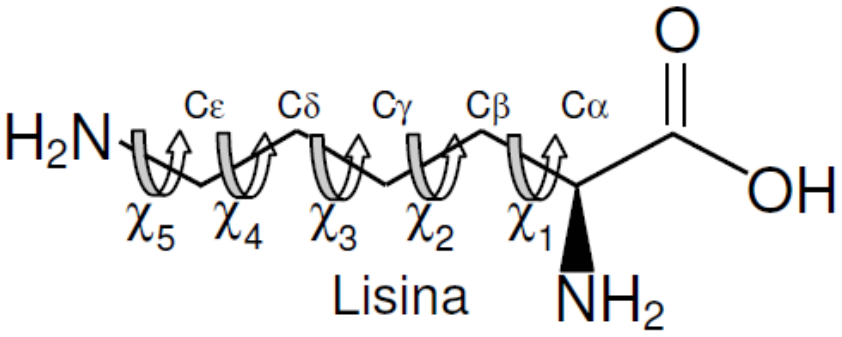

Figura 2.12: Ângulos de torção diedras da cadeia lateral do aminoácido lisina.

2.6.
Gráfico de Ramachandran

Os ângulos ϕ e ψ podem ter qualquer valor entre -180 e +180, porém, muitas combinações de ϕ e ψ são proibidas por interferências estéricas (interações não previsíveis, relacionadas à sobreposição de orbital molecular, distribuição de densidade eletrônica) entre átomos no esqueleto principal da cadeia polipeptídica e entre átomos da cadeia lateral dos aminoácidos. Os valores permitidos e proibidos para os ângulos de torção ϕ e ψ são graficamente demonstrados pelo mapa de Sasisekharan-Ramakrishnan-Ramachandran (RAMACHANDRAN e SASISEKHARAN, 1968), ou simplesmente mapa de Ramachandran.

A figura 2.14 apresenta o mapa de Ramachandran, destacando as regiões permissíveis para a combinação dos ângulos ϕ e ψ. Conforme será discutido nas seções seguintes, as regiões no mapa de Ramachandran representam, em termos de enovelamento, padrões de torção da cadeia polipeptídica (folhas β, hélices α).

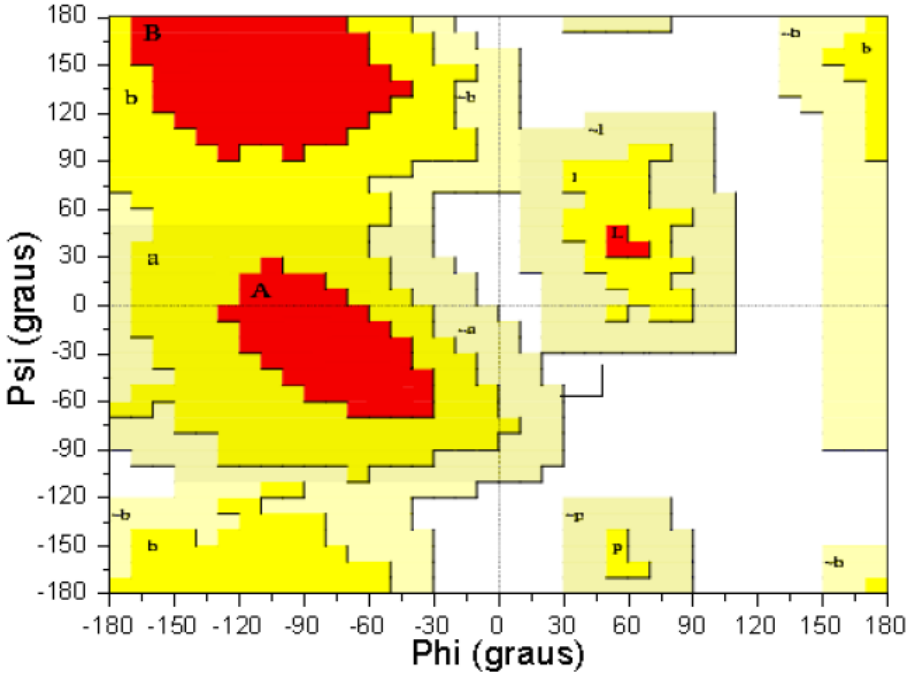

Figura 2.13: Mapa de Ramachandran: a região mais favorável é apresentada em vermelho, região permitida é apresentada em amarelo, região ainda aceitável é apresentada em amarelo claro e a região não permitida em branco. A região em vermelho no canto superior esquerdo representa a região de folhas β paralelas e antiparalelas. A região em vermelho no centro esquerdo, e no centro direito representa a região de hélices α à direita e a esquerda respectivamente. Fonte : (LASKOWSKI, MACARTHUR, *et al.*, 1993).

2.7.
Níveis estruturais nas Proteínas

Uma proteína é estudada em quatro níveis hierárquicos, cada um destes níveis será apresentado a seguir.

2.7.1.
Estrutura primária

De forma muito simplificada a estrutura primária de uma proteína consiste na sequência lineal de seus aminoácidos que formam a cadeia figura 2.15.

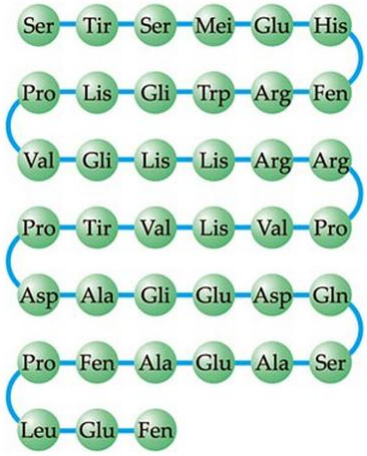

Figura 2.14: Estrutura primaria de uma Proteína.

Fonte: http://www.profpc.com.br/Química_das_células.htm

2.7.2.Estrutura Secundária

As proteínas podem se dobrar ou se alinhar de tal forma que certos padrões se repetem. Esse padrão de repetição é conhecido como estrutura secundária. Existem três classes neste tipo de estruturas: α-hélice, folha-β, voltas e alças. (PAULING e COREY, 1951).O padrão α-hélice se assemelha a uma mola, nesta estrutura as ligações de hidrogênio mantem as tiras das folhas vizinhas unidas, na folha-β o alinhamento ordenado das cadeias das proteínas é mantido por ligações de hidrogênio intermoleculares ou intramoleculares, em quanto às alças e voltas são estruturas irregulares. As voltas são formadas em regiões onde a proteína muda a sua direção, ou seja, após uma estrutura secundária regular em forma de α-hélice e folha-β na figura 2.16 são amostradas estas estruturas.

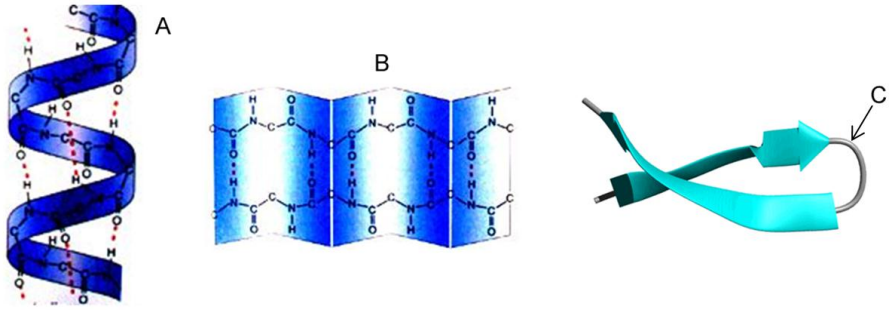

Figura 2.15: Estrutura Secundaria: A) α-Hélice, B) Folha-β, C) Vol ta Fonte: http://mrclay13bio.wikispaces.com/protein+structure

2.7.3. Estrutura terciária

A estrutura terciária de uma proteína é o arranjo de cada átomo no espaço tridimensional, ou como as suas estruturas secundárias estão distribuídas no espaço 3D. A estrutura terciária também é chamada de estrutura nativa.

Através da estrutura terciária de uma proteína é possível analisar ou prever a função que a mesma exerce no organismo. É possível, através de seu estudo, identificar o sítio ativo de enzimas, sítios de ligação em um receptor, ou um local de recombinação para a ação de outra proteína (LEHNINGER, NELSON e COX, 2002), A figura 2.17 apresenta um exemplo.

Figura 2.16: Representação Ribbon da estrutura terciária da Lisozima.

Fonte:
http://my.opera.com/tutoriabiologiaUBAXXI/archive/monthly/?day=20080329

2.7.4.
Estrutura quaternária

As proteínas podem apresentar diversas cadeias (ou subunidades) polipeptídicas formando uma estrutura quaternária. A estrutura quaternária de uma proteína é o arranjo de várias estruturas terciárias no espaço tridimensional figura 2.18.

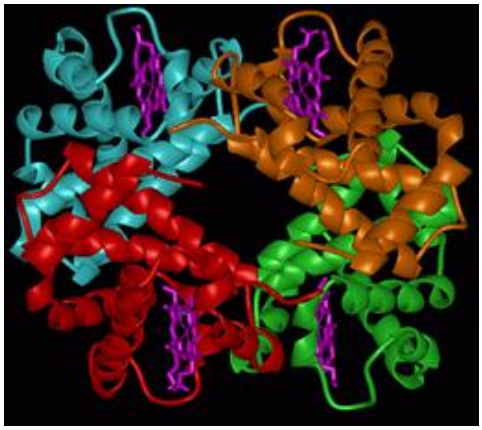

Figura 2.17: Estrutura quaternária numa representação tipo ribbon da Hemoglobina. Fonte: http://portaldoprofessor.mec.gov.br/fichaTecnicaAula.html?aula=1599

2.8.
Predição de estruturas de proteínas

2.8.1.
Predição por modelagem comparativa.

Na modelagem comparativa por homologia uma sequência de resíduos de aminoácidos de uma proteína (sequência alvo) é alinhada contra a sequência de aminoácidos de outra proteína com estrutura conhecida e armazenada no PDB (BERMAN, WESTBROOK, *et al.*, 2000) (sequência e estrutura-molde). Caso a sequência alvo seja bastante similar à sequência de estrutura conhecida, utiliza-se esta estrutura como um molde para a modelagem da estrutura da proteína (JONES, TAYLOR e THORNTON, 1992) (MARTIM-REMON, STUART, *et al.*, 2000).

Segundo Sternberg (STENBERG, 1997) um método de predição baseado em homologia possui seis passos básicos:

1. Encontrar um conjunto de proteínas, determinadas experimentalmente, que são homólogas à sequência da proteína-alvo.
2. Estabelecer um alinhamento de sequência entre a sequência-alvo e as proteínas com estruturas determinadas experimentalmente.
3. Identificar segmentos da cadeia principal da estrutura desconhecida que são conservados em relação a estruturas conhecidas.
4. Modelar as regiões variáveis (regiões de estruturas irregulares).
5. Modelar as cadeias laterais da proteína-alvo
6. Refinar a estrutura predita através de métodos de refinamento (após a construção do modelo da proteína-alvo é necessário otimizá-lo), tendo como base a energia potencial.

A técnica de modelagem comparativa por homologia pode ser aplicada toda vez que for possível detectar uma relação evolucionária entre a proteína-alvo e a proteína-modelo, a qual tem a estrutura 3D conhecida (BUJNICKI, 2006) (JONES, TAYLOR e THORNTON, 1992) (MARTIM-REMON, STUART, *et al.*, 2000).A relação evolucionária entre proteínas é um fator fundamental nos métodos de modelagem comparativa por homologia, pois, parte-se do preceito de que proteínas alvo podem ser moldadas a partir de proteínas homólogas com estrutura 3D determinada experimentalmente. A estrutura destas proteínas é similar no sentido de que aminoácidos com características físico-químicas idênticas ocupam posições iguais em proteínas homólogas. Os ângulos de torção da cadeia principal também preservam certo padrão em seus valores.

Segundo Tramontano (TRAMONTANO e LESK, 2006), a análise comparativa por homologia é o método mais utilizado para predição de estruturas de proteínas, e isto se deve a duas razões: primeiro à qualidade dos modelos previstos, que possuem uma razoável relação evolucionária, apresentam-se com uma acurácia maior do que aquelas produzidas com técnicas diferentes. O segundo motivo, se refere ao fato de que a confiabilidade do modelo pode ser estimada a priori. Desta forma, é possível estimar a qualidade da estrutura prevista.

2.8.2. Predição por reconhecimento de enovelamento – Threading

Estes métodos se baseiam na observação de que uma larga porcentagem de proteínas adota um número limitado de formas de enovelamento. Existem aproximadamente 10 diferentes formas de enovelamento em 50% das estruturas conhecidas (RUSSELL e BARTON, 1994). Através da detecção de similaridades estruturais, as quais não podem ser detectadas unicamente pela similaridade entre as sequências de aminoácidos, são construídos os modelos 3D.

Inicialmente, para uma dada sequência de resíduos de aminoácidos, é construída uma biblioteca de padrões de enovelamento. Se fragmentos da sequência da proteína-alvo se ajustam bem a estas formas de enovelamento, é possível deduzir um alinhamento, mesmo que não haja informação suficiente para construir um modelo 3D completo. Em um segundo momento, a partir das informações obtidas de proteínas com estruturas conhecidas, constrói-se modelos estruturais. Com base no valor retornado de uma função objetivo, estes modelos estruturais são pontuados. A partir da pontuação obtida por cada modelo estrutural todas as conformações construídas são classificadas e o modelo 3D final é escolhido. O alinhamento é frequentemente utilizado para identificar homologias que não podem ser descobertas por um alinhamento par a par de sequências de proteínas.

2.8.3. Predição ab initio (primeiros princípios)

A modelagem por métodos ab initio não depende do conhecimento prévio de estruturas de proteínas, sejam essas homólogas ou cadastradas no PDB, mas, sim, tenta determinar a conformação nativa tridimensional por meio de uma busca no espaço de possíveis conformações (VULLO, 2002). Esses métodos fazem a busca explorando o espaço de valores da energia livre das conformações, pois se sabe que a proteína apresenta a sua energia mínima no momento em que ela atinge a sua conformação nativa (KHIMASIA e COVENEY, 1997).

Para determinar a energia mínima de uma conformação durante o dobramento, utiliza-se uma função de minimização, também conhecida com função de fitness.

A função de minimização é baseada nas leis da física de movimentação em campos potenciais (dinâmica molecular), ou seja, nas interações entre os átomos presentes na sequência (VULLO, 2002). Sendo assim, a função deve conter parâmetros que reproduzam as propriedades energéticas, dinâmicas e estruturais das proteínas.

Esses métodos ab initio podem utilizar que representam a estrutura de uma proteína de forma simplificada (KOLINSKI e SKOLNICKB, 2004), ou seja, existem modelos em que o resultado final do dobramento pode representar fielmente uma proteína dobrada, mas outros podem apresentar resultados que não lembrem a proteína real.

A representação simplificada de uma proteína pode ser baseada em dois modelos: modelos discretos, também conhecidos como modelos baseados em grade (lattice) e modelos livres ou contínuos não-grade (off-lattice).

Os modelos baseados em grade (lattice) para a modelagem do PSP (*Protein structured Problem*) foram, inicialmente, propostos por (LAU e DILL., 1989), posteriormente seguido por diversos outros grupos de pesquisa (KOLINSKI e SKOLNICKB, 2004), (BRANDEN e TOOZ, 1998); (KHIMASIA e COVENEY, 1997).

Nos modelos em grade, os aminoácidos são posicionados em um ponto de uma grade (lattice). A grade utilizada é geralmente quadrada ou cúbica, sendo que um ponto dela pode ser ocupado somente por um único aminoácido. Como esses aminoácidos estão ligados a outros, os seus aminoácidos adjacentes estarão distribuídos na grade, por meio de um comprimento fixo (geralmente de valor unitário para cada eixo da grade).

Apesar de simplificados, os modelos discretos ainda preservam algumas características de uma proteína real, como, por exemplo, as interações entre os resíduos, estejam esses conectados ou não (KHIMASIA e COVENEY, 1997).

2.8.4. Predição de novo

Os métodos de novo são aqueles métodos de predição que, através de um conjunto de funções de classificação (scoring functions) e de funções especiais para cálculo de energia potencial (EP), derivadas de métodos puramente ab initio, buscam a predição de novas formas de enovelamento. Os métodos de novo são os que, atualmente, apresentam os melhores resultados nas predições realizadas no CASP (MOULT, 2005). São exemplos de métodos de novo: ROSETTA (ROBETTA) (ROHL, STRAUSS, *et al.*, 2004) (SIMONS, BONNEAU, *et al.*, 1999), LINUS (SRINIVASAN, FLEMING e ROSE, 2004) (SRINIVASAN e ROSE, 2002) e FRAGFOLD (JONES, TAYLOR e THORNTON, 1992).

Os métodos de novo buscam realizar a predição de novas formas de enovelamento baseando-se em moldes. Esta concepção surge a partir da observação de que quando um novo enovelamento é descoberto, este é composto por motivos estruturais comuns de fragmentos ou de estruturas super secundárias (por exemplo, uma fita beta separada de outra fita beta por uma alfa-hélice esse padrão é chamado de unidade beta-alfa-beta) de proteínas com estruturas conhecidas (TRAMONTANO e LESK, 2006). Desta forma, se existem fragmentos da proteína que se enovelam em estruturas similares, então é possível utilizar esta informação ou estes fragmentos na construção de novos modelos estruturais 3D de proteínas. Esta é a base dos métodos de novo baseados em fragmentos. Nestes métodos, a conformação de uma dada sequência alvo é construída com base em informações extraídas de fragmentos de proteínas com estruturas 3D conhecidas (TRAMONTANO e LESK, 2006). A conformação de uma proteína passa a ser vista como um conjunto de vários fragmentos da sequência de aminoácidos representando motivos estruturais diversos.

A estratégia adotada pelos métodos baseados em fragmentos é coletar as estruturas locais assumidas por pequenos segmentos de fragmentos em estruturas 3D já conhecidas e realizar a combinação destas estruturas locais para produzir um número de prováveis modelos 3D de uma proteína-alvo, onde o modelo final é selecionado levando-se em consideração a energia potencial mínima (derivada de métodos ab initio) (TRAMONTANO e LESK, 2006).

Devido ao fato de que sequências locais semelhantes nem sempre assumem a mesma estrutura 3D, por motivo do efeito do grande número de interações ocorrentes na estrutura terciária da proteína, os métodos de predição baseados em fragmentos não podem simplesmente fragmentar a sequência de aminoácidos da proteína alvo e através de consulta em bases de dados de estruturas de proteínas molde, obter as informações de enovelamento do fragmento em questão e realizar a junção destes fragmentos sem nenhum critério. É necessário que sejam estabelecidos critérios de relação entre fragmentos de forma que se possam determinar os fragmentos com maior probabilidade de inserção durante a construção do modelo final da sequência-alvo. Estes critérios surgem da ideia de que existem interações não covalentes entre os átomos da molécula e de que desta forma uma determinada região da proteína é influenciada por interações que ocorrem em outra região estrutural (TRAMONTANO e LESK, 2006).

Os métodos de predição de novo apresentam vantagens em relação aos outros métodos de predição. A primeira vantagem se refere à capacidade de predição de novas formas de enovelamento, o que não é possível de ser realizado pelos métodos baseados em análise comparativa por homologia. A segunda vantagem se refere à redução do espaço de busca conformacional, o que em métodos de predição ab initio é um grande problema e que demanda grande esforço computacional. Esta redução do espaço conformacional se deve ao fato que em uma simples substituição de um fragmento na proteína-alvo, está-se movendo de uma região de uma proteína um fragmento que já possui uma estrutura com mínima energia potencial. No entanto, apesar de reduzir o espaço de busca conformacional, os métodos de novo, que utilizam fragmentos de proteínas molde, ainda possuem duas principais limitações. A primeira está relacionada ao desafio de tratar o grande espaço de busca conformacional originado pelas diferentes formas de combinação de fragmentos molde. A segunda se refere ao desafio de reduzir a energia potencial da conformação nas regiões onde ocorre a combinação dos fragmentos.

De forma geral, um método de predição baseado em análise e combinação de fragmentos é composto por quatro etapas distintas onde, dada a sequência completa de aminoácidos de uma proteína, este procede:

1. Dividindo a sequência alvo em fragmento.
2. A partir de cada fragmento, realiza-se a busca por sequências (dos fragmentos) similares em um banco de dados de estruturas conhecidas.
3. Os fragmentos-molde são classificados (“ranqueamento”).
4. A partir dos fragmentos-molde e com a utilização de uma técnica de combinação, a estrutura tridimensional é construída.
5. A conformação é refinada.

2.9. Modelo Hidrofóbico-Polar

Este tipo de modelo pertence aos métodos de predição ab initio, este modelo foi desenvolvido por (LAU e DILL., 1989). Neste modelo, o alfabeto dos 20 aminoácidos é reduzido a um alfabeto de apenas duas letras, H e P, onde H representa os aminoácidos hidrofóbicos e P representa os aminoácidos polares ou hidrofílicos. Esse modelo tem como base o fato de existir uma tendência de resíduos hidrofóbicos agruparem-se no interior da molécula proteica. O chamado efeito hidrofóbico, ou força hidrofóbica, é considerada a força predominante no processo de dobramento proteico (LAU e DILL., 1989).

Seguindo a notação adotada por (CLOTE e BACKOFEN, 2000) e (BACKOFEN, 1999), no modelo HP uma sequência s de aminoacidos é um elemento $\{\mathrm{H, P}\}^{*}$. O i-esimo elemento de s é denotado por s_i. Uma conformação c de uma sequência s de comprimento n (o seja, uma sequência de n resíduos) passa a ser definida por:

$$c: [1..n] \rightarrow \mathcal{Y}^{d}$$

Onde $\mathcal{Y}^{d}$ uma malha de dimensão d=2 (para malhas bidimensionais ou quadráticas) ou d=3 (no caso de malhas tridimensionais ou cubicas). Além disso, duas condições são importantes:

1. $\forall 1 \leq i \leq n : \|c(i) - c(i+1)\| = 1,$ onde $\|.\|$ é a distancia euclidiana em $\mathcal{Y}^{d}$.
2. $\forall i \neq j : c(i) \neq c(j).$

A primeira condição é imposta pela estrutura da malha e determina que a distância entre dois resíduos consecutivos é 1 para qualquer configuração válida; a segunda condição evita a existência de ciclos e sobreposições (também chamadas de colisões) de resíduos dentro da malha.

Em uma conformação c, existe ainda a distinção entre monômeros conectados e monômeros vizinhos. Dois monômeros i e j são ditos de conectados em c se, e somente se, $j = i + 1$ ou $j = i - 1$. Deve-se observar que o número de monômeros conectados é fixo e independente da conformação c.

Por outro lado, dois monômeros são vizinhos quando i e j não são conectados e $\|w(i) - w(j)\| = 1$. A figura 2.19 ilustra um exemplo de conformação em Y^2 e Y^3.

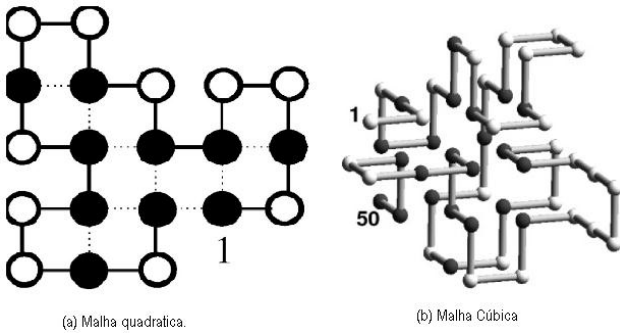

Figura 2.18: Conformações malha quadrática (2.19a) e malha cubica (2.19b). Fonte: (FIDANOVA e LIRKOV)

No modelo HP, dada uma conformação válida, a energia livre E é dada pelo número negativo de vizinhos de tipo H-H, Assim a conformação de energia mínima é a mesma que maximiza o número de contatos H-H. Formalmente, o estado nativo de uma molécula em malha é dada por:

$$E = \sum_{1 \le i+1 \le j \le n} B_{i,j}\delta(r_i, r_j)$$

Onde:

$$B_{i,j} = \begin{cases} 1 \; se \; os \; residuos \; i \; e \; j \; são \; do \; tipo \; H \\ 0 \qquad\qquad\qquad\qquad caso \; contrario \end{cases}$$

$$\delta(r_i, r_j) = \begin{cases} 1 \; se \; os \; residuos \; i \; e \; j \; são \; vizinhos \\ 0 \qquad\qquad\qquad\qquad caso \; contrario \end{cases}$$

Conceitualmente, o modelo HP é relativamente simples se comparado com outras abordagens, além de permitir algumas extensões de modo de tornar-se mais elaborado (BACKOFEN, 1999).

A pesar disso, no modelo HP, o problema de computar o estado nativo que maximiza o número de vizinho H-H é tido como um problema NP-completo tanto em duas dimensões, demonstrando formalmente por (CRESCENZI, GOLDMAN, *et al.*, 1998), quanto em três dimensões, como provado por (BERGER e LEIGHTON, 1998). Em ambos os casos, tal complexidade já era inferida, embora não completamente demonstrada por diferentes estudos (FRAENKEL, 1993) (HART e ISTRAIL, 2006) (UNGER e MOULT, 1993) (PATERSON e PRZYTYCKA, 1996).

2.10. Raio de giro

Raio de giro descreve a propagação global da molécula e é definido como a raiz quadrada de a distância média do conjunto de átomos do seu centro de gravidade comum. Esta medida que indica o quão compacto e encontram um conjunto de pontos (resíduos em uma grade). Conjuntos mais compactos possuem um menor valor R_g, é portanto as melhores conformações são aquelas que possuem menor R_g, o raio de giro é calculado da seguinte maneira:

$$R_g = \sqrt{\frac{\sum_{i=1}^{N}[(x_i - \bar{X})^2 + (y_i - \bar{Y})^2 + (z_i - \bar{Z})^2]}{N}}$$

Onde:

x_i, y_i, z_i : Coordenadas cartesianas do i-ésimo resíduo.

$\bar{X}, \bar{Y}, \bar{Z}$: Medias de todos os resíduos x_i, y_i, z_i.

N : Número de resíduos.

3
Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas no Modelo Hidrofóbico Polar (AGMO-HP)

Conforme exposto nos capítulos 1 e 2 dessa dissertação, o problema de predição de estruturas proteicas é um importante assunto de pesquisa multidisciplinar. Para resolvê-lo foram propostos diversos modelos com características próprias. Este trabalho baseia-se no modelo HP que usa átomos virtuais de dois tipos, hidrofóbicos e hidrofílicos. Trata-se de um modelo reduzido fundamentado na observação de que os aminoácidos hidrofóbicos tendem a ficar no centro da proteína e os aminoácidos hidrofílicos, em torno dos aminoácidos hidrofóbicos. Entre as técnicas aplicadas para resolver este tipo de problema estão algoritmos genéticos, que trabalham com conformações inválidas ou não. No primeiro caso, isto pode causar um detrimento do desempenho, enquanto que, no segundo caso, são usados mecanismos de reparo que podem, em etapas mais avançadas da evolução, destruir os contatos hidrofóbicos já formados. Por outro lado, mesmo que essas técnicas consigam conformações com estados baixos energia, as conformações resultantes podem não existir na natureza. Com o objetivo de melhorar o desempenho e encontrar estruturas proteicas mais naturais, este trabalho apresenta um modelo baseado em um algoritmo genético multiobjetivo. O referido modelo, denominado Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas no Modelo Hidrofóbico Polar AGMO-HP, tem as seguintes vantagens:

1. Menor número de avaliações da função objetivo.
2. Método pseudo adaptativo dos operadores genéticos.
3. Estruturas proteicas mais naturais, de baixa energia e compactas.
4. Fatores de medição da compactação dos aminoácidos hidrofóbicos e hidrofílicos baseados no raio de giro.

3.1. O modelo do Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas no Modelo Hidrofóbico Polar (AGMO-HP)

Uma vez que o problema de predição de proteínas no modelo HP em malhas 3D pode ser visto como um problema de otimização multiobjetivo conforme apresentado no capítulo anterior, o principal interesse deste trabalho é desenvolver um modelo que permita obter conformações de proteínas mais naturais e de mínima energia através da compactação e do número de contatos hidrofóbicos, respetivamente. A compactação será analisada nos aminoácidos hidrofóbico e hidrofílico das proteínas, usando o raio de giro.

As próximas seções deste capítulo apresentarão o algoritmo do modelo proposto neste trabalho, incluindo o a descrição de suas etapas, representação dos indivíduos, função de avaliação, população inicial, operadores genéticos, procedimento para conservação da diversidade e demais detalhes importantes referentes ao processo evolutivo.

3.1.1. Algoritmo Genético Multiobjetivo para Predição de Estruturas Proteicas Naturais (AGMO-HP)

A figura 3.1 mostra o diagrama de fluxo geral do algoritmo genético multiobjetivo que será usado neste trabalho. Em ordem de prioridade, este algoritmo tem os seguintes objetivos específicos:

1. Minimização do número de colisões das conformações obtidas.
2. Maximização do número de contatos hidrofóbicos.
3. Minimização da compactação dos aminoácidos hidrofóbicos das conformações obtidas. Em cada conformação (individuo) é calculada a compactação conforme detalhado na secção 2.11 dessa dissertação.
4. Minimização da compactação dos aminoácidos hidrofílicos das conformações obtidas. Esta compactação será calculada conforme apresentado na secção 2.11.

Nas próximas subseções, este algoritmo será explicado com maior detalhamento.

1. Gerar população inicial.
2. Repetir para todas as gerações.
3. Repetir ate gerar k novos indivíduos.
4. Selecionar dois pais pelo torneio.
5. Fazer Crossover o tipo de crossover será escolhido pela roleta.
6. Acumular créditos do operador de crossover executado em 5.
7. Fazer Mutação o tipo de mutação será escolhido pela roleta.
8. Acumular créditos do operador de mutação executado em 7.
9. Adicionar o individua gerado a uma população temporal.
10. Fim
11. Fim
12. Inserir os indivíduos da população temporal por um método de crowding.
13. Realocar os operadores da roleta de mutação e crossover segundo o seu desempenho em G gerações.

Figura 3.1: Algoritmo Genético Multiobjetivo proposto para predição de Estruturas Proteicas.

3.1.2. Representação dos indivíduos

Cada monômero da conformação possui coordenadas cartesianas (x, y, z) que precisam ser codificadas no genótipo. Para isto é usada a representação absoluta proposta por (UNGER e MOULT, 1993), na qual as direções são números inteiros com a seguinte disposição: 0 (esquerda), 1 (direita), 2 (acima), 3 (abaixo), 4 (atrás), 5 (frente).

Por exemplo, dada uma sequência de comprimento n = 10, HPPHPPHPPH, representada por um cromossomo [514130405] de comprimento n-1, tem-se um conjunto de coordenadas cartesianas, conforme ilustrado na figura 3.2, onde o primeiro aminoácido foi colocado na posição (0,0,0) e as demais coordenadas foram geradas somando ou subtraindo uma unidade na coordenada anterior, de acordo com a sua direção absoluta.

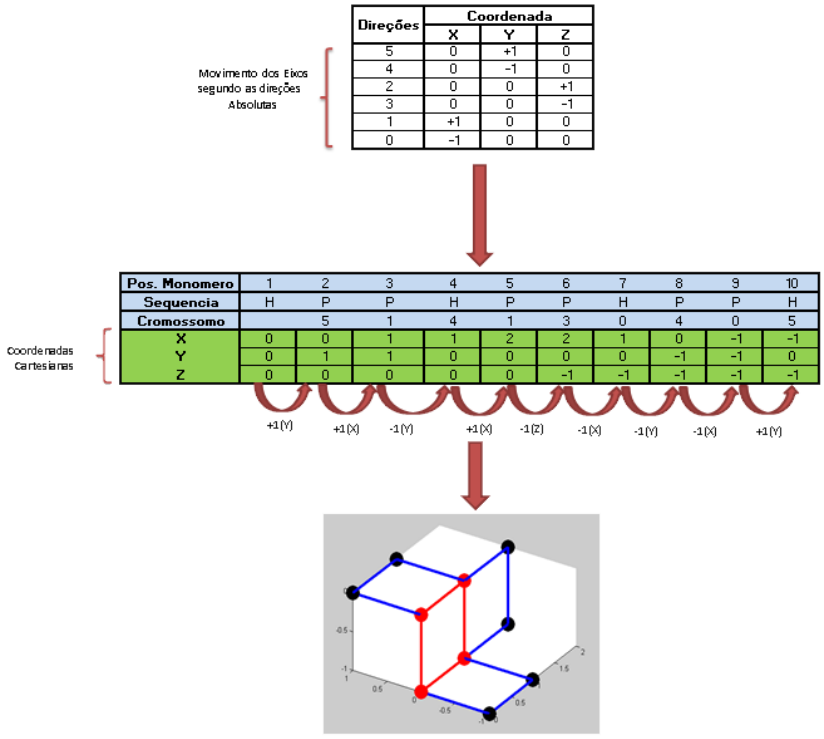

Figura 3.2: Exemplo de representação dos cromossomas no modelo HP. É amostrada a sequência absoluta de um cromossomo e a sua dinâmica para a obtenção da malha em 3D. Na malha, os aminoácidos hidrofóbicos em vermelho e os hidrofílicos em preto.

De acordo com a Apêndice, a função de aptidão é o valor da qualidade de um individuo e é calculada através da função de energia do modelo original para malhas 3D, no modelo HP (DILL., 1985) (LAU e DILL., 1989). Uma vez que, conforme discutido na seção 2.10, a energia livre é dada pelo número negativo de monômeros vizinhos do tipo H-H, a energia mínima em uma conformação é a mesma que maximiza o número de contatos H-H. De forma simplificada, uma conformação *C*, formalmente, pode ser escrita como:

$$f(C) = \max(NC_{HH})$$

Onde NC_{HH} é o número de contatos hidrofóbicos de uma conformação C.

3.1.4. População Inicial

Conformei o mecanismo de reparo proposto por (CUSTÓDIO, 2008) a população inicial é gerada aleatoriamente com indivíduos livres de colisões, através de uma distribuição normal de números inteiros. Os indivíduos são testados e, cada vez que um novo gene é adicionado, a estrutura toda é avaliada para verificar se este novo gene ocasiona colisões. Caso ocasione, o último gene é apagado e é gerado um novo, repetindo-se esse processo até formar um indivíduo válido ou atingir um número máximo determinado de iterações. Caso o número máximo de iterações tenha sido atingido, a sequência toda é apagada, sendo gerada nova sequência. Esta etapa do algoritmo é detalhada na figura 3.3.

1. Repetir ate completar o tamanho da população desejado
2. Repetir enquanto ate completar o tamanho do cromossomo desejado
3. Repetir enquanto o numero de tentativas seja menor o máximo permitido.
4. Gerar um gene aleatoriamente entre 0 e 5.
5. Adicionar o gene novo para o cromossomo atual.
6. Testar se o novo gene produziu colisões no cromossomo.
7. Se não produz colisões vai para o passo 2.
8. Eliminar gene do cromossomo.
9. Fim
11. Se o numero de máximo de tentativas foi atingido sem conseguir um cromossomo valido, este cromossomo será eliminado, voltar para o passo 2.
12. Fim
13. Adicionar o novo cromossoma na população.
14. Fim

Figura 3.3: Algoritmo para geração dos indivíduos da população inicial do algoritmo genético multiobjetivo proposto

3.1.5. Operadores genéticos

A seguir, os operadores genéticos de cruzamento e mutação são descritos, sendo descrito também um processo de atribuição de créditos para quantificar o desempenho de cada operador na obtenção de cromossomos de qualidade. Em algoritmos genéticos, quando são usados vários tipos de cruzamento ou mutação, costuma-se empregar uma roleta para fazer a escolha do operador específico que deverá ser utilizado. No caso deste trabalho, é proposta uma roleta pseudo adaptativa que apresenta duas etapas importantes: ajuste inicial dos pesos na roleta e troca dos pesos baseada no ranking de sucesso dos operadores em um número determinado de gerações. Além disso, também é considerado o custo de aplicação deste operador. Estes métodos serão detalhados nas próximas subseções.

3.1.5.1. Calculo dos créditos para a operação de cruzamento

Quando uma operação de cruzamento é realizada, a qualidade da nova solução é avaliada para determinar os créditos pelo sucesso ou não daquele operador da seguinte forma:

1. Se o filho gerado pelo cruzamento for melhor do que o melhor da população, o operador recebe créditos de $(2\ x\ fator_{ganho})/custo$, onde $fator_{ganho} = apt_{filho} - apt_{melhor}$, senão, pula-se para o passo 2.
2. Se o filho gerado pelo cruzamento for melhor que o melhor de seus pais, o operador recebe créditos de $(0{,}5\ x\ fator_{ganho})/custo$, onde $fator_{ganho} = apt_{filho} - apt_{pae}$
3. Se o filho não atende a quaisquer dos passos anteriores, o crédito é igual a zero.

O $custo$ é o número de vezes em que a função de avalição foi executada para obter o filho. A figura 3.4 apresenta o fluxograma deste processo

1. Pegar o melhor da população atual
2. Fazer um torneio entre o melhor da população e o novo individuo gerado pelo crossover
3. Se o individuo ganha o torneio
4. O credito deste operador será calculado como duas vezes a diferença entre suas aptidões
5. Senão
6. Pegar o melhor pai e fazer um torneiro com o individuo gerado
7. Se o individuo for melhor que o seu melhor pai.
8. O credito deste operador será calculado como 0.5 vezes a diferença entre suas aptidões.
9. Senão
10. O credito será igual a zero.
11. Fim
12. Fim
13. O credito final deste operador será calculado fazendo a divisão do credito obtido nos passo anteriores pelo custo(numero de vezes que foi executada a função de aptidão por este operador)
14. Fim

Figura 3.4: Algoritmo para calculo dos créditos quando é executado o crossover.

3.1.5.2. Cruzamento de um ponto

De acordo com a secção 2.2, o cruzamento de um ponto gera duas soluções filhas pela troca de informação dos pais, em um ponto aleatoriamente escolhido. De forma também aleatória é selecionada uma das soluções filhas para ser inserida na população, posteriormente, através do procedimento de substituição parental.

A função de aptidão é aplicada uma única vez para obter a avaliação do filho selecionado. Assim, o custo deste operador é *O(n)=1*, vide detalhes no exemplo da figura 3.5.

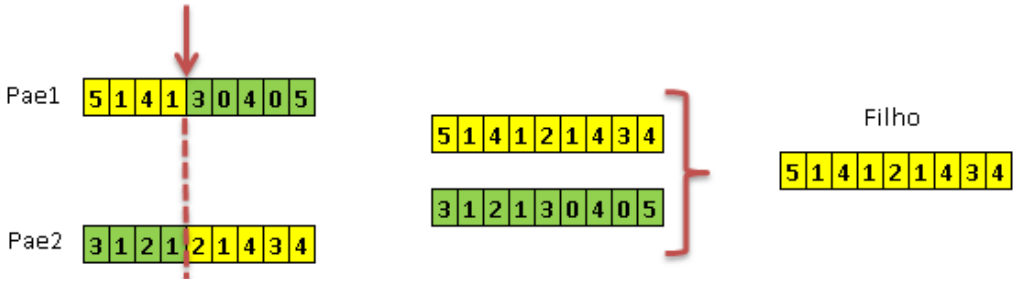

Figura 3.5 : Cruzamento de um ponto em dois cromossomas para sequência de comprimento dez.

Conforme visto na secção 2.2, o operador cruzamento de um ponto gera duas soluções filhas, a partir da troca de informação entre os pais, em pontos aleatoriamente escolhidos. Em seguida, um dos filhos é aleatoriamente escolhido para ser inserido na nova população. O custo deste operador é *O(n)=1*, como no caso anterior. Os detalhes são apresentados na figura 3.6.

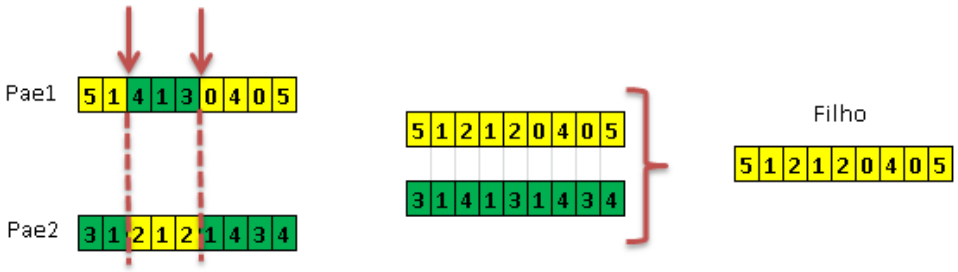

Figura 3.6: Cruzamento de dois pontos para em uma sequência de comprimento dez.

3.1.5.4. Cruzamento multiponto

O cruzamento multiponto gera duas soluções filhas a partir de dois pais, Por exemplo, para cromossomos de comprimento l:

1. O número de pontos de corte r, é gerado aleatoriamente entre os valores [2 $\frac{l}{2}$ >, o valor inferior deste intervalo fica a partir de 2, pois se fosse 1 geraria pontos de corte igual ao comprimento do cromossomo, isto pode ser prejudicial em estágios avançados da evolução onde poderia destruir os contatos H-H já formados, em quanto que o valor superior $\frac{l}{2}$ é para garantir que o número de pontos de corte não seja muito pequeno e em alguns casos equivalente aos operadores de cruzamento anteriores, por exemplo se $r = \frac{l}{2}$., o número de pontos de corte seria igual a $\left(l/\frac{l}{2}\right) = 2$ equivalente ao cruzamento de dois pontos.
2. Aleatoriamente, são gerados r pontos de corte, $p = [1\ l]$. Definidos os pontos de corte, procede-se com uma operação de cruzamento, conforme visto na secção 2.2.

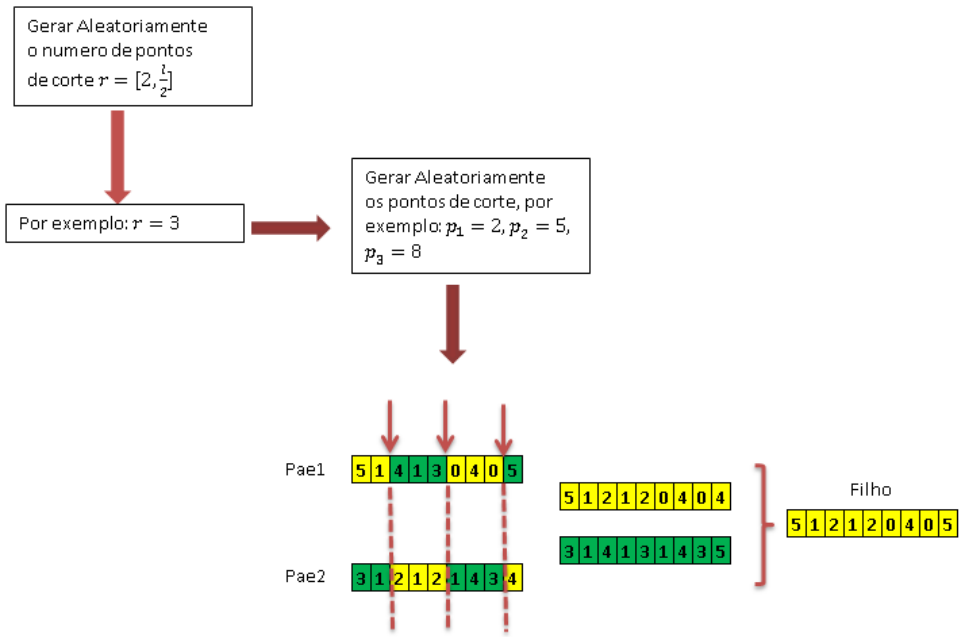

Figura 3.7: Cruzamento multiponto em sequências de comprimento dez

3.1.5.5. Cruzamento hidrofóbico – multiponto

Este operador gera duas soluções filhas a partir dos pais, mas com duas peculiaridades: as posições dos aminoácidos hidrofílicos não são alteradas, e o cruzamento é aplicado somente nos aminoácidos hidrofóbicos. Adicionalmente, este operador busca atuar diretamente nos aminoácidos hidrofóbicos da proteína. O cruzamento hidrofóbico-multiponto é executado de acordo com os seguintes passos:

1. Sejam os pais, $pai1$ e $pai2$. São gerados pais temporários usando os genes que pertencem num aminoácido hidrofóbico. Geram-se, então, $pai1^*$ e $pai2^*$.
2. O cruzamento multiponto simples é aplicado nos pais, $pae1^*$e $pae2^*$, sendo gerados os filhos, $filho_1^*$e $filho_2^*$.

3. Os genes do $filho_1^*$ substituirão os genes do $pai1$ da seguinte forma: o primeiro gene do $filho_1^*$ substitui o primeiro gene do $pai1$ que possui um aminoácido hidrofóbico, o segundo gene do $filho_1^*$ substitui o seguinte gene do pai que possui um aminoácido hidrofóbico, e assim por diante com os demais genes do filho. Depois de substituir todos os genes do $filho_1^*$ e do $filho_2^*$ nos seus respectivos pais, estes novos cromossomos são os filhos $filho_1$ e $filho_2$.
4. Um dos filhos ($filho_1, filho_2$) é escolhido aleatoriamente para ser inserido na nova população.

Um esquema do funcionamento deste operador é apresentado na figura 3.8.

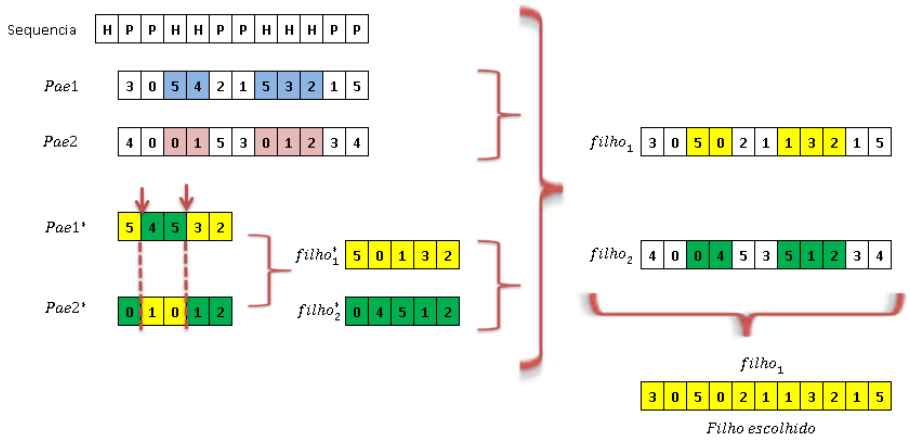

Figura 3.8: Operador de Crossover Hidrofóbico – Multiponto para sequências de comprimento doze.

3.1.5.6. Cruzamento hidrofílico – multiponto

Este operador gera duas soluções filhas a partir dos pais. Ao contrário do operador anterior, este operador não altera as posições iniciais dos aminoácidos hidrofóbicos, e o cruzamento é realizado apenas nos aminoácidos hidrofílicos. Este operador busca atuar diretamente nos aminoácidos hidrofílicos da proteína. O cruzamento hidrofílico-multiponto segue os passos abaixo:

1. Sejam os pais, $pai1$ e $pai2$. São gerados pais temporários usando os genes que possuem um aminoácido hidrofílico. Geram-se, então, $pai1^*$e $pai2^*$.
2. O cruzamento multiponto simples é aplicado nos pais $pai1^*$e $pai2^*$, sendo gerados os filhos, $filho_1^*$ e $filho_2^*$.
3. Os genes do $filho_1^*$ substituirão os genes do $pai1$ da seguinte forma: o primeiro gene do $filho_1^*$ substitui o primeiro gene do $pai1$ que possui um aminoácido hidrofílico, o segundo gene do $filho_1^*$ substitui o gene seguinte do pai que possui um aminoácido hidrofílico, e assim por diante com os demais genes do filho. Depois de substituir os genes do $filho_1^*$ e do $filho_2^*$ nos seus respetivos pais, esses novos cromossomos são os filhos $filho_1$ e $filho_2$.
4. Um dos filhos $(filho_1, filho_2)$ é escolhido aleatoriamente para ser inserido na população.

Um esquema do funcionamento deste operador é apresentado na figura 3.9.

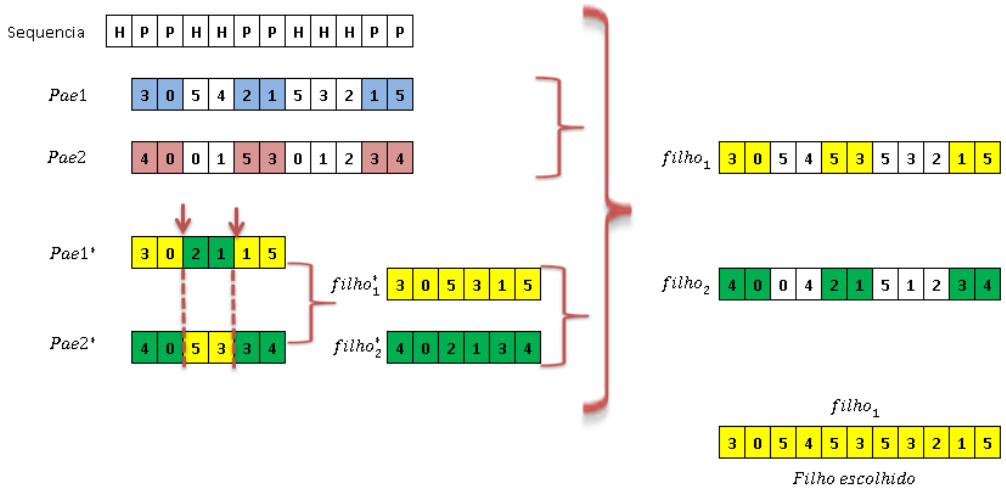

Figura 3.9: Operador de cruzamento hidrofílico–multiponto para sequências de comprimento doze.

3.1.5.7. Cálculo dos créditos para a operação de mutação

O cálculo dos créditos para o operador de mutação é feito de forma semelhante ao do cruzamento:

1. Se o filho gerado pela mutação for melhor do que o melhor da população o operador recebe créditos de $(2\ x\ fator_{ganho})/custo$, onde $fator_{ganho} = apt_{filho} - apt_{melhor}$, senão, pula-se para o passo 2.
2. Se o filho gerado pela mutação for melhor que o pai, o operador de mutação recebe créditos de $(0{,}5\ x\ fator_{ganho})/custo$, onde $fator_{ganho} = apt_{filho} - apt_{pai}$
3. Se o filho não atende a quaisquer dos passos anteriores, o crédito é igual a zero.

O $custo$ é o número de vezes em que a função de avalição foi executada para obter o filho. A figura 3.10 apresenta o diagrama de fluxo deste processo.

1. Pegar o melhor da população atual
2. Fazer um torneio entre o melhor da população e o novo individuo gerado pelo crossover
3. Se o individuo ganha o torneio
4.O credito deste operador será calculado como duas vezes a diferença entre suas aptidões
5. Senão
6.Pegar o pai e o novo individuo gerado e fazer um torneio
7.Se o individuo for melhor que o seu pai.
8. O credito deste operador será calculado como 0.5 vezes a diferença entre suas aptidões.
9.Senão
10. O credito será igual a zero.
11. Fim
12. Fim
13. O credito final deste operador será calculado fazendo a divisão do credito obtido nos passos anteriores pelo custo(numero de vezes que foi executada a função de aptidão por este operador)
14. Fim

Figura 3.10: Algoritmo para cálculo dos créditos quando é executado o mutação.

3.1.5.8. Mutação Simples

Este tipo de mutação é usada para gerar diversidade na população. Para cada um dos genes do cromossomo sorteia-se um valor v, entre zero e um. Se o valor sorteado for maior ou igual que um parâmetro chamado de '$bitwise$', o gene é trocado, aleatoriamente, por um número inteiro, entre zero e cinco. Os detalhes deste processo são apresentados na figura 3.11

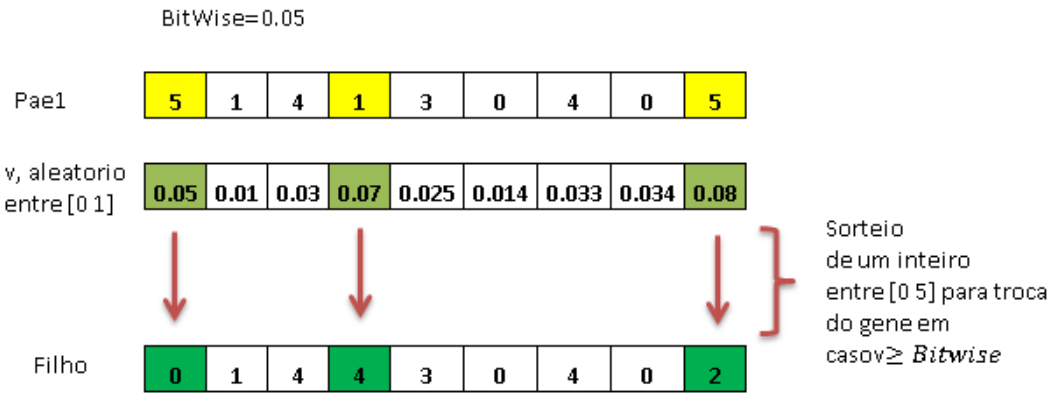

Figura 3.11: Operador de Mutação Simples

O custo deste operador é 1, pois o novo cromossomo é avaliado ao final, uma vez que todos os genes são trocados como pode ser observado na figura 3.11.

3.1.5.9. Mutação com busca exaustiva acumulada

Este operador é inspirado no de Mutação com Busca Exaustiva (EMUT), proposto por (CUSTÓDIO, 2008), no qual os novos indivíduos são gerados como se fossem resultantes de uma mutação simples, porém, com uma diferença: a escolha do novo valor do gene está condicionada ao teste das cinco direções possíveis e aquela que possui a melhor aptidão é mantida.

O operador proposto tem como objetivo realizar uma busca exaustiva, acumulando as melhorias em cada troca feita na cadeia do cromossomo, e

explorando regiões de maior aptidão. Os passos seguintes descrevem a atuação deste operador:

1. Dado um cromossomo de comprimento l e aptidão apt, geram-se l valores aleatórios v entre [0 1] para cada um dos genes do cromossomo. Estes valores são comparados com o *bitwise*, aplicando-se a mutação na posição em que $v \geq bitwise$.
2. Sejam x, $x+2$, $x+3$ as posições nas quais será aplicada a mutação. Para a posição x do cromossomo, são testados os seis valores possíveis. Se algum desses valores resulta em um cromossomo de melhor aptidão, $apt_1 > apt$, o cromossomo é atualizado pelo de melhor aptidão. Na posição $x+2$ também é feito o teste dos seis valores; se algum desses valores resulta em uma aptidão $apt_2 > apt_1$,o cromossomo é atualizado pelo último gerado. E para a posição $x+3$, da mesma forma, se for encontrado um valor para o cromossomo anterior, tal que $apt_3 > apt_2$, então, o cromossomo é atualizado pelo novo encontrado, e assim por diante, se houver mais posições para aplicar mutação. Este processo é apresentado na figura 3.12.

O custo deste operador é $k\ x\ 6$, onde k é o número de genes no qual o teste $v \geq bitwise$ é verdadeiro, e 6 é o número de vezes que a função de aptidão foi utilizada para avaliar um indivíduo.

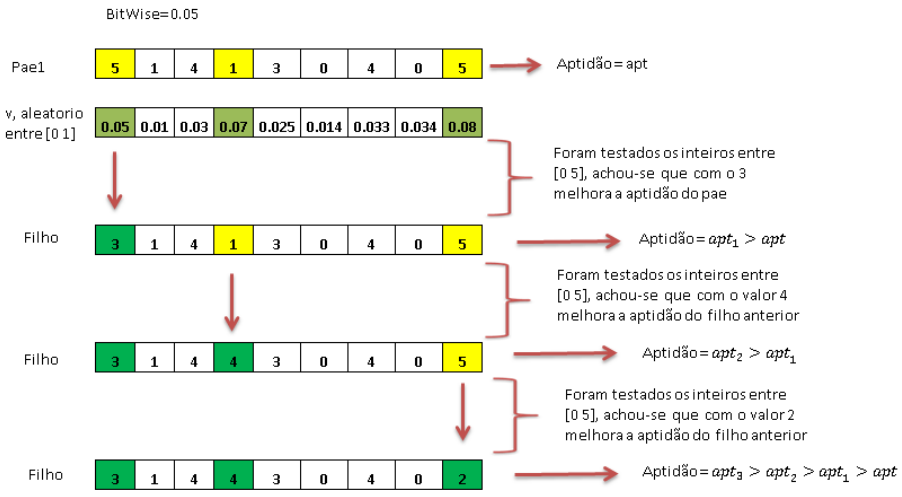

Figura 3.12: Operador de Mutação com busca exaustiva Acumulada

3.1.5.10. Mutação Troca de Segmento sem Colisões

No seu trabalho, (CUSTÓDIO, 2008) utiliza também uma mutação baseada em segmentos, no qual o segmento introduzido na cadeia do cromossomo é gerado aleatoriamente. Isto pode ocasionar um aumento da probabilidade de gerar soluções inválidas, sendo necessário o uso de uma função de reparo na aplicação deste operador.

Neste trabalho, não utilizamos funções de reparo de soluções invalidadas, porém é utilizado um operador mais conservador, diminuindo a probabilidade de obter soluções inválidas. Este operador é executado conforme segue:

1. O tamanho do segmento s é escolhido aleatoriamente na faixa de $[2;\ m]$, onde m é o tamanho máximo do segmento. O valor de m é fixado no início do algoritmo.
2. A posição inicial do segmento x_0 é gerada aleatoriamente. Este valor inteiro é escolhido no intervalo de posições $[1;\ l-s]$, onde l é o comprimento do cromossomo.
3. Aleatoriamente, é gerado um segmento de tamanho s. Esse segmento não apresenta colisões em sua cadeia. Além disso, as posições inicial e final não geram colisões locais com os genes nas posições x_0-1 e x_s+1.
4. Se um segmento for encontrado no passo anterior, o mesmo é usado em substituição ao segmento do cromossomo nas posições de x_0 até x_s. Caso não seja encontrado, volta-se para o passo 2 até um número máximo de tentativas $N_{tentativas}. = inteiro(^{l}/_{(2\ x\ m)})$ ser atingido.

   Na hipótese de não encontrar um segmento que possa ser utilizado, no limite máximo de tentativas, o cromossomo original será devolvido.

O custo deste operador é calculado da seguinte forma: $\left(^{s}/_{l}\right) x\ N_{intentos}$. Assim, para $s = l$ e $N_{tentativas} = 1$, observa-se que o $custo\ =\ 1$, igual o custo de avaliar um cromossomo de tamanho l. Na figura 3.13 o funcionamento deste operador é apresentado detalhadamente.

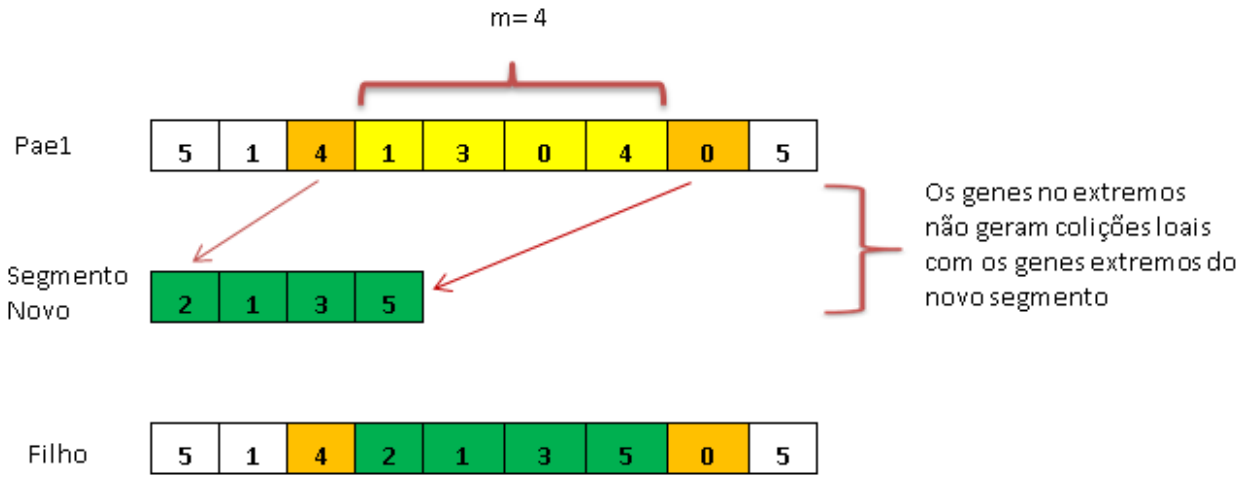

Figura 3.13: Operador de Mutação Troca de Segmento sem Colisões

3.1.5.11. Mutação Simples Acumulada

Esta mutação segue os mesmos passos que a Mutação com Busca exaustiva acumulada da secção 3.1.5.7, com a diferença que no passo 2 não são testadas as seis possibilidades de busca de uma melhor aptidão. Este operador, em vez disso, gera valores aleatórios no intervalo [0; 5] e faz o teste: se achar uma melhora esse caminho é mantido, seguindo com o próximo gene. Os detalhes deste operador são apresentados na figura 3.14.

3.1.5.12. Mutação 2-opt com Memoria

Este operador será utilizado como um otimizador local, verificando todos os pares de genes que existem dentro de um cromossomo e calculando, para cada par, a aptidão. Caso o par seja invertido, o operador procede a inversão que garante o maior ganho na aptidão do cromossomo pai.

Este operador explora a vizinhança do cromossomo pai para realizar a busca, sendo esta vizinhança definida por todos os cromossomos que diferem do pai por uma inversão simples.

Para um cromossomo de comprimento n, o número de pares existentes é dado por $C_2^n = n!/2!\,(n-2) = \frac{n(n-1)}{2}$, e corresponde ao número de avaliações da função de aptidão. Este operador é dispendioso computacionalmente, devendo sua aplicação seguir uma estratégia especial para não comprometer o desempenho do modelo. Esta estratégia consiste em aplicá-lo com taxas baixas e em utilizar um mecanismo de memória para não repetir a avaliação de um cromossomo. Este operador funciona de acordo com o seguinte procedimento:

1. Seja M o conjunto de todos os cromossomos nos quais este operador já foi aplicado. Verifica-se se o cromossomo pai existe no conjunto M. Caso não exista, procede-se a geração de cromossomos através de combinações (troca) de dois genes no cromossomo pai, guardando os mesmos em um conjunto T. Todavia, se o cromossomo pai já existir, executa-se a seleção por roleta mais uma vez para, em seguida, aplicar outro operador de mutação.
2. Calcula-se a aptidão de cada um dos cromossomos do conjunto T, sendo $filho_{melhor}$ o melhor de todos os cromossomos.
3. Faz-se um torneio entre o cromossomo pai e o cromossomo $filho_{melhor}$. O ganhador do torneio é o resultado deste operador. É importante ressaltar que este operador somente gera créditos se o filho fosse melhor que o pai.

A figura 3.15 apresenta um esquema de como são gerados os cromossomos por troca de pares de genes.

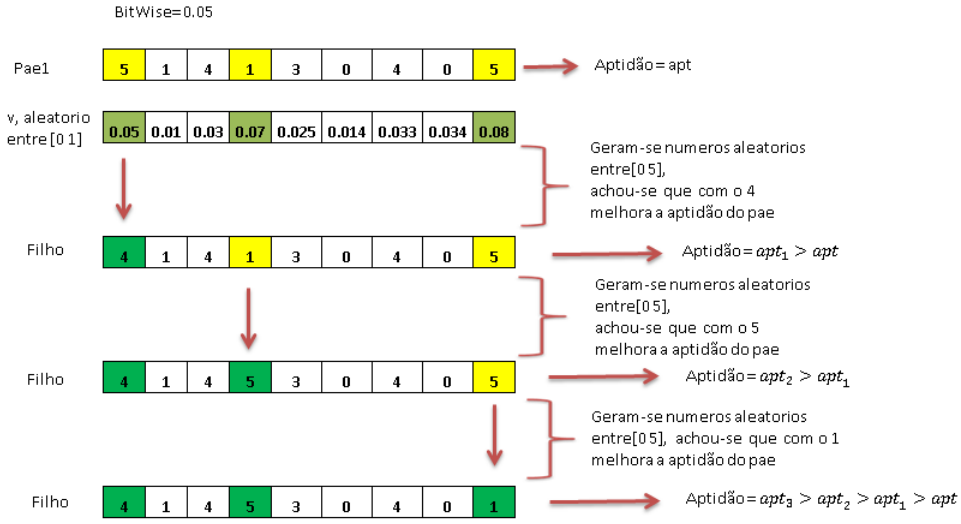

Figura 3.14: Operador de mutação simples acumulada

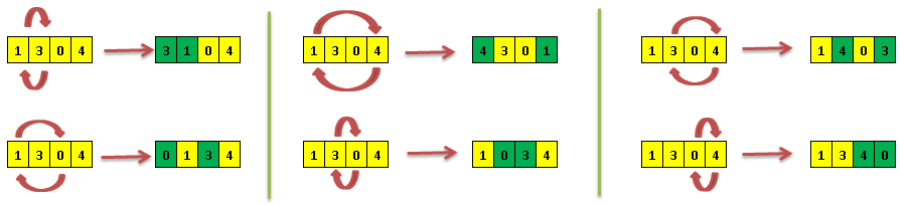

Figura 3.15: Mecanismo de geração dos cromossomos por troca de pares para no Operador de mutação 2-op com memoria.

3.1.5.13. Roleta baseada em créditos

Este trabalho utiliza diversos operadores de cruzamento e mutação. A escolha de qual deles será utilizado, em um dado momento da execução do algoritmo genético, será feita através de uma roleta.

O ajuste dos pesos para os operadores de cruzamento e mutação na roleta é feito, geralmente, através de métodos empíricos, mas não há garantias que valores ótimos serão obtidos para estes parâmetros. A roleta utilizada neste trabalho é executada com os seguintes passos:

1. Sejam os operadores $op1$, $op2$ e $op3$ e os pesos $w1 > w2 > w3$. Inicialmente, é feita uma análise do desempenho dos operadores para determinar a ordem de prioridade inicial de aplicação dos operadores na roleta. Esta análise é realizada através da comparação do desempenho dos operadores com a busca aleatória e entre si. Por exemplo, se o melhor foi $op2$, seguido de $op1$ e de $op3$, no final, então será atribuído o peso $w1$ para o operador $op2$, o peso $w2$ para $op1$ e $w3$ para $op3$, como seus pesos iniciais.
2. Os pesos $w1, w2, w3$ serão reposicionados, segundo o desempenho dos operadores, medido pelos créditos obtidos em G gerações, onde G é um parâmetro de entrada, definido no início do algoritmo. Com isto a roleta tem maior adaptabilidade, pois recebe informação do estado da evolução do algoritmo genético e reposiciona seus pesos dando prioridade para o melhor operador nesse estado do AG.

A figura 3.16 mostra a dinâmica de funcionamento da roleta proposta neste trabalho.

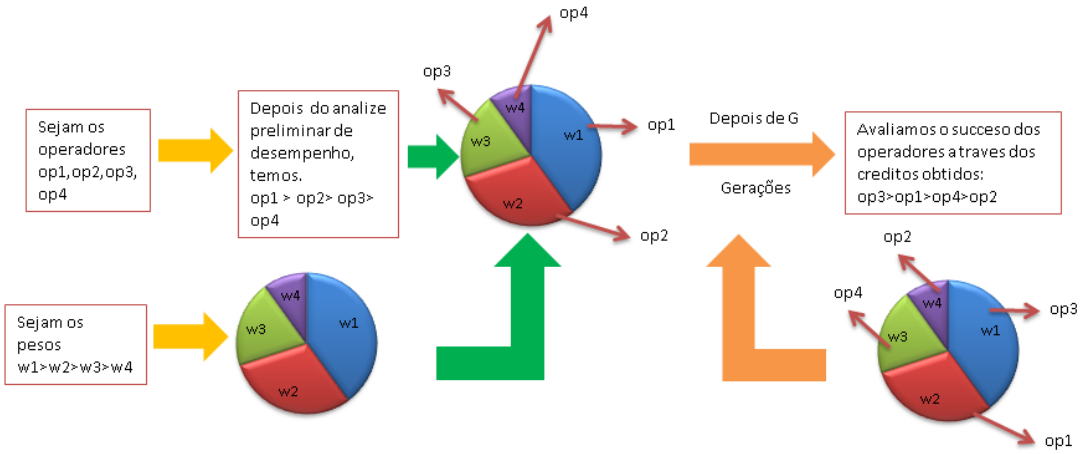

Figura 3.16: Dinâmica da roleta baseada em créditos

3.1.5.14. Método de Seleção dos Indivíduos para Reprodução

Como foi visto na secção 2.3, quando se utiliza AG's em problemas de otimização multiobjetivo, um aspecto importante é a seleção dos indivíduos. Esta deve conduzir a busca na direção da frente ótima de pareto. Portanto, o torneio proposto nesta dissertação é baseado no cumprimento dos objetivos listados abaixo, em ordem de dominância:

1. Minimização do número de colisões das conformações obtidas.
2. Maximização do número de contatos hidrofóbicos.
3. Minimização da compactação dos aminoácidos hidrofóbicos das conformações obtidas.
4. Minimização da compactação dos aminoácidos hidrofílicos das conformações obtidas.

Neste trabalho, o tamanho do torneio utilizado é 2. É importante destacar que o algoritmo do torneio proposto é dotado de flexibilidade para soluções inválidas que estão muito perto do espaço das soluções válidas. Assim, uma solução inválida pode competir com uma solução válida se cumprir a seguinte condição:

1. Seja uma solução válida *c1* com número colisões $c1\_col = 0$ e contatos hidrofóbicos igual a $c1\_HC$, e seja uma solução inválida *c2* com $c2\_col > c1\_col$ e $c2\_HC > c1\_HC$. Então, a solução $c2$ é melhor que $c1$, se:
   $c2\_col - c1\_col <= maxColTol$ e
   $c2\_HC - c1\_HC >= minContH$, onde
   $maxColTol$ e $minContH$ são parâmetros ajustáveis definidos no inicio do algoritmo. O primeiro é a tolerância máxima que duas soluções podem diferir para considerar a solução inválida *c2* como válida em relação a solução *c1* este parâmetro deve ser pequeno, senão as soluções inválidas podem aumentar na população. O segundo parâmetro, $minContH$, avalia a quantidade de contatos hidrofóbicos que uma solução inválida deve ter para ainda ser considerada como válida. É importante falar que somente é interessante conservar uma solução inválida se o parâmetro $minContH$ for suficientemente grande, sendo os esquemas com contatos hidrofóbicos destas soluções transmitidos na população. Os detalhes do algoritmo do torneio proposto neste trabalho são apresentados na figura 3.17.

3.1.5.15. Método de Substituição Parental

O método de substituição parental é muito importante em AG's multiobjetivos, pois contribui para a conservação da diversidade na população e, consequentemente, alcançar uma frente de pareto extensa e bem distribuída, de acordo com o exposto na secção 2.3..

Bazzoli e Tettamanzi (2004) observaram que, na predição de estruturas de proteínas do modelo HP pode existir diversos indivíduos com a mesma aptidão, mas estruturalmente diferentes. Por conseguinte, se a aptidão fosse utilizada como único critério de comparação, indivíduos que estivessem próximos a ótimos estruturalmente diferentes poderiam ser perdidos (CUSTÓDIO, 2008), podendo causar rápida perda de diversidade na população.

Para prevenir esse efeito, este trabalho utiliza um AG *steady state*, no qual, a cada geração, são gerados m novos indivíduos. Cada novo indivíduo é comparado com os s piores da população para determinar sua semelhança e substituir o mais semelhante encontrado "substituir o pior mais semelhante o novo indivíduo". Defina-se semelhança, neste trabalho, como a proximidade entre as compactações dos aminoácidos hidrofóbicos e hidrofílicos, uma vez que estes valores fornecem uma ideia da topografia da proteína. Dada duas conformações $c1$ e $c2$,o fator de semelhança pode ser calculado através da seguinte expressão:

$$fz = abs(CompH_{c1} - CompH_{c2}) + abs(CompP_{c1} - CompP_{c2}).$$

A figura 3.18 amostra um exemplo deste mecanismo.

```
1. Sejam duas conformações C1 e C2
2. Se o numero de colisões de C1 e igual do que C2
3.   Se o numero de contatos hidrofóbicos da C1 for maior do que a C2
4.                A conformação C1 é a melhor
5.   Senão
6.          Se o numero de contatos hidrofóbicos da C1 for igual do que C2
7.                   Se a compactação hidrofóbica de C1 for maior do que C2
8.                            A conformação C2 é a melhor
9.                   Senão
10.                           Se a compactação hidrofóbica de C1 for igual do que C2
11.                                  Se a compactação hidrofílica de C1 for maior do que C2
12.                                           A conformação C2 é a melhor
13.                                  Senão
14.                                     Se a compactação hidrofílica de C1 for igual do  que C2
15.                                        Pegar aleatoriamente uma conformação ente C1 e C2
16.                                     Senão
17.                                              A conformação C1 é a melhor
18.                                          Fim
19.                                  Fim
20.                           Senão
21.                                  A conformação C1 é a melhor
22.                           Fim
23.                  Fim
24.          Senão
25.                      A conformação C2 é a melhor
26.          Fim
27. Fim
28. Senão
29.       Se número de Colisões da conformação C1 é maior do que C2
30.               Se a diferença do número de colisões de C1 e C2 for menor que um limiar
                  e se a diferencia de suas compactações hidrofóbicas for maior que um limiar
31.                           A conformação C1 é a melhor
32                Senão
33.                           A conformação C2 é a melhor
34                Fim
35.       Senão
36.               Se a diferença do número de colisões de C2 e C1 for menor que um limiar
                  e se a diferencia de suas  compactações hidrofóbicas for maior que um limiar
37.                           A conformação C2 é a melhor
38.               Senão
39.                           A conformação C1 é a melhor
40.               Fim

41.       Fim
42. Fim
```

Figura 3.17: Método do torneio para seleção parental, para duas conformações C1, C2.

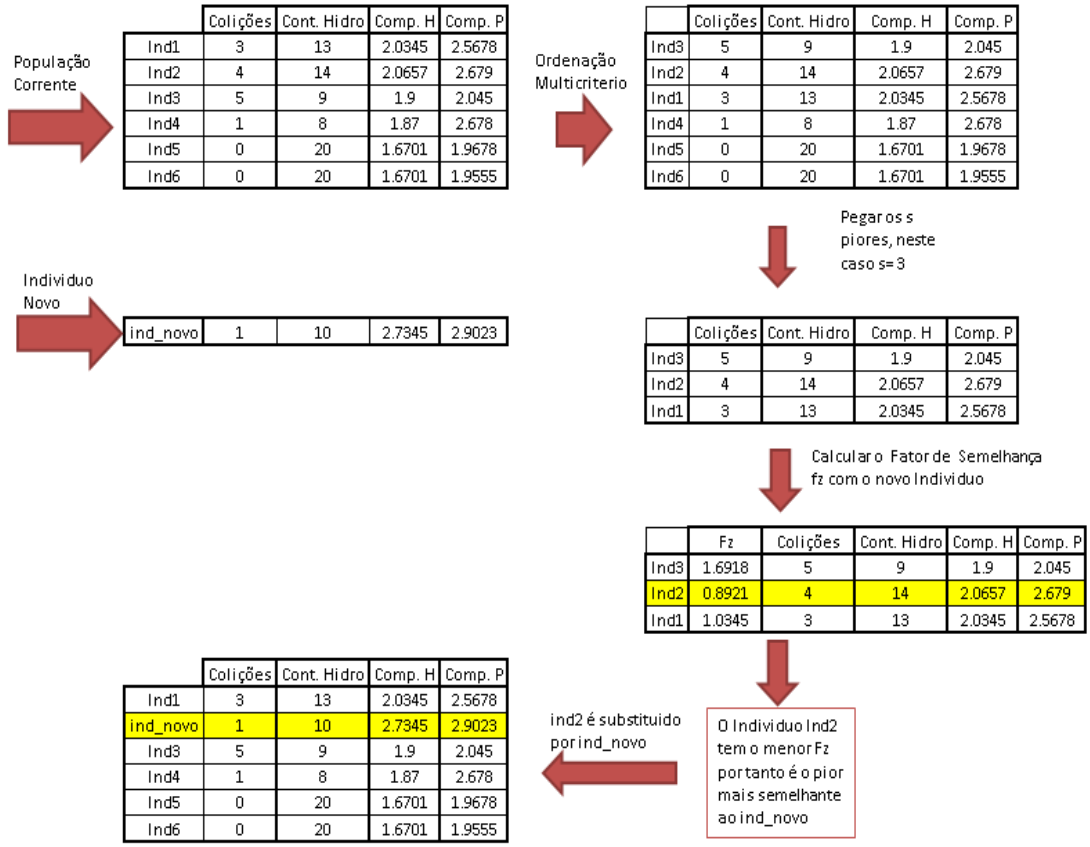

Figura 3.18: Método da substituição parental. Esta figura apresenta um esquema de como um novo indivíduo é inserido na população parental através do critério de semelhança baseado na compactação da proteína.

4
Experimento e Resultados

Este capítulo apresenta a validação do AGMO-HP através de um conjunto de teste composto de sequências extensamente estudadas na literatura. Os resultados serão apresentados em comparação com outras técnicas usadas para a predição de estruturas de proteínas no modelo hidrofóbico polar e assim determinar quão bem sucedida é a metodologia proposta.

4.1.
Conjunto de Teste

O conjunto de teste é composto por três subconjuntos:

1. Conjunto de teste formado por sequências de 27 monômeros: estas sequências foram geradas aleatoriamente por (UNGER e MOULT, 1993) de modo que aproximadamente 40% dos monômeros fossem hidrofóbicos. A tabela 4.1 apresenta este conjunto.
2. Conjunto de teste formado por sequências de 48 monômeros: estas sequências foram propostas por (YUE, FIEBIG, *et al.*, 1994). Uma característica importante desta sequência é que apresentam uma maior razão de contatos por comprimento da cadeia. A tabela 4.2 apresenta este conjunto
3. Conjunto de teste formado por sequências de diversos comprimentos abstraídas de proteínas reais tais como a *cambrina* (comprimento 46), *Bovine Protease Tripsin Inhibitor* (comprimento 56), *citocromo c* (comprimento 103). O fato de avaliar o modelo proposto com este conjunto de teste é interessante pois seria avaliado o modelo com proteínas reais.

Sequencia	Cadeia
27.1	PHPHPHHHPPHPHPPPPPPPPPPHHP
27.2	PHHPPPPPPPPPPHHPPHHPPHPPHPH
27.3	HHHHPPPPPHPPPPPHHHPPPPPPPPH
27.4	HHHPPHHHHPPPHPHPPHHPPHPPPHH
27.5	HHHHPPPPHPHHPPPHHPPPPPPPPPP
27.6	HPPPPPPHPHHHPPHHPPPHPPPPHPH
27.7	HPPHPHHPPPHPPPPPHPHHPHPHPHH
27.8	HPPPPPPPPPPPHPHPPPPPPPPHPHH
27.9	PPPPPPPHHHPPPHPHHPPPHPPHPPP
27.10	PPPPPHHPHPHPHPHPPHHPHHPHPPP

Tabela 4.1: Conjunto de teste com sequências de 27 monômeros

Sequencia	Cadeia
48.1	HPHHPPHHHHPHHHPPHHPPHPHHHPHPHHPPHHPPPHPPPPPPPPHH
48.2	HHHHPHHPHHHHHPPHPPHHPPHPPPPPPHPPHPPPHPPHHPPHHHPH
48.3	PHPHHPHHHHHHPPHPHPPHPHHPHPHPPPHPPHHPPHHPPHPHPPHP
48.4	PHPHHPPHPHHHPPHHPHHPPPHHHHHPPHPHHPHPHPPPPHPPHPHP
48.5	PPHPPPHPHHHHPPHHHHPHHPHHHPPHPHPHPPHPPPPPPHHPHHPH
48.6	HHHPPPHHPHPHHPHHPHHPHPPPPPPPHPHPPHPPPHPPHHHHHHPH
48.7	PHPPPPHPHHHPHPHHHHPHHPHHPPPHPHPPPHHHPPHHPPHHPPPH
48.8	PHHPHHHPHHHHPPHHHPPPPPPHPHHPPHHPHPPPHHPHPHPHHPPP
48.9	PHPHPPPPHPHPHPPHPHHHHHHPPHHHPHPPHPHHPPHPHHHPPPPH
48.10	PHHPPPPPPHHPPPHHHPHPPHPHHPPHPPHPPHHPPHHHHHHPPHH

Tabela 4.2: Conjunto de teste com sequências de 48 monômeros

Sequencia	Cadeia
46	PPHHHPHHHPPPHPHHPHHPPHPHHHHPHPPHHHHHPHPHHPPHHP
58	PHPHHHPHHHPPHHPHPHHPHHHPHPHPHHPPHHHPPHPHPPPPHPPHPPHHPPHPPH
103	PPHHPPPPPHHPPHHPHPPHPPPPPPPHPPPHHPHHPPPPPPHPPHPHPPHPPPPPHHH
	PPPPHHPHHPPPPPHHPPPPHHHHPHPPPPPPPPHHHHHPPHPP
136	HPPPPPHPPPPHPHHPHHPPPPHPHHHPPPPHPHPHHHHPPPPPPPPPPHPPHPPPHP
	HHPPPHHPPHPPHPHPHPPPPPPPPHPPPHHHHHHPPPHHPPHHHPPPHHPHHHHHPP
	PPPPPPPHPPPPHPHPPPP

Tabela 4.3: Conjunto de teste com sequências de proteínas reais.

4.2. Configuração do AGMO-HP

Nesta secção é apresentada a configuração básica do modelo proposto, esta configuração foi determinada a traves de experimentos preliminares, na tabela 4.4 é apresentada a configuração para sequências curtas menores de 48 monômeros, e na tabela 4.5 a configuração para sequências de 48 a mais monômeros:

Parâmetro		valor
Numero de Experimentos		50
Numero de Gerações		500
Tam População Inicial		500
Taxa CrossOver		95%
Taxa de Mutação		0.33%
Bitwise Mutação		0.005%
Pesos Inicias da Roleta	Cross. Um Ponto	90%
	Cross. Dois Pontos	8%
	Cross Multiponto	30%
	Cross Hidrofo. Multiponto	15%
	Cross Hidrofi. Multiponto	10%
	Mutação Simples	10%
	Mutação Busca Exaustiva (EMUT)	40%
	Mutaçõ Troca Segmentos	50%
	Mutção Simples Acumulativa	15%
	Mutação 2-op com Memoria	2%
Tamanho Segmento Mutação s Troca de Segmento (m)		7 genes
Steady State	Número de novos indivuos (m)	(Tam população) x 25%
	Número de piores (s)	(Tam população) x 30%

Tabela 4.4: Configuração para sequências curtas(SC) do AGMO-HP

Parâmetro		valor
Numero de Experimentos		50
Numero de Gerações		500
Tam População Inicial		700
Taxa CrossOver		95%
Taxa de Mutação		0.33%
Bitwise Mutação		0.005%
Pesos Inicias da Roleta	Cross. Um Ponto	90%
	Cross. Dois Pontos	80%
	Cross Multiponto	40%
	Cross Hidrofo. Multiponto	25%
	Cross Hidrofi. Multiponto	20%
	Mutação Simples	10%
	Mutação Busca Exaustiva (EMUT)	40%
	Mutaçõ Troca Segmentos	50%
	Mutção Simples Acumulativa	15%
	Mutação 2-op com Memoria	2%
Tamanho Segmento Mutação s Troca de Segmento (m)		10 genes
Steady State	Número de novos indivuos (m)	(Tam população) x 25%
	Número de piores (s)	(Tam população) x 30%

Tabela 4.5: Configuração para sequências longas(SL) do AGMO-HP

Para avaliar a capacidade do modelo de formar estruturas de proteínas mais compactas e globulares foi usada uma modificação do modelo AGMO-HP, sem os dois últimos objetivos que estão relacionados com esta característica, permitindo assim ter uma ideia do impacto na forma da proteína quando é levada em conta sua compactação. O modelo modificado foi chamado de Algoritmo genético multiobjetivo Simplificado no modelo Hidrofóbico Polar AGMOS-HP.

4.3. Resultados para sequências de 27 monômeros

Para estas sequências foi usado o modelo AGMO-HP com a configuração de parâmetros para sequências curtas (SC). Os resultados obtidos por este modelo foram comparados com o modelo desenvolvido por (KHIMASIA e COVENEY, 1997) baseado num algoritmo genético simples e os trabalhos de (PATTON, PUNCH III e GOODMAN, 1995) cujos resultados foram melhores que os de (UNGER e MOULT, 1993). Foi comparado também com o modelo de (KATIKIREDDY e JOHNSON, 2006) que usa um mecanismo de *backtracking*

(para correção de colisões) o qual apresentou resultados melhores do que os apresentados por (PATTON, PUNCH III e GOODMAN, 1995), finalmente o AGMO-HP foi comparado com o modelo proposto por (CUSTÓDIO, 2008) o qual usa mecanismos *crowding* baseados no fenótipo, para a conservação da diversidade e também utiliza um mecanismo adaptativo para o ajuste das probabilidades dos operadores genéticos, este modelo apresenta melhores resultados que os modelos de (PATTON, PUNCH III e GOODMAN, 1995) e (KATIKIREDDY e JOHNSON, 2006).

Segundo a tabela 4.6 o modelo AGMO-HP, para as sequencias 27.1, 27.2, 27.5, 27.7, 27.10 apresentou melhores resultados que o algoritmo genético simples desenvolvido por (KHIMASIA e COVENEY, 1997). No entanto encontrou o mesmo número de contatos hidrofóbicos que os outros algoritmos para todas as sequencias de 27 monômeros. No caso da sequência 27.6 o modelo proposto obteve um número de contatos hidrofóbicos superiores aos publicados por (PATTON, PUNCH III e GOODMAN, 1995) e (KATIKIREDDY e JOHNSON, 2006) igualando os resultados obtidos por (CUSTÓDIO, 2008).

Sequência	Custodio et al. (2008)		(AGS)Khimasia e al. (1997)		Katikireddy (2006)		Patton et al. (1995)		AGMO-HP (SC)			
									Contatos Hidrofobicos		Comp.NucleoH	Comp.NucleoP
27.1	9		8		9		9		9		1.00620	1.5090
	8.6	(0.5)	-	-	-	-	-	-	8.71	(0.08)		
27.2	10		8		10		10		10		0.96860	1.4870
	9.96	-0.2	-	-	-	-	-	-	9.43	(0.12)		
27.3	8		8		8		8		8		0.96864	1.48150
	7.78	(0.4)	-	-	-	-	-	-	7.97	(0.02)		
27.4	15		15		15		15		15		1.14730	1.47680
	14.70	(0.47)	-	-	-	-	-	-	13.96	(0.05)		
27.5	8		7		8		8		8		0.96864	1.50170
	8	(0)	-	-	-	-	-	-	7.7	0		
27.6	12		11		11		11		12		1.02470	1.48980
	10.76	0.80	-	-	-	-	-	-	11.18	(0.20)		
27.7	13		11		13		13		13		1.11180	1.46750
	12.48	(0.56)	-	-	-	-	-	-	12.94	(0.26)		
27.8	4		4		4		4		4		0.94280	1.55290
	4	(0)	-	-	-	-	-	-	4	(0)		
27.9	7		7		7		7		7		0.96010	1.50720
	6.94	(0.26)	-	-	-	-	-	-	6.84	(0)		
27.10	11		10		11		11		11		1.0985	1.5687
	10.62	(0.5)	-	-	-	-	-	-	10.78	(0.23)		

Tabela 4.6: Resultados obtidos com sequências de 27 monômeros em comparação com outros métodos. São amostrados o número de contatos hidrofóbicos media e desvio padrão. Também são apresentados os resultados obtidos da compactação dos aminoácidos hidrofóbicos (Comp. NúcleoH) e hidrofílicos (Comp. NúcleoP).

É importante destacar que o modelo proposto utiliza populações de 500 indivíduos como foi usado por (CUSTÓDIO, 2008) e (PATTON, PUNCH III e GOODMAN, 1995) enquanto (KATIKIREDDY e JOHNSON, 2006) variam o tamanho da população entre 1000 e 1600 indivíduos.

Nas figuras 4.1 a 4.5 são apresentadas conformações de proteínas obtidas com o modelo AGMO-HP e o modelo AGMOS-HP, os resultados evidenciam a importância da compactação de uma proteína na obtenção de estruturas mais naturais, isto pode ser observado com maior detalhe nas sequências 27.1, 27.5, 27.8 e 27.9 onde o AGMS-HP tende a formar fitas longas enquanto o AGMO-HP forma estruturas muito mais compactas e globulares. É importante destacar que todas as sequências mostradas nestas figuras correspondem aos valores máximos de energia obtidos pelo AGMO-HP amostrados na tabela 4.6.

Outro fato importante é a capacidade do modelo AGMO-HP de poder identificar proteínas com distintos graus de compactação nos seus aminoácidos hidrofóbicos mesmo que eles tenham o mesmo número de contatos hidrofóbicos, isto pode ser visto nas sequências 27.4, 27.7, 27.8 e 27.9.

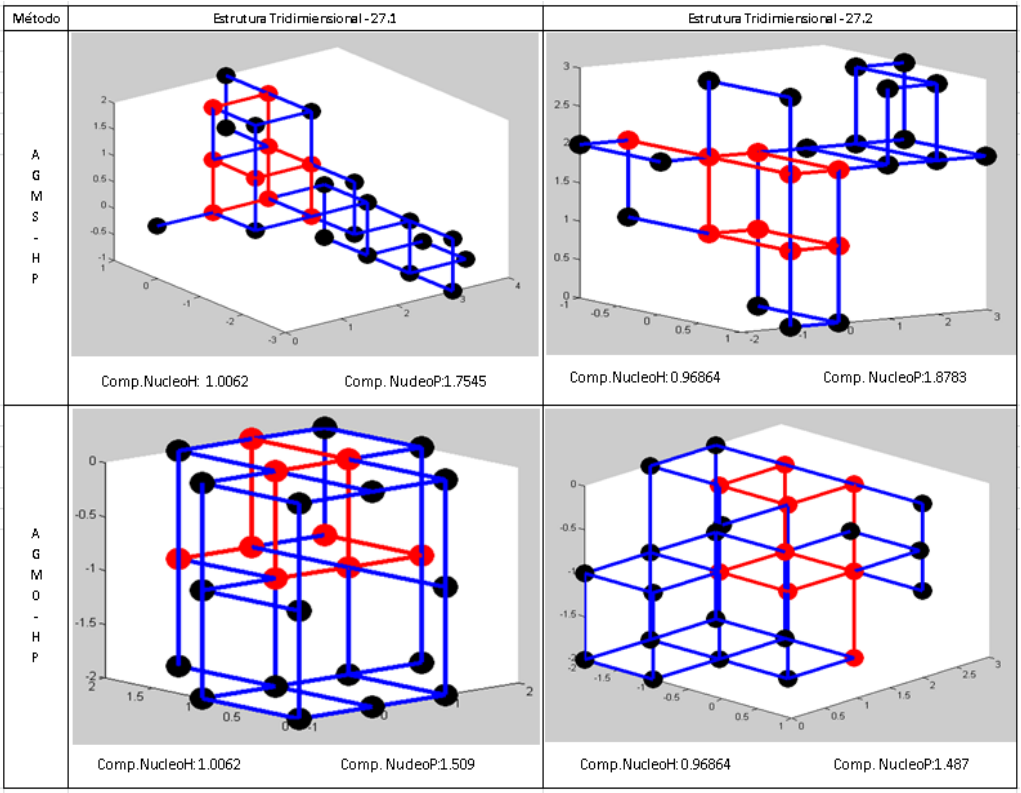

Figura 4.1: Estruturas tridimensionais preditas para as sequências 27.1 e 27.2, para os modelos AGMOS-HP e AGMO-HP.

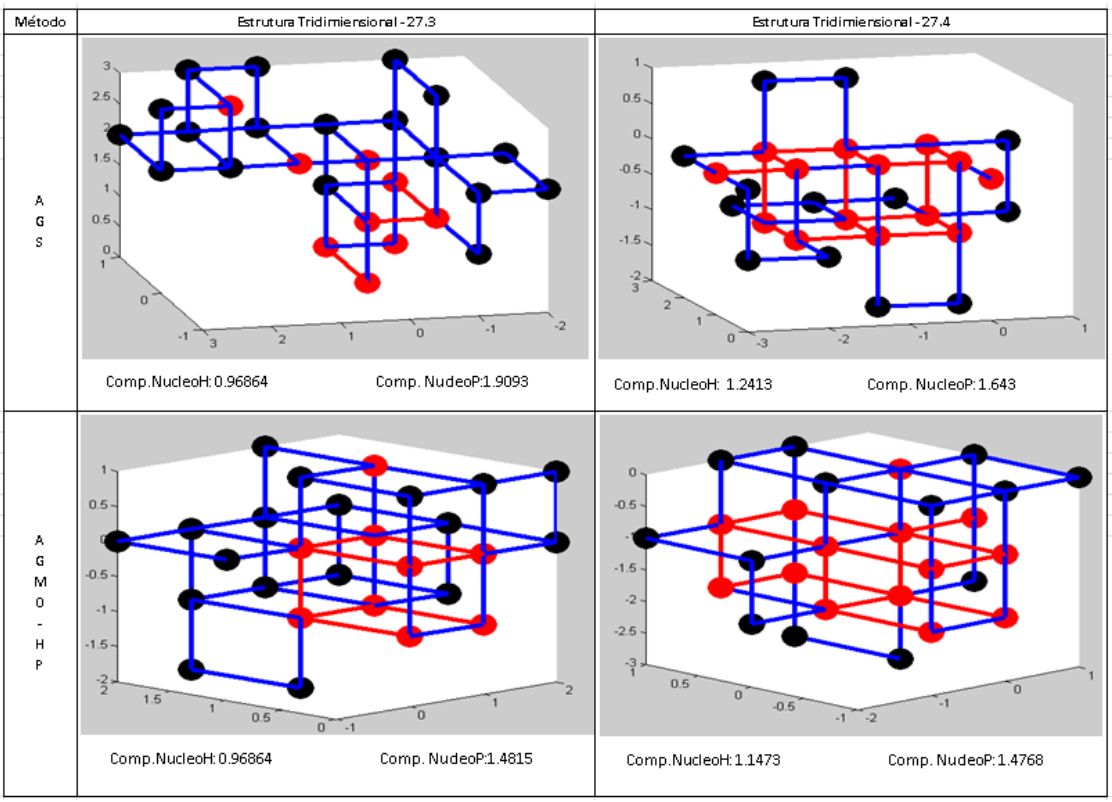

Figura 4.2: Estruturas tridimensionais preditas para as sequências 27.3 e 27.4, para os modelos AGMOS-HP e AGMO-HP.

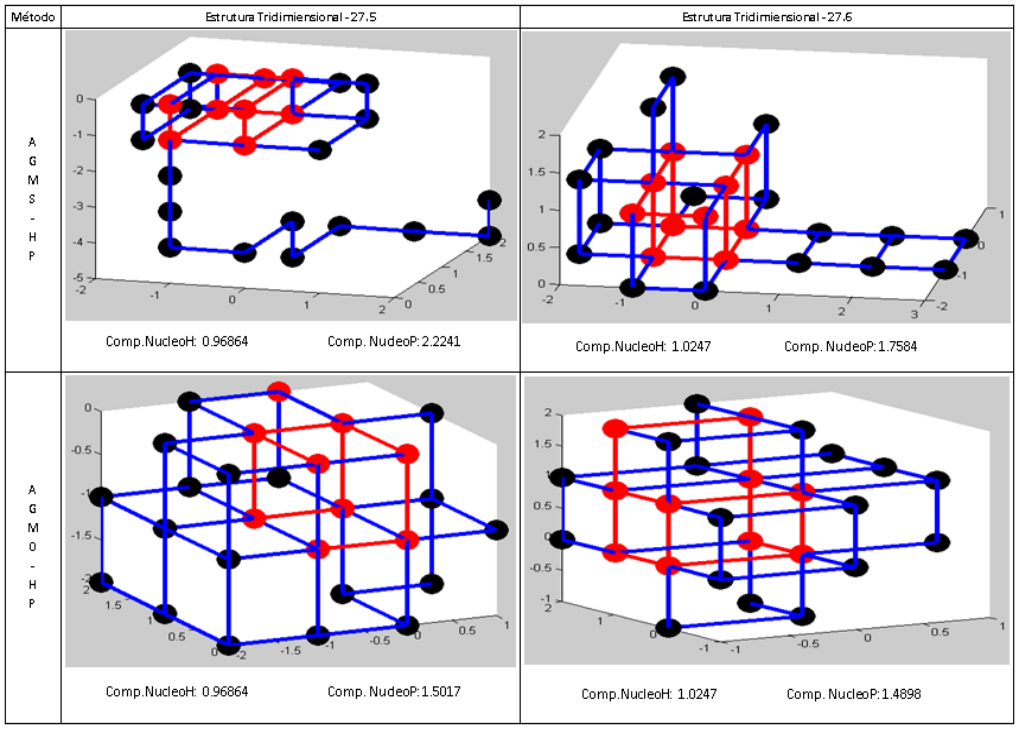

Figura 4.3: Estruturas tridimensionais preditas para as sequências 27.5 e 27.6, para os modelos AGMOS -HP e AGMO-HP.

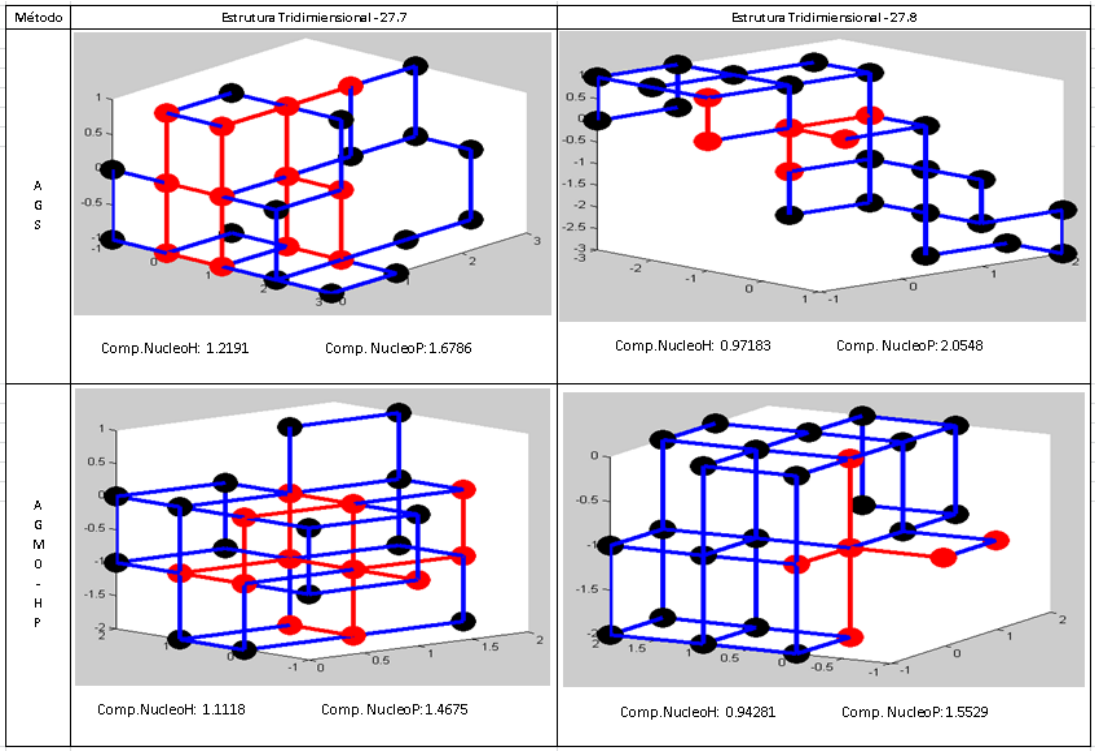

Figura 4.4: Estruturas tridimensionais preditas para as sequências 27.7 e 27.8, para os modelos AGMOS -HP e AGMO-HP.

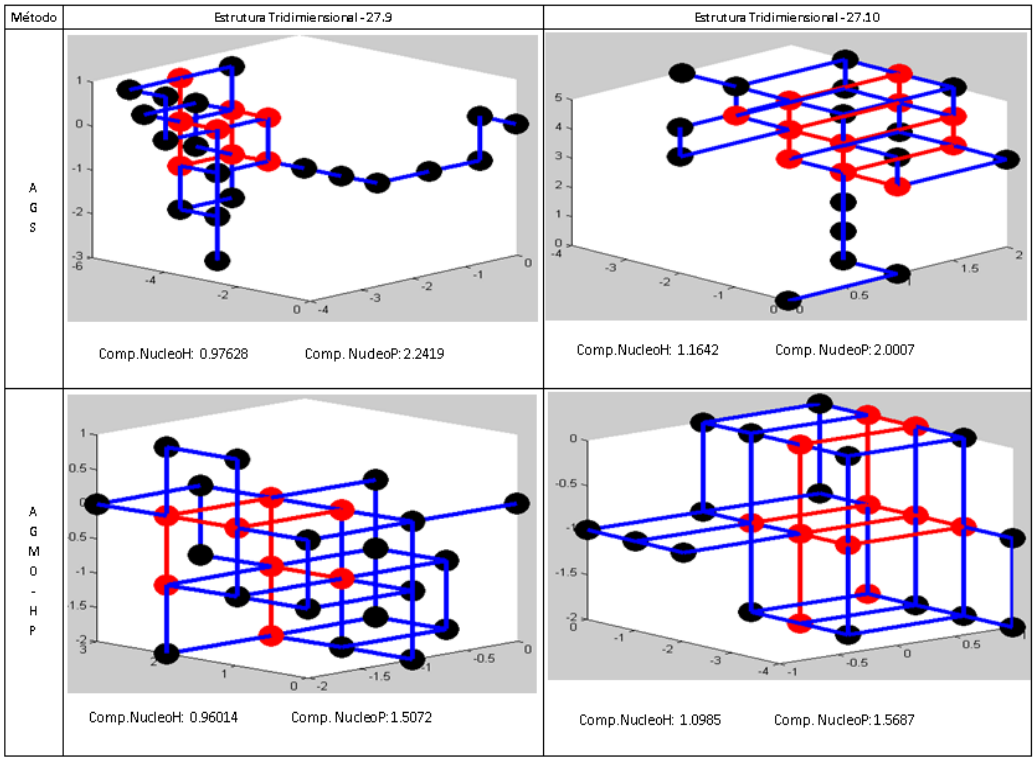

Figura 4.5: Estruturas tridimensionais preditas para as sequências 27.9 e 27.10, para os modelos AGMOS -HP e AGMO-HP.

4.4.
Resultados para sequências de 48 monômeros

Para estas sequências foi usado o modelo AGMO-HP com a configuração de parâmetros para sequências longas (SL). Os resultados obtidos por este modelo foram comparados com os modelos desenvolvidos por (KHIMASIA e COVENEY, 1997) algoritmo genético simples, (SHMYGELSKA e HOOS, 2005) baseado em otimização por colônia de formigas, (BAZZOLI e TETTAMANZI, 2004) baseado num algoritmo mimético, (DILL, FIEBIG e CHAN, 1993) desenvolveram um modelo chamado de *Hidrophobic Zipper*, finalmente o AGMO-HP foi comparado com o modelo proposto por (CUSTÓDIO, 2008).

Os resultados da tabela 4.7 mostram que o modelo AGMO-HP, em termos de contatos hidrofóbicos, somente consegui superar o algoritmo genético simples desenvolvido por (KHIMASIA e COVENEY, 1997). No entanto os seus resultados estão muito pertos dos achados por (DILL, FIEBIG e CHAN, 1993).

Sequência	Custodio et al. (2008)		(ACO) Shmygelska et al. (2005)		(HZ) Dill et al. (1993)		(AGS)Khimasia et al. (1997)		AGMO-HP (SL)			
									Contatos Hidrofobicos		Comp.NucleoH	Comp.NucleoP
48.1	32		32		31		24		30		1.4655	2.0981
	30.72	(0.67)	-	-	-	-	-	-	27.14	(0.26)		
48.2	34		34		32		24		31		1.4386	1.8543
	31.26	(0.59)	-	-	-	-	-	-	27.97	(0.50)		
48.3	34		34		31		23		31		1.4367	1.8449
	32.08	(0.80)	-	-	-	-	-	-	27.98	(0.39)		
48.4	33		33		30		24		30		1.4518	1.8061
	31.16	(0.81)	-	-	-	-	-	-	29.03	(0.36)		
48.5	32		32		30		28		29		1.4313	1.8817
	30.52	(0.73)	-	-	-	-	-	-	26.38	(0.14)		
48.6	32		32		29		25		29		1.4684	2.0798
	29.86	(0.78)		-		-	-	-	26.74	(0.27)		
48.7	32		32		29		27		29		1.4548	1.8741
	29.82	(0.56)	-	-	-	-	-	-	26.21	(0.21)		
48.8	31		31		29		26		29		1.4634	1.9806
	29.32	(0.58)	-	-	-	-	-	-	27.67	(0.22)		
48.9	34		34		31		27		31		1.4494	1.9410
	31.92	(0.66)	-	-	-	-	-	-	27.45	(0.34)		
48.10	33		33		33		26		32		1.4467	1.8119
	31.08	(0.56)	-	-	-	-	-	-	29.08	(0.11)		

Tabela 4.7: Resultados obtidos com sequências de 48 monômeros em comparação com outros métodos, são amostrados o número de contatos hidrofóbicos media e desvio padrão. Também são apresentados os resultados obtidos da compactação dos seus aminoácidos hidrofóbicos (Comp. Núcleo) e hidrofílicos (Comp. Núcleo).

Do mesmo modo que na seção anterior nas figuras 4.6 a 4.10 são apresentadas as conformações de proteínas obtidas com o modelo AGMO-HP e o modelo AGMS-HP. Os resultados obtidos mostram que o modelo AGMO-HP conseguiu estruturas proteicas mais compactas e globulares. Isto pode ser observado nas sequências 48.1 e 48.2 na Figura 4.6. Apesar dos dois algoritmos terem conseguido obter o mesmo número de contatos hidrofóbicos, somente o AGMO-HP conseguiu estruturas globulares. Podemos observar que a conformação obtida para a sequência de 48.1 pelo modelo AGMS-HP conseguiu proteínas com grau de compactação nos seus aminoácidos hidrofóbicos de 1.5383 e o modelo AGMO-HP conseguiu um grau de compactação de 1.4655.

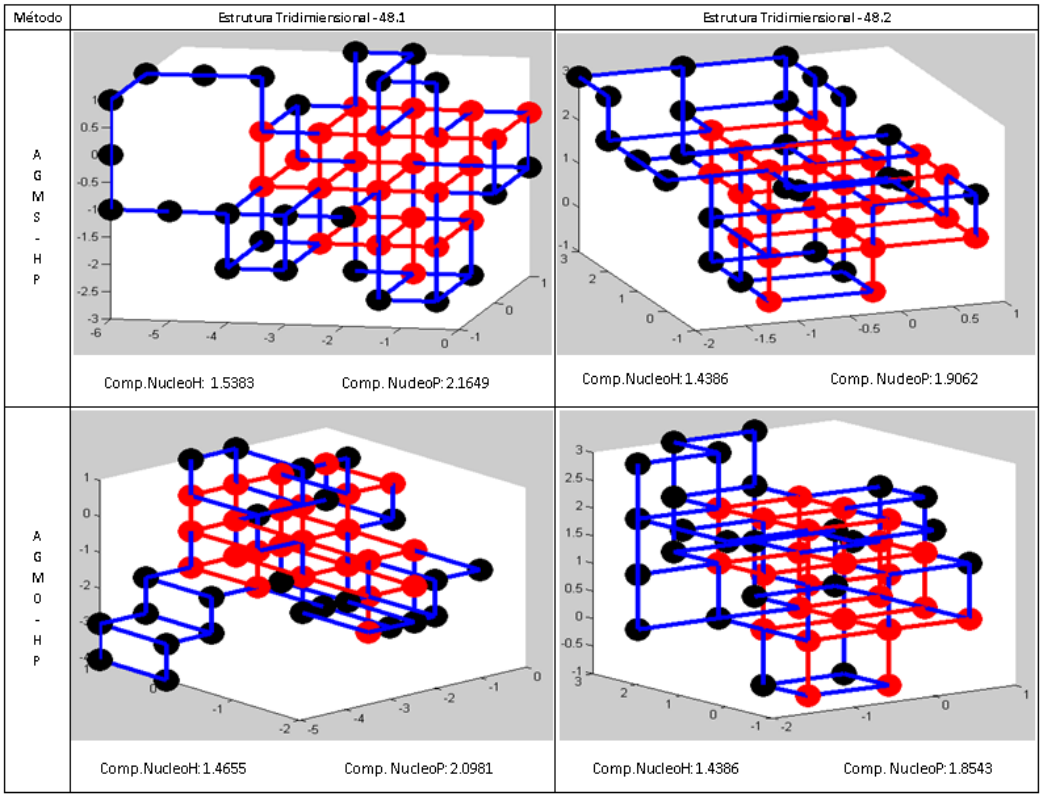

Figura 4.6: Estruturas tridimensionais preditas para as sequências 48.1 e 48.2, para os modelos AGMOS-HP e AGMO-HP.

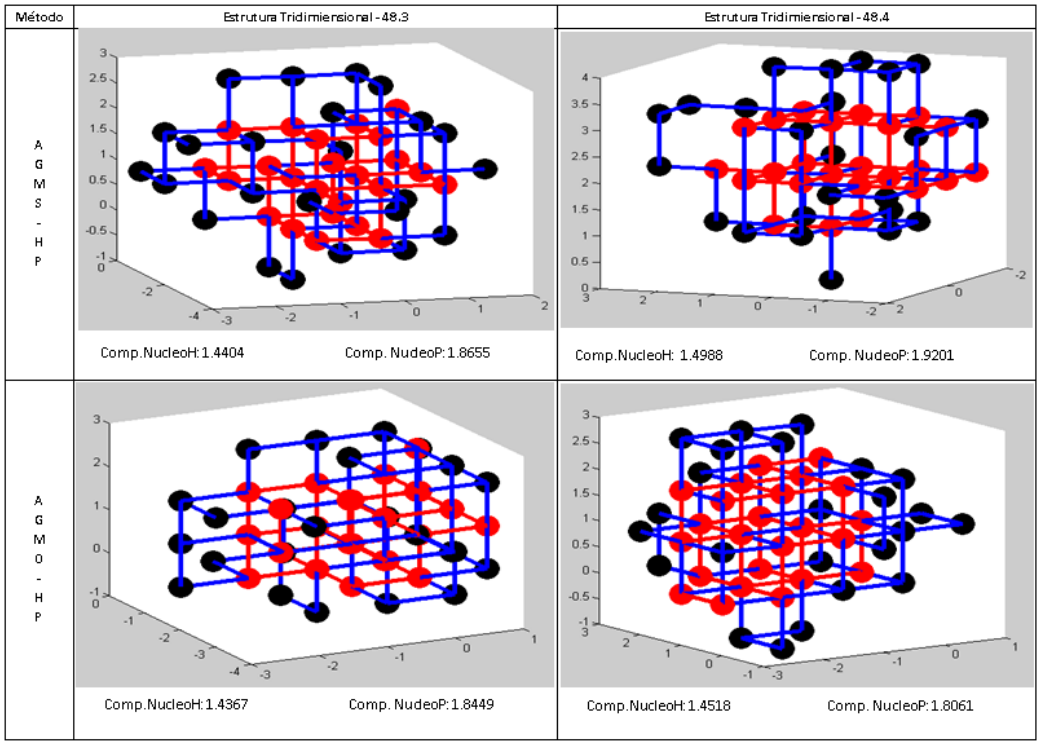

Figura 4.7: Estruturas tridimensionais preditas para as sequências 48.3 e 48.4, para os modelos AGMOS -HP e AGMO-HP.

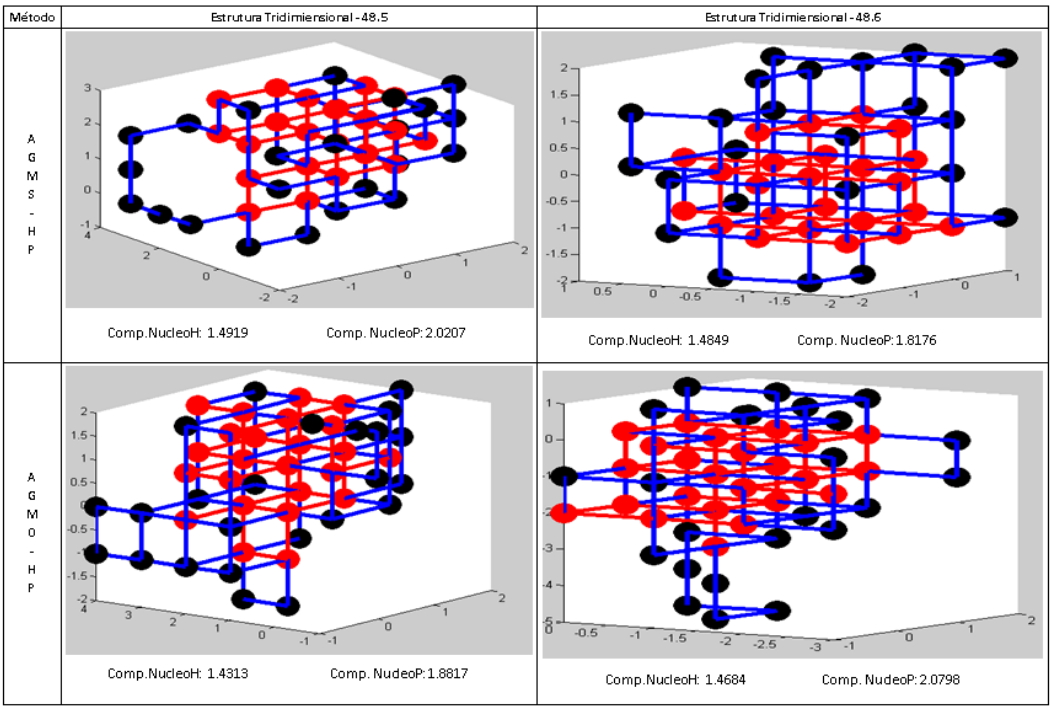

Figura 4.8: Estruturas tridimensionais preditas para as sequências 48.5 e 48.6, para os modelos AGMOS -HP e AGMO-HP.

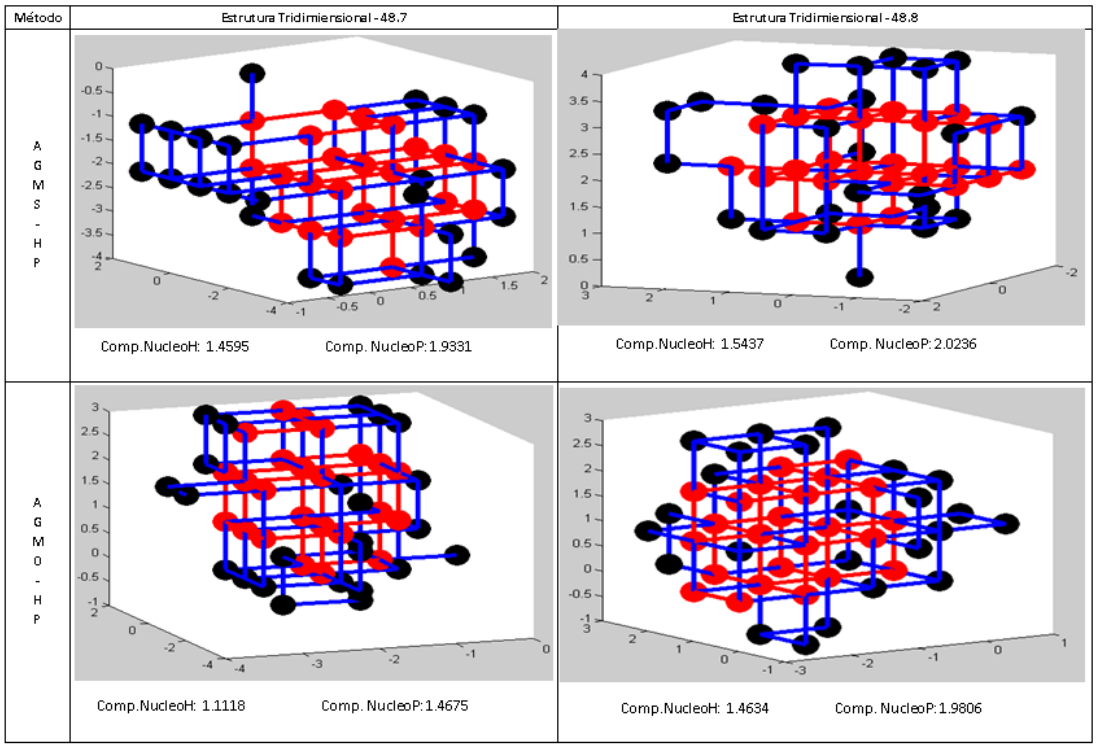

Figura 4.9: Estruturas tridimensionais preditas para as sequências 48.7 e 48.8, para os modelos AGMOS -HP e AGMO-HP.

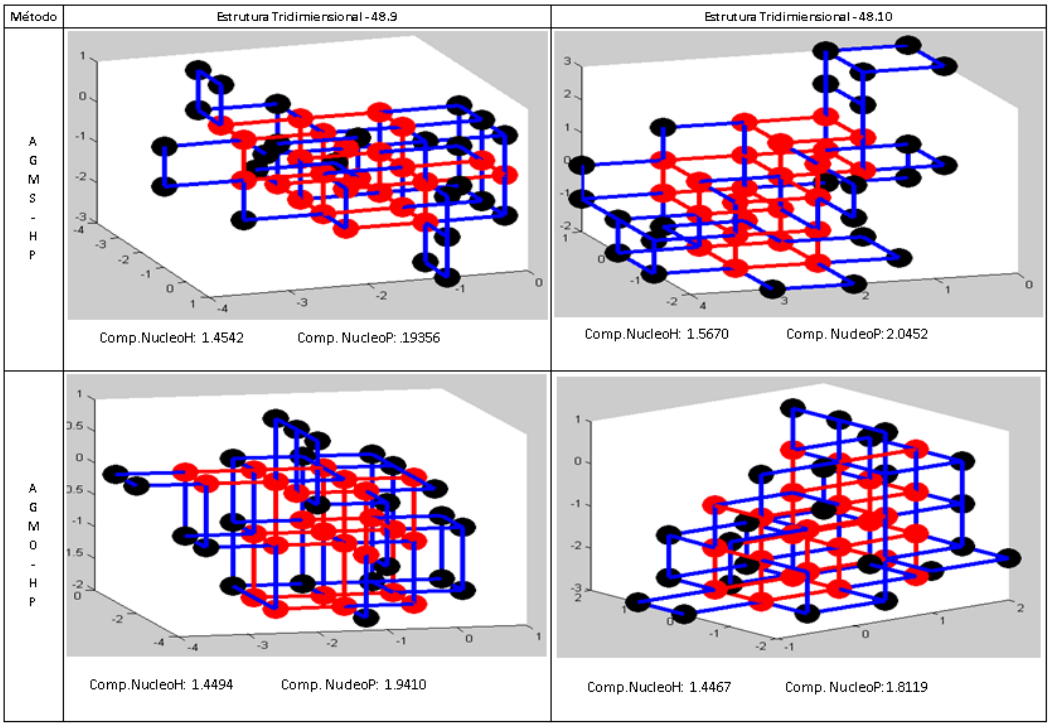

Figura 4.10: Estruturas tridimensionais preditas para as sequências 48.9 e 48.10, para os modelos AGMOS -HP e AGMO-HP.

4.5.
Resultados para sequências de proteínas reais.

Na tabela 4.8 são amostrados os resultados obtidos do modelo AGMO-HP como estas sequências, estes resultados foram comparados com os modelos de (CUSTÓDIO, 2008) e o modelo chamado de C*ontact Interactions* CI (TOMA e TOMA, 1996). Desta vez o modelo AGMO-HP logrou resultados comparáveis aos obtidos do CI. Nas sequências de 46 e 58 monômeros, na figura 4.11 observam-se aminoácidos hidrofóbicos com diferentes graus de compactação e o mesmo número de contatos hidrofóbicos este fato também foi identificado na secção anterior. Nas figuras 4.12, 4.13 destaca-se que para as sequências de 103 e 136 monômeros o AGMS-HP teve tendência a formar dois conjuntos de aminoácidos hidrofóbicos, no entanto o modelo AGMS-HP mostrou uma grande capacidade de compactação para este tipo de sequências formando um único conjunto de aminoácidos hidrofóbicos, isto pode ser evidenciado na forma da proteína e no valor das compactações de seus aminoácidos hidrofóbicos.

Sequência	Custodio et al. (2008)		Toma e Toma (1996)		AGMO-HP (SL)			
					Contatos Hidrofobicos		Comp.NucleoH	Comp.NucleoP
46	35		34		32		1.5673	1.7888
	33.04	(0.68)	-	-	29.06	(0.34)		
58	42		42		39		1.6687	2.0081
	40.04	(0.80)	-	-	35.10	(0.43)		
103	50		49		43		1.87747	2.570
	32.08	(1.51)	-	-	33.14	(0.24)		
136	70		65		61		2.1797	3.3433
	62.22	(1.94)	-	-	59.67	(0.15)		

Tabela 4.8: Resultados obtidos com sequências reais de 46, 58, 103 e 136 monómeros em comparação com outros métodos, são amostrados o número de contatos hidrofóbicos media e desvio padrão. Também são apresentados os resultados obtidos da compactação dos seus aminoácidos hidrofóbicos (Comp. NucleoH) e hidrofílicos (Comp. NucleoP).

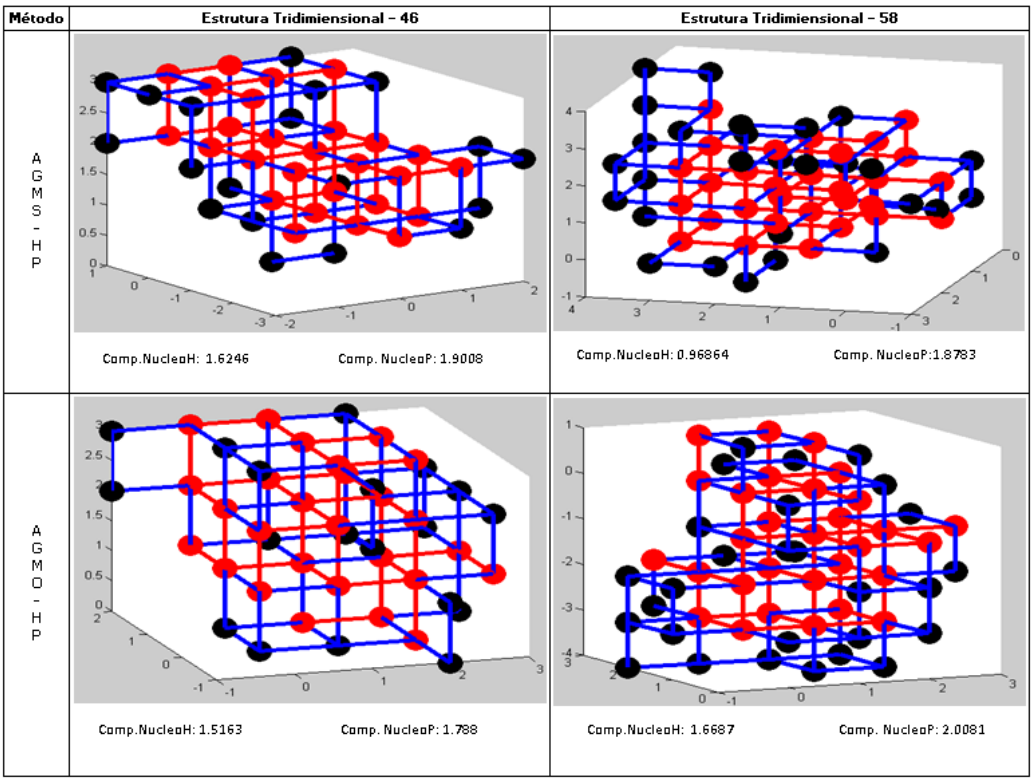

Figura 4.11: Estruturas tridimensionais preditas para as sequências 46 e 58, para os modelos AGMOS-HP e AGMO-HP.

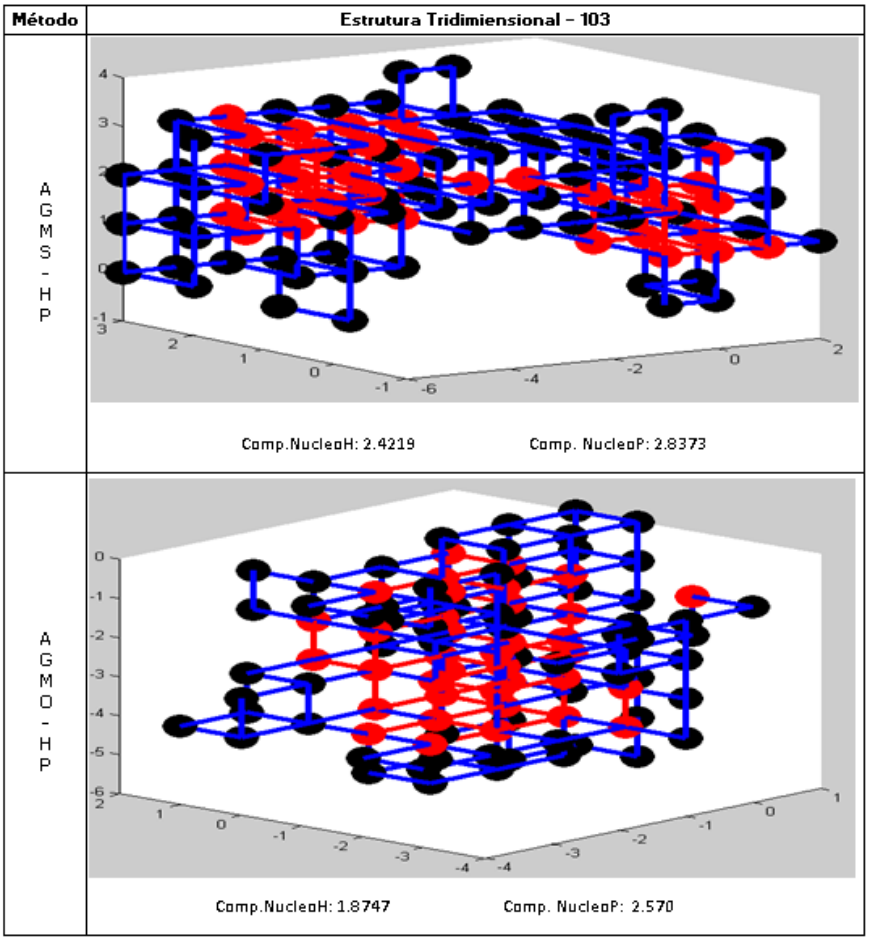

Figura 4.12: Estruturas tridimensional predita para a sequência de 103 monómeros, para os modelos AGMOS -HP e AGMO-HP.

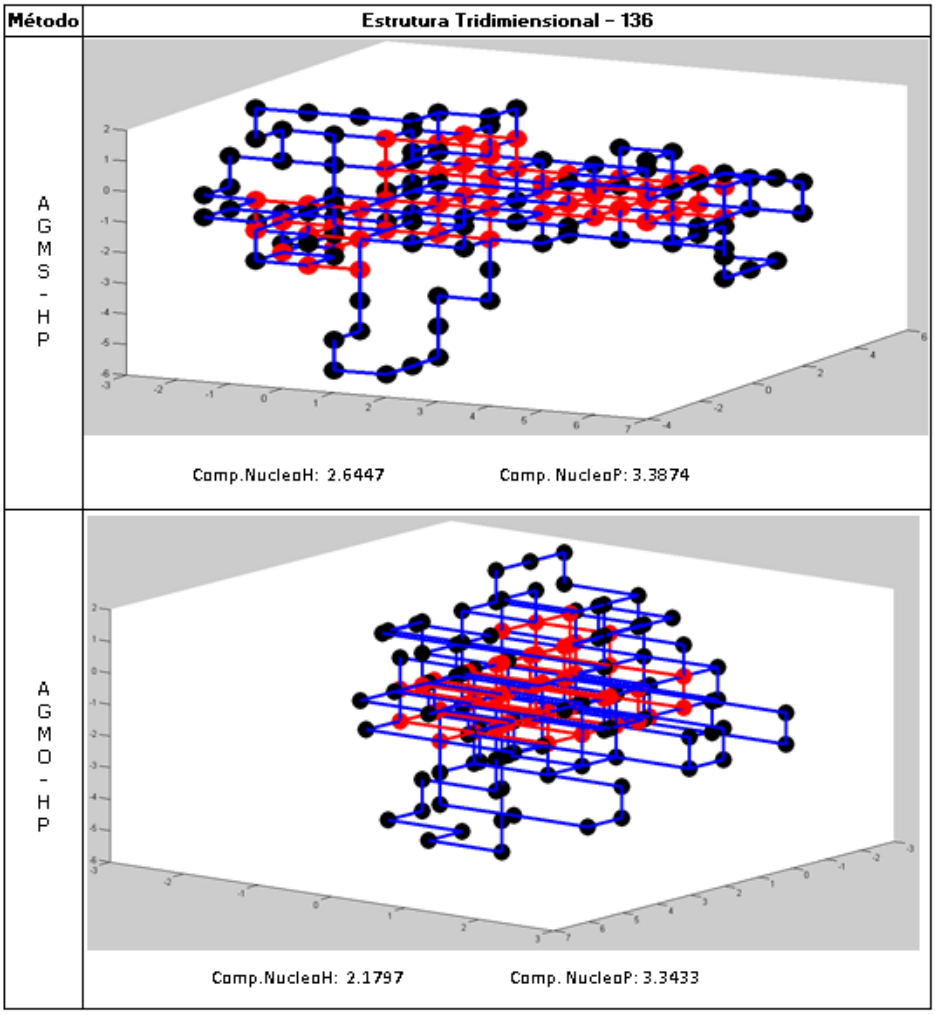

Figura 4.13: Estruturas tridimensional predita para a sequência de 136 monómeros, para os modelos AGMOS -HP e AGMO-HP.

5
Conclusões e trabalhos futuros

De acordo com a revisão da literatura apresentada neste trabalho, os métodos tradicionais não realizam tratamento adequado na compactação das proteínas. Em consequência disto, frequentemente são observadas estruturas proteicas com diferentes conformações espaciais, mas com a mesma quantidade de contatos hidrofóbicos (energeticamente iguais).

Baseado em um algoritmo genético multiobjetivo (AGMO-HP), esta dissertação propôs um modelo para predição de estruturas proteicas compactas e de mínima energia.

No referido modelo, a compactação considerou dois níveis de prioridade, a compactação dos aminoácidos hidrofóbicos e dos hidrofílicos. Isto permitiu obter estruturas proteicas mais compactas, verificadas através dos índices de compactação fornecidos por este modelo, os quais permitiram comparar também a qualidade das conformações num nível mais fino em relação aos modelos da literatura.

O AGMO-HP foi avaliado com 10 sequências de 27 e 48 monômeros, e em proteínas reais. Os resultados deste foram comparados com os de uma variante do modelo proposto, o AGMOS-HP, que não realiza a compactação dos aminoácidos hidrofóbico e polar. Verificou-se que as conformações preditas sem os critérios de compactação (AGMOS-HP) tendem a formar fitas de aminoácidos hidrofílicos, apresentando também níveis de energia similares aos de outros modelos. Por outro lado, ao fazer uso dos critérios de compactação, o modelo AGMO-HP foi capaz de gerar estruturas mais compactas dos aminoácidos hidrofóbicos e hidrofílicos. No caso de proteínas reais, o modelo AGMO-HP foi capaz de gerar conformações com uma única concentração de aminoácidos hidrofóbicos, o qual não foi observado no modelo AGMOS-HP, resultado mais consistente com proteínas achadas na natureza.

Outra característica relevante do modelo proposto foi a capacidade de obter estruturas proteicas de mínima energia com número muito menor de avaliações da função objetivo que o requerido pelos modelos tradicionais. Enquanto estes, em média, necessitaram de um número da ordem de 1 milhão, o AGMO-HP precisou de apenas 250.000 avaliações da função objetivo, para estruturas pequenas; e de 350.000, para estruturas maiores. A rápida evolução obtida com o modelo AGMO-HP justifica-se pelo emprego de um método de seleção desenvolvido para este trabalho, o torneio modificado, que aumenta as chances de seleção das conformações que sofrem colisões e com número elevado de contatos hidrofóbicos.

No que diz respeito à convergência, o modelo proposto foi bem sucedido devido à utilização da roleta pseudo-adaptativa, baseada nos créditos obtidos pelos operadores genéticos, evitando, assim, a estagnação prematura da evolução.

A continuidade deste trabalho deve observar as seguintes considerações:

1. Usar uma configuração de algoritmos genéticos master-slave para aperfeiçoar os parâmetros do modelo proposto; esta configuração pode fazer um melhor ajuste dos parâmetros de configuração.
2. O presente modelo pode ser usado como base para desenvolver um modelo baseado em átomos explícitos.
3. O modelo pode ser hibridizado com uma técnica de busca local mais sofisticada como, por exemplo, a busca tabu, o que permitiria explorar a vizinhança em busca de conformações de melhor qualidade.
4. O modelo pode ser testado com um número maior de uso da função objetivo na procura de melhores soluções.
5. A fim de melhorar o desempenho computacional, este modelo pode ser testado em ambientes de programação paralela baseados em GPU.

Referências Bibliográficas

ALTSCHUL, S. F. et al. Basic Local Alignment Search Tool. **Journal of Molecualar Biology**, v. 215, p. 403-410, 1990.

ANFINSEN, C. B. et al. **The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain**. Proceedings of the National Academy of Sciences of the United States of America. [S.l.]: [s.n.]. 1961. p. 1309-1314.

ARNOLD L. PATTON, W. F. P. I. A. E. D. G. A standard GA approach to native protein conformation prediction. **Morgan Kaufmann Publishers Inc**, San Francisco CA., 1995. ISSN ISBN 1-55860-370-0.

BÄCK, T.; FOGEL, D. B.; MICHALEWICZ, Z. **Evolutionary computation 2 Advanced algorithms and operators**. Institute of Physics Publishing. [S.l.]. 2000.

BACKER, J. E. **Adaptative selection methods for genetic algorithms**. Proceedings of the First International Conference on Genetic Algorithms and their Aplications. [S.l.]: J.J Grefenstette. 1985. p. 101-111.

BACKER, J. E. **Reducing bias and inefficiency in the selection algorithm.** Proceedings of the First International Conference on Genetic Algorithms and their Aplications. [S.l.]: Grefenstette. 1987. p. 14-21.

BACKOFEN, R. **Optimization techniques for the protein structure prediction problem**. Ludwig-Maximilians-Universität,Müchen, Institut für Informatik. München, Alemanha, p. 143. 1999.

BARBOSA, H. J. C. **Algoritmos genéticos para otimização em engenharia: uma introdução In: IV SEMINÁRIOS SOBRE ELEMENTOS FINITOS E MÉTODOS NUMÉRICOS EM ENGENHARIA,1996.** Juiz de Fora,MG. 1996.

BAXEVANIS; , A. D.; OUELLETTE, B. F. F. **Bioinformatics:** a practical guide to the analysis of genes and proteins. 3. ed. New Jersey, EUA: John Wiley and Sons, Inc, 2005.

BAZZOLI, A.; TETTAMANZI, A. G. B. A memetic algorithm for protein structure prediction in a 3d-lattice hp model. **Applications of evolutionary computing**, Milano, v. 3005, p. 1-10, 2004.

BERGER, A.; LEIGHTON, T. Protein Folding in the hydrophobic-hydrophylic (hp) model is np-complete. **J. Comp. Bio**, p. 27-40, 1998.

BERGER, B.; LEIGHTON, T. Protein Folding in the hydrophobic-hidrophilic (HP) model is NP-complete. **Journal of computational Biology**, v. 5, p. 27-40, 1998.

BERMAN, H. M. et al. The protein data bank. **Nucliec Acids Research**, 2000. 235-242.

BETTELHEIM, F. A.; BROWN, W. H. **Introdução à Bioquimica**. 9. ed. São Paulo: Cengage Learning, v. 1, 2012.

BRANDEN, C.; TOOZ, J. **Introduction to protein structure**. 2. ed. New York, EUA: Garlang Publishing Inc., 1998.

BRANDEN, C.; TOOZE, J. Introduction to Protein Structure. 2. ed. New York: Garland Publishing, 1999. p. 410.

BUI, T. N.; SUNDARRAJ,. An Efficient Genetic Algorithm for Predicting Protein Tertiary Structures in the 2D HP Model. **Proceedings of the second annual for finding low energy structures of model proteins**, New York, p. 30-39, 2005.

BUJNICKI, J. M. Protein structure prediction by recombination of fragments. **Chembio-chem: a European Journal of Chemical Biology**, 2006. 19-27.

CHORRO S. B, C. M. **Optimização Evolucionária Multi-Objectivo em Ambientes Incertos**. Universiade de Coimbra. Coimbra, p. 30-35. 2007.

CLÍMACO, J. N.; ANTUNES, C. H.; ALVES, M. J. Programação linear multiobjectivo. Do modelo de programação linear clássico à consideração explícita de várias funções objectivo., 2003.

CLOTE, P.; BACKOFEN, R. **Computational molecular biology:** An introduction. San Francisco. EUA: John Wiley & Sons, Inc. 2000.

COHEN, F.; KELLY, J. Therapeutic approches to protein-misfolding diseases. **Nature**, p. 426:905-909, 2003.

CRESCENZI, P. et al. On the complexity of the protein folding, v. 5, p. 423-446, 1998.

CUSTÓDIO, F. L. **Ab Initio Protein Structure Prediction with Genetic Algorithms**. Rio de Janeiro. 2008.

DAVIS, L. **Handbook of genetic algorithms**. [S.l.]: London International Thomson Computer Press, Boston, 1996.

DE JONG, K. A. **Analysis of the Behavior of a Class of Genetic Adaptative**. University of Michigan. Viginia. 1975.

DEB, K. Evolutionary algorithms for multi-criterion optimization in engineering design. **Evolutionary Algorithms in Engineering and Computer Science: Recent Advances in Genetic Algorithms, Evolution Strategies, Evolutionary Programming, Genetic Programming, and Industrial Applications**, 1999.

DEB, K. **Multi-objective optimization using evolutionary algorithms**. [S.l.]: John Willey and Songs, 2001. ISBN 978-0-471-87339-6.

DILL, K. A.; FIEBIG, K. M.; CHAN, H. S. Cooperativity in protein-folding kinetics. **Proc Natal Acad Sci USA**, p. 1942-1946, Mar 1993.

DILL., K. A. Theory for the folding and stability of globular proteins. **Biochemestry**, n. 24, p. 1501-1509, 1985.

FIDANOVA, S.; LIRKOV, I. Ant Colony System Approach for Protein Folding. **Proceedings of the International Multiconference on Computer Science and Information Technology**, Sofia, Bugaria, p. 887-891. ISSN 978-83-60810-14-9.

FOGEL, G.; CORNE, D. Evolutionary Computation in Bioinformatics. **Morgan Kaufmann Publishers**, Amsterdam, p. 393, 2003.

FRAENKEL, A. S. Complexity of protein folding. **Bulletin of mathematical Biology**, v. 55, p. 1199-1210, 1993.

GIBAS, G.; JAMBECK, P. **Desenvolvendo bioinformática**. 1. ed. Rio de Janeiro, BR: Editora Campus- O'Reilly, 2001.

GOLDBERG, D. E. **Genetic algorithms in search, optimization & machine learning.** Addison-Wesley: [s.n.], 1989.

HART, W. E.; ISTRAIL, S. Robust proof of NP-hardness for protein foldin: General lattices and energy potencials. **Journal of computational Biology**, v. 4, p. 30, 2006.

HART, W. E.; NEWMAN, A. **Handbook of Computational Molecular Biology**. Iowa State University: CRC Computer & Information Sicience, v. IX, 2005. 30-35 p.

HINTERDING, R.; MICHALEWICZ, Z.; EIBEN, A. E. **Adaptation in Evolutionary Computation:** A Survey. IEEE Transtions on Evolutionary Computation. [S.l.]: IEE Press, EUA. 1997. p. 124-141.

HOLLAND, J. Adaptation in natural and artificial systems. **Ann Arbor: Univ. of Michigan Press**, 1975.

JONES, D. T.; TAYLOR, W. R.; THORNTON, J. M. A new approch to protein fold recognition. **Nature**, p. 86-89, 1992.

JULSTROM, B. A. What have you done for me lately? Adapting operator probalities in steady-state genetic algorithm. **In Proceeding of the 6th International Conference on Genetic Algorithms.**, San Francisco, CA, USA, p. 81-87, 1995. ISSN 1-55860-370-0.

K YUE, K. M. F. P. D. T. H. S. C. E. I. S. A. K. A. D. A test of lattice protein folding algorithms. **Proc. Natl. Acad. Sci. USA**, San Francisco, p. 325-329, Jan 1995.

K.M. , M. Nonlinear multiobjective optimization, 1999.

KATIKIREDDY, A.; JOHNSON, C. M. A genetic algorithm with backtracking for protein structure prediction. **Keijzer, M: Cattolico, M., eds. GECCO 2006: Proceedings of the 8th Annual Conference on Genetic And Evolutionary Computation**, Washington,DC. EUA, p. 299-300, 2006.

KHIMASIA, M. M.; COVENEY, P. V. Protein structure prediction as a hard optimization problem: the genetic algorithm approach. **Molecular Simulation**, v. 19, p. 205-226, 1997.

KOLINSKI, A.; SKOLNICKB, J. Reduced models of proteins an their applications. **Polymer**, v. 45, p. 511-524, 2004.

KRASNOGOR, N. Studies on the Theory and Design Space of Memetic Algorithms. Phd Thesis, Bristol, United Kingdom, 2002.

KRASNOGOR, N. et al. Protein Structure Prediction With Evolutionary Algorithms, 1999.

LASKOWSKI, R. A. et al. Procheck: a program to check the stereochemical quality of protein structures. Journal of Applied Crystalography. **Journal of Applied Crystallography**, 1993. 283-291.

LAU, K. F.; DILL., K. A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. **Macromolecules**, p. 3986-3997, October 1989.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. **Principios de bioquimica**. 3. ed. [S.l.]: Savier, São Paulo, BR, 2002.

LINDEN, R. **Algoritmos Genéticos**. Rio de Janeiro: Abreu's System Ltda, 2008. ISBN 978-85-7452-373-6.

MARTIM-REMON, M. A. et al. Comparative protein structure modelling of genes an genomes. **Annual Review of Biophysics and Biomolecular Structure**, p. 291-235, 2000.

MEHUL M., ; PETER V. , C. Protein structure prediction as a hard optimization problem: the genetic algorithm approach, Cambridge, 11 August 1997.

MIETTINEN, K. M. **Nonlinear multiobjective optimization**. [S.l.]: Kluwer Academic Publishers. 1999.

MITCHELL, M. **An introduction to genetic algorithms**. London, England: MIT Press, 1997.

MOULT, J. A. Decade of CASP: progress, bottlenecks an prognosis in protein structure prediction. **Current Opinion in Structural Biology**, v. 15, p. 285-289, 2005.

OLIVARES, L.; GARCIA, L. Plegamiento de las proteinas: Un problema interdisciplinario. **Sociedad Quimica de Mexico**, p. 95-105, 2004.

PATERSON, A. L.; PRZYTYCKA, T. On the complexity of string folding. **Discrete Applied Mathematic**, v. 71, p. 217-230, 1996.

PATTON, A. L.; PUNCH III, W. F.; GOODMAN, E. D. A standard GA approach to native protein conformation preditcion. **Proceeding of international Conference on Genetic Algorithms**, p. 547-581, 1995.

PAULING, L.; COREY, R. B. Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains. **Proceesdings of the National Academy of Sciences**, v. 37, p. 235-240, 1951.

PAULING, L.; COREY, R. B. The pleated sheet, a new layer configuration of polypeptide chains. **Proceedings of the National Academy of Sciences of the United States of America**, 1951. 251-256.

PEREZ SERRADA, A. 1. **Introducción a la computación evolutiva**. [S.l.]. 1996.

PROTEIN Data Bank. Disponivel em: <http://www.rcsb.org/pdb/home/home.do>. Acesso em: 03 fev. 2012.

RAMACHANDRAN, G. N.; SASISEKHARAN, V. Conformation of polypetides and proteins. **Advances in Protein Chemistry**, 1968. 238-437.

ROHL, C. A. et al. Protein structure prediction using rosseta. **Methods in Enzimology**, 2004. 66-93.

RUSSELL, R. B.; BARTON, G. J. Structural feastures can be unconserved in proteins with similar folds. An analisys of side-chain to side-chain contacts secondary structure and accessibility. **Journal of Molecular Biology**, 1994. 332-350.

SCHONBRUN, J.; WEDMEYER, W. J.; BAKER, D. Protein Structure Prediction in 2002. **Current Opinion in Structural Biology**, v. 12, p. 348-354, 2002.

SCHULZE-KREMER, S. Genetic Algorithms and Protien Folding Protein Structure Prediction Methods and Protocols. [S.l.]: Humama Press Inc, v. 143, 2000. Cap. 9, p. 175-221.

SHMYGELSKA, A.; HOOS, H. H. BMC Bioinformatics. **An ant Colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem.**, 2005. Disponivel em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC555464/>. Acesso em: 25 mar. 2012.

SIMONS, K. T. et al. Ab initio protein structure prediction of CASP III targets using Rosseta. **Protein Suppl**, 1999. 171-176.

SRINIVASAN, R.; FLEMING, P. J.; ROSE, G. D. Ab protein folding using LINUS. **Methods in Enzymology**, 2004. 48-66.

SRINIVASAN, R.; ROSE, G. D. Ab initio prodiction of protein stucture using LINUS. **Proteins**, 2002. 489-495.

STENBERG, M. **Protein structure prediction:** a practical approch. New York, NY, USA: Oxford University Press, Inc, 1997.

TETTAMANZI, A. B. A. A. G. B. A Memetic Algorithm for Protein Structure, Milano, v. 3005 of LNCS, p. 1-10.

TOMA, L.; TOMA, S. Contact interactions method: a new algorithm for protein folding simulations. **Protein Sci**, p. 147-153, Jan 1996.

TRAMONTANO, A.; LESK, A. M. **Protein structure prediction**. 1. ed. Weinheim, Germany: John Wiley and Sons, Inc, 2006.

TRAMONTANO, A.; LESK, A. M. **Protein structure prediction**. 1. ed. Weinheim, Germany: John Wiley and Sons, Inc, 2006.

UNGER, R.; MOULT, J. Genetic algorithms for protein folding simulations. **J Mol Biol**, p. 75-81, 5 May 1993.

VULLO, A. On the role of machine learning in protein structure determination. **Journal of the Italian Association for Artificial Intelligence**, v. 1, p. 22-30, 2002.

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids: A structure for desoxyribose nucleic acid. **Nature**, New York, p. 737-738, 1953.

WHITLEY, D. **The GENITOR Algorithm and Selection Pressure:** Why Rank-Based Allocation of Reproductive Trials is Best. Proceedings of the third international conference on Genetic algorithms. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1989.

YUE, K. et al. A test of lattice protein folding algorithms. **Proc. Natl. Acad. Sci. USA**, p. 325-329, 1994.

YUE, K.; DILL, K. Sequence-structure relationships in proteins and copolymers. **Phys. Rev. E**, p. 2268-2278, 1993.

YUE, K.; DILL, K. A. Forces of tertiary structural organization in globular proteins. **Proc Natl Acad Sci USA**, p. 146-150, Jan 1995.

Apêndice

Algoritmos Genéticos

Os Algoritmos genéticos são um ramo dos algoritmos evolucionários e como tal são técnicas baseadas nos princípios da evolução para abordar uma ampla serie de problemas. Robustos, genéricos e facilmente adaptáveis, consistem de uma técnica amplamente estudada e utilizada em diversas áreas. Inspirado na maneira como o darwinismo explica o processo de evolução das espécies, (HOLLAND, 1975) decompôs o funcionamento dos AGs nas etapas de inicialização, avaliação, seleção, cruzamento, mutação, atualização e finalização estas etapas são detalhadas a continuação:

[Inicialização] Gere uma população aleatória de n cromossomos (soluções adequadas para o problema)

[Avaliação] Avalie a aptidão f(x) de cada cromossomo x da população

[Nova população] Crie uma nova população repetindo os passos seguintes até que a nova população esteja completa

[Seleção] Selecione de acordo com sua aptidão (melhor aptidão, mais chances de ser selecionado) dois cromossomos para serem os pais.

[Cruzamento] Com a probabilidade de cruzamento cruze os pais para formar a nova geração. Se não realizar cruzamento, a nova geração será uma cópia exata dos pais.

[Mutação] Com a probabilidade de mutação, altere os cromossomos da nova geração nos lócus (posição nos cromossomos).

[Atualização] Coloque a nova descendência na nova população

[Substitua] Utilize a nova população gerada para a próxima rodada do algoritmo

[Teste] Se a condição final foi atingida, pare, e retorne a melhor solução da população atual.

[Repita] Vá para o passo 3.

É importante ressaltar que os GA não são técnicas de *Hill climbing ,e,* por tanto, eles não ficam estagnados simplesmente pelo fato de terem encontrado um máximo local. Se um GA encontrou um bom individuo que é o melhor de todas as soluções achadas o GA segue na procura de melhores soluções (LINDEN, 2008).

Os algoritmos genéticos são uma técnica de busca e otimização que têm as seguintes características:

- Paralelos: O algoritmo genético trabalha com um conjunto de possíveis soluções que são avaliadas simultaneamente, a partir do teorema dos esquemas de (HOLLAND, 1975), tem-se que ao fazer uma busca por populações, a evolução de um algoritmo genético tende a favorecer indivíduos que compartilhem determinadas características, sendo assim capaz de avaliar implicitamente determinadas combinações ou esquemas como mais ou menos desejáveis, efetuando o que chamamos de uma busca por hiperplanos, de natureza paralela (GOLDBERG, 1989);
- Facilidade no uso de restrições: Mesmo que as restrições apresentem diversos graus de importância o trabalho com este tipo de problemas é facilitado por os AGs (BARBOSA, 1996). Por exemplo, se dois indivíduos violam restrições é considerado mais apto àquele que viola as mais flexíveis (*soft constraints*) em detrimento do que viola as mais graves (*hard constraints*).
- Podem trabalhar com funções discretas e continuas: Os AGs tem a capacidade de lidar com funções reais, discretas, booleanas e até mesmo categóricas, podendo inclusive misturar as representações sem prejuízo para a habilidade do AGs de resolver os problemas (LINDEN, 2008).
- Busca estocástica: Os AGs têm componentes aleatórios, mas também usam informação da população atual para determinar o estado seguinte da busca. As probabilidades de aplicação dos operadores genéticos fazem com que estes operem de forma previsível estatisticamente, apesar de não permitirem que se determine com exatidão absoluta o comportamento do sistema.
- Busca codificada: "os AGs não trabalham sobre o domínio do problema, mas sim sobre representações de seus elementos" (PEREZ SERRADA, 1996). Tal fator impõe ao seu uso uma restrição: para resolver um problema é necessário que o conjunto de soluções viáveis para este possa ser de alguma forma codificado em uma população de indivíduos.

Para aplicar um AG a uma classe especifica de problemas, precisamos definir alguns aspectos críticos sobre a implementação:

a) Representação: Um cromossomo num AG representa uma solução candidata, este critério refere-se à estrutura dos cromossomos (codificação). Deve ser definida uma codificação a mais natural possível para o problema, no melhor dos casos as representações não deveriam gerar soluções inválidas de esta forma ajudar a evolução. Os operadores genéticos devem ser adaptados à codificação.

   Os cromossomos podem ser ter, tipicamente, representações binarias, reais, e inteiras.

b) Função de Avaliação: Tecnicamente, a função de avaliação é uma função ou procedimento que atribui uma medida de qualidade aos indivíduos (genótipos) (CHORRO S. B, 2007).

   Tipicamente, esta função é composta por uma medida de qualidade no espaço das soluções possíveis (dos fenótipos) e pela representação inversa (decodificação).

c) Seleção: Os mecanismos de seleção baseiam-se no princípio Darwiniano da "sobrevivência dos mais aptos": os indivíduos com melhores valores de aptidão têm maiores probabilidades de serem escolhidos para reprodução. Portanto, além destes mecanismos permitirem determinar que indivíduos (progenitores) são escolhidos para cruzamento, também permitem escolher que indivíduos (descendentes) devem sobreviver para a geração seguinte.

   Os principais mecanismos de seleção são:

   - O método da roleta (GOLDBERG, 1989) é um método estocástico que consiste em associar os indivíduos a porções contíguas de uma roleta, em que cada porção é proporcional à aptidão do indivíduo que lhe está associado (os indivíduos com maior valor de aptidão têm maiores probabilidades de serem escolhidos). São então realizados vários lançamentos da roleta, sendo selecionados os indivíduos associados às porções atingidas por cada um destes lançamentos.
   - A ordenação linear (BACKER, 1985), os indivíduos da população são ordenados de acordo com os seus valores da função de aptidão, sendo atribuído a cada indivíduo um

valor que corresponde à sua posição na população ordenada. Desta forma, ao pior indivíduo é atribuído o valor 1 e ao melhor o valor N (em que N é o tamanho da população). Depois, a cada indivíduo é atribuída uma probabilidade de seleção calculada com base numa dada distribuição (as mais usuais são a linear e a exponencial). Esta técnica trava a convergência prematura do algoritmo (não há favorecimento dos melhores indivíduos) e evita, em gerações avançadas, a estagnação da população.

- A seleção por torneio (GOLDBERG, 1989) consiste em escolher aleatoriamente certo número de indivíduos da população (designado por dimensão do torneio) e fazer um torneio entre eles. Cada torneio consiste em comparar os valores de aptidão dos indivíduos envolvidos, sendo o vencedor (e o selecionado) aquele com melhor valor de aptidão. O número de torneios realizados é igual ao número de indivíduos a serem selecionados, ou seja, igual ao tamanho da população.

  Esta técnica não conduz à convergência prematura (desde que a dimensão dos torneios seja pequena), combate a estagnação da população, é simples de implementar e não requer grande esforço computacional. Este é talvez o mecanismo de seleção mais utilizado na resolução de problemas de otimização.

d) Substituição Parental

Um destes mecanismos é o elitismo, que foi introduzido por (DE JONG, 1975) e consiste em reter na população os seus melhores indivíduos, os quais passam diretamente para a próxima geração. Muitos investigadores têm encontrado no elitismo vantagens significativas para o desempenho dos AG (MITCHELL, 1997). Com esta técnica, pretende-se, por um lado, garantir que os melhores indivíduos de cada geração não sejam destruídos pela ação dos

operadores genéticos de cruzamento e mutação (a definir mais à frente), e por outro, acelerar a convergência do algoritmo.

Outro mecanismo é conhecido por AG geracional onde os indivíduos de uma população se reproduzem (pela aplicação dos operadores genéticos) até que uma nova população, do tamanho da original, esteja formada. Os indivíduos dessa nova população substituem os da antiga, chamada de parental.

Finalmente temos um AG não geracional, chamado de *steady-state*, onde cada novo individuo gerado é imediatamente avaliado e testado para inserção na população parental, para este tipo de AG precisamos estabelecer uma estratégia de remoção para escolher qual individuo será substituído (CUSTÓDIO, 2008), por exemplo: substituir baseado no posto ou a substituição do pior individuo na população (WHITLEY, 1989).

e) Operadores Genéticos

1. Mutação

   O operador genético de mutação consiste em alterar randomicamente um bit do cromossomo, este operador é de muita importância, pois a utilização deste operador genético nos AGs serve para, por um lado, fazer regressar à população os valores dos genes perdidos durante o processo de seleção, de modo a que possam ser testados num novo contexto, e por outro, proporcionar a entrada de novos genes que não estavam presentes nas populações anteriores. O valor da probabilidade de mutação deve ser baixo, o suficiente para diversificar os indivíduos da população e não prejudicar a convergência do algoritmo (CHORRO S. B, 2007).

2. Cruzamento

   O operador genético cruzamento (ou recombinação) consiste em efetuar trocas de genes entre dois indivíduos. Neste processo são gerados dois novos indivíduos (descendentes), resultantes da combinação de informação contida num par de indivíduos (progenitores). O sucesso do AG está apoiado na

expectativa de que o resultado do cruzamento entre indivíduos (progenitores) com melhores valores de aptidão gere novos indivíduos (descendentes) ainda de melhor qualidade (relativamente aos progenitores). Existem vários tipos de cruzamento, os quais dependem do tipo de representação usada na codificação dos indivíduos, para um AG simples temos os seguintes operadores de cruzamento:

- Cruzamento de um ponto de corte: Sejam os cromossomos

  $X=\{X_1,X_2\}$, $B=\{B_1,B_2\} \Rightarrow X+B =\{(X_1,B_2), (B_1,X_2)\}$
- Cruzamento de dois pontos de corte: Se

  $X=\{X_1,X_2,X_3\}$, $B=\{B_1,B_2,B_3\} \Rightarrow$

  $X+B = \{ (X_1,B_2,X_3), (B_1,X_2,B_3) \}$
- Cruzamento multiponto: Dados os cromossomos:

  $X=\{X_1,X_2,X_3,X_4,X_5,X_6...\}$, $B=\{B_1,B_2,B_3,B_5,B_6...\} \Rightarrow$

  $X+B=\{(X_1,B_2,X_3,B_4,X_5,B_6...), (B_1,X_2,B_3,X_4,B_5,X_6)\}$

f) Parâmetros

O sucesso de um AG depende, em grande parte, da escolha dos seus parâmetros de configuração. Entre os parâmetros a ser definidos temos as taxas de aplicação dos operadores de mutação e cruzamento, tamanho da população, número de gerações.

Muitos pesquisadores procuram descobrir qual seria o número de parâmetros mais adequado para resolver um problema especifico. O ajuste dos parâmetros em quase todas as pesquisas é feita numa etapa chamada de *tunnig* dos parâmetros. Mais resulta difícil de imaginar que um conjunto de parâmetros possa ser adequado para resolver um problema em todos os estágios da evolução sendo os AGs um processo dinâmico (LINDEN, 2008).

Segundo (HINTERDING, MICHALEWICZ e EIBEN, 1997) as técnicas de adaptação dos parâmetros podem ser classificadas assim:

1. Determinística: Quando a mudança nos valores dos parâmetros ocorre seguindo alguma regra determinística.

2. Adaptativa: Ocorre quando existe um *feedback* por parte do AG que é usada para determinar o valor do parâmetro na próxima geração (HINTERDING, MICHALEWICZ e EIBEN, 1997).
3. Adaptativa: Ocorre quando existe um *feedback* por parte do AG que é usado para determinar o valor do parâmetro na próxima geração (HINTERDING, MICHALEWICZ e EIBEN, 1997).