Algoritmo evolutivo de muitos objetivos para predição ab initio de estrutura de proteínas

Christiane Regina Soares Brasil

Algoritmo evolutivo de muitos objetivos para predição ab initio de estrutura de proteínas

Christiane Regina Soares Brasil

Orientador: **Prof. Dr. Alexandre Cláudio Botazzo Delbem**

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA.*

USP – São Carlos
Julho de 2012

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Brasil, Christiane Regina Soares
B823a Algoritmo Evolutivo de Muitos Objetivos para Predição Ab Initio de Estrutura de Proteínas / Christiane Regina Soares Brasil; orientador Alexandre Cláudio Botazzo Delbem. -- São Carlos, 2012.
125 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012.

1. Predição de estruturas de proteínas. 2. Otimização Multiobjetivo. 3. Modelo puramente ab initio. I. Delbem, Alexandre Cláudio Botazzo, orient. II. Título.

Agradecimentos

Agradeço, primeiramente, a Deus por tudo, a começar pelo dom da vida. A Ele, toda minha gratidão.*"O Senhor vai acendendo as lâmpadas diante de nós, à medida que delas necessitamos" - T. Alberione.*

A minha família, pelo profundo amor e intenso incentivo aos estudos desde a infância, meus pais Reginaldo e Fátima, e irmãs Cintia, Tatyane e Keyna. Desde meus primeiros erros, é o amor deles que me encoraja na busca pelo melhor de mim mesma. *"Sim, todo amor é sagrado, e o fruto do trabalho é mais que sagrado" - B. Guedes.*

Ao meu namorado Rafael, por todo amor, estímulo e compreensão, amizade e cumplicidade. Pelo amor que é para mim. *"Com minhas letras e canções, com o perfume das manhãs, com a chuva dos verões, com o desenho das maçãs, com você me sinto bem. Eu, pensando em você..." - P. Moska.*

Aos meus amigos, grandes tesouros na minha vida, Karen, Valéria(s), Rodrigo, Bruno, Mayron, Ludmila, Erinaldo, Daniel, Marcilyanne, colegas de laboratório, companheiros de música, irmãos de fé. *"Amigo é coisa pra se guardar debaixo de sete chaves dentro do coração" - M. Nascimento.*

Ao meu professor e orientador Alexandre Cláudio Botazzo Delbem, que tanto me ensina. *"Você pode encarar um erro como uma besteira a ser esquecida, ou como um resultado que aponta uma nova direção" - S. Jobs.*

À banca examinadora, Hélio J. C. Barbosa, Ricardo H. C. Takahashi, Fernando L. B. da Silva e Ivan N. da Silva, pelas sugestões enriquecedoras.

Aos funcionários do ICMC, em especial ao serviço de pós-graduação e aos vigilantes do instituto.

Por fim, agradeço à FAPESP, pelo suporte financeiro essencial a este trabalho.

"Há duas formas para se viver a vida:
uma é acreditar que não existe milagre.
A outra é acreditar que todas as coisas
são um milagre" - Albert Einstein

Resumo

Este trabalho foca o desenvolvimento de algoritmos de otimização para o problema de PSP puramente *ab initio*. Algoritmos que melhor exploram o espaço de potencial de soluções podem, em geral, encontrar melhores soluções. Esses algoritmos podem beneficiar ambas abordagens de PSP, tanto o modelo *ab initio* quanto os baseados em conhecimento a priori. Pesquisadores tem mostrado que Algoritmos Evolutivos Multiobjetivo podem contribuir significativamente no contexto do problema de PSP puramente *ab initio*. Neste contexto, esta pesquisa investiga o Algoritmo Evolutivo Multiobjetivo baseado em Tabelas aplicado ao PSP puramente *ab initio*, que apresenta interessantes resultados para proteínas relativamente simples. Por exemplo, um desafio para o PSP puramente *ab initio* é a predição de estruturas com folhas-β. Para trabalhar com tais proteínas, foi desenvolvido procedimentos computacionalmente eficientes para estimar energias de ligação de hidrogênio e solvatação. Em geral, estas não são consideradas no PSP por abordagens que combinam métodos de otimização e conhecimento a priori. Considerando somente van der Waals e eletrostática, as duas energias de interação que mais contribuem para a definição da estrutura de uma proteína, com as energias de ligação de hidrogênio e solvatação, o problema de PSP tem quatro objetivos. Problemas combinatórios (tais como o PSP), com mais de três objetivos, geralmente requerem métodos específicos capazes de lidar com muitos critérios. Para resolver essa limitação, este trabalho propõe um novo método para a otimização dos muitos objetivos, chamado Algoritmo Evolutivo Multiobjetivo com Muitas Tabelas (AEMMT). Esse método executa uma amostragem mais adequada do espaço de funções objetivo e, portanto, pode mapear melhor as regiões promissoras deste espaço. A capacidade de lidar com muitos objetivos capacita o AEMMT a utilizar melhor a informação oriunda das energias de solvatação e de ligação de hidrogênio, e então predizer estruturas com folhas-β e algumas proteínas relativamente mais complexas. Do ponto de vista computacional, o AEMMT é um novo método que lida com muitos objetivos (mais de dez) encontrando soluções relevantes.

Abstract

This work focuses on the development of optimization algorithms for the purely *ab initio* Protein Structure Prediction (PSP) problem. Algorithms that better explore the space of potential solutions can in general find better solutions. Such algorithms can benefit both *ab initio* and template-based PSP, that uses priori knowledge. Researches have shown that Multiobjective evolutionary algorithms can contribute significantly in the context of purely *ab initio* PSP. In this context, this research investigates the Multiobjective Evolutionary Algorithm based on Tables applied to purely *ab initio* PSP, which has shown interesting results for relatively simple proteins. For example, one challenge for purely *ab initio* PSP is the prediction of structures with β-sheets. To work with such proteins, this research has developed computationally efficient procedures to estimate hydrogen bond and solvation energies. In general, they are not considered by PSP approaches combining optimization methods with priori knowledge. Only by considering van der Waals and electrostatic, the two interaction energies that mostly contribute to defining a protein structure, and the hydrogen bond and solvation energies, the PSP problem has four objectives. Combinatorial problems (such as the PSP) with more than three objective usually require specific methods capable of dealing with many goals. To address this limitation, we propose a new method for many objective optimization, called Multiobjective Evolutionary Algorithm with Many Tables (MEAMT). This method performs a more adequate sampling of the space of objective functions and, therefore, can better map the promising regions of this space. The ability of dealing with many objectives enables the MEANT to better use information generated by solvation and hydrogen bond energies, and then predict structures with β-sheets and some relatively complex proteins. From the computational point of view, the MEAMT is a new method for dealing with many objectives (more than ten) finding relevant solutions.

Lista de Abreviaturas

AE	Algoritmos Evolutivos
AEMO	Algoritmos Evolutivos MultiObjetivo
AEMT	Algoritmo Evolutivo Multiobjetivo baseado em Tabelas
AEMMT	Algoritmo Evolutivo Multiobjetivo com Muitas Tabelas
ASP	*Atomic Solvation Parameters*
GDT-TS	*Global Distance Test - Total Score*
MOEA-D	*MultiObjective Evolutionary Algorithm based on Decomposition*
MOGA	*Multiple Objective Genetic Algorithm*
MOOP	*MultiObjective Optimization Problem*
NPGA	*Niched-Pareto Genetic Algorithm*
NSGA	*Non-dominated Sorting Genetic Algorithm*
PAES	*Pareto Archived Evolution Strategy*
PDB	*Protein Data Bank*
ProtPred	*Protein Predictor*
PSP	*Protein Structure Prediction*
RMSD	*Root-Mean-Square Deviation*
RNM	Ressonância Nuclear Magnética
SASA	*Solvent Accessible Surface Area*
SPEA	*Strentgh Pareto Evolutionary Algorithm*
VEGA	*Vector Evaluated Genetic Algorithm*
VOES	*Vector Optimized Evolution Strategy*
WBGA	*Weighted Based Genetic Algorithm*

Lista de Figuras

Lista de Tabelas

Sumário

CAPÍTULO

1

Introdução

Um grande bem para a humanidade é o avanço científico no tratamento de doenças consideradas incuráveis, o que implica em um trabalho árduo para milhares de pesquisadores do mundo todo. No anseio de encontrar a cura, ou pelo menos, a amenização de muitas doenças, pesquisadores de diversas áreas vêm unindo esforços e conhecimentos a fim de aumentar as chances de descoberta de novos fármacos. Nas áreas de Biologia Molecular, Bioquímica e Farmácia, muito se tem feito para o desenvolvimento de novas drogas e, mais do que isso, para o conhecimento profundo sobre as funcionalidades das proteínas e suas estruturas tridimensionais. Vale ressaltar que a compreensão da origem de diversas doenças depende diretamente de um profundo entendimento da funcionalidade das proteínas, que por sua vez está relacionada às estruturas tridimensionais com que estas são encontradas na natureza [23, 44, 193, 85, 52, 116, 189].

Nesse contexto, a predição de estruturas de proteína (PSP, do inglês *Protein Structure Prediction*) por simulação computacional pode ser um caminho relevante, fomentando o desenvolvimento de pesquisas em áreas que requerem a determinação de estruturas de novas proteínas ou de complexos moleculares envolvendo proteínas. O problema de PSP consiste em determinar a configuração tridimensional a partir de uma sequência de aminoácidos [19] de uma dada proteína.

Os métodos computacionais para PSP podem ser classificados em dois grupos: *ab initio* e baseados em conhecimento. Estes priorizam estruturas (conformações) de uma proteína que se assemelham a de sequências de proteínas obtidas por meio de métodos não computacionais, como a cristalografia de raio X e a Ressonância Nuclear Magnética (RNM). Esses

métodos baseiam-se então em conhecimento a priori de estruturas conhecidas de proteínas e têm produzido resultados relevantes. Por exemplo, em [49], a formação de complexos entre proteína e diferentes antígenos foi avaliada utilizando uma estrutura de proteína construída em computador com base em trechos de proteínas da mesma família com alto grau de similaridade.

Por outro lado, na predição *ab initio* é simulado computacionalmente a ação de campos de força no sistema de forma a determinar a sua estrutura mais estável, com o menor valor de energia de interação. A PSP *ab initio* pode ser vista como um problema de otimização, em que as variáveis são os ângulos torsionais (relativos às ligações entre resíduos de aminoácidos e desses com suas cadeias laterais) e se busca a conformação da proteína de menor energia. Esse problema é combinatório no espaço dos ângulos torsionais [158], sendo difícil obter um método com eficiência computacional suficiente para resolver estruturas de proteínas que não sejam muito pequenas e simples. Por isso, os métodos *ab initio* mais eficazes têm combinado métodos de otimização com o conhecimento sobre os ângulos torsionais mais frequentes em subestruturas de proteínas (em geral envolvendo sequências com três e nove aminoácidos) como, por exemplo, os métodos Rosetta [19, 53, 52, 116], I-TASSER [193, 149, 189] e QUARK [190].

Este trabalho foca no desenvolvimento de algoritmos de otimização melhores buscando contribuir em PSP puramente *ab initio*. Acredita-se que algoritmos que melhor explorem o espaço de fase posssibilitarão obter melhores soluções tanto em abordagens puramente *ab initio* quanto em abordagens que considerem conhecimento a priori. Em outras palavras, investigam-se aspectos fundamentais de otimização presentes no problemas de PSP e exploram-se avanços por meio da modelagem de métodos computacionais de otimização. Dessa forma, esses estudos computacionais podem contribuir em PSP sem necessitar sobrepropor trabalhos que têm sido realizados por Biólogos, Físicos, Químicos, Farmacêuticos e pesquisadores de áreas correlacionadas. De fato, há diversos aspectos envolvendo as bases de conhecimento sobre proteínas que não poderiam ser tratados por cientistas da computação. Por exemplo, Contreras [131] explora questões relativas ao esgotamento de informações que se pode extrair das bases de dados de proteínas em prol do desenvolvimento de melhores predições.

No contexto de PSP puramente *ab initio*, um conjunto de trabalhos tem mostrado que Algoritmos Evolutivos Multiobjetivos (AEMOs) podem contribuir significativamente [79, 81, 80, 120, 122, 123, 121, 72, 71, 29, 28, 25, 27, 7, 18, 26]. A otimização multiobjetivo mostra que, além do espaço de busca (todas as combinações de valores possíveis das variáveis de um problema), há o espaço das funções objetivo. Ao se considerar cada critério (objetivo) de

avaliação de uma solução (cada energia de interação que influência a estrutura da proteína) de forma independente, pode-se mapear as soluções no espaço da funções objetivo. As relações (proximidade ou distância) entre as soluções nesse espaço revelam em geral surpresas em relação à distribuição dessas no espaço de busca de problemas complexos, como problemas combinatórios em que as funções objetivo apresentam muito ótimos locais (multimodais). Os AEMOs usam esse tipo de informação para melhor orientar o processo de busca em direção a melhores soluções.

Este trabalho busca avançar em relação ao projeto de um AEMO que gerou o método ProtPred [120, 122, 123, 121, 72, 71, 29, 28, 25, 27, 7, 18, 26]. O ProtPred, inclui uma abordagem mono-objetivo (mono-ProPred) e outra multi-objetiva (NSGA-ProtPred) baseada no AEMO denominado NSGA-II [122] (do inglês, *Non-dominated Sorting Genetic Algorithm*). Em [152] foi proposto um novo AEMO, chamado Algoritmo Evolutivos Multiobjetivo por Tabelas (AEMT) que foi capaz de gerar soluções mais interessantes que o NSGA-II para o problema combinatório de projeto de redes de distribuição de energia elétrica [152]. Esse problema e o problema de PSP possuem certas semelhanças segundo aspectos computacionais, como utilizam mais que dois objetivos, as suas instâncias envolvem em geral centenas (ou mesmo milhares) de variáveis, além de funções objetivo serem multimodais.

A pesquisa mostra uma contribuição significativa do AEMT para PSP em relação ao NSGA-ProtPred. Porém, um dos desafios para PSP puramente *ab initio* é a predição de estruturas com folhas-β. Para esse fim, é preciso considerar modelos computacionalmente eficientes para energias de ligação de hidrogêncio e solvatação que, em geral não precisam ser consideradas em técnicas que combinam métodos de otimização com conhecimento a priori. Dessa forma, uma contribuição importante deste trabalho é o desenvolvimento de um modelo que descreva a interação proteína-solvente.

Para investigar tal interação, poderiam ser considerados os diversos modelos de água (modelos explícitos) que a literatura apresenta [185, 77]. No entanto, cômputos seriam realizados átomo a átomo (ou sítio a sítio), envolvendo moléculas de água e proteína, gerando alto custo computacional. Nesta pesquisa em questão, o solvente é tratado como meio dielétrico, isto é, não se considera explicitamente sua interação átomo a átomo. A partir de um modelo implícito ou contínuo [103], a contribuição energética proteína-solvente (solvatação) é calculada, considerando a superfície acessível ao solvente (SAS) [117]. A SAS é a área de superfície de uma biomolécula (proteína, DNA, etc) que é acessível a um solvente e pode ser calculada usando o algoritmo '*rolling ball*' de Shrake e Rupley[162]. Neste trabalho, utiliza-se o cálculo de SAS proposto por Gaudio e Takahata [74, 70]. Diferentes

conjuntos de parâmetros foram avaliados neste trabalho para se concluir quais apresentaram resultados adequados.

Estudo similar foi realizado para modelagem da energia de ligações de hidrogênio. Primeiramente, as ligações de hidrogênio foram modeladas por meio da modificação do potencial 10–12 Lennard-Jones [69]. A fim de obter melhores resultados, foram efetuadas modificações, alterando o modelo original [69], para que a função privilegie os pontos com distâncias menores entre os átomos de H e O. Não sendo suficiente tais mudanças para melhorias significativas nas predições, houve a necessidade do desenvolvimento de um novo modelo para ligação de hidrogênio, baseado no trabalho de Frishman e Argos, que trabalha com três componentes de energia, refinando os resultados [70].

Somente ao considerar as duas energias de interação que mais contribuem para definição da estrutura protéica (van der Waals e eletrostática [133, 13, 135]) e as energias de ligação de hidrogênio e solvatação tem-se um problema com quatro objetivos. Problemas com mais que três objetivos têm sido classificados com prolemas de muitos objetivos [99]. O aspecto de muitos objetivos pode reduzir drasticamente o desempenho de um AEMO desenvolvido para até três objetivos. Isso deve-se basicamente ao fato de o tamanho do espaço da funções objetivo (seu hipervolume) aumentar de forma exponencial com o número de objetivos.

Os desafios enfrentados em PSP puramente *ab initio* investigando o AEMT para PSP resultaram na proposição de um novo AEMO de muitos objetivos, denominado "Algoritmo Evolutivo Multiobjetivo com Muitas Tabelas" (AEMMT). Esse método realiza uma amostragem mais adequada do espaço das funções objetivo e, com isso, consegue melhor mapear as regiões promissoras nesse espaço. Essa amostragem é gerada utilizando critérios adicionais para avaliar uma solução. A partir das quatro contribuições energéticas consideradas, funções ponderações combinando dois a dois, três a três, e quatro a quatro produzem um conjunto de quinze critérios adicionais. Devido à estratégia de seleção de soluções do AEMMT, a inclusão de novos critérios não aumenta a complexidade para se investigar o espaço de funções, pelo contrário, benefia significativamente uma melhor amostragem desse espaço. Essa capacidade para lidar com muitos objetivos possibilitou ao AEMMT usufruir melhor das informações adicionais geradas pelas energias de solvatação e ligação de hidrogênio modeladas neste trabalho e, então, predizer estruturas com folhas-β de forma puramente *ab initio*. Na prática, outros critérios e combinações dos potenciais podem também serem incluídos, conforme mostrado no Capítulo 5.

O ineditismo desses resultados, motivaram a análise de desempenho do

AEMMT para proteínas mais complexas com mais de um domínio (Seção 5.3) e também da contribuição de cada energia de interação durante o processo evolutivo do AEMMT. Por fim, verificou-se que o AEMMT consegue predizer estruturas mais complexas desde que seus domínios não possuam interações particulares, como, por exemplo, ligações de dissulfeto (não modelada no AEMMT).

Esse resultado mostra a importância de desenvolver modelos computacionalmente eficientes considerando aspectos ainda não tratados no AEMMT. Dada a capacidade de o AEMMT em considerar muitos objetivos, os novos modelos devem beneficiar a qualidade das predições para proteínas mais complexas sem reduzir sua capacidade de exploração dos objetivos. Na verdade, a inclusão de novos objetivos no AEMMT tem auxiliado na amostragem do espaço de objetivos, aumentando a qualidade das soluções encontradas.

Do ponto de vista computacional deve-se ressaltar a contribuição científica por meio de um novo algoritmo capaz de apresentar vantagens na exploração de soluções em um problema combinatório complexo. Mostra-se a superioridade dessa abordagem em relação a um dos principais AEMOs, o NSGA-II. Além disso, o AEMMT consegue trabalhar adequadamente com mais de dez critérios de avaliação de uma solução, superando nesse aspecto o MOEA-D [192] (Subseção 4.3.5), que está entre os mais importantes AEMOs que lidam com muito critérios.

1.1 Tese e Contribuições

Dentro do contexto apresentado, essa tese desenvolvida neste trabalho pode ser sintetizada da seguinte forma:

- É possível desenvolver um AEMO capaz de predizer estruturas de proteínas relativamente complexas em predição puramente *ab initio*. Mais especificamente, um AE de muitos objetivos eficiente é capaz de atingir esse nível de predição.

O AEMMT desenvolvido e os resultados obtidos com ele confirmam essa tese. Além disso, as análises e experimentos realizados revelam outras contribuições importantes deste trabalho:

- O AEMMT é capaz de lidar com mais de dez critérios;
- Esse método pode lidar com proteínas de várias classes. Para isso, podem ser necessários potenciais distintos, para melhor modelar aspectos de cada classe;

- O AEMMT é também capaz de lidar com problemas combinatórios em que incertezas ou imprecisões nos modelos de critérios de avaliação podem envolver um conjunto de variações de cada critério processados simultaneamente durante a otimização;
- O método proposto mostra-se capaz de predizer folhas-β de forma puramente *ab initio*;
- O AEMMT obtem facilmente a contribuição relativa entre as energias ao longo do processo de predição. Tal informação é de interesse na investigação de relações entre o dobramento de uma proteína e sequências de configurações geradas pelo AEMMT ao longo de suas gerações.

1.2 Organização do texto

Os demais capítulos da tese estão organizados da seguinte maneira:

- **Capítulo 2:** apresenta os conceitos sobre proteínas e o problema de predição de estrutura das mesmas, além de explicar a modelagem *ab initio* usada neste trabalho e os modelos de energia *full-atom* utilizados;
- **Capítulo 3:** explica as modelagens de energias, como ligação de hidrogêncio e solvatação, considerando reformulações visando PSP por meio de AEMOs;
- **Capítulo 4:** apresenta conceitos básicos de algoritmos evolutivos multiobjetivos, e suas aplicações ao problema de predição de estruturas de proteínas, destacando o algoritmo evolutivo multiobjetivo baseado em tabelas (AEMT);
- **Capítulo 5:** propõe o algoritmo evolutivo multiobjetivo baseado em tabelas, descrevendo a metodologia utilizada e os resultados preliminares;
- **Capítulo 6:** descreve os experimentos realizados, assim como os resultados obtidos em cada etapa de desenvolvimento, apresentando os avanços alcançados gradativamente com o algoritmo AEMT;
- **Capítulo 7:** apresenta as conclusões deste trabalho, suas contribuições, suas limitações e propostas para trabalhos futuros.

CAPÍTULO

2

Problema de predição de estruturas terciárias das proteínas

2.1 Considerações iniciais

Este Capítulo visa descrever o problema de predição de estruturas terciárias das proteínas (PSP), apresentando na Seção 2.2 os conceitos sobre as estruturas das proteínas. A Seção 2.3 discute os diferentes métodos para a modelagem de PSP: por homologia, *threading*, *ab initio* e semi *ab initio*. O método utilizado neste trabalho é o puramente *ab initio* em algoritmos evolutivos. Na Seção 2.4 são descritos os modelos *ab initio* usados no desenvolvimento dos algoritmos evolutivos para PSP.

2.2 Estrutura de Proteínas

Toda célula, seja animal, vegetal ou até mesmo microbiana, envolve pelo menos uma proteína. As proteínas representam cerca do 50 a 80% do conteúdo (peso seco) da célula sendo, portanto, o composto orgânico mais abundante de matéria viva [133]. A própria denominação dessas moléculas indica o quão indispensáveis estas são para manutenção e reprodução da vida, uma vez que o termo proteína é originado do grego *proteio*, e significa "que tem prioridade, o mais importante".

As proteínas são formadas pela repetição de pequenas e simples unidades químicas (monômeros), ligadas covalentemente. Esses monômeros são

denominados aminoácidos[1]. Os principais elementos de um aminoácido são carbono (C), hidrogênio (H), oxigênio (O) e nitrogênio (N) (alguns aminoácidos contém enxofre (S)). Cada aminoácido possui uma estrutura básica comum, formada por um átomo de carbono central, denominado carbono-α (**C**$_\alpha$), um grupo amino **-NH**$_2$, um grupo carboxila **-COOH** e um radical **R**, também chamado de cadeia lateral do aminoácido [133], como pode ser observado na Figura 2.1. Na natureza, existem vinte e dois tipos de aminoácidos na natureza, mas apenas vinte são apresentados no código genético universal (Apêndice A), que são distinguidos uns dos outros pelo radical **R**.

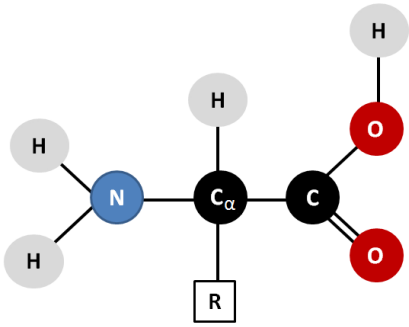

Figura 2.1: Estrutura molecular de um aminoácido.

Uma cadeia polipeptídica é composta por vários aminoácidos, conectados por meio de ligações peptídicas [24]. Uma ligação peptídica é a união do grupo amino de um aminoácido com o grupo carboxila de outro aminoácido, com a formação de um dipeptídio e a perda de uma molécula de água [39], como se pode observar na Figura 2.2.

A diferença entre as proteínas está na sequência de aminoácidos que constitue cada uma [24]. Para que duas proteínas sejam consideradas iguais é necessário que a sequência de aminoácidos seja a mesma, sabendo que cada sequência de aminoácidos corresponde a uma organização molecular única. A organização molecular é a maneira em que os aminoácidos interagem entre si e/ou com o meio. Existem níveis de organização molecular, que são chamadas de estruturas [24].

Essas estruturas podem ser classificadas em quatro tipos [24]: primária, secundária, terciária e quaternária, conforme pode ser observado na Figura 2.3.

A **estrutura primária** é a sequência de aminoácidos que diferencia uma proteína da outra, sendo o nível de organização molecular mais simples e mais importante, pois dele se origina o arranjo espacial da molécula. A **estrutura secundária** é a conformação tridimensional, na qual os aminoácidos estão

[1]Aminoácidos são os monômeros considerados isoladamente, enquanto que resíduos são os aminoácidos ligados na cadeia peptídica, pois no processo de formação da proteína ocorre a perda de átomos que compunham a estrutura original do aminoácido.

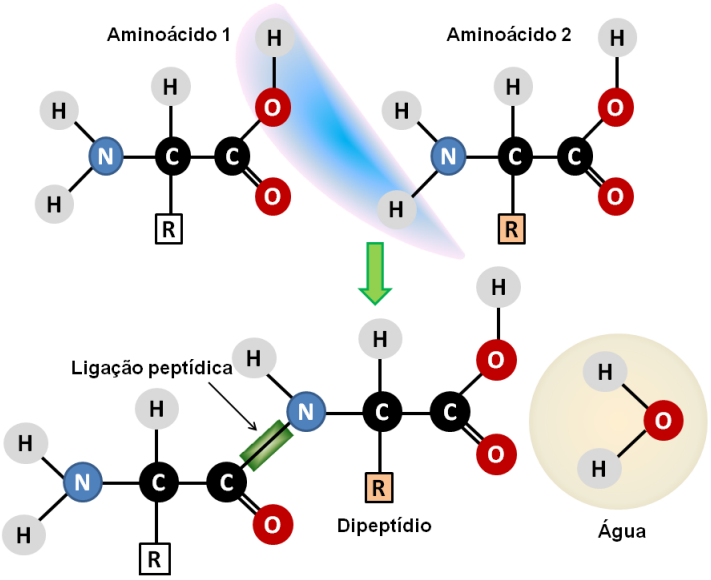

Figura 2.2: Processo de formação de uma ligação peptídica entre dois aminoácidos.

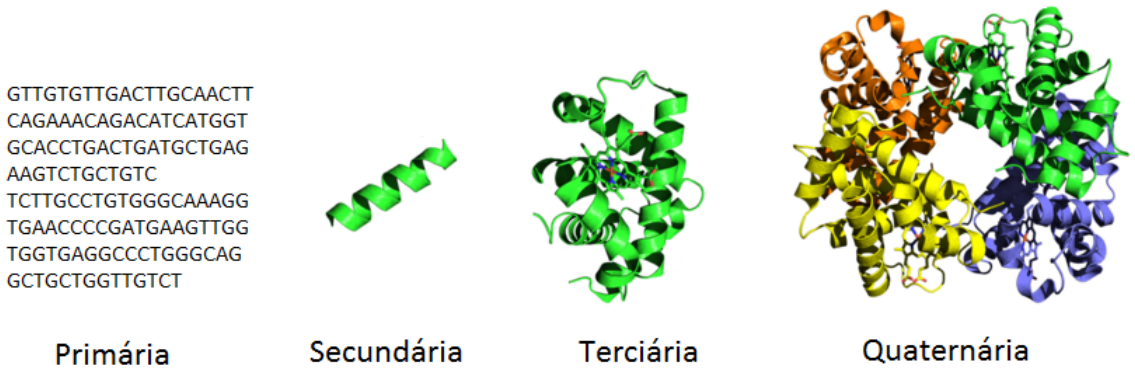

Figura 2.3: Estruturas de representação de proteínas.

dispostos interagindo entre si. São três os tipos principais de estruturas secundárias [24]:

Hélices-α são estruturas que assumem uma forma helicoidal formadas por 3.6 resíduos de aminoácidos por volta, como uma corda enrolada em torno de um tubo imaginário [141], como se pode observar na Figura 2.4.

Figura 2.4: Estrutura secundária de hélice-α.

Folhas-β são estruturas formadas entre 5 a 10 resíduos, em que um segmento (fita) da cadeia interage com outro, paralelamente, resultando em uma estrutura achatada e rígida [141], como se pode observar na Figura 2.5.

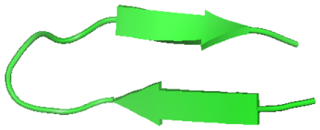

Figura 2.5: Estrutura secundária de folha-β.

As folhas-β podem ser paralelas ou antiparalelas, e diferem pelo ângulo formado nas ligações de hidrogênio, como mostra a Figura 2.6.

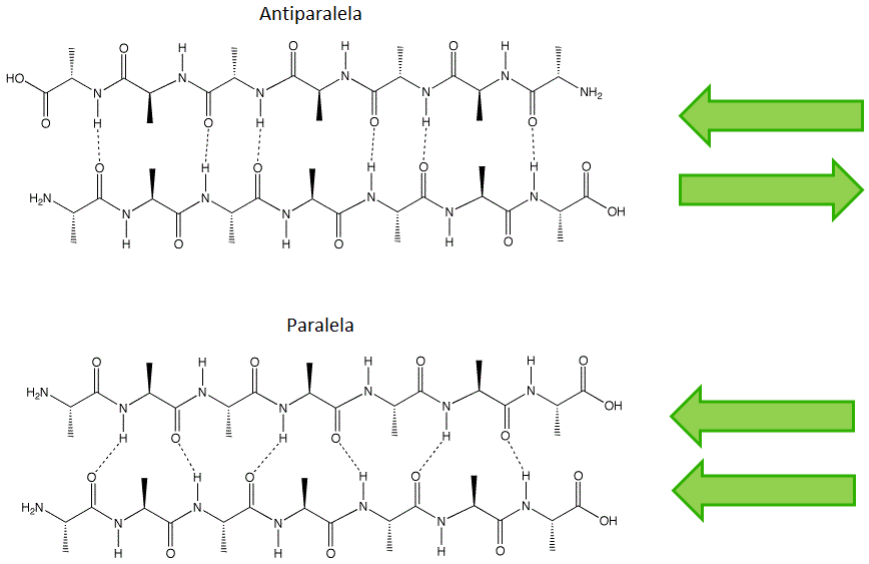

Figura 2.6: Folhas-β antiparalela e paralela.

Voltas são as estruturas responsáveis pela inversão da direção da cadeia polipeptídica, muito presente em proteínas globulares, e geralmente envolvem de três a quatro resíduos [141] (Figura 2.7).

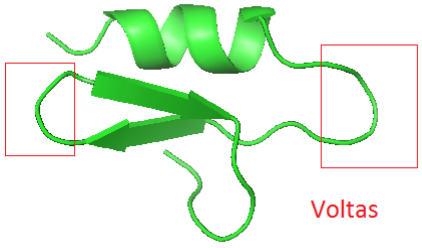

Figura 2.7: Estrutura secundária de voltas.

A **estrutura terciária** da proteína pode ser definida como a maneira com que as estruturas secundárias estão arranjadas tridimensionalmente em uma cadeia polipeptídica, dobrando-se ou não [107]. Enquanto a estrutura secundária é determinada pela interação estrutural de curta distância, a terciária é caracterizada pelas interações de longa distância entre aminoácidos, tais como interações hidrofóbicas, eletrostáticas, pontes de hidrogênio, entre outras [24]. Essa estrutura confere a atividade biológica às proteínas, ou seja, suas funções. Geralmente, as proteínas desempenham nos seres vivos as seguintes funções, fundamentais para a manutenção da vida: estrutural, enzimática, hormonal, de defesa, nutritivo, coagulação sangüínea e transporte, as quais estão intimamente ligadas a estrutura tridimensional que cada proteína apresenta.

A estrutura terciária de uma proteína é determinada pelas cadeias laterais dos aminoácidos; algumas cadeias são tão longas e hidrofóbicas[2] que perturbam a estrutura secundária helicoidal, gerando a dobra da proteína [8]. Portanto, as partes hidrofóbicas da proteína aglomeram-se no interior da proteína dobrada, por estar longe da água ou do ambiente onde a proteína está imersa, enquanto que as partes hidrofílicas ficam expostas na superfície da estrutura da proteína. Na estrutura terciária podem ser observados mais de um **domínio** nas moléculas, uma vez que domínio protéico é uma parte da cadeia polipeptídica que pode se enovelar independentemente para formar uma estrutura compacta e estável [24].

Algumas proteínas podem apresentar duas ou mais cadeias polipeptídicas, sendo que essas moléculas são estabilizadas pelas mesmas interações das estruturas terciárias. A conformação espacial dessas cadeias, interagindo entre si, é que determina a **estrutura quaternária** [24]. O composto de cadeias polipeptídicas pode produzir funções diferentes, do que se considerarmos cada estrutura terciária separadamente. A Figura 2.3 mostra a estrutura quaternária com um exemplo de molécula da Hemoglobina humana, composta por quatro cadeias polipeptídicas distintas.

Muitos são os pesquisadores que têm investigado o dobramento protéico, ou seja, a formação das estruturas terciárias, porém é um tema muito debatido, ainda com muitas questões não resolvidas, justamente pelo seu alto grau de complexidade [60, 104, 61, 137, 182, 21, 54, 146, 14, 41, 42, 147, 164]. Devido à dificuldade de compreensão desse processo, procura-se determinar a estrutura da proteína já dobrada, e posteriormente, obter características da mesma. A Secão 2.3 apresenta as diversas pesquisas sendo desenvolvidas para o problema de PSP.

[2]Cadeias hidrofóbicas tendem a se afastar da água, indo para o interior da molécula, enquanto que cadeias hidrofílicas situam-se na parte em contato com a água.

2.3 Problema de predição de estruturas

Há dois métodos experimentais disponíveis para determinação das estruturas terciárias de uma proteína: Cristalografia de Raio X, e Ressonância Nuclear Magnética (RNM) [24]. No entanto, esses métodos não são suficientes, porque além de serem processos caros e lentos, apresentam limitações em relação ao tamanho das proteínas. Desse modo, necessita-se de um método computacional que seja rápido e confiável para predizer estruturas de proteínas a partir de sequências protéicas, uma vez que o número dessas sequências vem crescendo a cada dia.

Segundo o paradoxo de Levinthal, uma cadeia polipetídica tem uma quantidade suficiente de estados de conformações possíveis [51]. Devido ao grande número de graus de liberdade da cadeia principal, relativos aos ângulos principais Φ e Ψ (Figura 2.8), para cada valor atribuído a um desses ângulos de um aminoácido, uma nova estrutura tridimensional é obtida. Na cadeia polipetídica, o dobramento da proteína depende dos ângulos de torção Φ e Ψ, tal que Φ é o ângulo de torção entre C_α e N, e Ψ, entre C_α e C. Embora a ligação peptídica seja plana, há rotação ao redor das ligações ao C_α de cada resíduo, permitindo o enovelamento da proteína. A análise da rotação desses ângulos identificou as regiões permitidas, onde não há sobreposição entre os átomos, e regiões não-permitidas, onde há essa sobreposição (choque entre os átomos). A partir dos ângulos Φ e Ψ temos um diagrama bidimensional, onde as regiões permitidas e proibidas estão nitidamente separadas. Tal diagrama é chamado de **diagrama de Ramachandran** [143] (Figura 2.9).

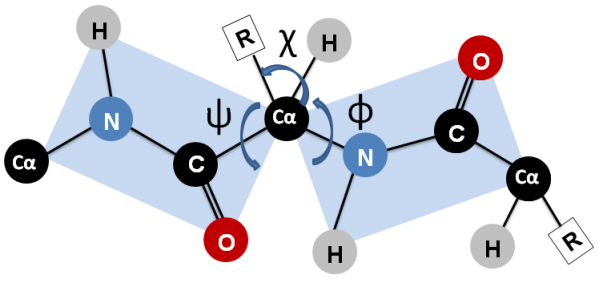

Figura 2.8: Ângulos dihedrais da proteína que constituem a cadeia principal (Φ e Ψ) e a cadeia lateral (χ)[40].

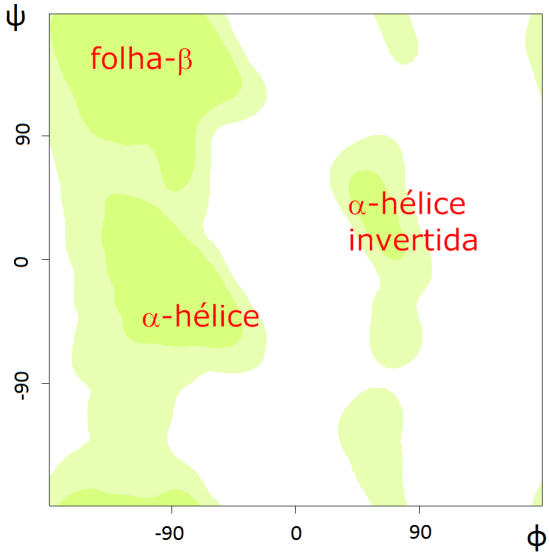

Figura 2.9: Diagrama de Ramachandran que mostra as conformações preferidas nas cadeias polipeptídicas, e suas zonas permitidas/proibidas.

Considerando que os ângulos Φ e Ψ podem assumir m valores cada um e que, para cada par desses valores em um resíduo resulta uma configuração espacial, cada resíduo pode estar em somente uma de m^2 configurações espaciais possíveis. Assim, para uma proteína com n resíduos têm-se da ordem de m^{2n} possíveis configurações espaciais [109]. Portanto, a determinação da estrutura de uma proteína por meio de busca pelo espaço de soluções é um problema intratável, ou seja, não-polinomial [158, 109, 41, 14].

Desse modo, são necessárias heurísticas e metaheurísticas computacionalmente eficientes que buscam aproximações da solução exata de um determinado problema, entre as quais se destacam os algoritmos evolutivos (AEs) [75, 46, 45]. Os AEs apresentam soluções consideradas adequadas para muitos problemas de grande complexidade. O conjunto de indivíduos investigados em uma iteração do AE é chamado de população. Os indivíduos da próxima população são gerados a partir de indivíduos da população atual. O operador de mutação de um indivíduo gera novos indivíduos semelhantes a um já conhecido (com pequenas alterações); enquanto que o operador de recombinação (*crossover* ou cruzamento) explora novas regiões do espaço de busca, combinando as coordenadas de dois indivíduos conhecidos (Apêndice B). Outro ponto positivo dos AEs é a relativa facilidade com que trabalham com os múltiplos critérios, satisfazendo a necessidade que o problema de predição tem ao buscar a minimização de várias energias.

As representações da proteína nas quais os AEs têm utilizado são: lattice [182], *off*-lattice [32] e *full-atom* [127, 42]. A representação utilizada neste trabalho é *full-atom*. Dentre os métodos de otimização para o problema de PSP, destacam-se as seguintes modelagens: baseados em homologia, *threading*, *ab initio* [184, 165] e semi *ab initio* [181], descritos nas subseções seguintes.

2.3.1 Modelagem por homologia

A modelagem por homologia busca predizer a estrutura terciária de uma proteína desconhecida com base em uma estrutura tridimensional conhecida (molde) de uma outra proteína. A primeira etapa do método é a identificação de, pelo menos, uma proteína com estrutura tridimensional conhecida, que serve de molde para a determinação da estrutura da proteína-problema. Definido o molde, passa-se ao alinhamento da sequência-problema com a sequência-molde. Após o alinhamento, é possível reconhecer regiões das sequências conservadas e regiões variáveis. As primeiras correspondem às regiões de máxima similaridade, isto é, em que as conformações devem ser muito semelhantes. Nas regiões variáveis não há correspondência estrutural, em geral, encontram-se principalmente voltas [90]. O melhor alinhamento

de várias sequências de estrutura conhecida é obtido por sobreposição das moléculas ou a partir de restrições espaciais. Uma melhoria na qualidade do alinhamento das sequências pode ser obtida utilizando outras informações de especialistas.

A modelagem por homologia é restringida pelo universo relativamente pequeno de estruturas terciárias conhecidas e pela necessidade de similaridades de sequências. Por outro lado, é frequente encontrar proteínas com baixa similaridade na sequência, mas que possuem estrutura terciária e funções similares. Essa característica motivou o desenvolvimento da modelagem por *threading*.

2.3.2 *Modelagem por* "threading"

As abordagens de *threading* são baseadas no fato de que muitas estruturas de proteínas no *Protein Data Bank* (PDB) [15] possuem configuração espacial similares, mesmo com sequências relativamente menos similares. As investigações indicam que muitas proteínas de natureza (sequência) distinta dobram-se da mesma forma produzindo estruturas semelhantes. Portanto, uma outra estratégia para a predição de estrutura terciária de proteínas é determinar a estrutura de uma nova proteína pela busca de seu melhor ajuste a alguma estrutura tridimensional particular na biblioteca de estruturas.

A abordagem de *threading* é utilizada quando a proteína não tem sequência com alta similaridade, mas pode ter uma estrutura tridimensional semelhante [118]. O alinhamento da sequência de busca com o modelo de estrutura pode ocorrer por alinhamento sequência-sequência ou sequência-estrutura. O alinhamento sequência-sequência visa encontrar o melhor alinhamento entre a sequência-problema e a sequência-molde por meio de inserções e remoções. No alinhamento sequência-estrutura, a sequência de busca é movimentada sobre a estrutura tridimensional sujeita às restrições físicas pré-determinadas referentes ao tamanho dos elementos da estrutura secundária, às regiões de volta que podem ser fixas ou variáveis dentro de um intervalo, entre outras restrições. As interações de pareamento e hidrofóbicas entre resíduos não locais são determinadas para cada posição da sequência contra a estrutura. Esses cálculos são usados para determinar o alinhamento mais favorável da sequência questionada contra o modelo de estrutura selecionado [12].

2.3.3 *Modelagem* "ab initio"

Nas abordagens *ab initio* não é necessário que exista qualquer tipo de homologia na sequência ou similaridade de estruturas tridimensionais em relação às proteínas de estrutura conhecida. Nesse contexto, algumas

técnicas computacionais têm sido utilizadas para mapear os modelos de sequência em uma estrutura, tais como: modelos de cadeia de Markov, RNAs, Inteligência Artificial baseada em regras, Monte Carlo, Algoritmo de Estimação de Distribuição [165, 11, 137, 91, 188, 36]. As dinâmicas moleculares podem ser usadas como parte de um algoritmo *ab initio*, envolvendo simulações de forças que atuam na proteína para reproduzir seu dobramento [107]. Cui (1998) desenvolveu pesquisas com modelos *ab initio*, utilizando uma função de energia potencial que considera as interações hidrofóbicas e as interações de forças de van der Waals.

As abordagens computacionais *ab initio* padrões encontram a estrutura tridimensional realizando buscas no espaço de conformações adequado, de acordo com campos de força [184]. Esses modelos computacionais são baseados em métodos de otimização, que envolvem dois pontos importantes: (1) a especificação da função de minimização e (2) a escolha do algoritmo de busca. As funções de minimização são baseadas em leis físicas envolvidas na estabilização do sistema, isto é, movimentação em campos potenciais (campos de força). Normalmente, a função visa minimizar a energia livre da molécula, posto que a estrutura nativa das proteínas apresenta energia mínima [109]. Portanto, os principais desafios para esses métodos *ab initio* são: a minimização da função de avaliação para proteínas complexas (por exemplo, com vários domínios), assim como o crescimento exponencial do espaço de busca conforme o aumento da quantidade de resíduos da proteína.

2.3.4 *Modelagem semi* "ab initio"

Na modelagem semi *ab inito* são acessados bancos de dados de estruturas de proteína para realizar uma busca conformacional baseada em conhecimento. Essas abordagens utilizam o fato que estruturas podem ser reconstruídas utilizando bibliotecas relativamente pequenas de estrutura-modelos de segmentos curtos [181]. As investigações evidenciam a existência de unidades de dobras autônomas em domínios de proteínas, que apresentam um papel importante no processo de dobramento da proteína [95]. As abordagens que utilizam a técnica de modelagem semi *ab initio* propõem a predição da estrutura tridimensional das proteínas baseada na recuperação, a partir de bases de dados, de pequenos segmentos [181] ou por meio do alinhamento de sequências [21] selecionando os segmentos consecutivos dos peptídeos que compõem a proteína. A próxima etapa da técnica semi *ab initio* combina as subunidades estruturais a fim de obter a estrutura da proteína completa. Esse processo ocorre no sentido do amino-terminal[3] para

[3]N-terminal (amino-terminal) é uma das extremidades da cadeia polipeptídica. A outra extremidade é chamada "C-terminal" ou "carboxi-terminal". As cadeias peptídicas são escritas

o carboxi-terminal realizando a combinação das subestruturas. Rosetta [19, 53, 52, 116], I-TASSER [193, 149, 189] e QUARK [190] são os algoritmos mais relevantes para predição de proteínas que também podem ser classificados como semi *ab initio*.

A Seção 2.4 descreve os modelos *ab initio* de representação da energia (lattice, *off*-lattice *full-atom*) e as funções de energia implementadas baseadas nos campos de força usados no pacote de modelagem molecular TINKER [142]. Deve-se observar que os métodos de cálculo de campo de força do TINKER foram implementados em linguagem Fortran. Os cálculos de potenciais utilizados neste trabalho foram portados para linguagem C (utilizada no desenvolvimento dos algoritmos desta tese). Além disso, outros potenciais também foram implementados: cálculo da energia de solvatação (Seção 3.2) e energia de ligação de hidrogênio (Seção 3.4).

2.4 Os modelos "ab initio" em algoritmos evolutivos

O problema de PSP é NP-completo [158, 109, 41, 14]. Assim, pelo princípio, necessita-se de um tempo exponencial para investigar completamente o espaço de busca procurando por uma estrutura de energia mínima total. Algoritmos de busca para lidar com o PSP precisam ser computacionalmente eficientes.

Embora muitos métodos diferentes para reconhecimento de dobra tenham sido desenvolvidos, pesquisadores têm encontrado dois fatores importantes e limitantes: as atuais funções de energias não são precisas o suficiente para calcular a energia de uma dada conformação; e o outro fator é a inexistência de um método computacional direto que possa reconhecer a conformação, posto que o espaço de busca de conformações é enorme. Muitas técnicas computacioanis têm sido experimentadas no problema de PSP, a fim de enfrentar a dificuldade encontrada. Tais técnicas são: Monte Carlo, Dinâmica Molecular, Rede Neural, Algoritmos Evolutivos, Algoritmos de Estimação de Distribuição, entre outras [165, 11, 137, 91, 188, 36]. Os AEs têm apresentado resultados relevantes para diversos problemas complexos. Esses algoritmos merecem grande atenção das pesquisas em predição de proteínas, uma vez que a determinação de uma estrutura terciária requer a minimização de várias energias, bem como a avaliação da interação da proteína com o meio.

Neste trabalho as seguintes funções de energia baseadas nos códigos disponíveis no sistema de modelagem molecular TINKER [142] foram utilizadas: energia de comprimento de ligação, energia de ligação, energia Urey-Bradley, energia imprópria, energia de torção, energia de van der

da esquerda para a direita, partindo do N-terminal em direção ao C-terminal, pois é esta a ordem em que os ribossomos sintetizam as proteínas.

Waals e energia de carga. As principais ligações não-covalentes são: van der Waals, eletrostática e ligações de hidrogênio. Estas diferem-se pela geometria, tamanho e especificidade. Por essa razão, foram escolhidas essas ligações para os experimentos realizados. Além dessas energias, também foi desenvolvida a energia de solvatação, um novo critério que considera a interação proteína-solvente. A seguir são apresentados as representações de energia e as características de cada função de energia investigados nesta pesquisa.

2.4.1 Modelos de representação de energia

Os modelos de representação de energia– lattice [159, 182], *off*-lattice [32] e *full-atom* [127, 42] – são descritos ao longo desta Subseção.

Modelo lattice

O modelo lattice[4] foi proposto por Shakhnovich (1991) [159] e aplicado em seguida por Unger e Moult (1993) [182]. As lattices são reticulados (*grids*) regulares quadrados ou cúbicos. A justificativa da utilização de modelos lattice está na simplificação no processo de busca no AE para predição de estruturas de proteínas. Contudo, é provado que mesmo utilizando modelos lattice, o problema de predição de estruturas é intratável [111]. A simplificação consiste em uma representação mais simples da proteína no espaço de forma que cada resíduo seja um elemento rígido posicionado um vértice do reticulado, e que cada resíduo é um elemento rígido posicionado em um ponto de uma lattice. Cada resíduo vizinho na sequência de aminoácidos deve estar em uma posição vizinha no lattice [22]. Além disso, o modelo não representa a estrutura interna de cada resíduo. Essas simplificações reduzem significativamente os cálculos computacionais. Por outro lado, esse modelo pode preservar características relevantes do sistema real como, por exemplo, as interações polares entre os resíduos [109]. Existem modelos lattice mais elaborados como pode ser visto em [31, 161].

Nos AEs para PSP com modelo lattice o cromossomo (representação computacional de uma solução) corresponde às posições de cada resíduo na lattice. Estas podem ser representadas por coordenadas internas ou por coordenadas cartesianas. Nas coordenadas internas, armazena-se a posição do resíduo atual, em relação ao resíduo anterior. Em outras palavras, o cromossomo armazena qual o deslocamento dentro da lattice para o resíduo atual em relação ao anterior [109]. Nas coordenadas cartesianas, armazena-se a posição de cada resíduo, de forma absoluta, não dependendo de cálculos em

[4] A palavra lattice, em geral, corresponde a uma malha quadricular.

relação à posição do resíduo anterior [111].

O modelo HP (hidrofóbico) [115], no qual H indica resíduos hidrofóbicos e P, aminoácidos polares pode apresentar uma representação lattice cúbica da cadeia polipeptídica. A simplicidade desse tipo de modelo acarreta em vantagens e torna possível o estudo de dobramento de proteínas preservando vários aspectos relevantes. A sequência de aminoácidos é dobrada em uma malha tridimensional de pontos, determinando-se as posições relativas dos mesmos. O modelo é flexível na obtenção de intervalos dos possíveis dobramentos que a proteína pode assumir, a partir da análise da distribuição de vértices referentes aos aminoácidos dentro dos seus respectivos tetraedros na malha.

Depois do trabalho de Unger e Moult (1993), Khimasia e Coveney (1997) elaboraram uma abordagem evolutiva que trabalha com um AG canônico[5] e uma função objetivo, que apresenta um modelo de campo de forças combinados com o modelo HP. O cromossomo informa, com sua codificação, o deslocamento do resíduo atual em relação ao anterior dentro da lattice [109], de acordo com a sequência de aminoácidos. Uma molécula de comprimento n é representada por uma sequência de $n-1$ deslocamentos do tipo C, B, E, D, F, T que correspondem, respectivamente, aos seus possíveis movimentos em um cubo para cima, baixo, esquerda, direita, frente e trás [111].

Os AEs que utilizam modelos lattice têm apresentado bons resultados, com tempo computacional relativamente baixo para proteínas pequenas (com uma sequência entre dez e cem aminoácidos). Para proteínas maiores, o desempenho desses AEs significativamente reduzido [72].

Modelo off-lattice

O modelo *off*-lattice [108] aproxima-se de uma configuração mais realista, pois permite que os ângulos Φ e Ψ adotem valores dentro da região de Ramachandran [143]. Desse modo, esses modelos produzem configurações próximas da estrutura nativa para polipeptídeos pequenos.

As melhorias obtidas pelo realismo desse modelo implicam em um alto custo computacional por causa de sua complexidade. Os modelos *off*-lattice incluem os C_α, todos os átomos da cadeia principal e, algumas vezes, da cadeia lateral. A conformação da cadeia principal são geralmente representados pelos ângulos Φ e Ψ de cada átomo C_α. Se as cadeias laterais forem incluídas, podem ser rígidas, semi-flexíveis ou flexíveis. Para as cadeias laterais rígidas, suas conformações em estruturas cristalográficas em raio-X são observadas

[5]Nos AGs canônicos [45], as soluções candidatas são codificadas em vetores binários de tamanho fixo.

e, para cada aminoácido, adota-se aquela que for mais comum. Para cadeias semi-flexíveis, é usado um método empírico que também utiliza um conjunto de estruturas obtidas por raio-X e particiona esse conjunto em grupos com formas similares. A conformação média de cada grupo é chamada **rotâmero**. Assim, para cada cadeia lateral é permitido adotar alguns dos rotâmeros mais comuns. Nesse caso, uma cadeia lateral pode ter muitas conformações possíveis [108].

Diversas pesquisas são desenvolvidas nessa área, introduzindo essas ideias aos modelos lattice e *off*-lattice, sendo tratados por simulações Monte Carlo [146][147][32][164]. Nos trabalhos de Caliri et al. (2004) [32], os resultados configuracionais e termodinâmicos são comparados com as propriedades de uma proteína real. Observa-se, a partir dos resultados, que o problema de dobramento pode ser analisado por dois processos separados: a busca pela estrutura nativa e a verificação da estabilidade dessa estrutura. Primeiro, um modelo *off*-lattice é desenvolvido para estimar a efetividade de forças entrópicas (efeito hidrofóbico) [146][147] produzindo um glóbulo flexível (dobrado) e direcionando a cadeia através de configurações que se aproximam da conformação nativa. Posteriormente, o modelo lattice é usado para verificar se a energia de contato baseada em potenciais hidrofóbicos puros é, de fato, eficaz, empacotando a cadeia e buscando a estrutura nativa.

Esse tipo de modelagem de interação energética falha em prover estabilidade configuracional para o glóbulo. Por outro lado, a adição de um conjunto heurístico de especificidades estéricas (forma e tamanho da proteína) ao potencial hidrofóbico auxilia a selecionar os caminhos de dobramento e a melhorar a condição de estabilidade do glóbulo em estruturas nativas. A partir de comparações entre dois conjuntos de resultados de simulação Monte Carlo, é mostrado que especificidades estéricas adequadas mudam drasticamente a atividade do sistema configuracional. Outros trabalhos nessa linha de modelagem podem ser vistos em [84, 97, 96, 98].

Modelo full-atom

Nos modelos *full-atom*, o cromossomo representa cada resíduo que constitui a proteína por seus ângulos diedrais (Φ e Ψ) e por seus ângulos da cadeia lateral (χ). As abordagens mais simples, em geral, utilizam somente os ângulos diedrais (Φ e Ψ) [157]. Quando se utilizam esses modelos, geralmente, são necessárias adaptações nos operadores de mutação e recombinação, devido às seguintes características: o número de ângulos das cadeias laterais não é fixo [42], podendo então cada gene possuir um tamanho variável; os ângulos possuem intervalo de valores fixos [-180, 180]. Esses modelos utilizam várias funções de avaliação, baseadas em modelos de

energia potencial. Por exemplo, Cui (1998) [42] utiliza uma função de energia potencial que considera as interações hidrofóbicas e as interações de forças de van der Waals.

As funções de energia estudadas neste trabalho foram: energia de comprimento de ligação, energia de ligação, energia Urey-Bradley, energia imprópria, energia de torção, energia de van der Waals e energia eletrostática; e entre as quais foram testadas as energias de van der Waals e eletrostática. Observe que, apesar de todas essas energias estarem disponíveis nos algoritmos desenvolvidos neste trabalho, nem todas são efetivamente empregadas no processo de predição, conforme pode ser visto nos Capítulos 3, 5 e 6.

Essas funções foram implementadas a partir dos códigos disponíveis no sistema de modelagem molecular TINKER [142], que é um pacote geral e completo para dinâmicas e mecânicas moleculares, com algumas especificidades para biopolímeros. Esse pacote foi desenvolvido para ser compatível com diversos outros, como os pacotes de campos de força CHARMM [30, 125], AMBER [34] e AMBERPLUS [88]. Outra vantagem é que reconhece os seguintes formatos de arquivo:

1. Coordenadas internas, em que são apresentados os ângulos de ligação, torção e comprimentos das ligações dos átomos;

2. Formato XYZ, com as coordenadas cartesianas de cada átomo;

3. Formato PDB, amplamente utilizado para representar as estruturas protéicas, ácidos nucléicos e nucleotídeos [142].

A seguir, são apresentadas as características de cada função de energia (Hamiltonianos) implementadas neste trabalho. Vale ressaltar a diferença entre **Hamiltonianos** e **potencial de energia**. Uma expressão matemática que descreve o potencial de energia de um sistema como uma função da distância de separação entre as espécies [6] é chamado um modelo Hamiltoniano efetivo (EHM, do inglês *Effective Hamiltonian Model*) [68]. Esse modelo é chamado "efetivo" por, apesar de não ser idêntico à realidade, supõe-se que representa os principais aspectos de interações reais. Logo, para um mesmo potencial de energia pode-se ter mais de um Hamiltoniano possível, pois depende da situação real que se deseja modelar como, por exemplo, diferenças entre interações de curto ou longo alcance. O modelo de ligações de hidrogênio desenvolvido neste trabalho resultou em mais de um Hamiltoniano (Seção 3.4).

[6] Neste contexto de PSP, espécies são átomos.

Energia de comprimento de ligação

As interações de comprimento de ligação são analisadas de acordo com a variação do comprimento da ligação [154]. A energia de ligação é menor quando se tem um comprimento de referência (r_0). Quando a ligação é comprimida, a nuvem de elétrons dos dois átomos é sobreposta. Por outro lado, se a ligação é afastada do equilíbrio, a energia aumenta gradualmente. Algumas vezes, a ligação é rompida, deixando de existir.

A Equação 2.1 é a expansão de Taylor da energia potencial de ligação em função da distância real r, em torno de uma distância de referência r_0.

$$E(r) = E(r_0) + \left.\frac{dE}{dr}\right|_{r=r_0}(r-r_0) + \frac{1}{2}\left.\frac{d^2E}{dr^2}\right|_{r=r_0}(r-r_0)^2 + \frac{1}{6}\left.\frac{d^3E}{dr^3}\right|_{r=r_0}(r-r_0)^3 + \dots \quad (2.1)$$

Na forma mais simples da Equação 2.1, o termo $(r-r_0)^2$ é conhecido como aproximação harmônica. Considerando $E(r_0) = 0$ e que em $r = r_0$ a energia é nula, a primeira derivada da energia é zero. Assumindo $k_r = \left.\frac{d^2E}{dr^2}\right|_{r=r_0}$ tem-se:

$$E_{bond}(r) = \frac{1}{2}k_r(r-r_0)^2. \quad (2.2)$$

O comprimento de ligação de referência r_0 é denominado de comprimento de ligação de equilíbrio. As forças entre átomos ligados são muito fortes, comparadas a outras forças relativas às interações entre os átomos. Por isso, utiliza-se uma aproximação harmônica. Vale ressaltar que a Equação 2.2 é uma aproximação para valores em torno de r_0 e, portanto, não mostra o comportamento real do potencial de comprimento de ligação para grandes desvios de r_0. Para situações com o comprimento de ligação longe de r_0, é necessário passar pelo aproximação harmônica e incluir termos de ordem maior, geralmente até $(r-r_0)^4$. Mesmo aumentando o intervalo de validação da Equação 2.1, o polinômio de Taylor ainda tenderá a infinito, quando $r \to \infty$, que não corresponde ao resultado físico. Um potencial que satisfaz as condições descritas é o potencial de Morse [132]:

$$E_{bond}(r) = D(1 - \exp(-\alpha(r-r_0)))^2, \quad (2.3)$$

em que D é a energia de disassociação da ligação e $\alpha = \sqrt{\frac{k}{2D}}$.

A Figura 2.10 mostra como a energia varia em relação ao comprimento da ligação. A linha tracejada mostra o comportamento da energia usando o potencial de Morse (Equação 2.3), que aproxima melhor o comportamento real da energia potencial de ligação, ocorrendo a disassociação (rompimento) da ligação após um dado afastamento do comprimento de ligação ideal. A linha

cheia ilustra a aproximação harmônica a partir da expansão de Taylor [175] para pequenas variações no comprimento da ligação em relação ao valor de referência (sem rompimento da ligação).

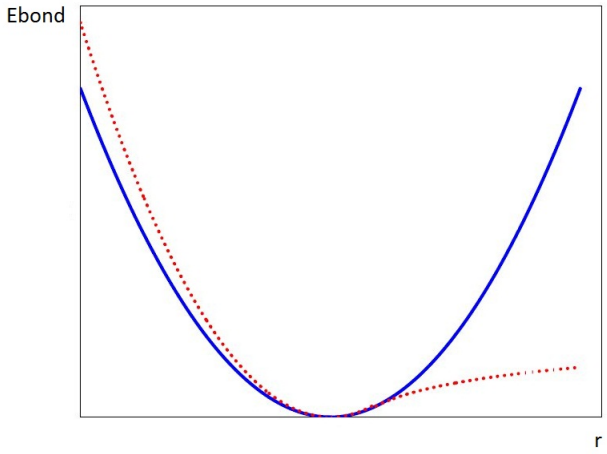

Figura 2.10: Função de energia potencial de comprimento de ligação [154]: Equação 2.2 (linha cheia) e Equação 2.3 (linha tracejada).

Energia de ângulo de ligação

O ângulo de ligação θ resulta da interação de três átomos (A, B e C) com ligações A–B e B–C [154], formando um ângulo θ com vértice em B. A energia necessária para a mudança do ângulo θ, que estabelece um equilíbrio, é significativamente menor do que aquela para a distorção do comprimento de ligação. Assim, as constantes de força de ângulo de ligação são proporcionalmente menores do que as constantes de força de comprimento de ligação. O termo de energia de ângulo de ligação na função objetivo é o somatório de todas as interações de ângulo de ligação da proteína em avaliação. Essa energia pode ser modelada pela Equação 2.4.

$$E_{angle}(\theta) = \frac{1}{2} k_\theta (\theta - \theta_0)^2, \tag{2.4}$$

em que θ_0 é o valor de ângulo típico, k_θ é a constante de força de ângulo de ligação e θ é o ângulo de ligação atual.

Energia de ângulo de torção

Os ângulos de torção são importantes em relação às interações de comprimento de ligação e de ângulos de ligação devido a dois importantes

aspectos [154]. O primeiro é que as barreiras internas de rotação são baixas em relação às outras interações, isto é, as mudanças no ângulos diedrais podem ser grandes. O segundo aspecto é que o potencial de torção é periódico por meio de uma rotação de 360^o.

Em geral, as interações de torção são modelas utilizando uma série de Fourier:

$$E_{tors}(\phi) = \sum_n \frac{1}{2} V_n \cos(n\phi), \tag{2.5}$$

em que n é o número de fases utilizadas, V_n são as constantes de força de rotação de torção e ϕ é o ângulo de torção atual. Geralmente, move-se o zero do potencial e inclui-se fatores de fase obtendo-se a Equação 2.6:

$$E_{tors}(\phi) = \sum_n \frac{1}{2} V_n (1 + \cos(n\phi - \gamma_n)). \tag{2.6}$$

Além disso, os ângulos de fase γ_n são escolhidos de maneira que V_n positivo corresponda a energia mínima em 180^o. A Figura 2.11 mostra as três primeiras fases da Equação 2.6. A linha cheia apresenta o gráfico da equação para $n = 1$, a linha pontilhada para $n = 2$ e a linha tracejada para $n = 3$.

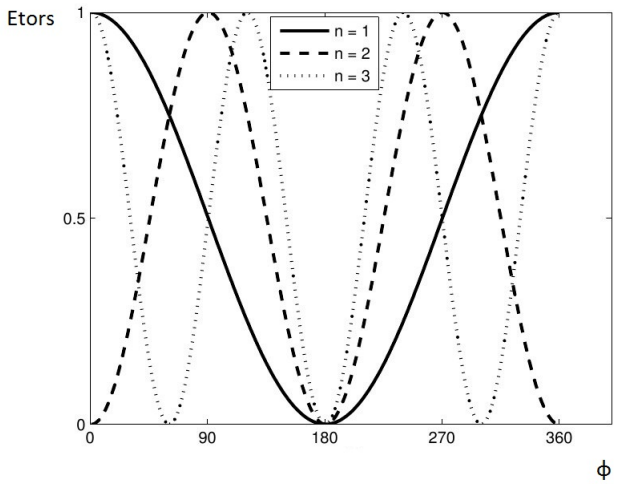

Figura 2.11: Termos da função de energia potencial de torção [154].

Energia Urey-Bradley

O termo de energia Urey-Bradley refere-se às interações entre pares de átomos separados por duas ligações atômicas, conhecida como interação 1:3 átomos [154]. Tais interações são calculadas usando um termo de aproximação harmônica da distância s entre os átomos i e j, como na energia de comprimento de ligação e energia de ângulo de ligação.

A energia de interação Urey-Bradley pode ser descrita como na Equação 2.7:

$$E_{urey}(s) = \frac{1}{2}k_{urey}(s - s_0)^2, \tag{2.7}$$

em que k_{urey} é a constante de força da interação Urey-Bradley e s_0 é a distância ideal entre os átomos i e j.

Energia imprópria

A energia imprópria é um potencial artificial que avalia as deformações dos ângulos de torção referentes a posições relativas entre átomos (configurações) consideradas não adequadas (em termos de sua geometria), mas que podem ocorrer em uma simulação. Por exemplo, a assimetria (quiralidade) de C-α no diedro pode ser mensurada pela energia imprópria [154]. O cálculo da energia imprópria pode ser obtido pelo somatório de todas as interações de energias impróprias da molécula. O cálculo da energia imprópria em geral utiliza uma aproximação harmônica dada pela Equação 2.8:

$$E_{improper}(\omega) = \frac{1}{2}k_{improper}(\omega - \omega_0)^2, \tag{2.8}$$

em que $k_{improper}$ é a constante de força imprópria e ω_0 é o ângulo de torção impróprio ideal.

Energia de van der Waals

A força de van der Waals é fraca em relação a energias relacionadas a ligações covalentes, mas importante na estabilidade de macromoléculas. Por meio da interação de dois átomos, esse potencial realiza o balanceamento entre forças de atração e repulsão [155, 154]. A força de repulsão aparece em curtas distâncias, com a interação elétron-elétron forte. A força de atração surge das flutuações na distribuição de carga da nuvem de elétrons. A flutuação na distribuição de elétrons em um átomo ou molécula gera um dipolo instantâneo que, ao seu redor, origina um dipolo induzido em um segundo átomo ou molécula. Dessa forma, origina-se uma interação atrativa. A interação atrativa apresenta um alcance maior do que a repulsão, mas a medida que a distância torna-se relativamente curta, a interação repulsiva torna-se dominante.

Geralmente, a interação de van der Waals é modelada com o potencial de Leonnard-Jones 6–12, que representa a energia de interação usando constantes A e C dependentes do tipo do átomo. A Equação 2.9 é a forma geral do potencial de Leonnard-Jones, em que $r = \frac{r_{i,j}}{R_i + R_j}$, $r_{i,j}$ indica a distância

Euclidiana entre os átomos i e j, e R_i e R_j correspondem aos raios de van der Waals de cada átomo. A Figura 2.12 ilustra esse potencial.

$$E_{vdw} = \sum_{i,j} \frac{A_{i,j}}{r^{12}} - \frac{C_{i,j}}{r^6}. \tag{2.9}$$

Neste contexto, as interações de van der Waals têm grande importância para a estabilidade de macromoléculas biológicas. Essas interações são calculadas para pares de átomos, em que todos os pares deveriam ser avaliados, à princípio. Nesse caso, a quantidade de interações aumenta com o quadrado do número de átomos para o modelo da Equação 2.9.

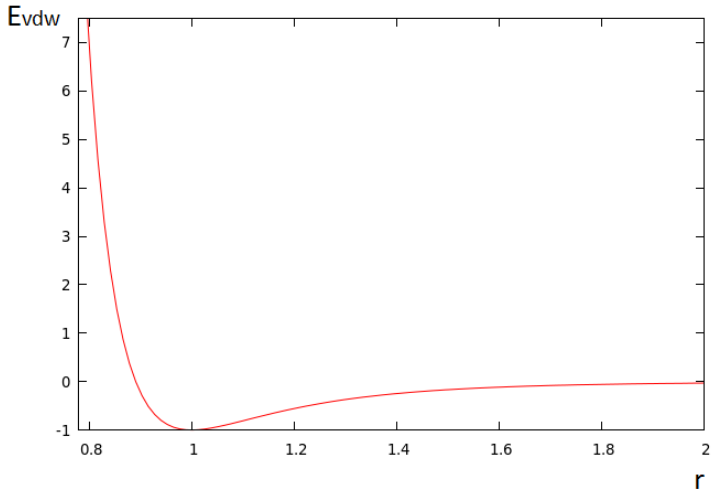

Figura 2.12: Função de van der Waals na forma padrão.

Para evitar que os valores da energia aumentem infinitamente, foi efetuado o corte de diminuição, para $r \geq 0.8$ matendo o valor constante de E_{vdw} para $r = 0.8$, como se pode observar na Figura 2.13. A função de van der Waals com valor de diminuição impede que o potencial de van der Waals cresça até o infinito quando a distância entre os dois átomos for pequena (neste caso, 25% menor do valor da soma dos raios de van der Waals dos átomos).

Em geral, também defini-se um valor de corte de distância entre dois átomos, por exemplo $r_{i,j} > 8$. Nesse caso, o valor da energia de van der Waals é multiplicada a partir de $r_{i,j} > 8$ por um polinômio de grau cinco $p(r) = c_5r^5 + c_4r^4 + c_3r^3 + c_2r^2 + c_1r^1 + c_0r^0$, tal que c_k ($k = 0, 1, \ldots, 5$) são constantes que dependem do valor de corte, a fim de suavizar o efeito da energia nas interações de longas distâncias em direção à energia zero. Na Figura 2.14 a linha verde representa o produto da energia de van de Waals com o polinômio, gerando um aumento significativo nas interações com distâncias maiores qu 8, afetando positivamente nas predições de estruturas protéicas.

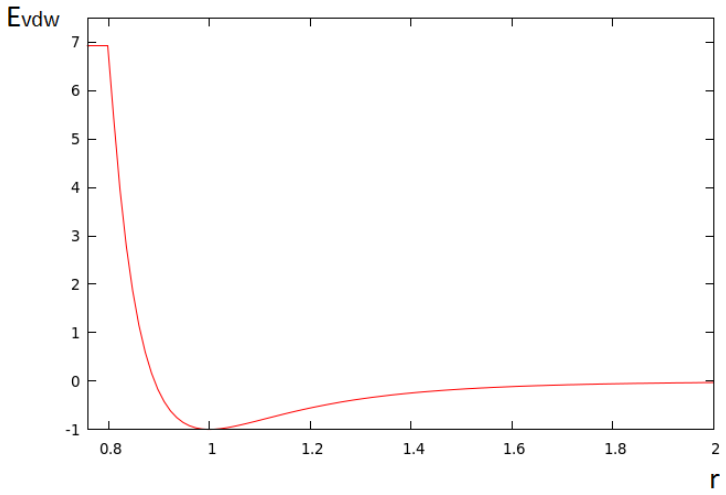

Figura 2.13: Função de van der Waals com corte de diminuição para $r \geq 0.8$.

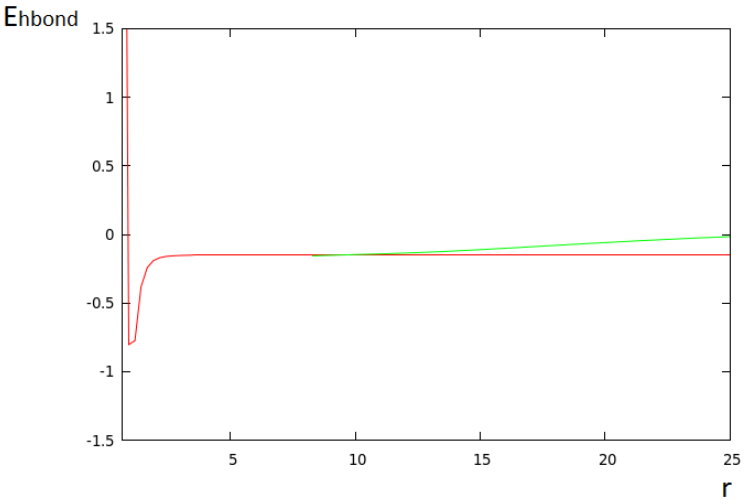

Figura 2.14: Função de van der Waals com polinômio de suavização. A curva em verde destaca o efeito do polinômio de suavização.

Energia eletrostática

A interação eletrostática entre dois átomos é indicada pelo potencial de Coulomb [155, 154]. Considerando que as cargas dos átomos não variam, tem-se que a energia eletrostática muda conforme a distância entre os átomos, conforme mostra a Equação 3.9.

$$E_{charge} = \sum_{i,j} \frac{q_i q_j}{D r_{i,j}}, \tag{2.10}$$

em que q_i e q_j são as cargas dos átomos, $r_{i,j}$ é a distância entre os átomos e D é a constante dielétrica.

Para tratar das interações eletrostáticas de longa distância, tem-se um método *tapering*, baseado na implementação do TINKER [142]. Diversas

simulações biomoleculares [155] têm aplicado este método, como AmberPlus [88]. No método *tapering*, o qual é similar ao método *switching* [168], o potencial eletrostático é dado pela Equação 2.11.

$$E_{charge} = \begin{cases} \frac{q_i q_j}{D r_{i,j}} - \frac{q_i q_j}{D r_c}, & \text{se } r_{i,j} < r_{tap} \\ \sum_{k=0}^{5} c_k r_{i,j}^k (\frac{q_i q_j}{D r_{i,j}} - \frac{q_i q_j}{D r_c}) + \frac{q_i q_j}{D} \sum_{k=0}^{7} f_k r_{i,j}^k, & \text{se } r_{tap} < r_{i,j} < r_{cut}, \end{cases} \tag{2.11}$$

em que $r_{tap}(< r_{cut})$ é uma distância de *tapering*, o início do efeito *tapering* é ajustado em $r_c = \frac{1}{2}(r_{tap} + r_{cut})$ (em geral, defini-se r_c em 13 Å), c_k e f_k são coeficientes calculados a partir de r_{tap} e r_{cut}, determinados para continuar suavemente a função eletrostática nas posições $r_{i,j} = r_{tap}$ e $r_{i,j} = r_{cut}$. Desse modo, assumindo o produto das cargas com valor negativo e variando a distância entre os átomos, obtém-se a curva mostrada na Figura 2.15.

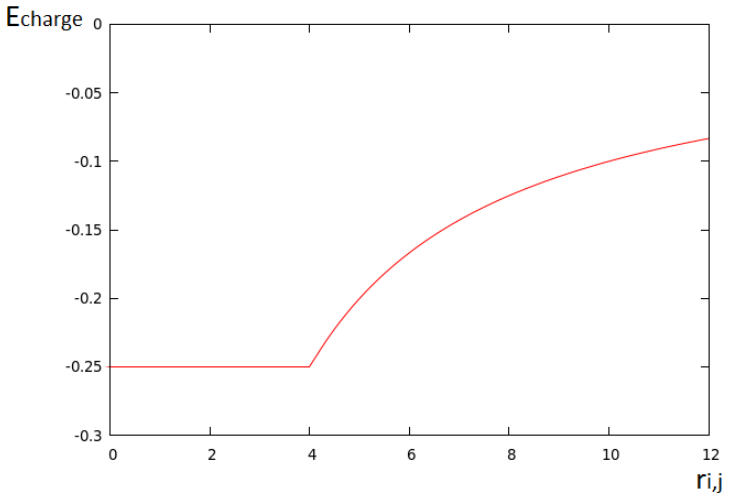

Figura 2.15: Função de energia eletrostática.

2.5 Considerações finais

Este Capítulo descreveu o problema de PSP e, para isso, foi importante apresentar alguns conceitos básicos sobre proteína, tais como sua composição e quais estruturas podem representar sua organização molecular. Essas estruturas podem ser: primária, secundária, terciária e quaternária. O foco deste trabalho está na predição de estruturas secundárias, usando um técnica computacional de otimização, que tenta resolver as questões não resolvidas pelos métodos experimentais tradicionais: Cristalografia de Raio-X e RNM. Devido ao problema de PSP ser considerado um problema NP-completo, diversos métodos computacionais estão sendo implementados por cientistas

da área, como Redes Neurais, Monte Carlo, Algoritmos Evolutivos, entre outros. Neste trabalho, a técnica dos AEs foi considerada mais apropriada, utilizando modelagem *ab initio* com modelo de representação *full-atom*. Os AEs assumem um conjunto de soluções (população) ao final da evolução. Desse modo, ao invés de ter que executar várias vezes o mesmo algoritmo para se obter várias soluções diferentes, como é realizado no programa de predição de estrutura QUARK [190] (que precisa de diversas execuções do Monte Carlo para obter um conjunto de possíveis soluções). Com AE requer apenas uma execução paea gerar diversas predições, sendo adequado para o problema de PSP, que necessita de extrema diversidade de soluções justamente por ser um problema computacionalmente complexo.

Dentre todas as energias apresentadas, os potenciais de van der Waals e de eletrostática são os que mais contribuem em PSP, responsáveis por mais de 60% da estrutura de uma proteína [133, 13, 135]. As demais energias envolvem ligações covalentes (energias Urey-Bradley, de comprimento de ligação, de ângulo de ligação, de ângulo de torsão) ou outros aspectos dessas ligações (energia imprópria). Apesar das energias envolvendo tais ligações serem importantes para estabilidade da molécula, são relativamente pouco afetadas por mudanças nos ângulos diedrais. Assim, para essas ligações podem-se considerar os valores de parâmetros usuais da literatura, não requerendo, à princípio, o cálculo de tais energias no processo de predição.

Outro aspecto importante do AE é a facilidade com que este trabalha com múltiplos critérios (Capítulo 4), o que caracteriza realisticamente o problema de predição, por necessitar de diversas energias (e/ou modelos dessas) para a estabilização das moléculas protéicas. No Capítulo 3 são descritas as modelagens de energias utilizadas neste trabalho, bem como as alterações que foram realizadas.

CAPÍTULO

3

Outras energias e aspectos reformulados considerando AEMOs

3.1 Considerações iniciais

Um dos grandes desafios do problema de PSP puramente *ab initio* é a predição de estruturas com folhas-β ou moléculas mais complexas com mais de um domínio. Para se conseguir modelar essas estruturas é necessário a adição de modelos de energia adequados para esse contexto. Este Capítulo tem como objetivo apresentar duas modelagens adicionais de energias desenvolvidas para o problema de PSP neste trabalho: ligação de hidrogênio e solvatação. A fim de obter melhores predições, sugere-se também a remodelagem de função de energia eletrostática. Na Seção 3.2 são apresentados os modelos de interação da proteína com o meio (solvente), destacando a energia de solvatação, a qual utiliza o cálculo da área de acessibilidade da superfície da molécula ao solvente. A Seção 3.3 mostra as modificações feitas na energia eletrostática, considerando o solvente na constante dielétrica. A Seção 3.4 apresenta o processo de modelagem da energia de ligações de hidrogênio, apresentando as etapas de desenvolvimento.

3.2 A interação da proteína com o meio

Na natureza, as proteínas encontram-se imersas por um solvente, que comumente é a água, tornando imprescindível o estudo da interação da proteína com o solvente [144]. Os principais critérios para avaliação

da interação da proteína com o solvente são: hidrofobicidade, modelos água/solvente, cálculo de energia de solvatação e área de acessibilidade.

A hidrofobicidade é um fator absolutamente relevante na estabilidade de uma proteína. Desse modo, os resíduos hidrofóbicos tendem a se afastar da água, dobrando-se para o interior da molécula; enquanto que os resíduos hidrofílicos são envolvidos por meios aquosos [148, 35]. Portanto, se uma proteína possui resíduos hidrofóbicos nas extremidades das cadeias laterais, muito provavelmente existem dobramentos nessas cadeias, a fim de se afastarem do solvente, ocasionando mudanças em sua organização tridimensional usual. Dessa forma, a hidrofobicidade é uma característica fundamental no problema de PSP [71].

Durante este trabalho, realizou-se pesquisas sobre solventes e como esses podem ser modelados, a partir de duas abordagens: o modelo contínuo (implícito) e o modelo discreto (explícito) (Figura 3.1). O modelo contínuo de solvatação assume que o solvente é um meio dielétrico, em que o soluto situa-se no interior de uma cavidade.

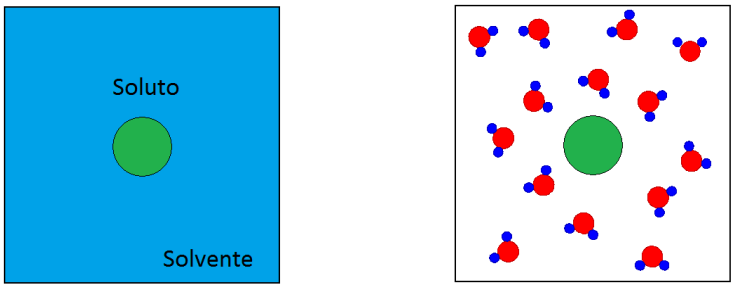

Figura 3.1: Representação do modelo contínuo (à esquerda) e modelo discreto (à direita).

Diversos métodos baseados nessa abordagem [65, 66, 172, 171, 178, 173] podem ser encontrados na literatura. A desvantagem é que os modelos contínuos não descrevem as interações explícitas entre soluto e solvente, particularmente, as ligações de hidrogênio entre a proteína e a água [4]. Nesse contexto, o modelo discreto de solvatação considera as moléculas individuais do solvente, no caso a água, que interagem com o soluto através de métodos clássicos [4] ou quânticos. Portanto, esse modelo tende a solucionar a desvantagem do modelo contínuo em relação às pontes de hidrogênio, descrevendo apropriadamente as interações explícitas entre soluto e solvente, átomo a átomo.

Neste trabalho são usados os modelos implícitos para avaliação da interação da proteína com o meio (por meio do cálculo da energia de solvatação e a área de acessibilidade) uma vez que esses requerem menos

custo computacional. Na Subseção 3.2.1 é explicado o cálculo da área de acessibilidade ao solvente, que foi utilizado neste trabalho.

3.2.1 Área de acessibilidade

Langmuir foi o primeiro a sugerir que a área da superfície molecular está diretamente relacionada aos estudos de solubilidade, como foi constatado em 1925 [113]. Posteriormente, com o avanço das pesquisas nessa área [86], concluiu-se que nem toda a superfície do soluto é acessível ao solvente, dependendo do tamanho da molécula de solvente. Isto pode ser notado pela superfície de cavidade criada no solvente. A cavidade acomoda a molécula de soluto, como se pode observar na Figura 3.2. A fim de calcular essa área de superfície da cavidade, Hermann propôs um modelo no qual a molécula do soluto é representada por um conjunto de esferas com o raio van der Waals, e centros localizados nos centros nucleares de cada átomo dessa molécula. Sendo assim, uma esfera com o raio de van der Waals, referente à molécula de solvente, rola sobre a molécula do soluto. A área da superfície determinada por esse rolamento é chamada de área de cavidade ou área de superfície acessível ao solvente (SASA, do inglês *Solvent Accessible Surface Area*). Diversas pesquisas destacam a importância da área de superfície molecular em relação às características físico-químicas das proteínas [86, 183, 5, 56, 57, 33].

O algoritmo para o cálculo da área de superfície acessível ao solvente (A_{acc}) e para o cálculo da área de superfície de van der Waals (A_{vdw}) (Figura 3.2) foi baseado no trabalho de Gaudio e Takahata [74]. A princípio, Higo e Go (1989) apresentaram um procedimento para o cálculo da área de superfície de macromoléculas, no entanto, este era inadequado para moléculas pequenas. O trabalho de Gaudio e Takahata resolveu, desta maneira, esse problema particular de moléculas pequenas, médias e grandes. Posteriormente, outros algoritmos para o cálculo da área de superfície molecular foram desenvolvidos [86, 89, 106, 163, 145]. Contudo, em relação à simplicidade, o trabalho de Gaudio e Takahata é o que sintetiza melhor a idéia, em uma implementação direta e descomplicada.

Para iniciar o cálculo, precisa-se de uma caixa retangular de maneira que esta enquadre a molécula em questão (Figura 3.3). O comprimento de cada aresta deve ser um múltiplo de 2 Å. Desse modo, cada aresta é dividida em 2 Å, a fim de se obter 2K cubos, K sendo uma constante qualquer. Estes são chamados cubos nível 1. Cada cubo é analisado e classificado em classes: interna, externa ou superfície para molécula. Se o cubo for interno ou externo não é considerado, caso contrário, será dividido em 8 novos cubos, cada um com aresta de 1 Å, e pertencentes ao nível 2. Os cubos desse nível são

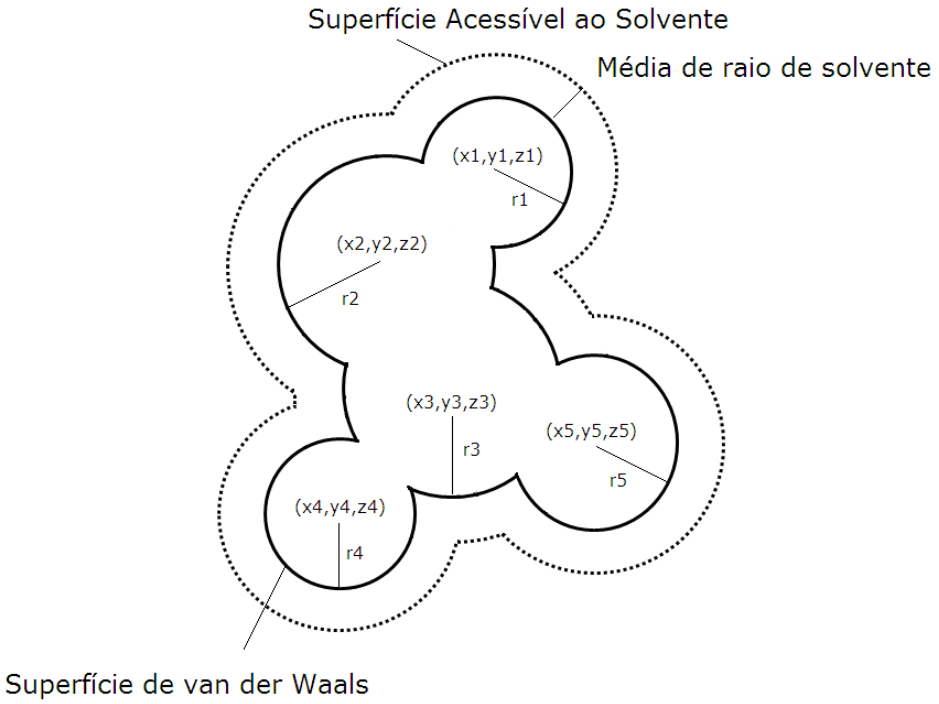

Figura 3.2: Área de superfície de van der Waals (A_{vdw}) e área de superfície acessível ao solvente (A_{acc}) [74].

classificados iguais aos de nível 1. Os cubos da superfície serão novamente divididos em 8 novos cubos com aresta de 0.5 Å, pertencentes ao nível 3 (Figura 3.3).

A cada passo que se avança um nível, os cubos diminuem, delimitando a área de superfície molecular com maior precisão. Para níveis menores que 5, os cubos são grandes demais para descrever a superfície molecular. Com esse cálculo, os valores aproximam-se dos reais. Quanto maior o nível, menor os cubos das superfícies, tal que esses cubos começam a particionar a área da superfície molecular que os intercepta (Figura 3.4).

A média da área que intercepta ($A_{intercept}$) é linearmente proporcional à área de uma face da superfície do cubo ($A_{cubeside}$) [74].

$$A_{intercept} = f A_{cubeside}, \tag{3.1}$$

em que f é o fator de proporcionalidade, determinado com alguns átomos típicos de cada nível (5, 6 e 7). A média de interseção depende do raio da esfera em questão. Portanto, é preciso apresentar um fator numérico para cada tipo de esfera. A área da superfície total (A_{calc}) pode ser calculada como a soma de todas as frações computadas ($A_{intercept}$) como mostrada na Equação 3.2.

$$A_{calc} = \sum A_{intecept}(i), i = 1 \ldots Nc, \tag{3.2}$$

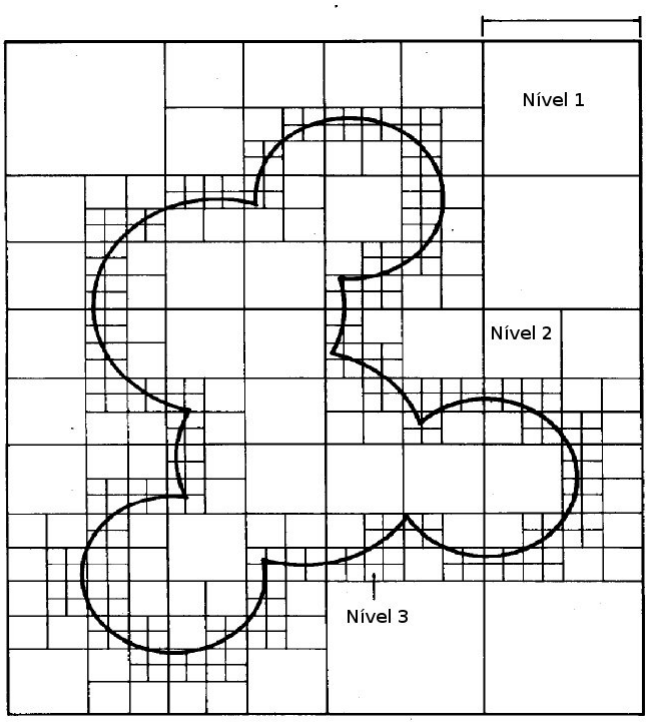

Figura 3.3: O procedimento adotado para computar a área da superfície molecular. A molécula é acomodada em uma caixa retangular, a qual é dividida em cubos de 2 Å, no nível 1. A aresta do cubo é dividida pela metade quando se avança cada nível [74].

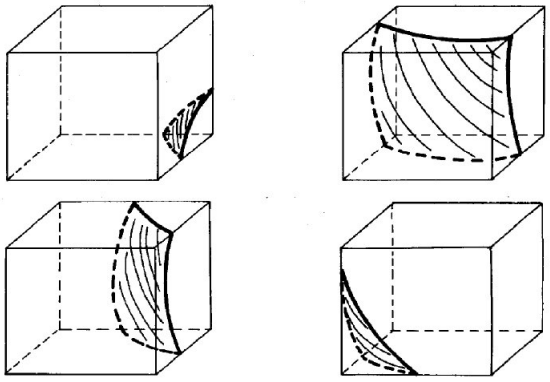

Figura 3.4: Alguns exemplos da interseção entre a superfície molecular e os cubos [74].

em que Nc é o número de cubos que interceptam a superfície molecular. Substituindo a Equação 3.1 na Equação 3.2, temos a Equação 3.3. A Equação 3.4 é a consequência do fato que todos os valores, $A_{cubeside}$, são os mesmos na Equação 3.3.

$$A_{calc} = f \sum A_{cubeside}(i), i = 1 \ldots Nc, \tag{3.3}$$

$$A_{calc} = fNcA_{cubeside}. \tag{3.4}$$

Os fatores para cada tipo de esfera são definidos a partir dos seguintes passos: construindo-se um sistema composto de duas esferas iguais, isoladas e não sobrepostas, localizadas sobre coordenadas fixas, com cada raio correspondendo ao raio de van der Waals de um determinado átomo (por exemplo, para H tem-se r = 1.2 Å). Esse sistema tem uma área de superfície conhecida (A_{true}), cujo valor real é $8\pi r^2$.

Uma vez que a área da superfície real (A_{true}) é igual a A_{calc} na Equação 3.4, segue que, da Equação 3.4 obtém-se Equação 3.5.

$$A_{true} \approx fNcA_{cubeside}, \tag{3.5}$$

f é calculado com a Equação 3.5, isto é,

$$f = \frac{A_{true}}{NcA_{cubeside}} = \frac{8\pi r^2}{Ncl^2}, \tag{3.6}$$

tal que l é o tamanho da aresta do cubo. Posto que no nível 1, o comprimento da aresta do cubo é l = 2 Å, no nível 2, l = 1 Å, no nível 3, l = 0.5 Å, e assim sucessivamente, no maior nível L tem-se que $l = 2^{(2-L)}$ Å. Essa relação pode ser, assim, deduzida como na Equação 3.6:

$$f = \frac{8\pi r^2}{Nc2^{(2-L)^2}}. \tag{3.7}$$

A Equação 3.7 é usada para obtenção do fator de proporcionalidade para todos os átomos para os níveis 5, 6 e 7. Os raios de van der Waals, em Angstron (Å), foram H = 1.2, C_{aliph} = 1.6, C_{arom} = 1.7, N = 1.5, O = 1.4, F = 1.35, Cl = 1.8, Br = 1.95 e I = 2.15 [183, 139]. Como os níveis 6 e 7 não aumentam significativamente a precisão do cálculo de área de acessibilidade, os cálculos são limitados ao nível 5. Os fatores foram determinados por esses mesmos átomos com os raios deles adicionados a 1.5 Å, a fim de calcular A_{acc} e A_{vdw}. Esse valor foi determinado por experimentos.

Há diversos pacotes de modelagem molecular com funções para o cálculo da área de superfície de acessibilidade ao solvente estão disponíveis, podendo-se citar: CHARMM [30, 125], TINKER [142], GROMACS [87], Naccess [94] e STRIDE [70]. Neste trabalho, foram usados os parâmetros de campos de força do CHARMM e as funções de energia baseadas no TINKER [72, 122]. Para o cálculo da área SASA, adaptou-se o código do STRIDE em nossa implementação.

3.2.2 Energia de solvatação

Para desenvolver uma descrição atômica explicitamente da iteração da água com a proteína foi estendida a idéia de Langmuir, Cohn e Edsall, e outros. A ideia básica é que a energia livre da interação de um soluto com água pode ser considerada como uma soma de energias de grupos de átomos. Langmuir usa a área da superfície exposta de um grupo como medida de sua interação com o solvente. Em nosso método, usamos a área de superfície de acessibilidade de Lee e Richards e outros pesquisadores, que foi explicada na seção anterior. É definida como a área a qual o centro de uma molécula de água de raio 1.4 Å pode mover-se em contato com a área livre da molécula. Em nosso método, o cálculo é baseado pelo parâmetro atômico de solvatação (ASP, do inglês *Atomic Solvation Parameters*) de cada átomo acessível à água [117]. Esses valores são determinados a partir da energia livre de transferência.

A proposta é multiplicar os parâmetros de cada átomo (C, N, O) com a SASA (área da superfície de acessibilidade ao solvente) correspondente. Podemos ignorar o átomo de Hidrogênio nesses cálculos, pois são quase inexpressivos (muito pequenos), sendo menores que os erros. Calcula-se, portanto, as SASAs dos átomos da cadeia principal da proteína. Os testes usaram dois artigos com ASPs baseados no trabalho de [186, 60]. A ideia é comparar a influência desses parâmetros na predição das estruturas terciárias de proteínas simples, compostas de duas ou mais hélices-α. Calcula-se as SASAs de cada átomo da cadeia principal, e multiplica-se por seus ASPs correspondentes. Com os testes, pode-se perceber que os parâmetros ASPs baseados em (Eisenberg e MachLachlan, 1986)[60] apresentaram melhores resultados. Neste artigo, mostram-se os seguintes ASPs: $ASP(C) = 16 \pm 2cal/A^2mol$; $ASP(N/O) = -6 \pm 4cal/A^2mol$.

Pode-se notar que os ASPs são negativos, excetos do Carbono. Isso significa que a hidratação é favorável aos átomos de Nitrogênio e Oxigênio, e desfavorável aos átomos de Carbono. A energia livre de solvatação é calculada nestes experimentos com a Equação 3.8:

$$E(S) = \sum_{i=1}^{n} ASP(a_i)SASA(a_i - a_r), \tag{3.8}$$

tal que n é o número de átomos da molécula, a_r é a área de referência de cada átomo, relativa à acessibilidade estática, e a_i é a área da superfície acessível ao solvente de um átomo no estado dobrado. Os valores dessas áreas foram obtidos do trabalho de Lee e Richards [117]. Desse modo, calculamos de fato a energia necessária para ir de um estado sem solvente para outro com solvente (sendo assim, a energia de solvatação é a energia requerida para

transferir uma proteína do vácuo para um solvente)[60]. O parâmetro atômico de solvatação depende do tipo de átomo (N, C e O), e sua unidade de medida é $cal/molA^2$.

3.3 Remodelagem na energia eletrostática

Foi sugerida uma modificação na constante dielétrica da energia eletrostática, usando um parâmetro relacionado à área de acessibidade ao solvente, a fim de lidar com a interação da proteína com o solvente. De acordo com a Equação 3.9, D é a constante dielétrica (a qual representa a influência do meio).

$$E_{charge} = \sum_{i,j} \frac{q_i q_j}{D r_{i,j}}. \tag{3.9}$$

Esse parâmetro relacionado à área de acessibidade ao solvente é calculado pela Equação 3.10:

$$\frac{A_{max} - A_{min}}{1 - 80} = \frac{A_{max} - A_i}{1 - D}, \tag{3.10}$$

em que:

- A_{max} é a área máxima de acessibilidade ao solvente que uma determinada proteína pode alcançar. É obtida a partir da população inicial, onde os indivíduos ainda não têm evolução alguma;
- A_{min} é a área mínima de acessibilidade ao solvente que uma determinada proteína apresenta. A estimativa para A_{min} será descrita a seguir;
- A_i é a área de acessibilidade ao solvente da proteína na interação i;
- A constante dielétrica é 1 quando a proteína está em contato com elementos dela mesma;
- A constante dielétrica é 80 quando a proteína está imersa na água;
- D é a variável do dielétrico relacionada à área de acessibilidade ao solvente.

Desenvolvendo a Equação 3.10, tem-se:

$$D = 1 + \frac{(A_{max} - A_i)(80 - 1)}{A_{max} - A_{min}}. \tag{3.11}$$

Para estimar a área mínima de acessibilidade ao solvente, foi proposta uma configuração mínima para a posição dos aminoácidos de uma proteína. Suponha que a menor configuração de uma proteína é aquela em que os

aminoácidos estão dispostos de acordo com o cubo da Figura 3.5, em que cada sub-cubo representa um aminoácido. Desse modo, os aminoácidos estão o mais próximo possível, sem que haja sobreposição entre eles.

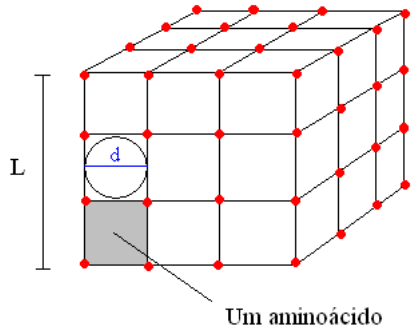

Figura 3.5: Conformação compactada dos aminoácidos de uma proteína.

Seja N_{aa} o número de aminoácidos que constituem a proteína. Segue que para cada lado (L) do cubo existem $N_{aa}^{1/3}$ aminoácidos. Assim, a área de um cubo é calculada pela seguinte equação:

$$A = 6L^2 = 6(dN_{aa}^{1/3})^2, \tag{3.12}$$

em que assume-se que $d \approx 5$ Å.

Portanto, tem-se uma equação para estimar a área mínima de uma proteína, baseada somente no número de aminoácidos. Essa proposta foi sugerida para que os cálculos fossem realizados de tal maneira que dependessem apenas da variável da quantidade dos aminoácidos, sem necessitar de outras simulações para estimar a área mínima de cada proteína.

Esse parâmetro substitui $D = 1$, a constante dielétrica (a qual representa a influência do meio, que se for a água, o valor é 80). Pela Equação 3.11, pode-se observar que a energia eletrostática considera a área de superfície acessível ao solvente em seu cálculo, mostrando que quanto menor a área, maior o valor do dielétrico considerando o meio e a área acessível a esse meio. Quanto maior o valor do dielétrico, menor o valor da energia eletrostática.

Portanto, conclui-se que com menor área acessível ao solvente, tem-se menor energia (satisfazendo a necessidade de minimizar a energia, visando a estabilidade da molécula). Nos trabalhos futuros (Capítulo 7), a intenção é aperfeiçoar a modelagem desse parâmetro que modela a interação da proteína com o meio, buscando obter melhores resultados nas predições, em especial, para proteínas globulares.

3.4 Modelagem da energia das ligações de hidrogênio

As ligações de hidrogênio são classificadas como interações eletrostáticas [13, 50]. As ligações de hidrogênio são, apesar de relativamente fracas, essenciais em macromoléculas, tais como DNA e proteínas. Essas interações são também responsáveis pelas propriedades específicas da água, o que a torna o solvente universal. O átomo de Hidrogênio numa ligação de hidrogênio (H) é parcialmente compartilhado entre dois átomos relativamente eletronegativos, como o nitrogênio (N) ou o oxigênio (O). O doador da ligação é o grupo que inclue ambos os átomos para o qual H é mais fortemente ligado e o próprio átomo H em si, enquanto que o receptor é o átomo menos fortemente ligado ao átomo H (Figure 3.6).

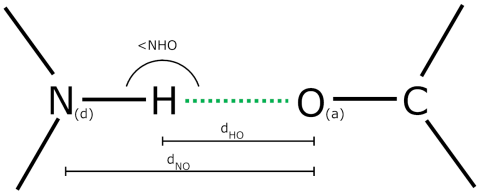

Figura 3.6: Representação da ligação de hidrogênio.

Normalmente, a interação entre as moléculas de O–H...O é considerada como o modelo de todas as ligações de hidrogênio. Assim, sugere-se que uma ligação de hidrogênio, descrita por: X–H...A, é formada por grupos fortemente polares X^{δ}–$H^{\delta+}$ de um lado, e átomos $A^{\delta+}$ do outro (X=O, N *halogens*; A=O, N, S, *halide*,etc.). No entanto, atualmente a ligação de hidrogênio não se resume a essa simples definição, pois foi constatado que possui regiões de transição contínuas para efeitos diferentes tais como a ligação covalente, o iônico puramente e as interações de van der Waals [73, 166]. Portanto, para um grupo X–H ser capaz de formar ligações de hidrogênio, X não precisa ser "muito eletronegativo", é somente necessário que X–H seja, no mínimo, levemente polar.

Tais ligações apresentam energias no intervalo $1 - 4kcalmol^{-1}$ [73, 62]. Ligações de hidrogênio são também mais longas que ligações covalentes; as distâncias das ligações dessas (medidas a partir do átomo H) alcançam desde 1.5 até 2.6 Å; enquanto que distâncias que estão no intervalo de 2.0 - 3.5 (Å) separam os dois átomos diferentes dos H em uma ligação de hidrogênio (ou seja, N e O). As mais fortes ligações de hidrogênio têm uma tendência de serem quase uma reta (situação ideal), aproximando o ângulo $N\widehat{H}O$ de $180°$, como aquela existente entre o doador da ligação (N), o átomo H e o receptor da ligação(O), conforme Figure 3.6 [13].

Pode-se calcular a energia das ligações de hidrogênio usando modelos Coulombicos, por meio da modificação do potencial 10–12 Lennard-Jones, como mostrado na Equação 3.13 [69].

$$E_{hbond} = \sum_{i,j} 5\frac{A_{i,j}}{r^{12}} - 6\frac{C_{i,j}}{r^{10}}. \tag{3.13}$$

3.4.1 Melhorias realizadas na energia das ligações de hidrogênio

Sobre a energia das pontes de hidrogênio, o interesse dessa pesquisa é focar nas ligações no interior das folhas-β antiparalelas. O desenvolvimento dessa energia foi baseada na função de van der Waals, posto que a implementação é semelhante a esta. No entanto, para pontes de hidrogênio há alguns detalhes que fazem toda diferença entre essas energias. Primeiramente, procura-se pelo Hidrogênio que está ligado no Nitrogênio da cadeia principal. Armazena-se, então, as coordenadas destes átomos assim que encontrados. A seguir, calcula-se a distância entre eles (N e H). Um ponto importante é que as ligações de hidrogênio atuam entre uma fita e outra, por isso, faz-se necessário estimar onde a fita antiparalela começa, como se pode observar na Figura 3.7. É importante destacar que em nossa abordagem, trabalhamos somente com as folhas antiparalelas.

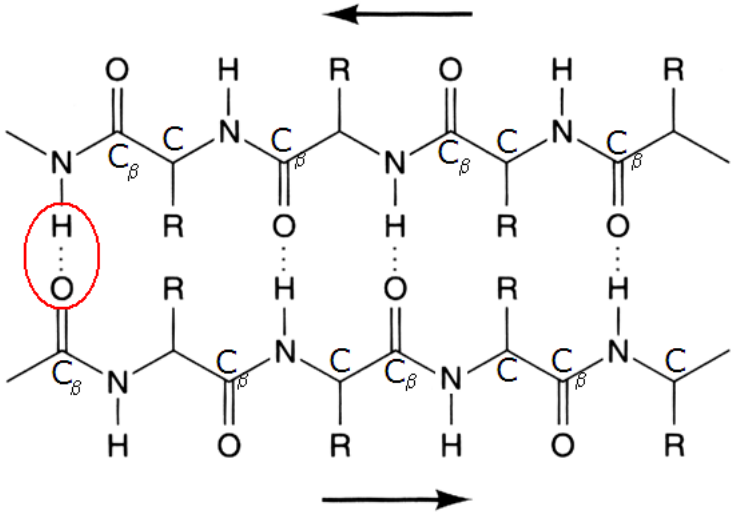

Figura 3.7: Ligação de hidrogênio na folha-β antiparalela.

O modo para estimarmos foi "saltar" cinco átomos de Nitrogênio da cadeia principal. Como nesta etapa inicial, os testes foram feitos com pequenas proteínas, foi uma maneira conveniente de estimação. A partir deste ponto da sequência de átomos, procura-se pelo O ligado ao C_β. Armazena-se as coordenadas do O, calculando-se a distância entre O e H.

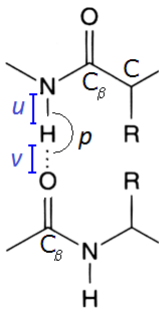

Figura 3.8: Modelagem da ligação de hidrogênio na folha-β antiparalela.

Então, calcula-se o cosseno do ângulo α formado pelos vetores (N–H) e (H...O), sendo chamados u e v, de acordo com a Figura 3.8. Para tal, utiliza-se o produto escalar $u.v$ e os módulos $|u|$ e $|v|$ com as Equações 3.14, 3.15 e 3.16:

$$u.v = ((dx_{NH})(dx_{HO})) + ((dy_{NH})(dy_{HO})) + ((dz_{NH})(dz_{HO})), \tag{3.14}$$

$$|u| = \sqrt{((dx_{NH})^2 + (dy_{NH})^2 + (dz_{NH})^2)}, \tag{3.15}$$

$$|v| = \sqrt{((dx_{OH})^2 + (dy_{OH})^2 + (dz_{OH})^2)}, \tag{3.16}$$

em que di_{aa} é a distância entre os átomos aa (aa = NH e OH) e i corresponde às coordenadas cartesianas x, y e z.

Desse modo, calcula-se o cosseno com a Equação 3.17.

$$\cos(\alpha) = \frac{u.v}{|u|.|v|}. \tag{3.17}$$

O ângulo ideal seria $180°$, entre N–H...O. Isso é quase impossível em sistemas dinâmicos em fase condensada. Talvez apenas no vácuo e sem repulsão estérica (ou seja, quando se tem átomos muito perto um do outro, deformando o peptídeo e favorecendo uma ligação do hidrogênio direcional). Logo, prefere-se os ângulos mais próximos do ideal. Para tal, foi proposta a multiplicação de E_{hbond} pelo termo E_α, quando E_{hbond} fosse negativa, e por $\frac{1}{E_\alpha}$ quando fosse positiva, em que $E_\alpha = \cos^2(\alpha)$ e α é o ângulo entre o Hidrogênio ligado ao Nitrogênio da cadeia principal e o Oxigênio ligado ao Carbono α, de acordo com a Equação 3.18. O ângulo pode ser maior que $135°$ e menor que $225°$. Com o termo E_α, estão sendo priorizados os ângulos mais próximos a $180°$.

$$E_{hbond} = \begin{cases} E_{hbond}E_{\alpha} & E_{hbond} < 0, \\ E_{hbond}\frac{1}{E_{\alpha}} & E_{hbond} \geq 0. \end{cases} \tag{3.18}$$

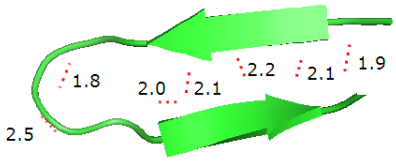

Figura 3.9: Proteína 1NIZ nativa com as ligações de hidrogênio.

Além disso, a Equação 3.13 (baseada em [69]) é alterada visando que a função privilegie os pontos cuja distância entre os átomos de H e O esteja mais próxima de 2 Å. Como mostrado na Figura 3.9, é em torno dessa distância que a ligação de hidrogênio tem maior estabilidade. Nesse sentido, foi realizada uma sequência de alterações (Equações 3.13, 3.19 e 3.20, até determinar a Equação 3.21, que privilegia os ligações formadas num intervalo 2–4 Å). Tal efeito é comprovado experimentalmente, conforme ilustra a Figura 3.10.

$$E_{hbond} = \sum_{i,j} 5\frac{A_{i,j}}{r^4} - 7\frac{C_{i,j}}{r^2}, \tag{3.19}$$

$$E_{hbond} = \sum_{i,j} 5\frac{A_{i,j}}{r^{12}} - 7\frac{C_{i,j}}{r}, \tag{3.20}$$

$$E_{hbond} = \sum_{i,j} 5\frac{A_{i,j}}{(r-1)^{12}} - 7\frac{C_{i,j}}{r-1}. \tag{3.21}$$

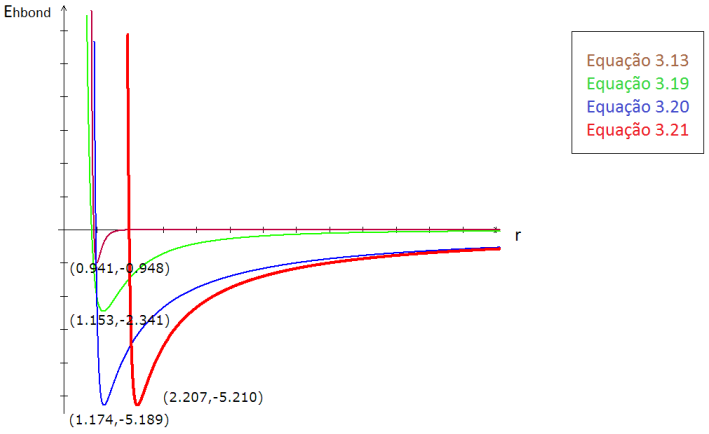

Figura 3.10: Funções experimentais para ligações de hidrogênio em folhas-β antiparalelas.

3.4.2 Remodelagem da energia das ligações de hidrogênio

Resultados obtidos em testes preliminares mostraram a necessidade de um modelo melhor para ligação de hidrogênio. Frishman e Argos [70] também verificaram tal necessidade e modelaram a energia de ligação de hidrogênio com três componentes como mostra a Equação 3.22.

$$E_{hb} = E_p E_t E_r, \tag{3.22}$$

em que a componente E_p está relacionada com o ângulo p formado entre a ligação do H e N da cadeia principal e a ligação do O com o C_α, e pode ser representada pela Equação 3.23.

$$E_p = \cos^2 p. \tag{3.23}$$

A componente E_t está relacionada com os ângulos t_0 e t_i, tal que t_0 é o ângulo formado entre os planos NHO e $CC_\alpha O$. O ângulo t_i é o suplementar do ângulo $C\widehat{O}H'$, em que H' é a projeção ortogonal de H no plano $CC_\alpha O$, $K_1 = 0.9\cos^6 110°$ e $K_2 = \cos^2 110°$. Essa componente é representada pela Equação 3.24.

$$E_t = \begin{cases} (0.9 + 0.1\sin 2t_i)\cos t_0, & 0 < t_i < 90° \\ K_1(K_2 - \cos^2 t_i)^3 \cos t_0, & 90° < t_i < 110° \\ 0, & t_i > 110°. \end{cases} \tag{3.24}$$

A componente E_r pode ser representada pela função $f(r)$, com a Equação 3.25, ilustrada na Figura 3.11:

$$f(r) = -2.8(4((3^6)/(r^6)) - 3((3^8)/(r^8))). \tag{3.25}$$

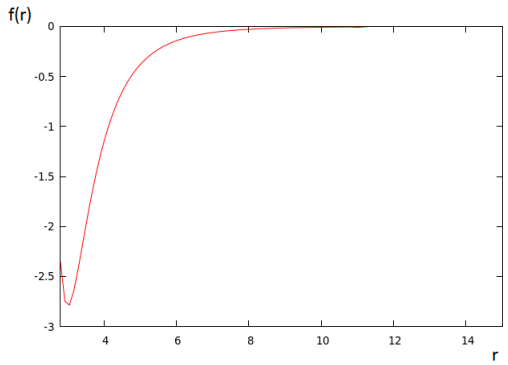

Figura 3.11: Modelo de energia de ligação de Frishman e Argos.

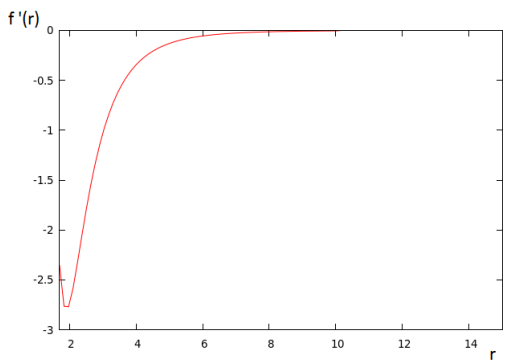

Figura 3.12: Modelo de Frishman e Argos transladado para mínimo 1.9 Å.

Note na Figura 3.11 que o mínimo de energia no artigo de Frishman e Argos encontra-se com aproximadamente 3 Å. Na proteína 1NIZ (Figura 3.9) e

em geral em ligações de hidrogênio em folhas-β, as ligações são formadas em sua maioria entre 1.9 e 2.2 Å. Logo, foi preciso transladar a função para que o ponto mínimo se encontrasse em 1.9 Å (Figura 3.12), obtendo a Equação 3.26.

$$f'(r) = -2.8(4((3^6)/((r+1.1)^6)) - 3((3^8)/((r+1.1)^8))). \tag{3.26}$$

Por meio de experimentos, observou-se que esse modelo limitava a exploração do espaço de busca por um algoritmo de otimização na formação de folhas-β, uma vez que $f'(r)$ é praticamente zero para átomos distantes mais que 8 Å. Esse modelo foi modificado para ampliar o raio de alcance de ligações de hidrogênio, de forma que o algoritmo de otimização pudesse privilegiar modificações em direção à formação de ligações de hidrogênio. Como se pode observar na Figura 3.13, o modelo adaptado tem mínimo com raio aproximadamente em 3.6 Å, com a Equação 3.27. Para se aproximar da distância encontrada nas ligações de hidrogênio, transladou-se a função, obtendo o mínimo em 2.5 Å (Figura 3.14), conforme a Equação 3.28.

$$f_{adap}(r) = -2.8(3.4((3^3)/((r^3)) - 3((3^6)/(r^6))), \tag{3.27}$$

$$f'_{adap}(r) = -2.8(3.4((3^3)/((r+1.1)^3)) - 3((3^6)/((r+1.1)^6))). \tag{3.28}$$

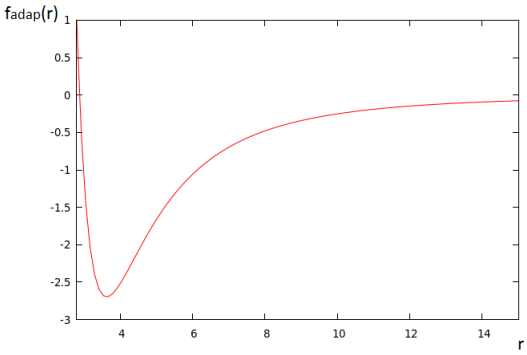

Figura 3.13: Modelo adaptado de Frishman e Argos, com mínimo em 3.6 Å.

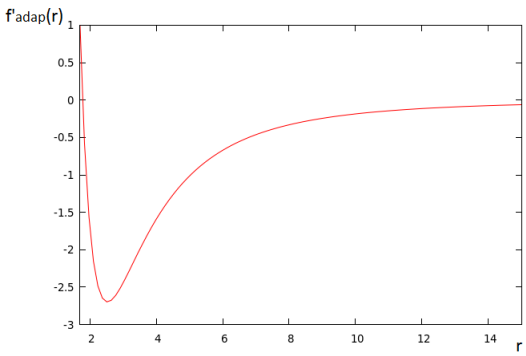

Figura 3.14: Modelo adaptado transladado com mínimo em 2.5 Å.

Pretendendo preservar tanto efeitos de curto quanto longo alcance no modelo de ligação de hidrogênio, realizou-se a soma do modelo original transladado (Figura 3.12) ao modelo adaptado transladado (Figura 3.14), gerando um novo modelo mais abrangente, como mínimo 2.2 Å (Figura 3.15) conforme mostra a Equação 3.29. A partir desse modelo, pode-se obter valores mínimos de energia de aproximadamente -1.85 kcal/mol, em torno do raio 2.2 Å.

$$fs(r) = f'(r) + f'_{adap}(r). \tag{3.29}$$

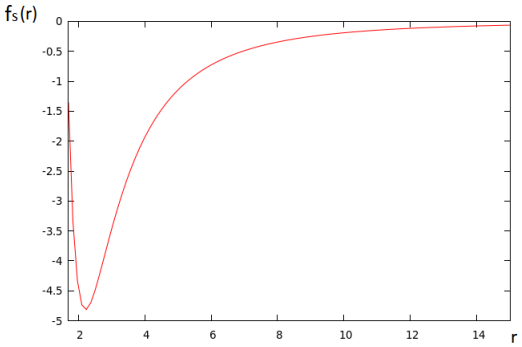

Figura 3.15: Modelo proposto de energia de ligação de hidrogênio para PSP.

Os resultados obtidos a partir da remodelagem e refinamento da energia de ligação de hidrogênio podem ser verificados na Seção 5.5.

3.5 Considerações finais

A predição de estruturas com folhas-β com métodos puramente *ab initio* é uma das motivações para o desenvolvimento de modelos computacionalmente eficientes para energias de ligação de hidrogênio e solvatação. Nesta pesquisa, a energia de solvatação foi desenvolvida, assumindo o solvente como meio dielétrico. A partir de um modelo implícito, a energia de solvatação é calculada, considerando a superfície acessível ao solvente (SASA).

As ligações de hidrogênio foram modeladas por meio da modificação do potencial 10–12 Lennard-Jones. As modificações nesse potencial buscam privilegiar os pontos com distâncias menores entre os átomos de H e O. Não sendo suficiente tais mudanças para melhorias significativas nas predições, houve a necessidade do desenvolvimento de um novo modelo para ligação de hidrogênio, baseado no trabalho de Frishman e Argos, que trabalha com três componentes de energia, melhorando significativamente as soluções preditas.

Nesse contexto, vale ressaltar que modelar interações moleculares é uma tarefa complexa, requerendo um estudo profundo na área de Biologia Molecular, como foi realizado neste trabalho para as ligações de hidrogênio. Essa modelagem envolveu três etapas, a princípio: entendimento do modelo, desenvolvimento do modelo e validação do mesmo, que foram realizadas de maneira incremental, ou seja, a partir dos resultados experimentais em PSP (usando os AEs desenvolvidos neste trabalho) modelo foi gradualmente melhorado.

Ao considerar as duas energias que mais contribuem para definição da estrutura protéica (van der Waals e eletrostática [133, 13, 135]) e as energias de ligação de hidrogênio e solvatação pode-se ter um problema com quatro objetivos. Logo, para o problema de PSP necessita-se de um algoritmo que seja

capaz de trabalhar eficientemente com dois ou mais objetivos, uma vez que cada objetivo representa um Hamiltoniano a ser considerado. Para estruturas mais complexas, com folhas-β ou mais de um domínio, supõe-se precisar de mais de duas energias envolvidas, uma vez que quanto mais complexa a estrutura, mais critérios (Hamiltonianos modelando aspectos adicionais das interações) precisam ser considerados. No Capítulo 4 são descritos algoritmos evolutivos adequados para problemas multiobjetivos, que tratam de dois ou mais critérios simultaneamente.

CAPÍTULO

4

Otimização Multiobjetivo

4.1 Considerações Iniciais

Este Capítulo descreve o problema de otimização multiobjetivo, apresentando as características gerais na Seção 4.2. A seguir, a Seção 4.3 mostra um histórico dos principais algoritmos evolutivos multiobjetivo, com suas respectivas descrições. Os algoritmos descritos neste Capítulo são: NSGA [167], NSGA-II[47], SPEA [196], SPEA2 [195], MOEA-D [192] e AEMT [55, 152]. Na Seção 4.4 são apresentadas as características que definem um problema de otimização com muitos objetivos. Na Seção 4.5 são brevemente discutidas algumas investigações de algoritmos evolutivos multiobjetivo no contexto de PSP.

4.2 Descrição de um problema multiobjetivo

O problema de PSP é caracterizado como problema de otimização multiobjetivo ou MOOP (do inglês, *MultiObjective Optimization Problem*), pois pode ser definido por duas ou mais funções objetivo a serem otimizadas simultaneamente (maximizadas ou minimizadas) [46]. De fato, o conjunto de energias envolvidas na processo de predição de uma proteína possue mais de dois campos de força (Seção2.4.1), que precisam ser minimizados de forma que possam combinar o efeito dos mesmos [44]. Portanto, o problema de PSP envolve naturalmente múltiplos critérios. As funções objetivo aplicadas nos MOOPs são, geralmente, conflitantes entre si. Uma função objetivo f_1 é conflitante com uma outra função f_2 quando não há garantia de melhorar

o valor de uma função f_1 e também melhorar outra f_2. Isso ocorre com as energias consideradas em um proteína, uma vez que minimizar a energia de van der Waals não implica em também minimizar a energia eletrostática, por exemplo.

O MOOP possui também restrições para que uma solução seja factível para o problema. O enunciado geral de um MOOP é o seguinte [46]:

$$\left.\begin{array}{rll} \text{maximizar/minimizar} & f_m(\mathbf{x}), & m = 1, 2, \ldots, N_{obj} \\ \text{restrita a} & g_j(\mathbf{x}) \geq 0, & j = 1, 2, \ldots, NR_{des}; \\ & h_k(\mathbf{x}) = 0, & k = 1, 2, \ldots, NR_{igu}; \\ & x_i^{(inf)} \leq x_i \leq x_i^{(sup)}, & i = 1, 2, \ldots, N_{var}. \end{array}\right\} \quad (4.1)$$

em que x é um vetor de N_{var} variáveis de decisão $x = (x_1, x_2, ..., x_{Nvar})^T$, denominado solução. Os valores x_i^{inf} e x_i^{sup} representam os limites inferior e superior, respectivamente, para a variável x_i. O espaço de variáveis de decisão ou espaço de decisão S_{dec} é limitado por esses valores. Existem as funções de restrição para as desigualdades ($g_j(\mathbf{x}) \geq 0$) e para as igualdades ($h_k(\mathbf{x}) = 0$). Uma solução x factível é aquela que satisfaz ambas as funções de restrições e os $2N_{var}$ limites. Se não satisfizer essa condição, x não é factível. O conjunto de todas as soluções factíveis geram a região factível ou espaço de busca S_{fact} [176].

O vetor funções objetivo $f(x) = [f_1(x), f_2(x), ..., f_{Nobj}(x)]$, em que $Nobj$ é o número de objetivos, representa um espaço multidimensional chamado espaço de objetivos S_{obj}. Esse espaço é a principal diferença entre otimização multiobjetivo e mono-objetivo [176]. Na otimimização mono-objetivo o espaço de busca é unidimensional, realizado pela aplicação direta da função ponderação dos objetivos.

A noção de otimalidade foi originalmente introduzida por Francis Ysidro EdgeWorth em 1881 [58], e generalizada por Vilfredo Pareto em 1896 [138], que apresentou o conceito de dominância de Pareto. Esse conceito é aplicado para comparar duas soluções factíveis do mesmo problema. Dadas duas soluções x e y, pode-se dizer que x domina y (denotado como $x \preceq y$) se as seguintes condições são satisfeitas:

- A solução x é pelo menos igual a y em todas as funções objetivo;
- A solução x é superior a y em pelo menos uma função objetivo.

O conjunto Pareto-Ótimo é aquele composto por soluções não-dominadas do conjunto de soluções. A fronteira de Pareto é o modo gráfico de mostrar o conjunto de valores das funções objetivo das soluções do conjunto Pareto-Ótimo. A Figura 4.1 ilustra alguns exemplos

de conjuntos Pareto-Ótimos, apresentando as muitas combinações de maximização/minimização de duas funções f_1 e f_2. A curva indica onde está situado o conjunto. Essa figura também mostra a possibilidade de existirem conjuntos Pareto-Ótimos formados por uma região contínua ou pela união de regiões descontínuas.

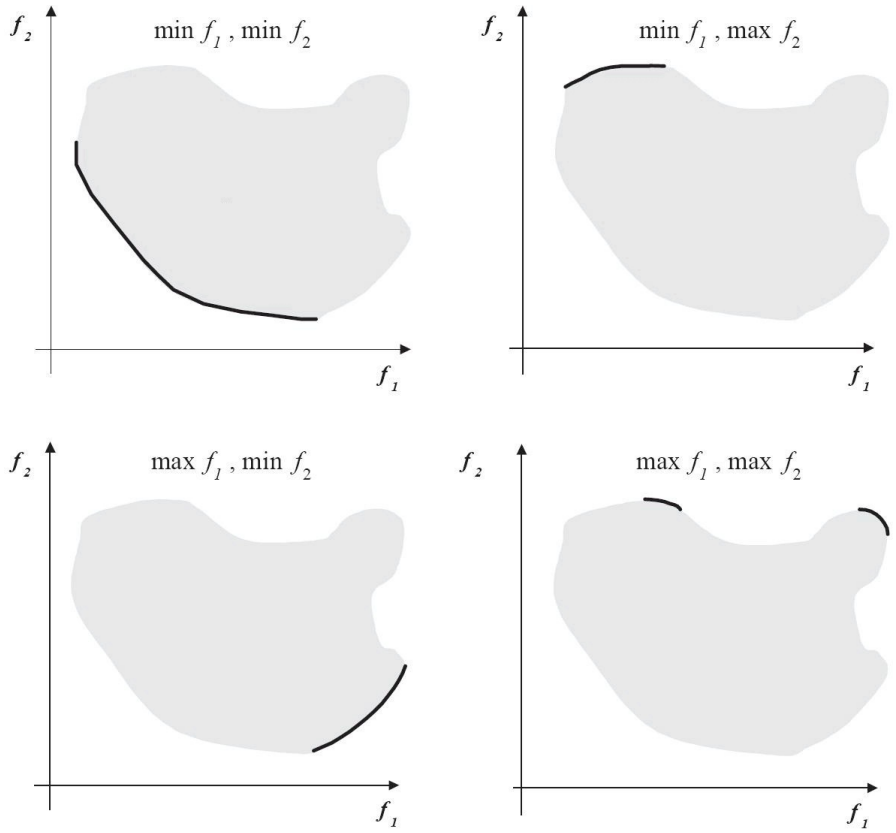

Figura 4.1: Exemplos variados de conjuntos Pareto-Ótimos no espaço de objetivos.

Dois conjuntos Pareto-Ótimos não-dominados localmente são ilustrados na Figura 4.2, apresentando a sua vizinhança no espaço de objetivos e no espaço de variáveis. Essa figura pretende mostrar que a análise de dominância para uma vizinhança B pode resultar em conjuntos Pareto-ótimos locais. Esses conjuntos possivelmente dificultam a busca pelo conjunto Pareto-ótimo global, uma vez que as soluções encontradas pelo algoritmos evolutivos multiobjetivos (AEMO) podem estar concentradas em conjuntos Pareto-ótimos locais, retardando, ou mesmo evitando a convergência para o conjunto ótimo.

Diversos métodos de otimização não baseados em AEs têm sido desenvolvidos para lidar com problemas multiobjetivos [130, 59, 126, 1, 37, 170]. No entanto, muitos deles apresentam limitações quando a fronteira de Pareto é côncava ou desconectada. Outros métodos requerem diferenciabilidade das funções objetivo e suas restrições, além da maioria das técnicas gerarem somente uma solução por execução. Portanto, nessas abordagens várias execuções (usando pontos de partida diferentes) são

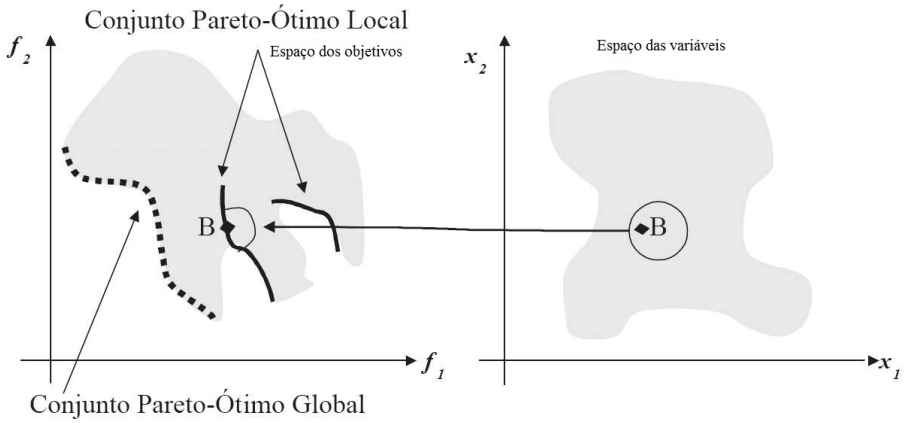

Figura 4.2: Soluções Pareto-Ótimas locais e globais [176].

necessárias, a fim de se obter um conjunto de soluções para o conjunto de Pareto-Ótimo.

Em contrapartida, os AEs [75, 92, 38] (Apêndice B) trabalham com um conjunto de possíveis soluções (população), o qual permite encontrar diversos elementos para o conjunto de Pareto-Ótimo em uma única execução do algoritmo. Além disso, AEs conseguem lidar sem dificuldade com a fronteiras de Pareto descontínuas ou côncavas [37]. Desse modo, os AEs representam uma técnica adequada para os problemas multiobjetivo, sendo simples e efetiva. Na Seção 4.3 são descritos os principais algoritmos evolutivos multiobjetivos (AEMOs), seguindo a ordem cronológica em que foram desenvolvidos.

4.3 História dos algoritmos evolutivos multiobjetivo

O desenvolvimento de AEMOs pode ser organizado em duas gerações: primeira geração (período em que se enfatizou a simplicidade dos algoritmos) e segunda geração (período em que se enfatizou a eficiência) [37]. Essas gerações são descritas nas Subseções 4.3.1 e 4.3.2.

4.3.1 A primeira geração dos AEMOs

O primeiro AEMO foi desenvolvido por Schaffer (1985) [153], denominado VEGA (do inglês, *Vector Evaluated Genetic Algorithm*). No VEGA, Schaffer propôs uma modificação do algoritmo genético tradicional proposto por Holland [92], sendo que sua implementação é bastante simples. Este método trabalha com subpopulações para otimizar cada objetivo separadamente, uma vez que que cada solução é avaliada por uma única função-objetivo. O operador de cruzamento combina as melhores soluções individuais, tentando encontrar soluções próximas da região ótima de Pareto, porém eventualmente,

o método converge para a melhor solução de um dos objetivos, "prendendo-se" aos extremos de fronteira Pareto-ótimo. Outra desvantagem é que esse método não obtém diversidades suficiente nas soluções da fronteira de Pareto.

O conceito de Pareto em AEMOs foi introduzido por Goldberg (1989) [75], que criou um procedimento de ranqueamento dos indivíduos baseado no conceito de não-dominância. De acordo com esse procedimento, cada indivíduo não-dominado deve ser removido da população e receber um *ranking* 1; dos que restavam, os novos indivíduos não-dominados recebiam *ranking* 2 e também eram removidos. Este procedimento se repetia, até que todos os indivíduos recebessem seu *ranking*, atribuindo uma probabilidade de seleção tanto maior, quanto menor o valor do *ranking*. Goldberg enfatizou a necessidade de um mecanismo de compartilhamento da aptidão (*fitness sharing*) para garantir uma diversidade adequada nas soluções.

No entanto, essa ideia demorou alguns anos a ser implementada. Kursawe (1991) [112] apresentou o VOES (do inglês, *Vector Optimized Evolution Strategy*), que apresenta um mecanismo para reter indíviduos não-dominados, e para excluir as soluções excedentes, as quais são escolhidas principalmente considerando o critério de proximidade entre as soluções, diferente do que propôs Goldberg. Esse método foi pouco utilizado, devido a necessidade de se utilizar cromossomos diplóides, contendo um cromossomo dominante e um recessivo. Ainda na década de 1990, Hajela e Lin (1992) [78] propuseram um método que aplica pesos variáveis às funções-objetivo, denominado WBGA (do inglês, *Weighted Based Genetic Algorithm*) [46]. Esse método não utiliza o conceito de não-dominância, porém possui como vantagem a sua simplicidade de desenvolvimento, além de apresentar uma modelagem de uma função de aptidão para problemas de minimização e maximização simultâneos. Em contrapartida, como em outros métodos baseados em pesos, pode não encontrar soluções ótimas em espaços de busca grandes, predendo-se em subótimos.

Os AEMOs tradicionais foram desenvolvidos na década de 1990, considerando a não-dominância de Pareto, apresentando alguns métodos de compartilhamento da aptidão e utilizando elitismo[1] em suas gerações. Fonseca e Fleming (1993) foram os primeiros a implementarem a ideia de Goldberg de ordenar as soluções baseado no conceito de não-dominância, desenvolvendo o MOGA (do inglês, *Multiple Objective Genetic Algorithm*) [67]. Em MOGA, o ranqueamento de um dado indivíduo corresponde ao número de indivíduos dominados pelo mesmo na atual população. Todos os indivíduos não-dominados assumem os valores mais altos possíveis de aptidão (todos

[1]Elitismo é o método que copia os melhores indivíduos para a nova população, possibilitando aumentar rapidamente o desempenho do AE, pois previne a perda da melhor solução já encontrada [75].

eles obtém a mesma aptidão, de tal modo que eles podem ser amostrados na mesma taxa), enquanto que aqueles dominados são penalizados de acordo com a densidade populacional da região correspondente a que pertencem.

O NSGA (do inglês, *Non-dominated Sorting Genetic Algorithm*) [167] é semelhante ao MOGA, pois utiliza o procedimento de ranqueamento de indivíduos não-dominados, porém classifica os vetores solução de maneira diferente. Nesse método as soluções são subdivididas em classes, e todas as soluções não-dominadas de uma mesma classe recebem a mesma aptidão. Horn et al. (1994) [93] propuseram o algoritmo NPGA (do inglês, *Niched-Pareto Genetic Algorithm*), sendo que este não precisa calcular um valor de aptidão que priorize soluções não-dominadas, pois o conceito de dominância é introduzido no operador de seleção, denominado de Torneio de Pareto. A vantagem do NPGA é justamente não necessitar de um cálculo explícito para a função de aptidão e a complexidade não ser proporcional ao número de objetivos. A desvantagem é a introdução de novos parâmetros a serem configurados e a influência desses parâmetros nas soluções encontradas.

Os primeiros AEMOs a serem destacados pelo sucesso foram: MOGA, seguidos pelos métodos NPGA e NSGA, que são os principais representantes dessa geração, conforme Coello (2006) [37]. Desse modo, pode-se observar que a primeira geração foi caracterizada pelo simplicidade dos algoritmos propostos [37]. A seguir, são descritos os primeiros AEMOs, que surgiram na segunda geração.

4.3.2 A segunda geração dos AEMOs

A segunda geração dos AEMOs iniciou-se com o trabalho de Zitzler e Thiele (1998) [196], que desenvolveram o SPEA (do inglês, *Strentgh Pareto Evolutionary Algorithm*). O SPEA é caracterizado pela manutenção de uma população externa de soluções não-dominadas encontradas. Em 2001, foi desenvolvida um abordagem melhorada desse método, denominada SPEA2 [195]. Nessa geração, foi implementada também uma melhoria do NSGA, denominada NSGA-II [47]. No mesmo período, foi desenvolvido o PAES (do inglês, *Pareto Archived Evolution Strategy*) por Knowles e Corne (1999) [110]. Esse método é capaz de encontrar soluções diversas no conjunto Pareto-Ótimo, porque mantém um arquivo de soluções não-dominadas, apesar de utilizar uma simples estratégia (1+1)[2] de evolução de busca local, o que pode prejudicar seu desempenho para problemas mais complexos.

O MOEA-D (do inglês, *MultiObjective Evolutionary Algorithm based on Decomposition*) foi proposto por Zhang e Li (2007) [192], que propõe a utilização de funções ponderações entre os objetivos considerados, de maneira

[2]Um pai gera um único filho, que emprega busca local [16].

combinatória. Esse método está entre os mais importantes AEMOs que lidam com muito critérios (três ou mais objetivos). Em 2005, foi proposto o AEMT (Algorimo Evolutivo multiobjetivo baseado em Tabelas) [48], que se baseia na ideia do VEGA de representar objetivos em tabelas, utilizando como um critério de seleção o conceito de não-dominância.

Muitas abordagens multiobjetivo têm sido desenvolvidas na segunda geração (que se estende até os dias atuais). Durante essa geração, um dos aspectos mais enfatizados foi, sem dúvida, a eficiência tanto dos algoritmos quanto das estruturas de dados utilizadas, conforme [37].

A seguir, são descritos mais detalhadamente os algoritmos NSGA (e NSGA-II), SPEA (e SPEA2), MOEA-D e AEMT.

4.3.3 NSGA e NSGA-II

O NSGA foi proposto por Deb (2000) [47, 167] e utiliza o conceito de classificação em camadas (*ranking*), descrito em Goldberg (1989) [75]. Em tal procedimento, primeiramente as soluções são separadas em dominadas e não-dominadas. Cada solução não-dominada recebe um valor de aptidão, e o processo é repetido sucessivamente, utilizando somente os indivíduos dominados, até que toda a população seja classificada.

Desse modo, os valores de aptidão atribuídos aos indivíduos de uma camada são sempre maiores que os daqueles de camadas posteriores, garantindo uma quantidade maior de cópias dos melhores indivíduos em cada geração. Entretanto, essa abordagem é pouco eficiente, uma vez que a classificação é repetida diversas vezes. Neste contexto, surge uma abordagem melhorada o NSGA-II, uma abordagem melhorada do NSGA.

O NSGA-II [47] utiliza o procedimento de ordenação elitista por dominância (Pareto *ranking*). Essa ordenação classifica as soluções de um conjunto M em k fronteiras F_1, F_2, ..., F_k de acordo com o grau de dominância de tais soluções. Logo, a fronteira F_1 contém soluções não-dominadas de todo o conjunto M. A fronteira F_2 possui as soluções não-dominadas de $M - F_1$; F_3 contém as soluções de $M \subset (F_1 \bigcup F_2)$, e assim sucessivamente. Cada solução i em P tem dois valores calculados: nd_i (número de soluções que dominam a solução i); e U_i (conjunto de soluções que são dominadas pela solução i).

Observando o Algoritmo 1, primeiramente são calculados os valores nd_i e determinados os conjuntos U_i para as soluções em M. Além disso, as soluções com $nd_i = 0$ estão contidas na fronteira F_1. A seguir, o conjunto de soluções dominadas U_i é percorrido para cada solução i de F_1. O contador nd_j de cada solução j em U_i é decrementado de 1. Se nd_j = 0, a solução j pertence a fronteira posterior, neste caso, F_2. O último passo é classificar todas as soluções em uma fronteira.

O algoritmo NSGA-II gerencia duas populações, P e Q, ambas de tamanho N_{ind}. Na primeira iteração, os indivíduos iniciais da população P_1 geram soluções em Q_1 pelos operadores (seleção, recombinação e mutação). No próximo passo, realiza-se uma competição para escolher N_{ind} indivíduos para preencher as vagas na próxima população entre $2N_{ind}$ indivíduos contidos em $R_t = P_t \bigcup Q_t$. Esse passo é feito pela ordenação por dominância em R_t, direcionando as soluções não-dominadas contidas nas fronteiras diretamente para a próxima geração (elitismo).

Algoritmo 1: Algoritmo para Ordenação por Dominância.

Entrada: M, um conjunto de soluções.
Saída: F_1, F_2,..., F_k, as fronteiras que classificam as soluções de M.

```
para solucao_i ∈ M faça
  nd_i = 0
  U_i = ∅
  para (solucao_j ≠ solucao_i)&(solucao_j ∈ M) faça
    se i ⪯ j então
      U_p = U_p ∪ j
    fim se
    se j ⪯ i então
      nd_i = nd_i + 1
    fim se
  fim para
  se nd_i = 0 então
    F_1 = F_1 ∪ i
  fim se
fim para
k = 1
enquanto F_k ≠ ∅ faça
  Temp = ∅
  para solucao_i ∈ F_k faça
    para solucao_j ∈ U_i faça
      n_j = n_j − 1
      se n_j = 0 então
        Temp = Temp ∪ j
      fim se
    fim para
  fim para
  k = k + 1
  F_k = Temp
fim enquanto
```

O NSGA-II garante a diversidade na fronteira, com uma estimativa de densidade aplicada nas soluções que rodeiam cada indivíduo da população. Assim, a média da distância das duas soluções adjacentes a cada indivíduo em relação a todos os objetivos é calculada. Tal média é denominada de distância de multidão. O Algoritmo 2 é o pseudo-código do procedimento para calcular

o valor da distância de multidão (*crowdist*) do *n*-ésimo indivíduo do conjunto M e f_m, o valor da m-ésima função objetivo para tal indivíduo.

Algoritmo 2: Cálculo da Distância de Multidão.

Entrada: Um conjunto $|M|$ de soluções.
Saída: Valores de distância de multidão da n-ésima solução em M.
para i = 1 to $|M|$; **faça**
  $dist_i = 0$;
fim para
para m = 1 to $|M|$ **faça**
  Classificar M por f_m.
  $crowdist_1 = crowdist_{|M|} = \infty$;
  para i = 2 to $|M| - 1$ **faça**
    $crowdist_i = crowdist_i + f_m(M_{i+1}) - f_m(M_{i-1})$;
  fim para
fim para

O valor do aptidão (*fitness*) de cada solução i é calculado pelo $ranking_i = k$ (o valor de *ranking* i é igual ao número da fronteira F_k) e $crowdist_i$ (o valor de distância de multidão de i). A seleção é realizada por torneio. Logo, duas soluções são analisadas e comparadas para escolher aquela que gerará descendentes na nova população. Uma solução y é preferida em relação a solução z quando:

- y possui um *ranking* menor que z, ou seja, $ranking_y < ranking_z$;
- Ambas as soluções possuem o mesmo *ranking* e i possui um maior valor de distância de multidão.

O cálculo da distância de multidão possibilita que as soluções mais diversas ocupem as últimas vagas vazias de P_{t+1}, preservando a diversidade das soluções. A população Q_{t+1} é gerada por meio dos operadores de seleção por torneio, recombinação e mutação em P_{t+1}. O NSGA-II executa N_{iter} iterações e as soluções finais aparecem em $P_{t+1} \bigcup R_{t+1}$. O pseudo-código do algoritmo NSGA-II é descrito no Algoritmo 3.

De acordo com os experimentos (Seção 5.4), para o problema de predição de estruturas protéicas, este algoritmo não é o mais adequado, como se pode verificar também em outros trabalhos [43, 124, 55, 195, 192]. Apesar das fronteiras F_1, F_2, ..., F_k gerarem um conjunto de soluções não-dominadas distribuídas em relação às de outros AEMOs, o NSGA-II tende a convergir prematuramente em problemas complexos com mais de dois objetivos, ou mesmo em problemas combinatórios com somente dois objetivos. Isso deve-se ao fato de, nesses problemas, quase todas as novas soluções serem não dominadas a partir de uma certa geração.

Algoritmo 3: NSGA-II.

Entrada: Conjunto de parâmetros relevantes ao NSGA-II
Saída: Soluções na população P_{final} e Q_{final}
Inicialização
Criar uma população de soluções aleatórias P_1 de N_{ind} indivíduos.
Ordenar P_1 por dominância.
Aplicar operadores genéticos P_1 para gerar uma nova população Q_1 de tamanho N_{ind}.
para $geracao$ = 1 to N_{iter} **faça**
  Aplicar o Algoritmo 1 em $R_t = P_t Q_t$.
  $k = 1$
  enquanto $P_{t+1} + F_k \leq N_{ind}$; **faça**
    Aplicar o Algoritmo 2 em F_k.
    $P_{t+1} = P_{t+1} \bigcup F_k$;
    $k = k + 1$;
  fim enquanto
  Aplicar o Algoritmo 2 em F_k.
  Classificar a F_k pelo ranking e distância de multidão.
  Copiar as primeiras $N_{ind} - |P_{t+1}|$ soluções de F_k para P_{t+1}.
  Gerar a nova população Q_{t+1} aplicando os operadores genéticos em P_{t+1}.
fim para
$P_{final} = P_{t+1}$
$Q_{final} = Q_{t+1}$

4.3.4 SPEA e SPEA2

O SPEA [196] utiliza uma população externa contendo os melhores indivíduos encontrados até o momento e, a cada geração, essa população é atualizada. O cálculo da função de aptidão de um determinado indivíduo considera a sua distância para a fronteira Pareto e a distribuição das soluções em uma dada geração. A diversidade das soluções encontradas pelo SPEA é avaliada quando a população externa atinge um número máximo de indivíduos. Quando a população alcança essa quantidade, é aplicado um algoritmo de agrupamento (do inglês, *Clustering Method*), que visa eliminar as soluções que excedem o valor limite das populações externas, sem prejudicar a diversidade da fronteira desta população.

O SPEA2, proposto por Zitzler, Laumns e Thiele [195], apresenta melhorias sobre o SPEA, tais como [37]:

- Utilização de uma função de aptidão baseada no número de indivíduos que dominam ou são dominados por outro;
- Cálculo de densidade baseado na distância dos vizinhos mais próximos;
- Método de seleção que preserva os indivíduos das extremidades da fronteira Pareto.

O SPEA2 apresentou melhores resultados que seu antecessor em todos os testes realizados por Zitzler, Laumanns e Thiele [195], além de ser bastante competitivo com o NSGA-II.

4.3.5 MultiObjective Evolutionary Algorithm based on Decomposition

O MOEA-D [192] explicitamente decompõe o problema multiobjetivo em subproblemas escalares de otimização, os quais são tratados simultaneamente. Em cada geração, a população é composta das melhores soluções encontradas para cada subproblema. As relações de vizinhança entre esses subproblemas são definidas baseadas nas distâncias entre seus vetores de coeficiente de ponderação. As soluções ótimas para dois subproblemas vizinhos deveriam ser muito semelhantes. Cada subproblema é otimizado em MOEA-D utilizando a informação a partir da vizinhança dos subproblemas.

MOEA-D tem as seguintes características [192]:

- Uma maneira simples de introduzir aproximações de decomposição dentro das estratégias de computação evolutiva multiobjetivo;
- Uma complexidade computacional menor que NSGA-II, com aproximações de decomposição avançadas melhores que NSGA-II com três objetivos;
- Técnicas de normalização de objetivos podem ser incorporadas em MOEA-D para tratar objetivos escalados diferentemente;
- É muito comum usar métodos de otimização escalar em MOEA-D, desde que cada solução seja associada com um problema de otimização escalar.

A decomposição feita pelo MOEA-D consiste em transformar um problema multiobjetivo em subproblemas escalares e otimizá-los simultaneamente, conforme já fora dito. A técnica utilizada pelo MOEA-D é a Decomposição de Tchebycheff [130]. Por meio de pesos diferentes, pode-se encontrar pontos distintos na fronteira Pareto. Desse modo, o MOEA-D otimiza diversos subproblemas paralelamente, cada um com um vetor de pesos distintos. Uma desvantagem dessa abordagem é que a função de ponderação pode não ser contínua ou diferenciável. Entretanto, como o MOEA-D não requer cálculo de derivadas, assim, isso não se torna um problema crítico.

No MOEA-D, a vizinhança de um vetor de pesos é determinada pelos demais vetores de pesos próximos a ele. Como cada vetor de pesos define um subproblema diferente, pode-se dizer que os subproblemas vizinhos são aqueles com vetores de pesos com menores distâncias. Essa definição é

fundamental, pois esse algoritmo utiliza as informações da vizinhança para buscar melhores soluções. A cada geração, a população do MOEA-D é formada pela melhor solução de cada subproblema até o momento. Desse modo, subproblemas com vetores de pesos próximos tendem a ter soluções próximas. Pode-se concluir que soluções promissoras de um subproblema têm grande probabilidade de serem fortes soluções para outros subproblemas. O MOEA-D utiliza as soluções de subproblemas vizinhos para realizar as operações de reprodução (cruzamento e mutação). O pseudo-código do MOEA-D é descrito no Algoritmo 4, em que POP_{ext} é a população externa, V_p são vetores de peso, $SubPob$ representa os subproblemas e $Nsub$ indica o número de subproblemas envolvidos.

Algoritmo 4: MOEA-D

enquanto (Critério de parada não atingido) **faça**
  Cria POP_{ext}.
  $POP_{ext} = \emptyset$.
  Calcula a distância Euclidiana entre todos V_p.
  Seleciona os mais próximos para a vizinhança de um vetor V_k.
  para i = 1 to $Nsub$ **faça**
    Seleciona dois pontos da vizinhança do vetor V_k
    Gera uma nova solução S_i utilizando operadores de reprodução.
    se (S_i for melhor) **então**
      Atualizar os valores do vetor V_k.
    fim se
    Para cada $SubPob$ vizinho, verificar se S_i é melhor que a solução atual e, caso seja, substituí-la.
  fim para
  Atualizar a POP_{ext}, somente com pontos não-dominados.
fim enquanto

Neste trabalho, o MOEA-D não foi usado como método para comparação para o AEMT porque considerou-se mais relevante verificar a importância do critério de não-dominância do que a ponderação entre os objetivos, que é o diferencial do MOEA-D para aumento de desempenho do AEMT. Além disso, o aspecto de considerar diversas funções (no caso, funções ponderação) para melhor orientar a exploração no espaço dos objetivos do MOEA-D está presente nos AEs desenvolvidos neste trabalho. Isso ocorre por meio das tabelas de subpopulações, em que cada uma está associada a um objetivo ou função ponderação (Capítulo 5). Assim, pode-se dizer que, de certa forma, os AEs propostos nesta tese reproduzem também aspectos presentes no MOEA-D. A importância de utilizar várias funções ponderações em problemas multiobjetivos revelada neste trabalho é também verificada pela própria estratégia de exploração do espaço dos objetivos do MOEA-D.

4.3.6 Algoritmo Evolutivo Multiobjetivo baseado em Tabelas

O AEMT é capaz de trabalhar com mais de uma subpopulação paralelamente por meio de tabelas [48]. De modo geral, cada subpopulação (tabela) armazena indivíduos avaliados por uma função de aptidão, podendo então lidar com múltiplos critérios de seleção de indivíduos de maneira simultânea. O AEMT também apresenta uma subpopulação para indíviduos avaliados por uma função de ponderação entre os critérios. Estendendo a ideia do AEMT, pode-se adicionar outra subpopulação que armazena os indivíduos não-dominados de cada geração. É importante ressaltar que para cada tabela, existe um critério de seleção correspondente.

A seleção de indivíduos é realizada pela escolha aleatória de duas tabelas diferentes, e a partir de cada uma obtém-se um indivíduo pela seleção por torneio [75]. Dessa maneira, dois indivíduos provenientes de duas subpopulações distintas geram um novo indivíduo por meio de operadores de reprodução (Apêndice B), que é inserido em todas as tabelas, desde que sua adequação ao objetivo relativo a cada tabela seja melhor que pelo menos um dos indivíduos da mesma.

Essa é a proposta original do AEMT, e como foi comprovado no trabalho de Santos (2010) [55, 152], o AEMT encontra soluções adequadas que NSGA-II não encontra para o contexto de distribuição elétrica. Além disso, é capaz de aplicar o compartilhamento da aptidão por meio da migração de características de indivíduos entre as subpopulações, garantindo uma diversidade adequada de soluções na fronteira de Pareto. Brasil e Delbem (2011) [26] também mostram que o algoritmo AEMT apresenta melhores resultados que o NSGA-II para o problema de PSP.

O pseudo-código desse algoritmo multiobjetivo é apresentado no Algoritmo 5, em que $SubPop_i$ representa as subpopulações; $Subpop1$ e $Subpop2$, as subpopulações escolhidas; $Ind1$ e $Ind2$, os indivíduos pais; $NovoInd$, novo indivíduo gerado; $NSubpop$, o número de subpopulações.

4.4 Problemas de Otimização com Muitos Objetivos

Conforme descrito nas seções anteriores, algoritmos evolutivos multiobjetivos otimizam simultaneamente duas ou mais funções objetivo, conseguindo encontrar um conjunto de soluções em um única execução do algoritmo. Recentemente, há um crescente interesse na aplicação de AEMOs para resolver problemas de otimização com muitos objetivos [2, 102, 101, 99, 197], aqueles que tratam de quatro ou mais objetivos. No entanto, em geral, AEMOs não obtém soluções adequadas com grande número de critérios de seleção [102].

Algoritmo 5: AEMT

Inicialização do contador de gerações g.
Gera subpopulações iniciais $SubPop_i$.
Avalia os indivíduos das subpopulações iniciais.
enquanto critério de parada não atingido **faça**
  $Subpop1 = SelecioneSubpop()$;
  $Subpop2 = SelecioneSubpop()$;
  $Ind1 = SelecioneIndiv(Subpop1)$;
  $Ind2 = SelecioneIndiv(Subpop2)$;
  $NovoInd = Reproduz(Ind1, Ind2)$;
  $Avalia(NovoInd)$;
  para $i = 1$ to $NSubpop$ **faça**
    $AtualizaSubop(Subpop_i, NovoInd)$;
  fim para
  $g = g + 1$
fim enquanto

Neste contexto, dentre os AEMOs conhecidos, o MOEA-D tem destacado-se por lidar bem com problemas com até dez critérios. Atualmente, MOEA-D é um dos melhores AEMOs para muitos objetivos, apresentando alta capacidade de busca, bem como grande eficiência computacional [119]. Por exemplo, o MOEA-D trabalha adequadamente com quatro ou mais objetivos [101]. Esse algoritmo também pode obter soluções bem distribuídas na fronteira de Pareto usando um número de vetores com peso, com diferentes direções em funções de ponderação escalar [100, 101].

No entanto, no trabalho de Ishibuchi (2011)[99], foi verificado que o seu desempenho em problemas multiobjetivos com objetivos altamente correlacionados é prejudicado, enquanto que os algoritmos NSGA-II e SPEA2 apresentam poucos efeitos negativos quando tratam de objetivos com alta similaridade. Também na literatura não há resultados de sucesso do MOEA-D para problemas com mais de dez objetivos. Por fim, o MOEA-D requer o pré-conhecimento das soluções ótimas de cada problema mono-objetivo que compõe o problema de muitos objetivos para calcular distâncias de cada solução aos extremos da fronteira. Na Seção 4.5 são apresentados alguns AEMOs aplicados ao problema de PSP.

4.5 Aplicação de AEMOs para PSP

No contexto do problema de PSP, tem-se aplicado AEMOs usando diferentes modelos de energias (Seção 2.3), que é um ponto que distingue uma abordagem de outra, além da estratégia evolutiva utilizada. Cutello (2005) desenvolveu um AEMO para PSP, que considera as energias entre átomos ligados e não-ligados como dois critérios de seleção conflitantes que devem ser

minizados simultaneamente, portanto, bem caracterizado como um problema multiobjetivo. Esse trabalho de Cutello [44] aplicou uma abordagem do PAES (denominada I-PAES) para o problema de PSP *ab initio*, a qual propõe dois operadores de hipermutação específicos para PSP. Outros trabalhos também foram desenvolvidos muito semelhantes a esse [105, 43].

Handl e outros [79, 81, 82, 80] tem desenvolvido um trabalho de pesquisa sobre o problema de PSP sob o ponto de vista multiobjetivo. As funções de energia baseadas no conhecimento para a predição de estrutura de proteínas são tipicamente combinações lineares de um número de diferentes termos de energia ponderadas. Neste contexto, Handl investiga quais seriam as configurações ideais para a ponderação dessas energias, e se, de fato, existem essas configurações [82]. No entanto, esses trabalhos não apresentam resultados relevantes para predições puramente *ab initio* de estruturas.

O grande desafio no avanço da pesquisa de algoritmos evolutivos multiobjetivo para o problema de PSP, e no âmbito geral é, justamente, combinar características positivas dos algoritmos estudados até o momento, buscando diminuir as desvantagens encontradas.

4.6 Considerações Finais

Este Capítulo tem como objetivo descrever o problema de otimização multiobjetivo, assim como apresentar a história e o desenvolvimento dos AEMOs, desde a primeira geração até a segunda, destacando as vantagens e as desvantagens de cada proposta. Muitos algoritmos, especialmente na segunda geração, enfatizaram o uso do conceito de não-dominância e elitismo, priorizando a eficiência dos algoritmos. Isso justifica porque resultados melhores são encontrados na segunda geração, como diversos trabalhos de comparação para comprovação [43, 124, 26].

Os algoritmos da primeira geração que utilizaram o conceito de não-dominância (MOGA, NSGA e NPGA) foram mais bem sucedidos que os primeiros MOEAs implementados (VEGA, VOES e WBGA), o que comprova a importância de usar um conjunto de indíviduos não-dominados. Os algoritmos da segunda geração (SPEA e NSGA-II) que começaram a utilizar técnicas elitistas apresentam melhores resultados que os algoritmos da primeira geração, além de apresentarem maior eficência nos algoritmos.

No entanto, não se pode ignorar as ideias iniciais dos primeiros algoritmos multiobjetivos, pois delas ainda surgiram novas propostas melhores que foram aperfeiçoadas. Nesse contexto, pode-se citar o AEMT proposto na segunda geração, que foi inspirado pela ideia básica do VEGA da primeira geração, ambos utilizando tabelas para representar os objetivos. Com o

mesmo propósito do MOEA-D (proposto posteriormente ao AEMT), ambos usam funções de ponderação para melhor explorar o espaço de objetivos. Não por acaso, ambos os algoritmos, MOEA-D e AEMT, são capazes de trabalhar bem com muitos objetivos (três ou mais). O AEMMT possui certa flexibilidade em relação ao MOEA-D, uma vez que: não precisa lidar com vetores de peso e analisar vizinhança entre esses vetores; não requer conhecimento do ótimo de cada objetivo considerado independentemente (pontos extremos da fronteira) e, à princípio, não possui restrições relativas a desempenho para mais que dez objetivos.

O desenvolvimento do AEMT para o problema de PSP buscou aumentar sua capacidade de exploração do alcance nos espaços de busca e de objetivos, garantindo a diversidade da população e uma boa distribuição na fronteira de Pareto. A proposta original do AEMT foi apresentada neste Capítulo, sendo que no Capítulo 5 será descrito o desenvolvimentoo dessa abordagem para o problema de PSP.

CAPÍTULO

5

Desenvolvimento de AEMTs para PSP

5.1 Considerações iniciais

Neste Capítulo propõe-se novos AEMOs para PSP baseados no AEMT (Subseção 4.3.6). Primeiramente, a Seção 5.1.1 descreve o algoritmo ProtPred [120, 122, 72, 18], uma abordagem puramente *ab initio*. A Seção 5.2 propõe AEMTs para o problema PSP apresentando as etapas de desenvolvimento do método. Em seguida, na Seção 5.3, encontra-se uma descrição nas proteínas utilizadas para avaliar os AEMTs. Por fim as Seções 5.4 e 5.5, apresentam, respectivamente, análise dos resultados obtidos pelos AEMTs propostos e uma avaliação de predição de folha-β pelo AEMMT, uma nova abordagem multiobjetivo com muitas tabelas.

5.1.1 Descrição do ProtPred

O algoritmo ProtPred foi proposto em 2006 [120] e foi implementado inicialmente usando a abordagem mono-objetivo (chamada mono-ProtPred), isto é, avaliando as soluções a partir de uma função ponderação das energias. Em seguida, estendeu-se o mesmo utilizando um método multiobjetivo, o NSGA-II 4.3.3. Essa abordagem foi denominada NSGA-ProtPred.

A execução do ProtPred primeiramente inicializa de uma população de conformações randômicas. Os ângulos de torsão (Φ, Ψ e χ) são obtidos aleatoriamente a partir de regiões restritas de acordo com o diagrama de Ramachandram (Seção 2.3). Depois disso, a energia da conformação é avaliada. Primeiramente, a estrutura da proteína em coordenadas internas[1] é

[1] Coordenadas internas são representadas pelos ângulos de torsão (Capítulo 2)

transformada em coordenadas cartesianas.

Há três tipos de operadores de recombinação[2], implementados no ProtPred: BLX-α [64], o *crossover* uniforme e o *crossover* dois-pontos (Apêndice B). Além disso, há três tipos de operadores de mutação. O primeiro operador de mutação atua sobre uma cadeia peptídica, de forma que todos os valores dos ângulos da cadeia principal e cadeia lateral de um resíduo escolhido aleatoriamente são reamostrados a partir das regiões restritas. O segundo e o terceiro operadores de mutação modificam todos os valores dos ângulos da cadeia principal e da cadeia lateral de um resíduo selecionado usando uma distribuição uniforme para alterar os ângulos. O segundo operador provoca uma perturbação maior no valor dos ângulos e o terceiro uma perturbação menor [120]. A escolha de qual dos dois operadores será aplicado é feita aleatoriamente de acordo com a taxa de aplicação definida para cada um (a descrição mais detalhada desses operadores está no Apêndice B).

Esse algoritmo permite predizer satisfatoriamente algumas estruturas pequenas (com cerca de 20 aminoácidos) com uma hélice-α, por exemplo. No entanto, para estruturas um pouco maiores (com mais de uma estrutura secundária) ou envolvendo folhas-β, os resultados não são satisfatórios. Em [43, 124, 55, 195, 192] mostrou-se que o NSGA-II (que é o método multiobjetivo utilizado no ProtPred) não consegue evoluir adequadamente para problemas combinatórios com três ou mais objetivos. De fato, em [192] mostra-se que há problemas combinatórios com dois objetivos para os quais o NSGA-II também não tem sucesso.

Portanto, mostra-se necessário o desenvolvimento de um AEMO que supere as barreiras presentes no NSGA-II para se tratar adequadamente o problema de PSP de forma puramente *ab initio*. A proposta deste trabalho no contexto de otimização multiobjetivo foi desenvolver AEMTs adequados para PSP, conforme explicado no Capítulo 4. A Seção 5.2 explica o desenvolvimento de AEMTs para PSP.

5.2 Proposta de AEMTs para PSP

No desenvolvimento de AEMTs para PSP, primeiramente, verificou-se a contribuição desse método para PSP utilizando um AEMT com dois objetivos, considerando as energias de van der Waals e eletrostática, que são em geral as energias que mais contribuem para determinação da estrutura de uma proteína [3, 133]. Para lidar com dois objetivos, o AEMT requer três tabelas: uma para cada uma das energias e outra para a função de ponderação entre

[2]Operadores de recombinação também pode ser chamados de cruzamento ou recombinação.

as mesmas [26]. A Figura 5.1 ilustra a estrutura de subpopulações (tabelas) do AEMT para essa modelagem biobjetivo.

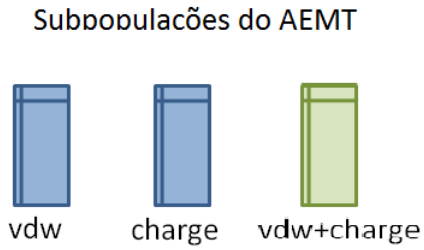

Figura 5.1: Subpopulações usadas no AEMT biobjetivo com três tabelas, em que vdw indica a energia de van der Waals; $charge$, a eletrostática e $vdw + charge$ representa a função ponderação de van der Waals e eletrostática.

A seleção de indivíduos é realizada pela escolha de duas tabelas não necessariamente diferentes, aleatoriamente, e a partir de cada uma escolhe-se um indivíduo (Subseção 4.3.6). Desse modo, tem-se dois indivíduos oriundos de duas subpopulações diferentes, sobre os quais aplica-se um operador de recombinação. Para esclarecer o procedimento, considera-se uma tabela A relativa a um objetivo, e uma tabela B relativa ao outro objetivo. A partir dessas duas tabelas, escolhem-se aleatoriamente dois indivíduos, a_i, b_i das tabelas A e B, respectivamente, e realiza-se a recombinação. As características de a_i e b_i são então cruzadas, obtendo um novo indivíduo ab_i que será inserido em todas as tabelas, desde que sua adequação ao objetivo relativo a tal tabela seja melhor que pelo menos um dos indivíduos da mesma (Seção 5.4.1). Resultados experimentais em PSP com o AEMT são apresentados na Subseção 5.4.1.

Por outro lado, conforme mostra o trabalho de Santos [152], a fronteira de Pareto obtida por esse método possui poucas amostras de soluções. No problema de PSP, há muitos ótimos locais, por isso é preciso uma amostragem relativamente grande na fronteira de Pareto. A fim de amostrar melhor a fronteira de Pareto (em termos de número de amostras, de sua extensão e de um mapeamento bem uniforme), é proposto neste trabalho o AEMT$_{ND}$. Esse método inclui uma tabela a mais de indivíduos não-dominados. Observe que o AEMT, em geral, obtém uma fronteira relativamente extensa, apesar de conter poucas amostras de forma mal distribuída. Por outro lado, o NSGA-II gera fronteira bem menos extensas, mas com número relativamente grande de amostras melhor distribuídas ao longo da fronteira. A estrutura de tabelas (subpopulações) do AEMT possibilita hibridizar esses dois métodos de forma eficaz por meio da inclusão de uma tabela adicional que considera o critério de seleção empregado no NSGA-II. A Seção 5.4.2 apresenta resultados mostrando vantagens do AEMT$_{ND}$ sobre o AEMT em PSP. A Figura 5.2 destaca

a nova tabela com indivíduos não-dominados (ND) na estrutura populacional do AEMT$_{ND}$.

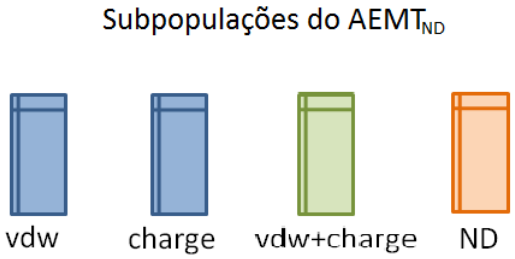

Figura 5.2: Subpopulações usadas no AEMT com quatro tabelas, em que ND indica a tabela de indivíduos selecionados pelo critério de não-dominância.

Apesar da melhoria na densidade das amostras de soluções na fronteira de Pareto obtida pelo AEMT$_{ND}$, verificou-se que esse método não consegue predizer estruturas com folhas-β. Esse desempenho do AEMT$_{ND}$ pode ser explicado, à princípio, pelo fato de energias de ligações de hidrogênio e solvatação terem uma contribuição mais significativa na definição da estrutura de uma folha-β. Buscando resolver esse aspecto, Hamiltonianos relativos a tais energias foram pesquisados para serem utilizados no mono-ProtPred, conforme mostrados na Seção 2.4.1. A estrutura do AEMT, agora AEMT$_{ND}$, possibilita novamente incluir mais critérios de seleção de indivíduos de forma simples, pela adição de mais tabelas. Nesse sentido, inseriu-se ao método duas novas tabelas correspondentes às energias de solvatação e ligações de hidrogênio, conforme ilustra a Figura 5.3.

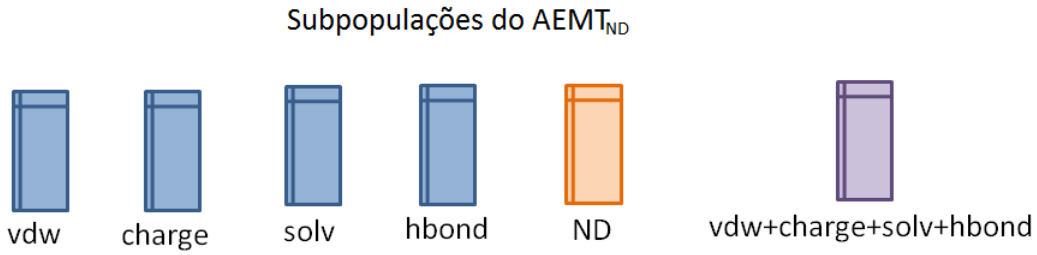

Figura 5.3: Subpopulações usadas no AEMT com quatro tabelas, em que $solv$ indica solvatação e $hbond$ refere-se a ligações de hidrogênio.

A adição destas duas novas subpopulações aumentou a diversidade de estruturas na população final. Porém, a estrutura de folha-β ainda não foi predita de forma puramente *ab initio* usando AEMT$_{ND}$ com quatro energias. Deve-se observar que a inclusão de mais dois objetivos em um AEMO pode prejudicar drasticamente o seu desempenho [55, 195, 192]. Isso se deve ao fato de a fronteira de Pareto estar em espaço de alta dimensão, requerendo

número de soluções bem maior para que se possa amostrar adequadamente a fronteira, no caso, em um espaço de funções quadridimensional (vdw, $charge$, $solv$ e $hbond$). Para mapear melhor a fronteira, investigou-se a inclusão de um número combinatório de critérios de seleção. A ideia básica é combinar diversos Hamiltonianos em diversas funções ponderações utilizadas como critérios adicionais de seleção do AEMO. Novamente, esses diversos critérios são facilmente tratadas no AEMT pela simples inclusão de mais uma tabela por critério.

Observe que a ideia de usar diversas funções de ponderação em AEMOs de muitos objetivos (em geral com três a dez objetivos) foi também empregada por Ishibuchi em [99] com o método MOEA-D, descrito no Capítulo 4, o qual também aplica a ponderação entre os objetivos. A diferença está justamente na ponderação dos objetivos no MOEA-D, baseada na Decomposição de Tchebycheff [130], utilizando as soluções de subproblemas vizinhos para realizar as operações de reprodução (cruzamento e mutação). Neste trabalho não foi desenvolvido este método, como já foi mencionado anteriormente. Aqui as combinações dos quatro objetivos (seis combinações duplas e quatro triplas), além das seis que já existiam no AEMT, resultam em um total de dezesseis subpopulações, conforme ilustra a Figura 5.4. Esse novo AEMO foi denominado AEMMT (Algoritmo Evolutivo multiobjetivo com Muitas Tabelas).

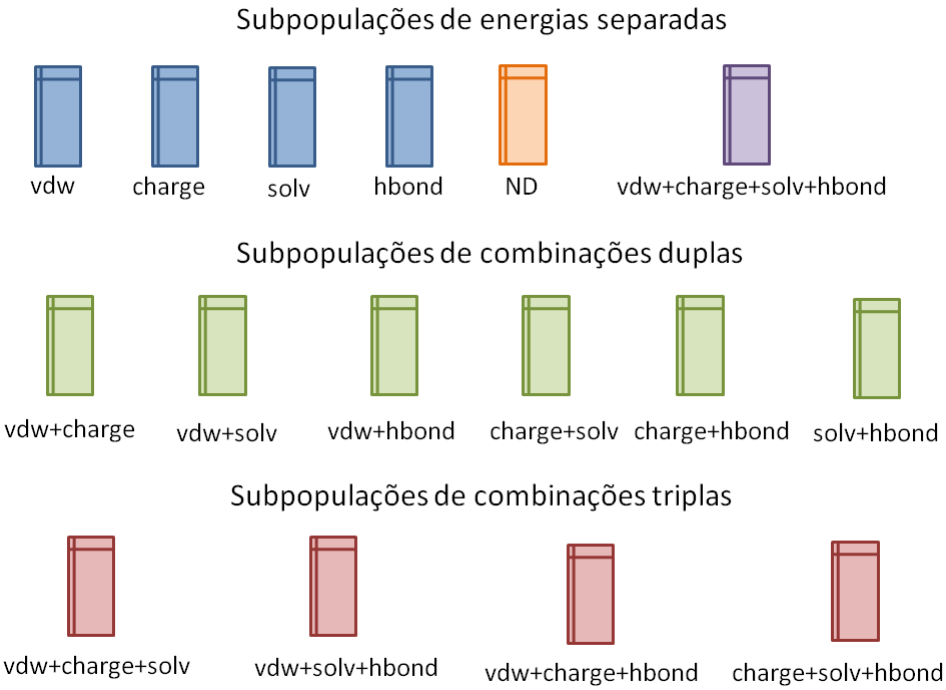

Figura 5.4: Subpopulações usadas no AEMMT com dezesseis tabelas.

Para o sucesso do AEMMT, foi necessário mudar o critério de escolha de tabelas. Tal escolha não ocorre mais aletoriamente, mas de maneira em

que se considera quais as subpopulações contribuíram mais até a geração atual. Na primeira iteração do AEMMT, nenhuma subpopulação (tabela) foi escolhida ainda, logo, nenhuma contribuiu no processo de predição. Uma subpopulação contribuir significa, nesse contexto, ser uma subpopulação que possui indivíduo escolhido a partir desta e que gere um novo indivíduo que consiga ser inserido em pelo menos uma tabela. Em outras palavras, uma subpopulação que seja capaz de contribuir com um ancestral que gera um indivíduo bem sucedido. Desse modo, para cada geração, duas subpopulações podem aumentar suas contribuições, que se acumulam ao longo da evolução.

A cada geração, deve-se escolher entre as tabelas com mais alta contribuição até aquela geração. Para isso, utilizou-se a seleção por torneio [75] aplicada na escolha de tabelas. Caso seja utilizado o torneio de dois, escolhem-se em geral tabelas entre as 50% que mais contribuíram até o momento. Se for o torneio de quatro, seleciona-se em geral entre os 25% de tabelas de maior contribuição. Quanto mais tabelas a serem escolhidas pelo torneio, mais atenua-se o efeito das subpopulações que contribuem mais. O tipo de torneio pode ser alterado, de forma dinâmica, à medida que se evolui, ou se manter fixo do início ao fim da execução. Nos experimentos deste trabalho foi utilizado o torneio de dois, sem alterá-lo durante a execução.

Na Seção 5.3 estão descritas as proteínas utilizadas nos experimentos para mono-ProtPred, NSGA-ProtPred, AEMT, AEMT$_{ND}$ e AEMMT, cujos desempenhos são analisados nas Subseções 5.4.1 e 5.4.2.

5.3 Proteínas utilizadas

Para avaliar o AEMT com dois objetivos, foram utilizadas quatro proteínas diferentes, obtidas no banco de proteínas *Protein Data Bank*[3] e compostas basicamente por hélices-α e voltas. As proteínas são: 1SOL (20 aminoácidos), 1A11 (25 aminoácidos), 2KOE (40 aminoácidos), 2K7Y (45 aminoácidos) e 1NIZ (16 aminoácidos). Essas proteínas foram escolhidas devido a simplicidade delas e por terem sido obtidas pelo procedimento RNM, o qual pode preservar efeitos do solvente no qual a proteína está imersa, diferentemente da estrutura de cristal utilizada em cristalografia de raio X (Seção 2.3). A seguir, essas proteínas são descritas com mais detalhes.

5.3.1 Hélice-α com 1 domínio

As proteínas escolhidas para serem trabalhadas com um domínio de estrutura hélice foram 1SOL, com 20 aminoácidos [179] (Figuras 5.5 e 5.6)

[3] http://www.pdb.org.

e 1A11, com 25 aminoácidos [134] (Figuras 5.7 e 5.8), obtidas do banco de proteínas *Protein Data Bank* - PDB.

Figura 5.5: Estrutura secundária de 1SOL, com 40% da estrutura em hélice (com 8 resíduos).

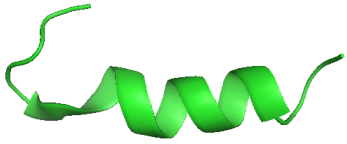

Figura 5.6: Proteína 1SOL com representação *cartoon* usando o visualizador molecular PyMol [156].

Figura 5.7: Estrutura secundária de 1A11, com 92% da estrutura em hélice (com 23 resíduos).

Figura 5.8: Proteína 1A11 com representação *cartoon* usando o visualizador molecular PyMol.

5.3.2 Hélice-α com 2 domínios

As proteínas escolhidas para serem trabalhadas com dois domínios de estrutura hélice foi 2KOE com 40 aminoácidos [180] (Figuras 5.9 e 5.10), e 2K7Y com 45 aminoácidos [187] (Figuras 5.11 e 5.12), ambas obtidas do banco de proteínas *Protein Data Bank* - PDB.

Observe o detalhe de que a volta entre as duas hélices da 2KOE forma uma pequena folha-β de dois resíduos (Figura 5.9). Sobre certo ponto de vista, a 2KOE pode ser considerada como tendo três domínios.

Figura 5.9: Estrutura secundária de 2KOE, com 52% de hélice (2 hélice envolvendo 21 resíduos) e 5% de folha-β (2 fitas com 2 resíduos).

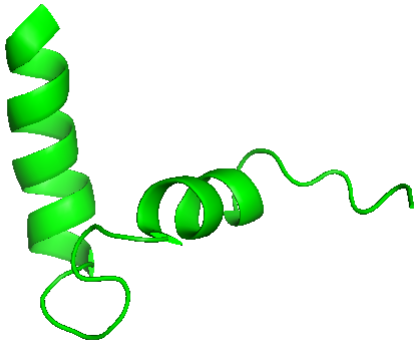

Figura 5.10: Proteína 2KOE com representação *cartoon* usando o visualizador molecular PyMol.

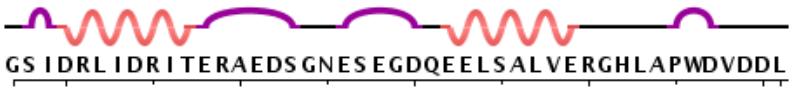

Figura 5.11: Estrutura secundária de 2K7Y, com 35% de hélice (2 hélices em 16 resíduos).

Figura 5.12: Proteína 2K7Y com representação *cartoon* usando o visualizador molecular PyMol.

5.3.3 Folha-β com 1 domínio

A proteína escolhida para ser trabalhada com folha-β de um domínio foi a 1NIZ [160], com 16 aminoácidos, obtida do banco de proteínas *Protein Data Bank* - PDB.

Figura 5.13: Estrutura secundária da proteína 1NIZ, com 50% de folha-β.

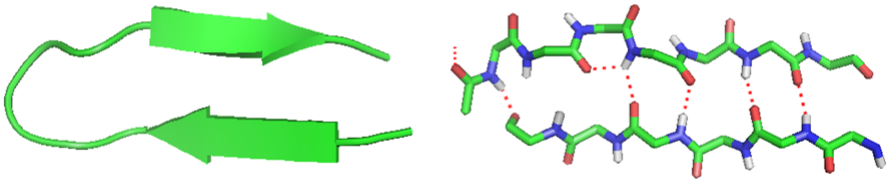

Figura 5.14: Proteína 1NIZ, mostrada pelo visualizador molecular PyMol: representação *cartoon* (à esquerda) e *stick* (à direita) mostrando as ligações de hidrogênio.

Os experimentos realizados durante o estudo e desenvolvimento do AEMT, AEMT$_{ND}$ e AEMMT estão descritos na Seção 5.4.

5.4 Análise preliminar do AEMT com dois objetivos

Primeiramente (Subseção 5.4.1) verificou-se o desempenho do AEMT na predição do conjunto de proteínas descritas na Subseção. Na Subseção 5.4.2 analisam-se as vantagens do AEMT$_{ND}$. Por fim, a Seção 5.5 mostra que o AEMMT é capaz de predizer estruturas com folhas-β.

Para configurar os parâmetros de campos de força, utiliza-se o pacote do CHARMM [30, 125][4].

5.4.1 Análise do AEMT para PSP

Os algoritmos usados nesses experimentos foram: mono-ProtPred [120], NSGA-ProtPred [122] e AEMT com dois objetivos. Recorde que o AEMT usa três tabelas neste caso, relativas às energias de van der Waals, eletrostática e a ponderação de ambas. Os resultados foram analisados por meio das métricas de similaridade RMSD [136], GDT-TS [191] e dif-PSS (dif-PSS é

[4]Arquivo editável charmm27.prm.

a diferença da porcentagem da estrutura secundária da estrutura predita em relação à estrutura nativa correspondente). O dif-PSS não é um índice usual na literatura sobre PSP, mas se mostrou interessante para orientar o desenvolvimento dos métodos, pois esse índice destaca aspectos também observados por uma inspeção visual da estrutura tridimensional de uma proteína. Analisando esse índice pode-se automatizar etapas de análise de soluções durante o desenvolvimento de AEMTs para PSP. O dif-PSS não é utilizado para avaliar as estruturas preditas no Capítulo 6. Para o cálculo desse índice foi utilizado o pacote STRIDE [70].

O RMSD (do inglês, *Root-Mean-Square Deviation*) [136] é a medida da distância média entre os átomos (geralmente da cadeia principal) de proteínas sobrepostas. O cáculo do RMSD é dado pela fórmula descrita na Equação 5.1.

$$RMSD(a,b) = \sqrt{\frac{\sum |r_{ai} - r_{bi}|^2}{n}}, \tag{5.1}$$

em que r_{ai} e r_{bi} são as posições dos átomos i das estruturas a e b respectivamente, as quais têm sido sobrepostas e n, o número de átomos considerados.

Conforme Cutello [44], apesar do RMSD ser a métrica de similaridade mais usada, esta apresenta alguns aspectos negativos, tais como: o melhor alinhamento nem sempre corresponde ao menor RMSD encontrado, a significância do RMSD depende do tamanho da estrutura a ser analisada, e a significância varia com o tipo da proteína a ser tratada. Observa-se que o cálculo do RMSD foi feito pelo programa PyMol [156], utilizando as opções padrões desse software.

O GDT-TS (do inglês, *Global Distance Test - Total Score*) [191] é uma medida de similaridade usada para comparar duas estruturas de proteínas com a mesma sequência de aminoácidos com estruturas terciárias diferentes. O valor GDT-TS de uma estrutura é um valor decimal (entre 0 e 1) que representa o quanto a predição convergiu para a estrutura nativa. O cálculo do GDT-TS é dado pela fórmula descrita na Equação 5.2.

$$GDT - TS = \frac{(p1 + p2 + p4 + p8)}{4}, \tag{5.2}$$

em que $p1$, $p2$, $p4$ e $p8$ são as porcentagens do número de pares de resíduos alinhados com distância menor que 1, 2, 4 e 8Å, respectivamente.

Um valor de GDT-TS igual a 1.0 significa que cada resíduo da estrutura predita é alinhado com a nativa com distância menor que 1 Å. Portanto, valores de GDT-TS acima de 0.5 podem indicar boas predições. Neste trabalho, o GDT-TS foi calculado pelo aplicativo TM-score [194].

Cada experimento foi exaustivamente repetido a fim de obter o tamanho

da população apropriada, bem como o número de gerações adequado para cada proteína usando o método da bisseção [83]. Os melhores resultados foram obtidos com 450 indivíduos por geração. No algoritmo AEMT, cada subpopulação tem 150 indivíduos, uma vez que há três subpopulações relacionadas aos três objetivos. A melhor ponderação das energias também foi investigada, obtendo-se pesos $p_{vdw} = 1.0$ (para energia de van der Waals) e $p_{charge} = 0.5$ (para energia eletrostática). Para o NSGA-ProtPred, as energias de van der waals e eletrostática são as próprias funções objetivo. As soluções encontradas são obtidas de dez execuções do AEMT e dos demais métodos avaliados. Consideram-se para análise as soluções da execução em que se conseguiu menor RMSD médio da população final.

As Figuras 5.15, 5.16, 5.17 e 5.18 ilustram as soluções não-dominadas da última geração, isto é, as fronteiras obtidas pelo AEMT comparadas às fronteiras dos métodos NSGA-ProtPred e mono-ProtPred. Uma simples inspeção visual revela que as fronteiras geradas por tais abordagens são diferentes, independentemente das proteínas. Embora as fronteiras obtidas pelo AEMT não apresentem um conjunto vasto de pontos, elas apresentam os menores valores para ambas energias em todos os casos. Esse resultado revela a maior capacidade de exploração do espaço de busca do AEMT, possibilitando predições melhores. A partir das soluções da fronteira, é possível calcular o RMSD, GDT-TS e dif-PSS das estruturas [136, 191] e, assim, escolher entre as estruturas preditas com os menores valores de RMSD, GDT-TS ou dif-PSS. Esse procedimento foi aplicado aos resultados obtidos pelos três métodos (mono-ProtPred, NSGA-ProtPred e AEMT) e então sintetizados na Tabela 5.1.

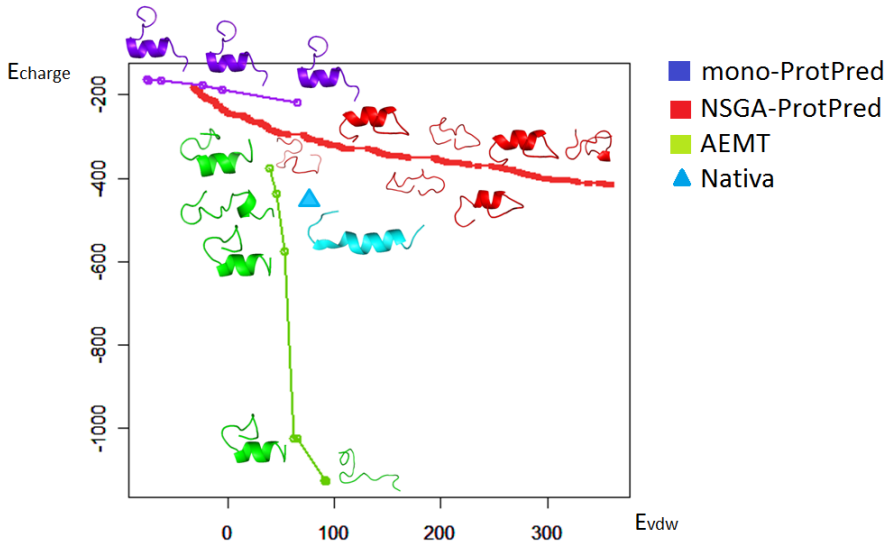

Figura 5.15: Fronteiras de Pareto para 1SOL com campos de força de van der Waals e eletrotática (Energia em kcal/mol).

Tabela 5.1: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 1SOL (os números em destaque (*) na tabela são os melhores valores da média obtida de cada índice).

Algoritmos	mono-ProtPred			AEMT			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,637	5,414	0,050	0,550	6,912	0,000	0,450	7,253	0,400
2	0,637	5,414	0,050	0,550	6,913	0,000	0,462	6,826	0,400
3	0,637	5,414	0,050	0,550	6,913	0,000	0,462	7,145	0,400
4	0,637	5,414	0,050	0,475	4,597	0,400	0,437	6,999	0,400
5	0,637	5,414	0,050	0,550	4,865	0,200	0,437	5,444	0,400
6				0,537	5,474	0,200	0,637	4,847	0,400
7				0,487	5,701	0,000	0,525	4,632	0,400
8				0,587	4,319	0,100	0,637	4,572	0,100
Média	**0,637***	**5,414***	**0,050***	0,544	5,712	0,113	0,525	5,949	0,357
Melhor indivíduo	**0,637***	5,414	**0,050***	0,587	**4,319***	0,100	0,637	4,572	0,100

De acordo com a Figura 5.15 e a Tabela 5.1, pode-se observar que as melhores predições foram encontradas pelo AEMT (apesar de os valores médios serem favoráveis ao mono-ProtPred), o qual apresenta maior diversidade de estruturas na fronteira comparada ao mono-ProtPred e menores energias que o mono-ProtPred e NSGA-ProtPred. Apesar do NSGA-ProtPred apresentar grande diversidade de soluções na população final, as melhores estruturas não predominam no conjunto da fronteira de Pareto. Em resumo, os melhores valores em energia foram encontrados pelo AEMT.

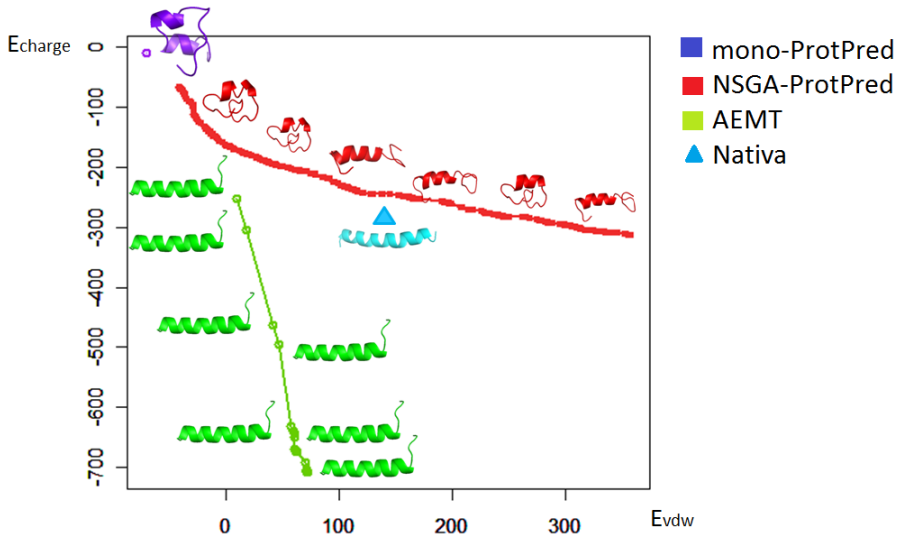

Figura 5.16: Fronteiras de Pareto para 1A11 com campos de força de van der Waals e eletrotática (Energia em kcal/mol).

A Figura 5.16 e a Tabela 5.2 evindeciam melhor por meio da proteína 1A11 que o AEMT pode melhorar as predições em relação ao outros dois métodos. Apesar das estruturas da fronteira obtida pelo AEMT não apresentarem grande diversidade de soluções, elas possuem valores de RMSD, GDT-TS e dif-PSS melhores que os obtidos pelo NSGA-ProtPred e mono-ProtPred. Visualmente, pode-se confirmar a semelhança das estruturas preditas com a nativa.

Tabela 5.2: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 1A11.

Algoritmos	mono-ProtPred			AEMT			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,541	4,902	0,312	0,790	1,288	0,120	0,360	7,380	0,920
2				0,790	1,013	0,120	0,420	7,461	0,680
3				0,790	1,288	0,120	0,420	7,461	0,680
4				0,790	1,288	0,120	0,420	7,461	0,680
5				0,830	1,286	0,120	0,360	5,591	0,920
6				0,790	1,289	0,120	0,410	7,436	0,680
7				0,800	1,299	0,120	0,420	5,487	0,920
8				0,790	1,288	0,120	0,410	5,614	0,920
9				0,790	1,119	0,120	0,420	5,487	0,680
10				0,790	1,288	0,120			
11				0,790	1,288	0,120			
12				0,790	1,287	0,120			
13				0,790	1,311	0,120			
Média	0,541	4,902	0,312	**0,794***	**1,256***	**0,120***	0,404	6,592	0,786
Melhor indivíduo	0,541	4,902	0,312	**0,790***	**1,013***	**0,120***	0,360	5,591	0,920

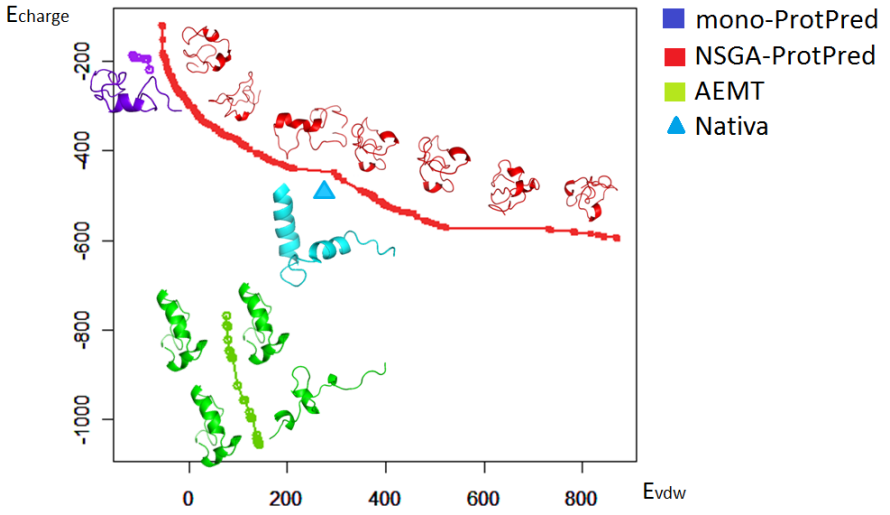

Figura 5.17: Fronteiras de Pareto para 2KOE com campos de força de van der Waals e eletrostática (Energia em kcal/mol).

Tabela 5.3: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 2KOE.

Algoritmos	mono-ProtPred			AEMT			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,331	7,495	0,400	0,325	7,713	0,295	0,438	6,799	0,045
2	0,331	7,495	0,400	0,325	7,713	0,295	0,400	7,878	0,295
3				0,325	7,713	0,295	0,325	10,635	0,295
4				0,325	7,713	0,295	0,325	10,515	0,295
5				0,325	7,713	0,295	0,331	10,407	0,295
6				0,325	7,713	0,295	0,325	10,543	0,15
7				0,325	7,713	0,295	0,425	6,175	0,295
8				0,325	7,713	0,295	0,331	11,142	0,045
9				0,388	5,425	0,240	0,431	6,754	0,045
10							0,406	7,066	0,295
11							0,325	10,810	0,295
12							0,325	10,810	0,070
Média	0,331	**7,495***	0,400	0,332	7,713	0,2981	**0,366***	9,128	**0,214***
Melhor indivíduo	0,331	7,495	0,400	**0,388***	**5,425***	**0,240***	0,425	6,175	0,295

Para a proteína 2KOE, considerada mais complexa que as anteriores, os experimentos também confirmam que o AEMT é capaz de obter melhores resultados. Devido a estrutura não ser tão simples (Seção 5.3.2) a melhor estrutura predita ocorre apenas uma vez na fronteira, repetindo outras

soluções menos adequadas na fronteira do AEMT. Pela Figura 5.17 e Tabela 5.3 pode-se perceber que a diversidade de estruturas na população final do AEMT foi comprometida, mas apesar disso consegue-se obter ainda menor valor de RMSD, GDT-TS e dif-PSS.

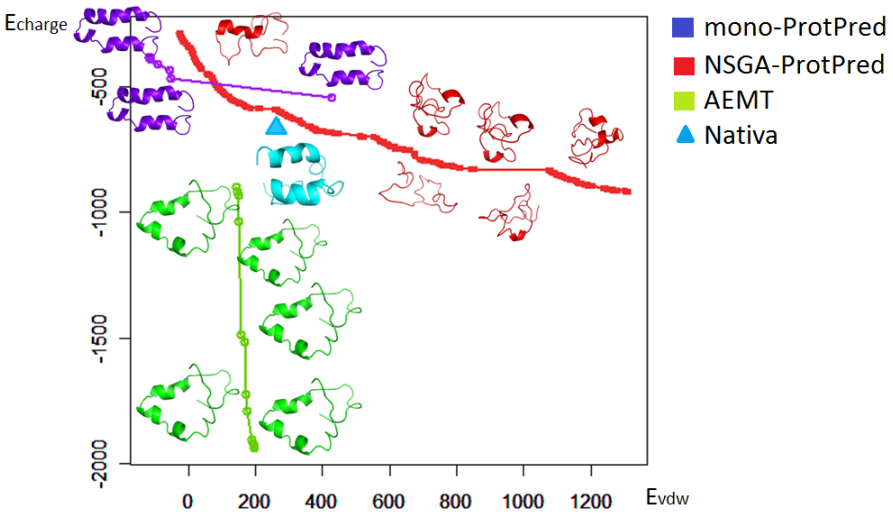

Figura 5.18: Fronteiras de Pareto para 2K7Y com campos de força de van der Waals e eletrostática (Energia em kcal/mol).

Tabela 5.4: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 2K7Y.

Algoritmos	mono-ProtPred			AEMT			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,156	8,984	0,130	0,322	7,093	0,130	0,144	9,236	0,240
2	0,156	8,984	0,130	0,333	7,620	0,130	0,139	8,110	0,220
3	0,156	8,984	0,130	0,333	7,624	0,130	0,128	8,369	0,350
4	0,156	8,984	0,130	0,261	9,297	0,20	0,117	9,175	0,240
5				0,333	7,620	0,130	0,133	8,847	0,240
6				0,261	9,422	0,130	0,122	9,294	0,240
7				0,261	7,093	0,350	0,117	8,235	0,350
8				0,333	10,944	0,35	0,128	8,431	0,350
9							0,128	8,506	0,350
10							0,128	9,759	0,240
11							0,128	8,192	0,350
12							0,133	7,801	0,240
Média	0,156	8,984	**0,130***	**0,305***	**8,342***	0,171	0,129	8,663	0,284
Melhor indivíduo	0,156	8,984	0,261	**0,261***	**7,093***	0,350	0,133	7,801	**0,240***

Conforme a Figura 5.18 e a Tabela 5.4, o AEMT obteve melhores resultados de RMSD e GDT-TS, assim como as energias conseguiram ser significativamente menores em relação aos outros métodos. No entanto, uma inpeção visual mostra que as predições não são satisfatórias em termos de domínios corretamente determinados. Isso é devido a complexidade da molécula e a interação que possivelmente existe entre as duas hélices, bem como à presença de uma pequena folha-β (com dois resíduos) entre tais hélices, aumentando a interação entre os domínios. Note que visualmente a solução encontrada pelo mono-ProtPred mostra-se melhor que as encontradas pelos outros métodos, apesar de os valores de energia indicarem o AEMT com

melhores resultados. É importante observar também que o mono-ProtPred obteve menor dif-PSS, confirmando a conclusão obtida visualmente.

Observe que as soluções no extremo esquerdo da fronteira obtida pelo mono-ProtPred (Figura 5.18) são não-dominadas em relação às soluções da fronteira do AEMT. Conclui-se então que se a capacidade de busca do AEMT fosse aumentada, este poderia talvez estender os extremos da fronteira encontrada, atingindo soluções de baixo dif-PSS, como obtidas pelo mono-ProtPred.

Pode-se dizer que as predições feitas pelo AEMT foram satisfatórias, exceto para proteínas mais complexas em que apenas o dif-PSS não é favorável ao AEMT. Buscando desenvolver um AEMT que produzisse uma fronteira de Pareto melhor amostrada de forma a obter uma melhor aproximação da fronteira ótima de Pareto, foi proposto o AEMT$_{ND}$.

Esse método trabalha com a característica de ponderação de energia do AEMT, ao mesmo tempo em que considera o conceito de não-dominância presente no NSGA-II (Seção 5.2). Essa mudança foi implementada inserindo-se uma nova tabela com os indivíduos não-dominados de cada geração. Uma abordagem semelhante foi desenvolvida com sucesso em trabahos no contexto de distribuição de energia elétrica [152, 150]. A Subseção 5.4.2 analisa as vantagens desse novo método para PSP puramente *ab initio*.

5.4.2 Avaliando o AEMT$_{ND}$

Essa Subseção apresenta experimentos que visam mostrar como o AEMT$_{ND}$ pode melhorar a fronteira de Pareto obtida em comparação com os algoritmos mono-ProtPred, NSGA-ProtPred e AEMT. Os parâmetros do AEMT$_{ND}$ foram os mesmos empregados para o AEMT, assim como também foram realizadas dez execuções de cada método (Subseção 5.4.1) e selecionou-se a execução com maior RMSD médio na população final para se analisar os resultados.

Tabela 5.5: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 1SOL.

Algoritmos	mono-ProtPred			AEMT$_{ND}$			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,637	5,414	0,050	0,675	3,362	0,000	0,450	7,253	0,400
2	0,637	5,414	0,050	0,7500	1,509	0,300	0,462	6,826	0,400
3	0,637	5,414	0,050	0,737	1,509	0,000	0,462	7,145	0,400
4	0,637	5,414	0,050	0,737	3,660	0,150	0,437	6,999	0,400
5	0,637	5,414	0,050	0,737	1,509	0,150	0,437	5,444	0,400
6				0,737	1,509	0,300	0,637	4,847	0,400
7				0,475	7,894	0,150	0,525	4,632	0,400
8				0,412	1,509	0,300	0,637	4,572	0,100
9				0,737	1,509	0,050			
10				0,412	3,362	0,300			
Média	0,637	5,414	**0,050***	**0,640***	**2,733***	0,191	0,525	5,949	0,357
Melhor indivíduo	0,637	5,414	**0,050***	**0,737***	**1,509***	0,150	0,637	4,572	0,100

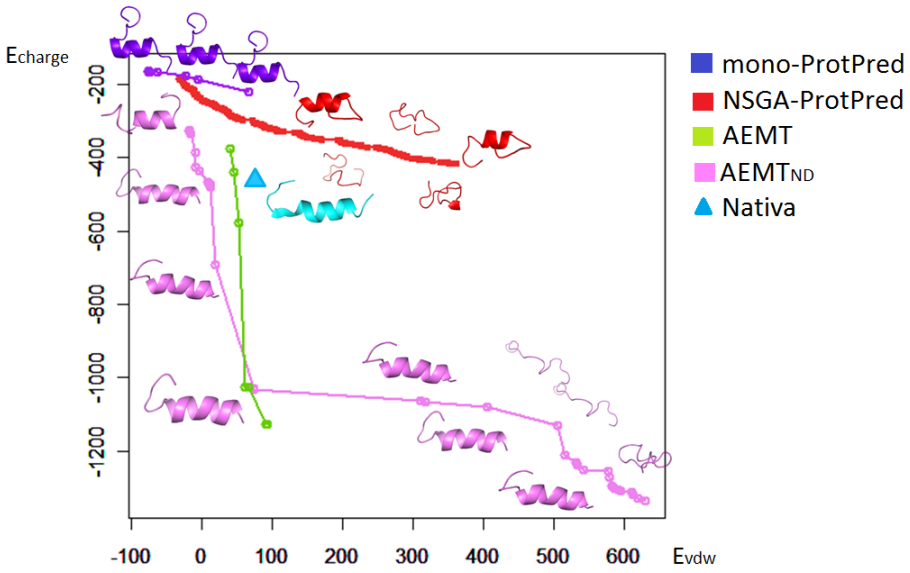

Figura 5.19: Fronteiras de Pareto para 1SOL com campos de força de van der Waals e eletrostática (Energia em kcal/mol).

Para a proteína 1SOL, a Figura 5.19 e a Tabela 5.5 mostram que o AEMT$_{ND}$ apresenta estruturas preditas mais próximas da nativa, uma vez que o AEMT$_{ND}$ gera uma diversidade maior de estruturas na fronteira de Pareto, inclusive quando comparada em relação ao AEMT. Pode-se notar que a fronteira de Pareto construída a partir da população final do mono-ProtPred produz uma diversidade muito limitada, apresentando estruturas com os mesmos resultados em termos de RMSD, GDT-TS e dif-PSS. A Figura 5.19 também apresenta resultados relativamente ruins obtidos pelo algoritmo NSGA-ProtPred. A Tabela 5.5 confirma que o AEMT$_{ND}$ é mais adequado para proteína 1SOL, apresentando melhores RMSD e GDT-TS, apesar do dif-PSS ser maior que o obtido pelo mono-ProtPred.

Tabela 5.6: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 1A11.

Algoritmos	mono-ProtPred			AEMT$_{ND}$			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,541	4,902	0,312	0,51	5,305	0,44	0,36	7,380	0,920
2				0,51	5,294	0,44	0,42	7,461	0,680
3				0,77	1,843	0,08	0,42	7,461	0,680
4				0,49	1,843	0,32	0,42	7,461	0,680
5				0,79	1,834	0,08	0,36	5,591	0,920
6				0,79	1,766	0,08	0,41	7,436	0,680
7				0,79	1,807	0,08	0,42	5,487	0,920
8				0,5	5,491	0,04	0,41	5,614	0,920
9				0,49	1,834	0,08	0,42	5,487	0,680
10				0,7	5,526	0,36			
Média	0,541	4,902	0,312	**0,634***	**3,254***	**0,228***	0,404	6,592	0,786
Melhor indivíduo	0,541	4,902	0,312	**0,790***	**1,807***	**0,080***	0,404	6,592	0,786

Para proteína 1A11, a Figura 5.20 mostra que AEMT$_{ND}$ produz as melhores soluções, com estruturas mais similares à nativa, enquanto que o mono-ProtPred apresenta menor diversidade e estruturas bem diferentes da nativa. A Figura 5.20 também mostra resultados inferiores obtidos pelo

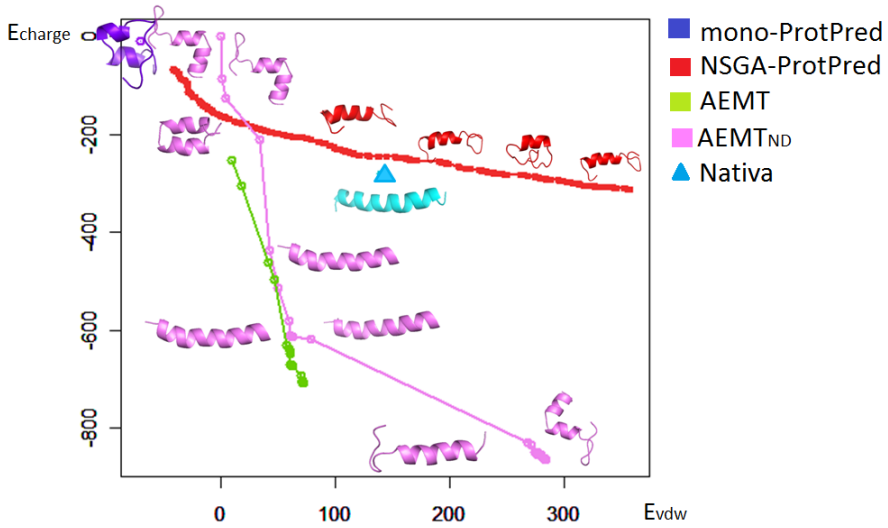

Figura 5.20: Fronteiras de Pareto para 1A11 com campos de força de van der Waals e eletrostática (Energia em kcal/mol).

NSGA-ProtPred. A Tabela 5.6 confirma numericamente que AEMT$_{ND}$ produziu mehores resultados para proteína 1A11, com menor de média de RMSD e dif-PSS, e mais alta média de GDT-TS.

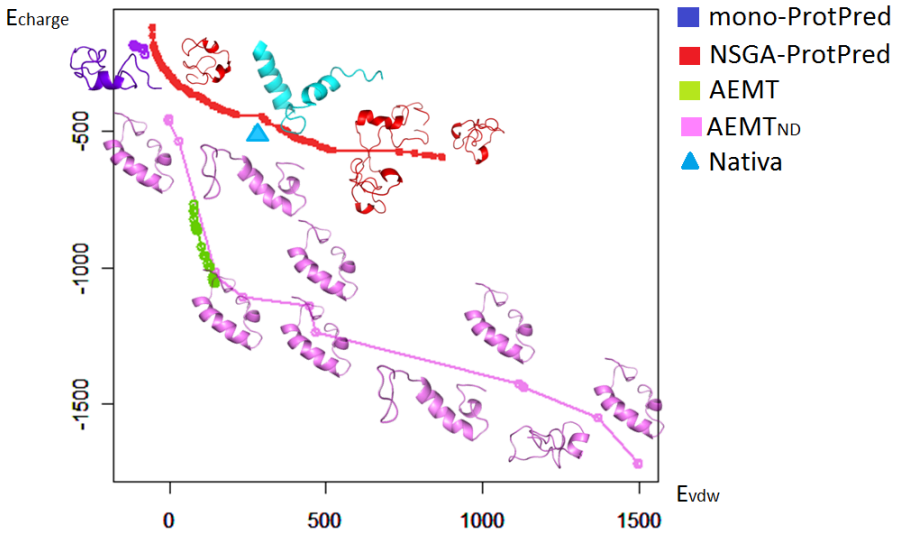

Figura 5.21: Fronteiras de Pareto para 2KOE com campos de força de van der Waals e eletrostática (Energia em Kcal/mol).

Considerando a proteína 2KOE, a Figura 5.21 revela que o AEMT$_{ND}$ gera uma diversidade maior de estruturas, que são também mais próximas à nativa, mostrando menor diversidade de estruturas na população final do mono-ProtPred, além de serem diferentes da nativa. A Tabela 5.7 confirma numericamente que AEMT$_{ND}$ é mais adequado para predição da proteína

Tabela 5.7: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 2KOE.

Algoritmos	mono-ProtPred			$AEMT_{ND}$			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,331	7,495	0,400	0,356	6,742	0,17	0,438	6,799	0,045
2	0,331	7,495	0,400	0,350	6,766	0,395	0,400	7,878	0,295
3	0,331	7,495	0,400	0,343	7,979	0,170	0,325	10,635	0,295
4				0,343	7,979	0,170	0,325	10,515	0,295
5				0,356	6,742	0,120	0,425	6,175	0,295
6				0,356	6,742	0,120	0,325	10,543	0,150
7				0,331	6,043	0,170	0,425	6,175	0,295
8							0,331	11,142	0,045
9							0,431	6,754	0,045
10							0,406	7,066	0,295
11							0,325	10,810	0,295
12							0,325	10,810	0,070
Média	0,331	7,495	0,400	0,348	**6,999***	**0,187***	**0,366***	9,128	0,214
Melhor indivíduo	0,331	7,495	0,400	0,331	**6,043***	**0,170***	**0,425***	6,175	0,295

2KOE, uma vez que esses valores de GDT-TS diferem pouco em relação aos obtidos pelo $AEMT_{ND}$.

Para proteína 2K7Y, a Figura 5.22 mostra que o $AEMT_{ND}$ apresenta maior diversidade na população final. O mono-ProtPred apresenta algumas predições mais satisfatórios visualmente, mesmo assim o dif-PSS obtido pelo $AEMT_{ND}$ ainda é melhor. A Tabela 5.8 mostra que os menores valores de média de RMSD e dif-PSS são obtidos pelo $AEMT_{ND}$, assim como a mais alta média de GDT-TS.

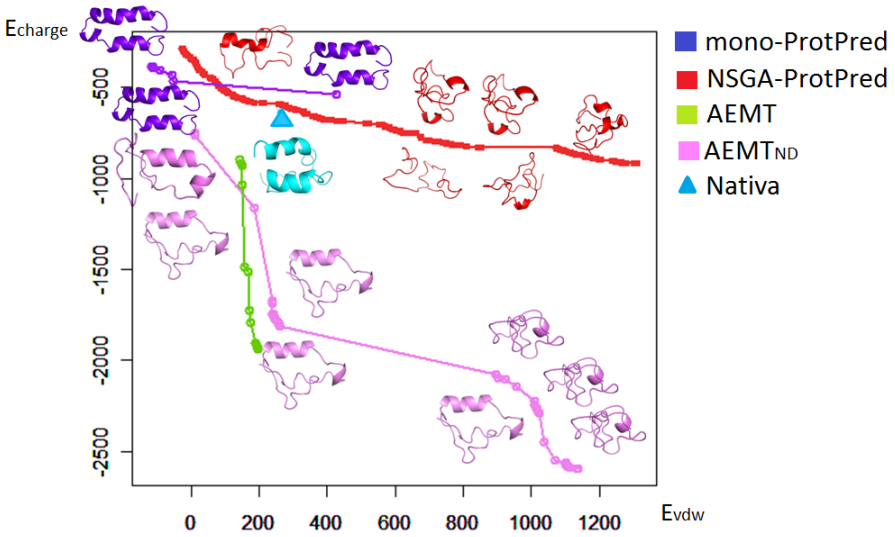

Figura 5.22: Fronteiras de Pareto para 2K7Y com campos de força de van der Waals e eletrostática (Energia em kcal/mol).

Por fim, a Tabela 5.9 sumariza os dados obtidos a partir dos quatro métodos – mono-ProtPred, NSGA-ProtPred, AEMT e $AEMT_{ND}$. Para isso, utiliza-se o hipervolume [9, 10], que tem sido o principal índice de desempenho na comparação de AEMOs. O hipervolume é a área (no caso biobjetivo para PSP) acima da fronteira no gráfico do espaço das funções. Por exemplo, a área obtida a partir da fronteira do $AEMT_{ND}$ é bem maior que a calculada

Tabela 5.8: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 2K7Y.

Algoritmos	mono-ProtPred			AEMT$_{ND}$			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,156	8,984	0,261	0,322	7,093	0,130	0,144	9,236	0,240
2	0,156	8,984	0,261	0,333	7,620	0,130	0,139	8,110	0,220
3	0,156	8,984	0,261	0,333	7,624	0,130	0,128	8,369	0,350
4	0,156	8,984	0,261	0,261	9,297	0,200	0,117	9,175	0,240
5				0,333	7,620	0,130	0,122	9,294	0,240
6				0,261	9,422	0,130	0,122	9,294	0,240
7				0,261	7,093	0,350	0,117	8,235	0,350
8				0,333	10,944	0,350	0,128	8,431	0,350
9							0,128	8,506	0,350
10							0,128	9,759	0,240
11							0,128	8,192	0,350
12							0,133	7,801	0,240
Média	0,156	8,984	0,261	**0,305***	**8,342***	**0,171***	0,129	8,663	0,284
Melhor indivíduo	0,156	8,984	0,261	**0,322***	**7,093***	**0,130***	0,133	7,801	0,240

Tabela 5.9: Síntese dos Hipervolumes calculados.

Medidas	Hipervolume (x10^3)			
Proteínas	NSGA-ProtPred	mono-ProtPred	AEMT	AEMT$_{ND}$
1SOL	59,852	2,081	22,515	**453,558***
1A11	68,264	0	11,307	**160,366***
2KOE	342,105	0,245	10,947	**1154,100***
2K7Y	594,830	40,811	32,877	**1966,427***

usando a fronteira do mono-ProtPred para todos os casos testados. Os hipervolumes obtidos com o AEMT$_{ND}$ são ordens de grandeza maiores que dos demais métodos investigados, conforme mostra a Tabela 5.9, evidenciando a superioridade do AEMT$_{ND}$ do ponto de vista de AEMOs.

No entanto, para predição de folhas-β notou-se um desempenho bastante inferior. A Figura 5.23 e a Tabela 5.10 revelam que os resultados dos quatro métodos não são muito diferentes entre si. Essa conclusão é corroborada pelo cálculo dos hipervolumes[5], mostrados na Tabela 5.11, em que se pode verificar que não há diferenças de ordem de grandeza entre o NSGA-ProtPred e o AEMT$_{ND}$, nem entre o mono-ProtPred e o AEMT. Uma interpretação possível para esse resultado é que a predição puramente *ab initio* de uma folha-β é tão mais complexa que nenhum dos métodos consegue amostrar adequadamente regiões interessantes do espaço das funções.

Buscando melhorar tal amostragem foi proposto o AEMMT (Seção 5.2). Verificou-se também a necessidade de modelar outras energias. Por exemplo, a energia de solvatação influencia significativamente no quanto duas hélices de uma estrutura se aproximam, como nos casos das proteínas 2KOE e 2K7Y. Por outro lado, as ligações de hidrogênio podem também ser fundamentais na formação de folhas-β, como no caso da proteína 1NIZ.

Esse desenvolvimento foi realizado utilizando a proteína 1NIZ (Subseção 5.3.3) para avaliação de desempenho do AEMMT.

[5]O cálculo de hipervolumes foi realizado usando o pacote *emoa* no aplicativo R.

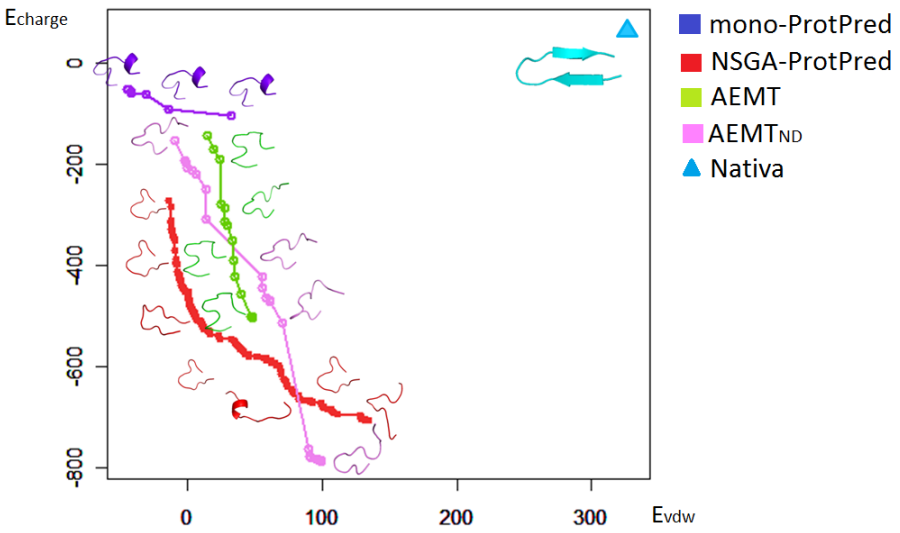

Figura 5.23: Fronteiras de Pareto para 1NIZ com campos de força de van der Waals e eletrostática (Energia em kcal/mol).

Tabela 5.10: GDT-TS, RMSD, dif-PSS calculados para estruturas preditas das fronteiras de 1NIZ.

Algoritmos	mono-ProtPred			AEMT			$AEMT_{ND}$			NSGA-ProtPred		
Estruturas	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS	GDT-TS	RMSD	dif-PSS
1	0,446	4,774	0,500	0,464	4,607	0,500	0,411	2,986	0,500	0,500	3,492	0,500
2	0,446	4,774	0,500	0,464	4,683	0,500	0,447	3,104	0,500	0,500	4,034	0,500
3	0,446	4,774	0,500	0,464	4,683	0,500	0,447	3,104	0,500	0,500	4,038	0,500
4	0,446	4,774	0,500	0,464	4,683	0,500	0,447	3,106	0,500	0,500	4,033	0,500
5				0,464	4,687	0,500	0,447	3,106	0,500	0,500	4,033	0,500
6				0,464	4,719	0,500	0,447	3,124	0,500	0,500	4,038	0,500
7				0,464	4,730	0,500	0,447	3,124	0,500	0,500	4,033	0,500
8				0,464	4,731	0,500	0,447	3,124	0,500	0,375	3,746	0,500
9				0,446	4,744	0,500	0,447	3,124	0,500	0,446	3,045	0,500
10				0,446	4,752	0,500	0,447	3,124	0,500	0,411	3,021	0,500
11										0,446	3,045	0,500
12										0,429	3,644	0,500
Média	0,446	4,774	0,500	0,460	4,702	0,500	0,443	**3,103***	0,500	**0,467***	3,683	0,500
Melhor indivíduo	0,446	4,774	0,500	**0,464***	4,607	0,500	0,411	**2,986***	0,500	0,411	3,021	0,500

Tabela 5.11: Hipervolumes calculados para 1NIZ.

Medidas	Hipervolume (x10^3)			
Proteína	NSGA-ProtPred	mono-ProtPred	AEMT	$AEMT_{ND}$
1NIZ	**47,082***	2,152	6,007	24,985

5.5 Avaliando o AEMMT em predição de folha-β

O AEMMT lida com quatro energias (van der Waals, eletrostática, solvatação e ligações de hidrogênio) por meio de dezesseis tabelas, que consideram todas as possibilidades de combinações dessas quatro energias mais uma tabela de soluções não-dominadas.

O AEMMT melhora significativamente a formação da estrutura da folha-β. A Figura 5.24 ilustra duas estruturas obtidas que, apesar de não possuírem ligações de hidrogênio entre as fitas da folha, assemelham-se à estrutura da

proteína nativa 1NIZ (Subseção 5.3.3).

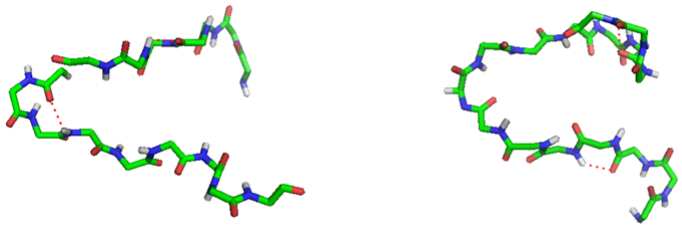

Figura 5.24: Exemplos de estruturas preditas para 1NIZ obtidas pelo AEMMT (quatro energias e dezesseis tabelas).

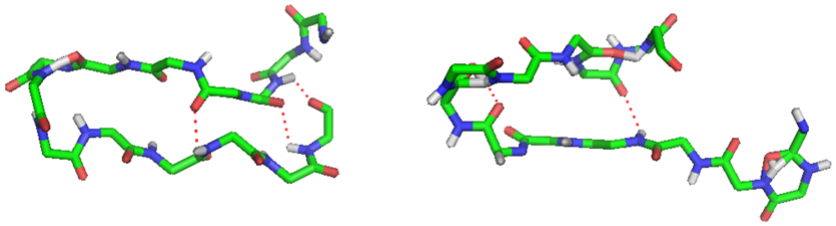

Figura 5.25: Exemplos de estruturas preditas para 1NIZ com AEMMT (quatro energias e dezessete tabelas.

Para se conseguir que o AEMMT "percebesse" que as fitas estavam relativamente próximas de formarem ligações de hidrogênio, adicionou-se um segundo Hamiltoniano para essa energia no modelo do AEMMT. Tal função considera que há energia de ligação em nível não desprezível entre átomos de O e H a mais de 6 Å (Seção 3.4). Para isso, é preciso incluir uma nova tabela na estrutura de subpopulações do AEMMT, totalizando dezessete tabelas. A Figura 5.25 mostra que a melhor estrutura obtida nos primeiros testes usando o AEMMT com dezessete tabelas (um algoritmo puramente *ab initio*) foi muito próxima à nativa, com RMSD igual a 2.646 Å.

5.6 Considerações finais

Baseado nos experimentos realizados com os algoritmos mono-ProtPred, AEMT, AEMT$_{ND}$ e AEMMT pode-se obter as seguintes conclusões:

(*i*) O mono-ProtPred consegue gerar predições adequadas para proteínas relativamente simples, considerando que o espaço para essas proteínas é

significativamente menor. Isso se explica pelo fato da simples ponderação das energias não ser suficiente para soluções saírem de regiões de ótimos locais;

(*ii*) O AEMT consegue aumentar a capacidade de busca. No entanto, mesmo podendo obter resultados melhores que o mono-ProtPred, observa-se que para estruturas mais complexas (2KOE e 2K7Y) as predições ainda não são satisfatórias e a diversidade de estruturas também não é tão signiticativa;

(*iii*) O AEMT$_{ND}$ consegue melhorar os resultados com as proteínas 1SOL, 2KOE e 2K7Y, enquanto que para proteína 1A11 não se obteve resultados melhores comparados à versão AEMT. Porém, visualmente as predições para 2KOE e 2K7Y ainda não apresentaram resultados interessantes, além de não conseguir predizer a folha-β da 1NIZ;

(*iv*) O AEMMT surge com o intuito de predizer proteínas mais complexas, com folhas-β. O AEMMT, de fato, consegue predizer a folha-β da proteína 1NIZ. Dada essa capacidade, supõe-se que o AEMMT pode melhorar as predições para uma diversidade de proteínas, como, por exemplo, proteínas com mais de um domínio.

O desempenho do AEMMT induz a conclusão de que a maneira mais adequada para trabalhar com os objetivos (energias) no problema de PSP é representá-los por meio de seus vários Hamiltonianos e as várias combinações desses. Por exemplo, no caso da proteína 1NIZ, não se sabia qual Hamiltoniano seria o mais relevante, o modelo proposto por Frishman e Argos [70] ou a variação proposta na Seção 3.4. Então, utilizou-se mais de um Hamiltoniano para modelar a mesma energia, obtendo a predição de uma folha-β. Tal predição puramente *ab initio* pode ser considerada inédita, posto que não foi encontrada outra na literatura. Os resultados preliminares obtidos com o AEMMT motivaram uma melhor investigação do quanto esse método pode contribuir para PSP. O Capítulo 6 explora vários aspectos do PSP por meio do AEMMT.

CAPÍTULO

6

Experimentos com AEMMT para PSP

6.1 Considerações iniciais

Este Capítulo mostra os experimentos realizados usando o AEMMT para PSP puramente *ab initio* e os resultados obtidos a partir dos mesmos. A Seção 6.2 descreve os parâmetros usados para execução do AEMMT e as proteínas escolhidas para os experimentos. A Seção 6.3 apresenta alguns casos bem sucedidos com o AEMMT, com um domínio (com hélice-α e folha-β) e dois domínios (com folhas-β) com pouca interação entre eles, enquanto que a Seção 6.4 descreve casos em que existe interação forte entre os domínios. Por fim, a Seção 6.5 mostra a contribuição relativa das subpopulações durante o processo de evolução por meio do AEMMT.

6.2 Descrição dos experimentos com AEMMT

Esse método trabalha com muitos objetivos, usando a combinação dos potenciais de energia representados por tabelas (subpopulações). Os potenciais de energia utilizados nesses testes são: van der Waals, eletrostática, solvatação e ligação de hidrogênio. Desse modo, o AEMMT considera dezesseis tabelas, onde quatro tabelas representam as quatro energias, uma tabela representa a ponderação total das quatro energias, uma tabela representa os indivíduos não-dominados de cada geração, e o restante (dez tabelas) representa a combinação das energias de interação (quatro combinações triplas e seis duplas), conforme fora descrito na Seção 5.2. Dessa forma, consegue-se melhor amostrar o espaço das funções objetivo envolvendo as

quatro energias modeladas e, consequentemente, melhor estimar a fronteira ótima de Pareto.

As proteínas foram selecionadas a partir do banco de proteínas *Protein Data Bank* - PDB, com o critério de suas estruturas terem sido determinadas com RNM, pois assim o processo de determinação da estrutura nativa considerou, a princípio, as proteínas imersas em solvente. Esse aspecto é de interesse, uma vez que AEMMT considera um modelo de energia de solvatação. Além disso, as proteínas escolhidas variam de tamanho, podendo ser bem pequenas e simples com apenas um domínio, como 2RLG [20], 1SOL [179], 2XL1 [17], 1NIZ [160] e 2EVQ [6], e outras um pouco maiores com mais de um domínio, como 1G26 [177], 2KOE [180], 2K7Y [187] e 1CRN [174]. Com isso, busca-se analisar como se comporta o AEMMT com diferentes estruturas, com um domínio ou mais. Esses experimentos demonstram melhorias alcançadas pelo método AEMMT em predição puramente *ab initio*, assim como mostram os desafios encontrados durante essa investigação com algumas proteínas.

Para todos os experimentos, o AEMMT foi executado com 640 indivíduos (40 indivíduos em cada subpopulação) e 8000 gerações (com exceção da proteína 1G26, que os melhores resultados foram obtidos com 100000 gerações). Os pesos relativos às energias na função ponderação foram: $p_{vdw} = 1.0$; $p_{charge} = 0.5$; $p_{solv} = 0.5$ e $p_{hbond} = 0.5$ quando as estruturas eram hélices-α e $p_{vdw} = 1.0$; $p_{charge} = 0.5$; $p_{solv} = 1.0$ e $p_{hbond} = 2.0$ para folhas-β. Foram realizadas dez execuções para cada caso, escolhendo a execução com o menor RMSD médio da população final. O AEMMT utiliza o pacote do CHARMM [30, 125][1] para configurar os parâmetros de campos de força.

6.3 Casos bem sucedidos com AEMMT

Observando os resultados apresentados na Tabela 6.2, pode-se notar que as soluções encontradas pelo método AEMMT foram melhores comparadas aos outros métodos, os menores valores de RMSDs obtidos combinados aos maiores valores de GDT-TS para quase todas as predições confirmam o desempenho.

Tabela 6.1: Exemplos de predições bem sucedidas com o AEMMT.

Proteínas	Nº de Aminoácidos	Domínios
2EVQ	12	1 Folha-β
1NIZ	16	1 Folha-β
2RLG	18	1 Hélice-α
1SOL	20	1 Hélice-α
2XL1	24	1 Hélice-α
1G26	31	2 Folhas-β

[1] Arquivo editável charmm27.prm.

Tabela 6.2: Melhores RMSDs e GDT-TS das proteínas comparando os métodos mono-ProtPred, NSGA-ProtPred, AEMT$_{ND}$ e AEMMT.

	mono-ProtPred		NSGA-ProtPred		AEMT$_{ND}$		AEMMT	
Proteínas	RMSD	GDT-TS	RMSD	GDT-TS	RMSD	GDT-TS	RMSD	GDT-TS
2EVQ	4,625	0,604	2,762	0,708	1,499	0,792	**0,758***	**0,8125***
1NIZ	3,864	0,3929	2,823	**0,411***	2,107	0,393	**1,643***	**0,411***
2RLG	4,281	0,556	3,569	0,611	1,866	0,653	**1,550***	**0,736***
1SOL	5,414	**0,637***	4,319	0,587	1,509	**0,637***	**0,977***	0,625
2XL1	4,694	0,437	2,055	0,437	1,281	0,531	**1,116***	**0,767***
1G26	9,544	0,355	6,202	**0,452***	4,202	0,315	**3,861***	**0,452***

Por meio da inspeção visual, pode-se confirmar que, de fato, as estruturas preditas pelo AEMMT estão bem próximas das conformações nativas correspondentes. Em alguns casos os menores valores de RMSD não representam as melhores predições, por isso, para algumas proteínas mostram-se mais de uma solução, por considerar outras estruturas preditas, com RMSDs um pouco maiores, também interessantes, como é o caso da 2EVQ, 1NIZ e 1G26.

As Figuras 6.1, 6.2, 6.3, 6.4, 6.5 e 6.6 contrapõe as estruturas nativa e preditas envolvendo hélices-α.

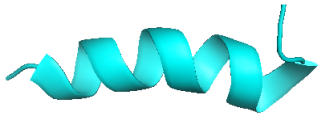

Figura 6.1: Estrutura nativa da proteína 2RLG.

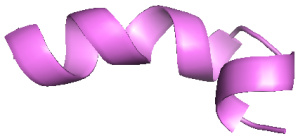

Figura 6.2: Estrutura predita para 2RLG pelo AEMMT.

Figura 6.3: Estrutura nativa da proteína 1SOL.

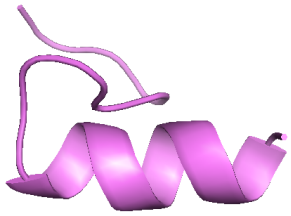

Figura 6.4: Estrutura predita para 1SOL pelo AEMMT.

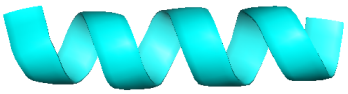

Figura 6.5: Estrutura nativa da proteína 2XL1.

Figura 6.6: Estrutura predita para 2XL1 pelo AEMMT.

As Figuras 6.7, 6.8, 6.9 e 6.10 mostram as estruturas nativa e preditas com uma folha-β. Por fim, as Figuras 6.11 e 6.12 possibilitam comparar as estruturas nativa e preditas para uma proteína com duas folhas-β.

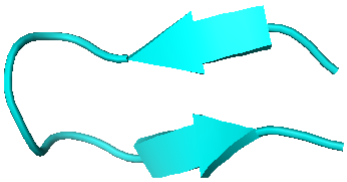

Figura 6.7: Estrutura nativa da proteína 2EVQ.

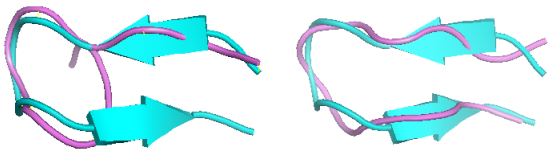

Figura 6.8: Estruturas preditas para 2EVQ pelo AEMMT.

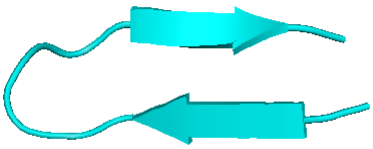

Figura 6.9: Estruturas nativas da proteína 1NIZ.

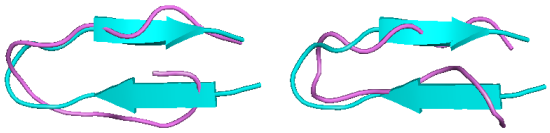

Figura 6.10: Estruturas preditas para 1NIZ pelo AEMMT.

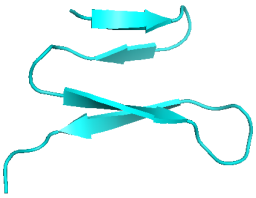

Figura 6.11: Estrutura nativa da proteína 1G26.

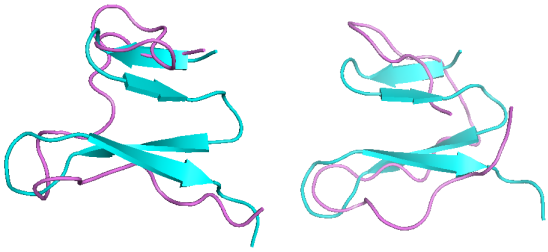

Figura 6.12: Estruturas preditas para 1G26 pelo AEMMT.

Pode-se observar que o AEMMT foi capaz de predizer as estruturas secundárias com apenas um domínio com resultados satisfatórios, mesmo sendo predições de forma puramente *ab initio*. É importante destacar também a contribuição deste trabalho ao conseguir determinar tanto hélices-α quanto folhas-β.

Do ponto de vista computacional, esses experimentos mostram a capacidade do AEMMT em alcançar regiões no espaço de busca, nunca antes atingidas por métodos puramente *ab initio*, regiões em que ocorre formação de folhas-β. Pode-se dizer que predizemos uma folha-β tanto pelos valores obtidos nas métricas de similaridade (como por exemplo, a proteína 2EVQ apresenta RMSD menor que 1 Å(0,758) e GDT-TS bem próximo de 1 (0,812),

que são valores encontrados em predições consideradas de boa precisão) quanto por inspeção visual, em que claramente pode-se observar a forma de "grampo" em torno da estrutura nativa.

O grande número de tabelas do AEMMT melhora a exploração do espaço de busca, bem como estima melhor a fronteira de Pareto no espaço das funções objetivo. Para a proteína com duas folhas-β (1G26), conseguiu-se estruturas interessantes, porém com RMSDs maiores (acima de 3 Å). Isso parece ser devido a aspectos de interações entre os domínios que precisam ser ainda modelados no AEMMT. Por exemplo, há duas ligações de dissulfeto na 1G26 e esse tipo de interação não é modelado nos Hamiltonianos de nenhuma das tabelas do AEMMT. A Seção 6.4 avalia o AEMMT considerando proteínas em que há interações significativas entre domínios.

6.4 Casos com interações entre domínios

Algumas proteínas apresentam características particulares que não foram consideradas ainda na implementação do AEMMT, posto que os Hamiltonianos trabalhados ainda não modelam alguns efeitos específicos da organização molecular de certas proteínas. Algumas proteínas desse tipo foram analisadas (Tabela6.3), e os resultados são sintetizados na Tabela 6.4.

Tabela 6.3: Proteínas com interações entre domínios analisadas.

Proteínas	N^o de Aminoácidos	Domínios
2KOE	40	2 Hélices-α
2K7Y	45	2 Hélices-α
1CRN	46	2 Hélices-α + 1 Folha-β

Tabela 6.4: Melhores RMSDs e GDT-TS das proteínas comparando os métodos mono-ProtPred, NSGA-ProtPred, AEMT$_{ND}$ e AEMMT.

	mono-ProtPred		NSGA-ProtPred		AEMT$_{ND}$		AEMMT	
Proteínas	RMSD	GDT-TS	RMSD	GDT-TS	RMSD	GDT-TS	RMSD	GDT-TS
2KOE	7,487	0,331	6,175	0,388	5,191	**0,431***	**4,987***	0,406
2K7Y	8,984	0,156	7,801	0,133	7,093	0,150	**5,333***	**0,161***
1CRN	6,959	0,293	7,635	0,310	6,908	0,294	**6,000***	**0,375***

Os aspectos de modelagem de interações entre elementos de uma proteína não considerados no desenvolvimento do AEMMT são os seguintes:

1. Ligações dissulfeto em sua organização molecular, como é o caso da 1CRN;

2. Ligações de hidrogênio entre domínios, como é o caso das proteínas 2KOE (com uma pequena folha-β com dois aminoácidos entre as hélices-α) e 1CRN;

3. Estruturas secundárias muito próximas na mesma molécula (entre 4Å e 7Å), como é o caso das proteínas 2K7Y e 1CRN.

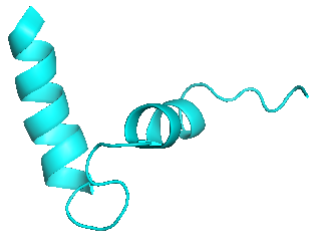

Figura 6.13: Estrutura nativa da proteína 2KOE.

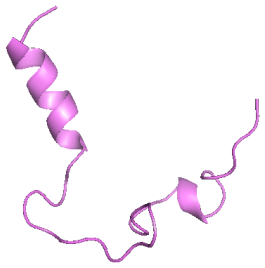

Figura 6.14: Estrutura predita para 2KOE pelo AEMMT.

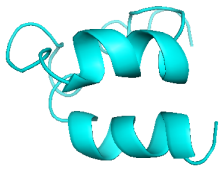

Figura 6.15: Estrutura nativa da proteína 2K7Y.

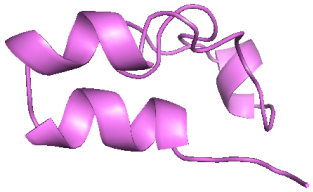

Figura 6.16: Estrutura predita para 2K7Y pelo AEMMT.

Figura 6.17: Estrutura nativa da proteína 1CRN.

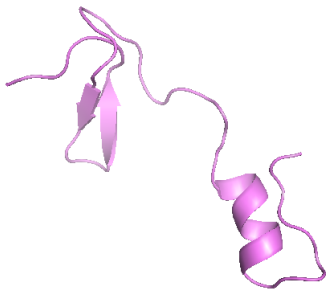

Figura 6.18: Estrutura predita para 1CRN pelo AEMMT.

Apesar de se obter RMSDs acima de 4Å para esse conjunto de proteínas, o AEMMT consegue melhores resultados quando comparados aos métodos mono-ProtPred, NSGA-ProtPred e AEMT$_{ND}$. A proteína 1CRN apresenta as três características que ainda não foram tratadas pelo AEMMT, justificando a predição menos satisfatória entre todas realizadas. Assim, a modelagem futura no AEMMT desses aspectos representa um desafio para nosso grupo de pesquisa e demais interessados em utilizar melhores métodos *ab initio*.

Neste contexto, pretende-se incluir Hamiltonianos que modelem interações como as ligações de dissulfeto. Outra proposta para solucionar a questão de proteínas com domínios com interações fortes é a implementação de

um algoritmo de estimação de distribuição [151, 114, 140, 76] que possa identificar durante a execução do método de predição os domínios que constituem a molécula. Com isso, poderia-se evitar que tais domínios (blocos construtivos do ponto de vista do algoritmo de estimação de distribuição) sejam danificados devido a ligações inadequadas com outros domínios. Naturalmente, outra proposta seria também combinar o AEMMT com conhecimento a priori obtido de homologia entre sequências, como tem sido realizado por abordagens semi *ab initio*, como Rosetta [19, 53, 52, 116] e I-TASSER [193, 149, 189].

6.5 Mimetização do dobramento da proteína

A partir dos experimentos com o AEMMT com dezesseis subpopulações, pode-se perceber a importância das subpopulações durante o processo de evolução. Desse modo, pode-se mimetizar, de certa forma, o processo de dobramento de uma proteína [8, 107], uma vez que as subpopulações representam os potenciais de energias, ou as combinações das mesmas, possivelmente envolvidas na estabilização das estruturas protéicas. Assim, pelo cálculo de quais subpopulações foram mais escolhidas a cada estágio evolutivo do AEMMT para predição de uma proteína, obtem-se a proporção das subpopulações que mais contribuíram com as gerações de novos indivíduos de sucesso (que sobrevivem na próxima geração).

A Figura 6.19 mostra a contribuição relativa (em porcentagem) de cada tabela durante o processo evolutivo do AEMMT para proteína 1SOL. Conforme ocorre essa evolução, a proteína tende a ficar mais enovelada, assim a contribuição relativa de cada tabela é uma estimativa da importância referente aos diferentes Hamiltonianos em cada etapa do dobramento da proteína.

Diferentemente da 1SOL (estrutura de hélice-α), as contribuições relativas para 2EVQ (estrutura de folhas-β) mostram outra distribuição dessas contribuições até a geração 4000. Por exemplo, pode-se notar no caso da 1SOL (Figura 6.19) que a contribuição do potencial de van der Waals é evidentemente mais acentuada que as outras energias, principalmente no início da evolução. Outros potenciais também são relevantes na predição por meio das combinações $vdw + hbond$ (combinação 8), e $vdw + solv$ (combinação 7), ressaltando que em ambas as combinações o potencial de van der Waals está presente.

Pode-se verificar que a energia de van der Waals tem grande importância no processo evolutivo do AEMMT, ocorrendo esse fenômeno tanto para hélice-α (1SOL) quanto para folha-β (2EVQ). No entanto, quando as proteínas são formadas por hélice, a contribuição da van der Waals predomina até ao final

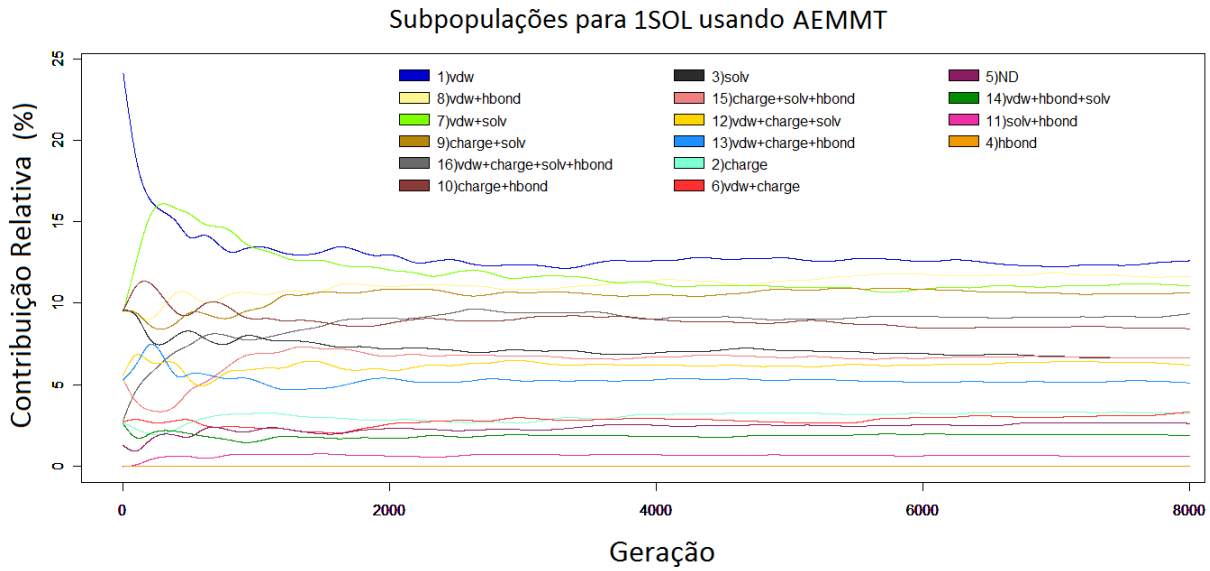

Figura 6.19: Contribuições relativas das subpopulações para o dobramento da proteína 1SOL durante a evolução do AEMMT.

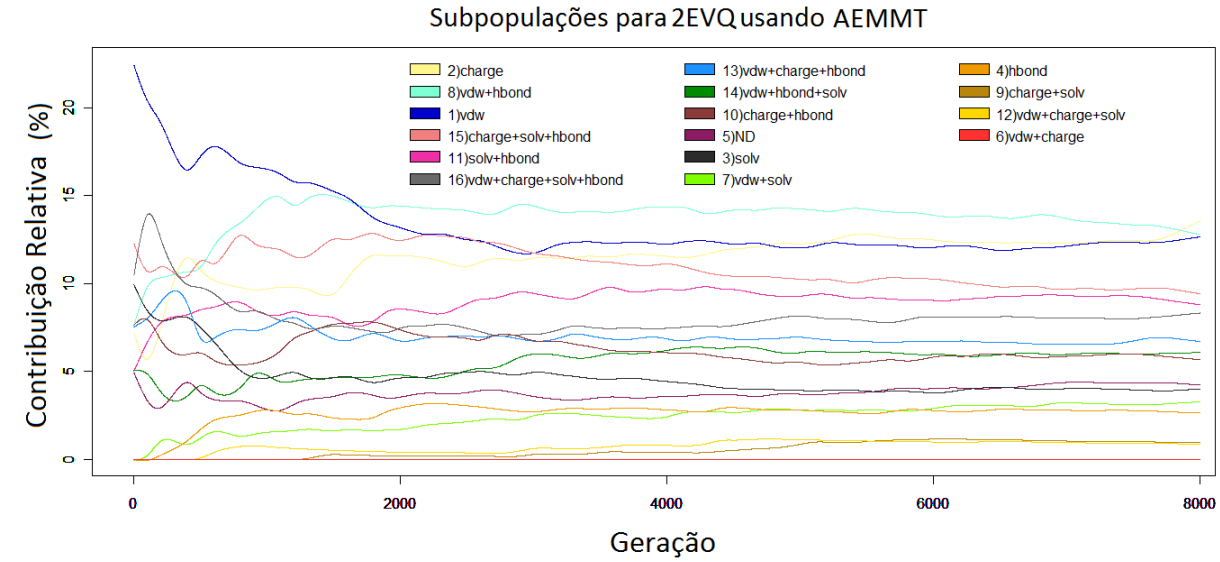

Figura 6.20: Contribuições relativas das subpopulações para o dobramento da proteína 2EVQ durante a evolução do AEMMT.

da execução (Figura 6.19), enquanto que para folha-β a sua contribuição varia mais ao longo do processo evolutivo (Figura 6.20).

Como se pode notar na Figura 6.20, à medida que o potencial de van der Waals decai sua contribuição, as combinações $vdw + hbond$ (combinação 8), $charge + solv + hbond$ (combinação 15) e $solv + hbond$ (combinação 11) crescem suas contribuições relativas na predição. A energia de solvatação acarreta o efeito do dobramento, contribuindo para que se gerem estruturas com forma de "grampos", enquanto que a energia de ligação de hidrogênio acentua o paralelismo de duas fitas. Logo, essas combinações mostram-se, de fato, importantes na formação de folhas-β. É importante notar que a energia de ligação de hidrogênio considerada separadamente não gera predições satisfatórias, confirmando que a contribuição dessa energia ocorre após forte influência do potencial de van der Waals, eletrostático e solvatação.

As contribuições relativas das subpopulações também explicam, de certa forma, porque trabalhar com AEMMT para o problema de PSP pode ser mais vantajoso do que o mono-ProtPred, AEMT$_{ND}$ ou NSGA-ProtPred, uma vez que:

1. O AEMMT possibilita estimar a importância de cada energia ao longo da evolução, seguindo um processo que se assemelha ao dobramento da proteína;

2. As contribuições relativas mudam também conforme o tipo de estrutura secundária (hélice-α e folha-β), constatando que as energias contribuem de maneira diferente para cada tipo de estruturas;

3. Mesmo sem saber a priori qual tipo de estrutura secundária (e portanto, a influência – peso – dessas energias ao longo do dobramento), o AMMT consegue predizer tanto hélices-α quanto folhas-β.

6.6 Considerações finais

Neste capítulo, foi possível mostrar que o AEMMT obtém resultados mais abrangentes quando comparados aos métodos mono-ProtPred, NSGA-ProtPred e AEMT$_{ND}$. O modo como foram tratados os objetivos no AEMMT, realizando a combinação entre eles, possibilita melhor amostrar o espaço de busca e a fronteira de Pareto no espaço de funções objetivo.

As proteínas precisam de um conjunto de combinações de energias para sua estabilização, ressaltando que esse conjunto pode ser específico para cada proteína. Além de não se saber qual o melhor conjunto dessas combinações para cada proteína, não se sabe a ordem das contribuições ao longo do processo de dobramento.

Os Hamiltonianos disponíveis na literatura para uma mesma energia também possuem um certo grau de conflito, conforme o caso dos modelos de ligações de hidrogênio (modelo proposto em [69] e as variações a partir de outro modelo desenvolvido em [70] - Seção 3.4). O AEMMT possibilita que os Hamiltonianos mais relevantes para predição de cada tipo de proteína sejam escolhidos ao longo do processo evolutivo. Dessa forma, o AEMMT contorna aspectos difíceis de modelagem em PSP por meio do uso de várias tabelas considerando uma ampla possibilidade de Hamiltonianos relevantes.

Outro ponto positivo é a capacidade de predizer folhas-β, sem nenhum conhecimento a priori. A partir desse resultado, trabalhos futuros podem ser desenvolvidos buscando predizer proteínas envolvendo folhas-β com mais domínios e com mais acurácia nos resultados.

CAPÍTULO

7

Conclusão

O problema de PSP apresenta diversos desafios que têm sido enfrentados por colaboradores de várias áreas envolvendo estudos sobre proteínas. Neste contexto, a predição por um método computacional puramente *ab initio* apresenta entraves adicionais, uma vez que o espaço de busca cresce exponencialmente com o tamanho da proteína. Além disso, este trabalho apresenta evidências que as diferenças entre Hamiltonianos presentes na literatura podem afetar significativamente a qualidade das predições. O impacto dessas diferenças mostra-se também dependente do tipo de domínios presentes em cada molécula. Esses aspectos sumarizam a complexidade do problema de PSP modelado de forma puramente *ab initio*.

Porém, esse tipo de abordagem do problema não possui somente desvantagens. Do ponto de vista computacional, sabe-se que métodos que exploram de forma mais inteligente o espaço de busca e, em geral, obtém soluções significativamente melhores ou de forma mais eficiente, podem melhor aproveitar conhecimento a priori sobre o domínio de um problema (como informações de base de dados de proteínas) e produzir abordagens mais relevantes.

Evidentemente, o nível de contribuição do método de otimização para PSP em relação a contribuição que uma base de conhecimento a priori sobre problema pode produzir é uma questão por si só relativamente complexa. Há estudos sobre este aspecto, conforme pode ser visto em [131]. Por outro, resultados bem estabelecidos em Ciência da Computação sobre complexidade de algoritmos envolvendo problemas combinatórios e também multimodais, mostram que mesmo modificações sutis na formulação do problema, por meio de relaxações em suas restrições ou mesmo pelo reprojeto de suas estruturas

de dados, podem reduzir significativamente sua complexidade em relação à formulação original.

Sobre essa óptica, este trabalho explorou a modelagem multiobjetivo para o problema de PSP. É trivial formular esse problema como multiobjetivo, pois basta considerar, por exemplo, cada potencial de interação molecular como um objetivo do problema. Porém, as vantagens dessa nova modelagem não são tão evidentes. Um estudo rápido sobre o assunto pode concluir apenas que se estaria aumentando a complexidade do problema com o aumento do número de objetivos, o que de fato não ocorre em relação ao AEMO desenvolvido neste trabalho: o AEMMT.

A Seção 7.1 descreve as contribuições obtidas neste trabalho, do ponto de vista computacional e de modelagem molecular.

7.1 Contribuições relevantes

A investigação realizada neste trabalho construiu um novo AEMO, o AEMMT, que se beneficia da inclusão de critérios adicionais de avaliação das soluções, produzindo melhores predições. O AEMMT revela também a contribuição relativa de cada critério no processo evolutivo do algoritmo. Por meio dessa análise, evidencia-se que proteínas com domínios distintos podem requerer as maiores contribuições de potenciais diferentes. Em outras palavras, um conjunto consistente de Hamiltonianos precisa ser construído de forma que se possa lidar com um espectro amplo de tipos de proteínas.

Nesse sentido, os modelos computacionalmente eficientes de energia de solvação e de ligação de hidrogênio desenvolvidos neste trabalho, mostram-se fundamentais para a predição de folhas-β de forma puramente *ab initio*. O AEMMT determina durante a execução a contribuição relativa entre as energias modeladas, possibilitando adequá-las de acordo com o tipo de estruturas secundárias da molécula, bem como com o estágio de enovelamento (dobramento) ao longo da evolução.

Essa capacidade possibilitou o AEMMT também encontrar mais de um domínio em proteínas mais complexas. Destaca-se nesse sentido a aproximação da estrutura da proteína 1G26 (que possui duas folhas-β) com RMSD de 3.861 Å, mesmo sem modelar o efeito de ligações de dissulfeto presentes nessa estrutura. Para proteínas relativamente menos complexas, o AEMMT também apresentou contribuições significativas, aumentando a qualidade das predições. Por exemplo, o AEMMT gerou estruturas com RMSD de cerca de 1 Å para 2RLG, 1SOL e 2XL1, que possuem uma hélice-α.

O AEMMT destaca-se por lidar com um número maior do que dez critérios. Em geral, análises para avaliar o desempenho de AEMOs de

muito objetivos [99] lidam com instâncias de problemas com número de objetivos variando de três a dez [101, 99]. Os experimentos com o AEMMT envolveram até dezessete critérios, beneficiando-se do número maior de critérios, e melhorando a qualidade das soluções encontradas com aumento relativamente pequeno de custo computacional.

Uma possível explicação para esse comportamento é que os dezessete critérios utilizados no AEMMT estão em um espaço de funções de dimensão dezessete, mas somente a nível computacional, pois o problema físico é formulado considerando potencias de van der Waals, eletrostático e as energias de solvatação e ligação de hidrogênio. Assim, os dezessete critérios de certa forma "projetam-se" no espaço das quatro funções objetivos do problema físico melhor amostrando esse espaço.

Essas características mostram que o AEMMT pode ser um método interessante para ser investigado em um diversidade de problemas combinatórios envolvendo muitos objetivos ou em problemas para os quais uma formulação de muitos objetivos gere informações relevantes, beneficiando processo de busca. Um caso mais evidente ocorre em problemas em que há incertezas ou imprecisões em relação aos critérios utilizados. O AEMMT pode trabalhar à princípio com os diversos critérios e variações desses buscando realizar uma análise mais robusta de cada solução, sem com isso prejudicar sua eficiência computacional.

7.2 Trabalhos futuros

A flexibilidade de formular e resolver problemas complexos gerada pelo AEMMT indica que sua contribuição pode ser estendida não somente para problemas de outras áreas, mas também pode ser um novo recurso para se investigar diversos aspectos interessantes de PSP. Dentre eles, podem-se destacar:

- Investigar conjuntamente em uma mesma execução do AEMMT as contribuições relativas de Hamiltonianos distintos para ligação de hidrogênio: propostas de [69], de [70] e a desenvolvida neste trabalho (Seção 3.4);

- Analisar variações de Hamiltonianos para outros potencias, por exemplo, considerando modelagens diferentes entre bibliotecas computacionais para dinâmica molecular;

- Considerar parâmetros de campos de força distintos em uma mesma execução do AEMMT. Com isso, poderia-se analisar o efeito de diferentes

valores de diéletrico ao longo do processo evolutivo do AEMMT e em relação a estágios de enovelamento de tipos diferentes de proteínas;

- Incluir modelos para ligações de dissulfeto e outras interações moleculares que se mostrem relevantes conforme o tipo de proteínas investigado. Vale ressaltar que a modelagem de interações moleculares é uma tarefa muito difícil, especialmente para pesquisadores com formação em Ciência da Computação, requerendo de um estudo profundo na área de Biologia Molecular, como foi realizado neste trabalho para as ligações de hidrogênio;
- Avaliar as estruturas durante a execução pelo índice GDT-TS, uma vez que para isso basta a inclusão de mais uma tabela no AEMMT e tal critério pode induzir o método a amostrar em regiões mais corentes segundo esse índice.

Referências Bibliográficas

[1] Multiobjective meta-heuristics: An overview of the current state-of-the-art. *European Journal of Operational Research*, 137(1):1–9, 2002. Citado na página 49.

[2] S. Adra and P. Fleming. Diversity management in evolutionary many-objective optimization. *Evolutionary Computation, IEEE Transactions on*, 15(2):183–195, 2011. Citado na página 59.

[3] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Fundamentos da Biologia Celular*. Artmed Editora, 2007. Citado na página 64.

[4] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford Clarendon Press, 1987. Citado na página 30.

[5] G. L. Amidon, S. H. Yalkowsky, S. T. Anik, and S. C. Vabni. Solubility of nonelectrolytes in polar solvents. v. estimation of the solubility of aliphatic monofunctional compounds in water using a molecular surface area approach. *J. Phys. Cm.*, 79:2239, 1975. Citado na página 31.

[6] N. H. Andersen, K. A. Olsen, R. M. Fesinmeyer, X. Tan, F. M. Hudson, L. A. Eidenschink, and S. R. Farazi. Citado na página 86.

[7] F. B. Andrade, A. C. B. Delbem, C. R. S. BRASIL, D. R. F. Bonetti, and V. V. de Melo. Visualization of the evolutionary process of protein structure prediction and its potential energy. *BIOMAT 2010 – Institute for Advanced Studies of Biosystems*, 2010. Citado nas páginas 2 e 3.

[8] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96):223–230, 1973. Citado nas páginas 11 e 91.

[9] J. Bader and E. Zitzler. HypE: An Algorithm for Fast Hypervolume-Based Many-Objective Optimization. *Evolutionary Computation*, 2009. Citado na página 80.

[10] J. Bader and E. Zitzler. Robustness in Hypervolume-based Multiobjective Search. TIK Report 317, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, 2010. Citado na página 80.

[11] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, , and W. Nadler. A new monte carlo algorithm for protein folding. *Physical Review Letters*, 80:3149, 1998. Citado nas páginas 15 e 16.

[12] A. Baxevanis and B. Ouellette. Bioinformatics: A practical guide to the analysis of genes and proteins. *Computers & Chemistry*, 26(5):549–551, 2001. Citado na página 14.

[13] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry, Fifth Edition : International Version*. W. H. Freeman, February 2002. Citado nas páginas 4, 28, 38 e 44.

[14] B. Berger and F. T. Leighton. Protein folding in the hydrophobic-hydrophilic (*p*) is np-complete. In *RECOMB*, pages 30–39, 1998. Citado nas páginas 11, 13 e 16.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000. Citado na página 14.

[16] H. G. Beyer. *The theory of evolution strategies*. Springer, 2001. Citado na página 52.

[17] S. Bhushan, H. Meyer, A. L. Starosta, T. Becker, T. Mielke, O. Berninghausen, M. Sattler, D. N. Wilson, and R. Beckmann. Structural basis for translational stalling by human cytomegalovirus and fungal arginine attenuator peptide. *Molecular Cell*, 40(1):138–146, 2010. Citado na página 86.

[18] D. R. F. Bonetti. Aumento da eficiência do cálculo da energia de van der waals em algoritmos genéticos para predição de estruturas de proteínas. *Mestrado em Ciências Matemáticas e de Computação - Universidade de São Paulo*, 2010. Citado nas páginas 2, 3 e 63.

[19] R. Bonneau and D. Baker. Ab initio protein structure prediction: Progress and prospects. *Annu. Rev. Biophys. Biomol. Struct*, 30:173–189, 2001. Citado nas páginas 1, 2, 16 e 91.

[20] S. Bourbigot, E. Dodd, C. Horwood, N. Cumby, L. Fardy, W. H. Welch, Z. Ramjan, S. Sharma, A. J. Waring, M. R. Yeaman, and V. Booth. Citado na página 86.

[21] J. Bowie and D. Eisenberg. An evolutionary approach to folding small alfa-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci.*, 91:4436–4440, 1994. Citado nas páginas 11 e 15.

[22] K. Braden. A simple approach to protein structure prediction using genetic algorithms. In J. R. Koza, editor, *Genetic Algorithms and Genetic Programming at Stanford 2002*, pages 36–44. Stanford Bookstore, Stanford, California, 94305-3079 USA, June 2002. Citado na página 17.

[23] P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005. Citado na página 1.

[24] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, 2 edition, 1999. Citado nas páginas 8, 9, 11 e 12.

[25] C. R. S. Brasil, D. R. F. Bonetti, E. F. Faria, A. C. B. Delbem, and F. L. B. Silva. Evolutionary algorithms approaches for the protein structure prediction with new criteria. In *BIOMAT 2009 - International Symposium on Mathematical and Computational Biology*, 2009. Citado nas páginas 2 e 3.

[26] C. R. S. Brasil, A. C. B. Delbem, and D. R. Bonetti. Investigating relevant aspects of moeas for protein structures prediction. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, GECCO '11, pages 705–712, New York, NY, USA, 2011. ACM. Citado nas páginas 2, 3, 59, 61 e 65.

[27] C. R. S. Brasil, A. C. B. Delbem, and E. F. Bonetti, D. R. F. ; Faria. Ab initio studies using hydrogen bond in evolutionary algorithm. In *CIFARP 2009 - 7th International Congress of Pharmaceutical Sciences*, 2009. Citado nas páginas 2 e 3.

[28] C. R. S. Brasil, A. C. B. Delbem, T. W. Lima, and D. R. F. Bonetti. Algoritmo evolutivo para o problema de predição de proteínas, considerando a interação soluto solvente no critério de avaliação. In *32 Reunião Anual da Sociedade Brasileira de Química*, 2009. Citado nas páginas 2 e 3.

[29] C. R. S. Brasil, T. W. Lima, P. H. R. Gabriel, and A. C. B. Delbem. Mo-protpred: An multiobjective evolutionary algorithm to protein structure prediction with area accessibility. In *BIOMAT 2008*

- *International Symposium on Mathematical and Computational Biology*, 2008. Citado nas páginas 2 e 3.

[30] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983. Citado nas páginas 20, 34, 71 e 86.

[31] T. N. Bui and G. Sundarraj. An efficient genetic algorithm for predicting protein tertiary structures in the 2d hp model. In *Proceedings of the 2005 conference on Genetic and evolutionary computation*, GECCO '05, pages 385–392, New York, NY, USA, 2005. ACM. Citado na página 17.

[32] A. Caliri, L. Rocha, and M. Tarrago Pinto. The water factor in the protein-folding problem. *Brazilian Journal of Physics*, 34(1):90 – 101, 2004. Citado nas páginas 13, 17 e 19.

[33] P. Camilleri, S. A. Watts, and J. A. Boraston. A surface area approach to determination of partition coefficients. *I. C/rem. Sue. Perkin Trans.*, 11:1699, 1988. Citado na página 31.

[34] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, Dec. 2005. Citado na página 20.

[35] M. Charton and B. I. Charton. The structural dependence of amino acid hydrophobicity parameters. *Biol. 99*, pages 629–644, 1982. Citado na página 30.

[36] B. Chen and J. Hu. Protein structure prediction based on hp model using an improved hybrid eda. In Y.-p. Chen, L. M. Hiot, and Y. S. Ong, editors, *Exploitation of Linkage Learning in Evolutionary Algorithms*, volume 3 of *Adaptation, Learning, and Optimization*, pages 193–214. Springer Berlin Heidelberg, 2010. Citado nas páginas 15 e 16.

[37] C. Coello. Evolutionary multiobjective optimization: a historical view of the field. *Computational Intelligence Magazine, IEEE*, 1(1):28–36, 2006. Citado nas páginas 49, 50, 52, 53 e 56.

[38] C. Coello, D. V. Veldhuizen, and G. Lamont. Evolutionary algorithms for solving multiobjective problems. *Genetic algorithms and evolutionary computation ; 5. New York: Kluwer Academic*, 2002. Citado nas páginas 50 e 123.

[39] R. Copeland. Methods for protein analysis - a pratical guide to laboratory protocols. *Chapman e Hall*, 1993. Citado na página 8.

[40] T. E. Creighton. *Proteins: Structures and Molecular Properties*. W. H. Freeman, second edition edition, Aug. 1992. Citado na página 12.

[41] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *Journal of Computational Biology*, 5:597–603, 1998. Citado nas páginas 11, 13 e 16.

[42] Y. Cui, R. Chen, and W. Wong. Protein folding simulation with genetic algorithm and supersecondary structure constraints. *Proteins*, 31:247–257, 1998. Citado nas páginas 11, 13, 17, 19, 20 e 124.

[43] V. Cutello, G. Narzisi, and G. Nicosia. A class of pareto archived evolution strategy algorithms using immune inspired operators for ab-initio protein structure prediction. In *Proceedings of the 3rd European conference on Applications of Evolutionary Computing*, EC'05, pages 54–63. Springer-Verlag, Berlin, Heidelberg, 2005. Citado nas páginas 55, 61 e 64.

[44] V. Cutello, G. Narzisi, and G. Nicosia. A multiobjective evolutionary approach to the protein structure predicition problem. *J. R. Soc. Interface*, 83:1–13, 2005. Citado nas páginas 1, 47, 61 e 72.

[45] K. De Jong. *Evolutionary computation: a unified approach*. Mass: MIT Press, 2006. Citado nas páginas 13 e 18.

[46] K. Deb. *Multiobjective Optimization using Evolutionary Algorithms*. 2001. Citado nas páginas 13, 47, 48, 51, 123 e 124.

[47] K. Deb, S. Agrawal, A. Pratab, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multiobjective optimization: Nsga-ii. *KanGAL report 200001*, 2000. Citado nas páginas 47, 52 e 53.

[48] A. C. B. Delbem, A. C. P. L. F. Carvalho, and N. G. Bretas. Main chain representation for evolutionary algorithms applied to distribution system reconfiguration. *IEEE Transactions on Power Systems*, 20:425–436, 2005. Citado nas páginas 53 e 59.

[49] A. C. B. Delbem, M. Tendler, C. A. Brito, M. M. Vilar, N. S. Freire, C. M. Diogo, M. S. Almeida, J. F. Silva, W. Savino, R. C. Garrat, N. Katz, and A. J. G. Simpson. A schistosoma mansoni fatty acid-binding protein,

sm14, is the potential basis of a dual-purpose anti-helminth vaccine. *Proceedings of the National Academy of Sciences of the United States of America*, 93:269–273, 1996. Citado na página 2.

[50] G. R. Desiraju. *A Bond by Any Other Name*, volume 50. 2011. Citado na página 38.

[51] K. A. Dill. Dominate forces in protein folding. *Biochemistry*, 29, 1990. Citado na página 12.

[52] F. DiMaio, T. C. Terwilliger, R. J. Read, A. Wlodawer, G. Oberdorfer, U. Wagner, E. Valkov, A. Alon, D. Fass, H. L. Axelrod, and et al. Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature*, 473(7348):540–543, 2011. Citado nas páginas 1, 2, 16 e 91.

[53] F. DiMaio, M. D. Tyka, M. L. Baker, W. Chiu, and D. Baker. Refinement of protein structures into low-resolution density maps using rosetta. *Journal of Molecular Biology*, 392(1):181–190, 2009. Citado nas páginas 2, 16 e 91.

[54] A. Dinner, A. Sali, M. Karplus, and E. Shakhnovich. The folding mechanism of larger model proteins: Role of native structure. *Proc. Natl. Acad. Sci. USA*, 93:8356–8361, 1996. Citado na página 11.

[55] A. C. dos Santos. *Algoritmo evolutivo computacionalmente eficiente para reconfiguração de sistema de distribuição*. Tese de Doutorado - EESC - Universidade de São Paulo, 2009. Citado nas páginas 47, 55, 59, 64 e 66.

[56] W. J. Doucette and A. W. Andren. Correlation of octanol/water partition coefficients and total molecular surface area for highly hydrophobic aromatic compounds. *Environ. Sci. Technol.*, 21:821, 1987. Citado na página 31.

[57] W. J. Dunn III, M. G. Koehler, and S. Grigoras. The role of solvent-accessible surface area in determining partition coefficients. *J. Med. Ckem.*, 30:1121, 1987. Citado na página 31.

[58] F. Y. Edgeworth. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*, volume 42. P. Keagn, 1881. Citado na página 48.

[59] M. Ehrgott. *Multicriteria optimization*. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, 2000. Citado na página 49.

[60] D. Eisenberg and A. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986. Citado nas páginas 11, 35 e 36.

[61] D. Eisenberg, M. Wesson, and M. Yamashita. Interpretation of protein folding and binding with atomic solvation parameters. *Chemica Scripta*, 29A:217–221, 1989. Citado na página 11.

[62] J. Emsley. Very strong hydrogen bonding. *Chem. Soc. Rev.*, 9:91–124, 1980. Citado na página 38.

[63] L. J. Eshelman and J. D. Schaffer. Real-coded genetic algorithms and interval-schemata. In *FOGA*, pages 187–202, 1992. Citado na página 124.

[64] L. J. Eshelman and J. D. Schaffer. Real-coded genetic algorithms and interval-schemata. *Foundations of Genetic Algorithms 2 (FOGA-2)*, pages 187–202, 1993. Citado na página 64.

[65] M. Feig and S. Tanizaki. Development of a heterogeneous dielectric generalized born model for the implicit modeling of membrane environments. *Modelling Molecular Structure and Reactivity in Biological Systems*, 2006. Citado na página 30.

[66] M. Feig and S. Tanizaki. Extending the horizon: Towards the efficient modeling of large biomolecular complexes in atomic detail. *Theoretical Chemistry Accounts*, 2006. Citado na página 30.

[67] C. Fonseca and P. Fleming. Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization. *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 416–423, 1993. Citado na página 51.

[68] H. L. Friedman. Electrolyte Solutions at Equilibrium. *Ann. Rev. Phys. Chem.*, 32:179–204, 1981. Citado na página 20.

[69] R. A. Friesner. *Computational methods for protein folding*, volume 120. Wiley, 2002. Citado nas páginas 4, 39, 41, 94 e 97.

[70] D. Frishman and P. Argos. Knowledge-based secondary structure assignment. *Proteins: structure, function and genetics*, 23:566–579, 1995. Citado nas páginas 3, 4, 34, 42, 72, 84, 94 e 97.

[71] P. H. R. Gabriel. Algoritmos evolutivos e modelos simplificados de proteínas para predição de estruturas terciária. *Mestrado em Ciências da Computação e Matemática Computacional – Universidade de São Paulo, USP, Brasil*, 2010. Citado nas páginas 2, 3 e 30.

[72] P. H. R. Gabriel, T. W. De Lima, R. A. Faccioli, I. N. Silva, and A. C. B. Delbem. Pure ab initio evolutionary approach to protein structure prediction. *International Symposium on Mathematical and Computation Biology (BIOMAT)*, 2007. Citado nas páginas 2, 3, 18, 34 e 63.

[73] J. Gao, D. S. Garner, and W. L. Jorgensen. Ab initio study of structures and binding energies for anion-water complexes. *Journal of the American Chemical Society*, 108(16):4784–4790, 1986. Citado na página 38.

[74] A. Gaudio and Y. Takahata. Calculation of molecular surface area with numerical factors. *Compurers Chem.*, 16:277–284, 1992. Citado nas páginas 3, 31, 32 e 33.

[75] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989. Citado nas páginas 13, 50, 51, 53, 59, 68 e 123.

[76] D. E. Goldberg. *The Design of Innovation*. Kluwer Academic Publishers, 2002. Citado na página 91.

[77] B. Guillot. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liquids*, 101:219–260, 2002. Citado na página 3.

[78] P. Hajela and C. Y. Lin. Genetic search strategies in multicriterion optimal design. *Structural and Multidisciplinary Optimization*, 4:99–107, 1992. Citado na página 51.

[79] J. Handl, D. Kell, and J. Knowles. Multiobjective optimization in bioinformatics and computational biology. *IEEE Transactions on Computational Biology and Bioinformatics*, 2006. Citado nas páginas 2 e 61.

[80] J. Handl, J. Knowles, R. Vernon, D. Baker, and S. C. Lovell. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. In *Proteins: Structure, Function, and Bioinformatics*, volume 80, pages 490–504, 2012. Citado nas páginas 2 e 61.

[81] J. Handl, S. C. Lovell, and J. Knowles. Investigations into the effect of multiobjectivization in protein structure prediction. In *Proceedings of the 10th int. conf. on Parallel Problem Solving from Nature*, pages 702–711. Springer-Verlag, 2008. Citado nas páginas 2 e 61.

[82] J. Handl, S. C. Lovell, and J. Knowles. Multiobjectivization by decomposition of scalar cost functions. In *Proceedings of the 10th international conference on Parallel Problem Solving from Nature: PPSN*

X, pages 31–40, Berlin, Heidelberg, 2008. Springer-Verlag. Citado na página 61.

[83] G. R. Harik. *Learning gene linkage to efficiently solve problems of bounded difficulty using genetic algorithms*. PhD thesis, Ann Arbor, MI, USA, 1997. UMI Order No. GAX97-32090. Citado na página 73.

[84] W. Hart and S. Istrail. Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86*J. Comput. Biol*, 1997. Citado na página 19.

[85] G. Helles. A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of the Royal Society Interface the Royal Society*, 5(21):387–396, 2008. Citado na página 1.

[86] R. B. Hermann. Theory of hydrophobic bonding ii. the correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Cm.*, 76:2754, 1972. Citado na página 31.

[87] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008. Citado na página 34.

[88] M. Higashi and D. G. Truhlar. Citado nas páginas 20 e 27.

[89] J. Higo and N. Go. Algorithm for rapid calculation of excluded volume of large molecules. *Journal of Computational Chemistry*, 10:376, 1989. Citado na página 31.

[90] M. Hilbert, G. Bohm, and R. Jaenicke. Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins*, 17:138–151, 1993. Citado na página 13.

[91] D. A. Hinds and M. Levitt. A lattice model for protein-structure prediction at low resolution. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 89:2536–2540, 1992. Citado nas páginas 15 e 16.

[92] J. H. Holland. *Adaptation in natural and artificial systems*. MIT Press, Cambridge, MA, USA, 1992. Citado nas páginas 50, 123 e 124.

[93] J. Horn, J. Horn, N. Nafpliotis, N. Nafpliotis, D. E. Goldberg, and D. E. Goldberg. A niched pareto genetic algorithm for multiobjective optimization. In *In Proceedings of the First IEEE Conference on*

Evolutionary Computation, IEEE World Congress on Computational Intelligence, pages 82–87, 1994. Citado na página 52.

[94] S. Hubbard and J. Thornton. Naccess: computer program. *Department of Biochemistry and Molecular Biology, University College London*, 1993. Citado na página 34.

[95] Y. Inbar, H. Benyamini, R. Nussinov, and H. Wolfson. Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics*, 19:158–168i, 2003. Citado na página 15.

[96] A. Irback, C. Peterson, and F. Potthast. Identification of amino acid sequences with good folding properties in an off-lattice model. *Physical Review*, 55:860–867, 1997. Citado na página 19.

[97] A. Irback, C. Peterson, F. Potthast, and O. Sommelius. Local interactions and protein folding: a three-dimensional off-lattice approach. *J. Chem. Phys.*, 107:273–282, 1997. Citado na página 19.

[98] A. Irback and F. Potthast. Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.*, 103:10298–10305, 1995. Citado na página 19.

[99] H. Ishibuchi, Y. Hitotsuyanagi, H. Ohyanagi, and Y. Nojima. Effects of the existence of highly correlated objectives on the behavior of moea/d. In *EMO'11*, pages 166–181, 2011. Citado nas páginas 4, 59, 60, 67 e 97.

[100] H. Ishibuchi and Y. Nojima. Optimization of scalarizing functions through evolutionary multiobjective optimization. *Lecture Notes in Computer Science*, pages 51–65, 2007. Citado na página 60.

[101] H. Ishibuchi, Y. Sakane, N. Tsukamoto, and Y. Nojima. Adaptation of scalarizing functions in moea/d: An adaptive scalarizing function-based multiobjective evolutionary algorithm. In *EMO'09*, pages 438–452, 2009. Citado nas páginas 59, 60 e 97.

[102] H. Ishibuchi, N. Tsukamoto, and Y. Nojima. Evolutionary many-objective optimization: A short review. In *IEEE Congress on Evolutionary Computation*, pages 2419–2426. IEEE, 2008. Citado na página 59.

[103] J. Israelachvili. Intermolecular and surface forces. *Academic Press, London*, 1991. Citado na página 3.

[104] R. Jaenicke and R. Rudolph. Refolding and association of oligomeric proteins. *Methods Enzymol*, 131:218–250, 1986. Citado na página 11.

[105] M. Judy, K. Ravichandran, and K. Murugesan. A multiobjective evolutionary algorithm for protein structure prediction with immune operators. *Computer Methods in Biomechanics and Biomedical Engineering*, 12(4):407–413, 2009. Citado na página 61.

[106] H. R. Karfunkel and V. Eyrand. An algorithm for the representation and computational of supermolecular surfaces and volumes. *J. Comp. Chern.*, 10:628, 1989. Citado na página 31.

[107] M. Karplus and E. Shakhnovich. *Protein Folding, chapter Protein Folding: Theorical Studies of Thermodynamics and Dynamics*. 1992. Citado nas páginas 11, 15 e 91.

[108] M. Karplus and E. Shakhnovich. *Fundamentals concepts of bioinfomatic*. Benjamin Cumming, 2002. Citado nas páginas 18 e 19.

[109] M. Khimasia and P. Coveney. Protein structure prediction as a hard optimization problem: the genetic algorithm approach. *In Molecular Simulation*, 1997. Citado nas páginas 13, 15, 16, 17 e 18.

[110] J. Knowles and D. Corne. The pareto archived evolution strategy: A new baseline algorithm for multiobjective optimisation. *Congress on Evolutionary Computation, IEEE Service Center*, pages 98–105, 1999. Citado na página 52.

[111] N. Krasnogor, W. Hart, J. Smith, and D. Pelta. Protein structure prediction with evolutionary algorithms. *Genetic and Evolutionary Computation Conference*, 1999. Citado nas páginas 17 e 18.

[112] F. Kursawe. A variant of evolution strategies for vector optimization. In *Parallel Problem Solving from Nature. 1st Workshop, PPSN I, volume 496 of Lecture Notes in Computer Science*, pages 193–197. Springer-Verlag, 1991. Citado na página 51.

[113] I. M. Langmuir. Coiloid. symp. monograph. 3:48, 1925. Citado na página 31.

[114] P. Larranaga and J. A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation (Genetic Algorithms and Evolutionary Computation)*. Kluwer Academic Publishers Group, 2002. Citado na página 91.

[115] K. F. Lau and K. Dill. Macromolecules. 22:3986–3997, 1989. Citado na página 18.

[116] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y.-E. A. E. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, 487:545–574, 2011. Citado nas páginas 1, 2, 16 e 91.

[117] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol*, 55:379–400, 1971. Citado nas páginas 3 e 35.

[118] C. Lemer, M. Rooman, and S. Wodak. Protein structure prediction by threading methods: evaluation of current techniques. *Proteins*, 23:337–355, 1995. Citado na página 14.

[119] H. Li and Q. Zhang. Multiobjective optimization problems with complicated pareto sets, moea/d and nsga-ii. *Trans. Evol. Comp*, 13:284–302, April 2009. Citado na página 60.

[120] T. W. Lima. Algoritmos evolutivos para predição de estruturas de proteínas. *Dissertação de mestrado, Usp - ICMC*, 2006. Citado nas páginas 2, 3, 63, 64, 71, 124 e 125.

[121] T. W. Lima, A. Caliri, F. L. B. Silva, R. Tinos, G. Travieso, I. N. Souza, E. Marques, A. C. B. Delbem, V. Bonatto, R. Faccioli, C. R. S. Brasil, P. H. R. Gabriel, V. T. Do O, and D. R. F. Bonetti. Some modeling issues for protein structure prediction using evolutionary algorithm. *Evolutionary Computation*, 2009. Citado nas páginas 2 e 3.

[122] T. W. Lima, R. A. Faccioli, P. H. R. Gabriel, A. C. B. Delbem, and I. N. Silva. Multiobjective evolutionary approach to ab initio protein tertiary structure prediction. *BIOMAT 2006*, pages 269–286, 2006. Citado nas páginas 2, 3, 34, 63 e 71.

[123] T. W. Lima, P. H. R. Gabriel, R. A. Faccioli, A. C. B. Delbem, and I. N. Silva. Evolutionary algorithm to ab initio protein structure prediction with hydrophobic interactions. *CEC 2007*, 2007. Citado nas páginas 2 e 3.

[124] H.-L. Liu and X. Li. The multiobjective evolutionary algorithm based on determined weight and sub-regional search. In *Proceedings of the Eleventh conference on Congress on Evolutionary Computation*, CEC'09, pages 1928–1934, Piscataway, NJ, USA, 2009. IEEE Press. Citado nas páginas 55, 61 e 64.

[125] A. D. MacKerel, C. L. Brooks, L. Nilsson, B. Roux, Y. Won, and M. Karplus. CHARMM: The energy function and its parameterization with an overview of the program. In P. v. R. Schleyer et al., editor, *The Encyclopedia of Computational Chemistry*, volume 1, pages 271–277. John Wiley & Sons: Chichester, 1998. Citado nas páginas 20, 34, 71 e 86.

[126] E. Mattias and G. Xavier. A survey and annotated bibliography of multiobjective combinatorial otimization. *OR Spectrum*, 22(4):425–460, 2000. Citado na página 49.

[127] D. McGarrah and R. Judson. Analysis of the genetic algorithm method of molecular conformation determination. *Journal of Computational Chemistry*, 14:1385–1395, 1993. Citado nas páginas 13 e 17.

[128] Z. Michalewicz. *Genetic algorithms + data structures = evolution programs (2nd, extended ed.)*. Springer-Verlag New York, Inc., New York, NY, USA, 1994. Citado na página 124.

[129] Z. Michalewicz and D. B. Fogel. *How to Solve It: Modern Heuristics*. Springer, enlarged 2nd edition, 2004. Citado nas páginas 123 e 124.

[130] K. Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research and Management Science*. Kluwer Academic Publishers, Dordrecht, 1999. Citado nas páginas 49, 57 e 67.

[131] C. B. Moreira, I. Ezkurdia, M. L. Tress, and A. Valencia. Empirical limits for template-based protein structure prediction: the CASP5 example. *FEBS Lett*, 579(5):1203–1207, Feb. 2005. Citado nas páginas 2 e 95.

[132] P. M. Morse. Diatomic molecules according to the wave mechanics. ii. vibrational levels. *Phys. Rev.*, 34:57–64, Jul 1929. Citado na página 21.

[133] D. L. Nelson and M. Cox. *Lehninger Principles of Biochemistry*. 2004. Citado nas páginas 4, 7, 8, 28, 44, 64 e 119.

[134] S. J. Opella, F. M. Marassi, J. J. Gesell, A. P. Valente, Y. Kim, M. Oblatt-Montal, and M. Montal. Structures of the m2 channel-lining

segments from nicotinic acetylcholine and nmda receptors by nmr spectroscopy. *America*, 6:374–379, 1999. Citado na página 69.

[135] B. R. Oren M. Becker, Alexander D. MacKerell Jr. *Computational Biochemistry and Biophysics*. 2001. Citado nas páginas 4, 28 e 44.

[136] C. Orengo, D. Jones, and J. Thornton. *Bioinformatics: Genes, Protein and Computers*. 2003. Citado nas páginas 71, 72 e 73.

[137] E. O'Toole and A. Panagiotopoulos. Monte carlo simulation of folding transitions of simple model proteins using chain growth algorithm. *J. Chem. Phys.*, 97, 1992. Citado nas páginas 11, 15 e 16.

[138] V. Pareto. *Cours d'économie politique*, volume 1,2. F. Rouge, 1896. Citado na página 48.

[139] L. Pauling. The nature of the chemical bond and the structure of molecules and crystals. *Cornell Univ. Press*, 1960. Citado na página 34.

[140] M. Pelikan, D. E. Goldberg, and F. G. Lobo. A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21:5–20, 2002. 10.1023/A:1013500812258. Citado na página 91.

[141] G. Petsko and D. Ringe. Proteins structure and function. *New Science Press Ltd*, 2004. Citado nas páginas 9 e 10.

[142] J. Ponder. Tinker: Software tools for molecular design. *Washington University, Saint Louis*, 2001. Citado nas páginas 16, 20, 26 e 34.

[143] G. N. Ramachandran and C. M. Venkatachalam. Conformation of polypeptide chains. *Annual Review of Biochemistry*, 38:45–82, 1969. Citado nas páginas 12 e 18.

[144] C. Reichardt. *Solvents and Solvent Effects in Organic Chemistry*. 1990. Citado na página 29.

[145] T. Richmond. Solvent accessible surface area and excluded volume in proteins. analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.*, 178:63, 1984. Citado na página 31.

[146] L. Rocha, M. da Silva, and A. Caliri. Entropic force and folding of macromolecules. *Physics Letters A*, 20:178–182, 1996. Citado nas páginas 11 e 19.

[147] A. Roosevelt, M. A. A. da Silva, and A. Caliri. Deterministic folding: The role of entropic forces and steric specificities. *Journal of Chemical Physics*, 114:9, 2000. Citado nas páginas 11 e 19.

[148] G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229:834–838, 1985. Citado na página 30.

[149] A. Roy, A. Kucukural, and Y. Zhang. I-tasser: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 2010. Citado nas páginas 2, 16 e 91.

[150] D. S. Sanches, M. R. Mansour, J. B. A. London, A. Delbem, and A. C. Santos. Integrating relevant aspects of moeas to solve loss reduction problem in large-scale distribution systems. *IEEE Trondheim PowerTech 2011*, 2011. Citado na página 77.

[151] R. Santana, Y. Saeys, J. Flores, J. Lozano, Y. Van de Peer, R. Blanco, C. Bielza, P. Larra, et al. A review of estimation of distribution algorithms in bioinformatics. *BioData Mining*, 1:6, 2008. Citado na página 91.

[152] A. Santos, A. Delbem, J. London, and N. Bretas. Node-depth encoding and multiobjective evolutionary algorithm applied to large-scale distribution system reconfiguration. *IEEE Transactions on Power Systems*, pages 1254–1265, 2010. Citado nas páginas 3, 47, 59, 65 e 77.

[153] J. Schaffer. Multiple objective optimization with vector evaluated genetic algorithms. *Genetic Algorithms and their Applications: Proceedings of the First International Conference on Genetic Algorithms*, pages 93–100, 1985. Citado na página 50.

[154] R. Schleif. Analysis of Protein Structure and Function: A Beginner's Guide to CHARMM. 2006. Citado nas páginas 21, 22, 23, 24 e 26.

[155] T. Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2002. Citado nas páginas 24, 26 e 27.

[156] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010. Citado nas páginas xii, 69 e 72.

[157] S. Schulze-Kremer. Genetic algorithms for protein tertiary structure prediction. *ECML '93: Proceedings of the European Conference on Machine Learning*, pages 262–279, 1993. Citado na página 19.

[158] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. 1997. Citado nas páginas 2, 13 e 16.

[159] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus. Protein folding bottlenecks: A lattice Monte Carlo simulation. *Physical Review Letters*, 67(12):1665–1668, Sept. 1991. Citado na página 17.

[160] M. Sharon, N. Kessler, R. Levy, S. Zolla-Pazner, M. Gorlach, and J. Anglister. Alternative conformations of hiv-1 v3 loops mimic beta hairpins in chemokines, suggesting a mechanism for coreceptor selectivity. *Structure*, 11:225–236, 2003. Citado nas páginas 70 e 86.

[161] A. Shmygelska and H. H. Hoos. An improved ant colony optimisation algorithm for the 2d hp protein folding problem. In *Proceedings of the 16th Canadian society for computational studies of intelligence conference on Advances in artificial intelligence*, AI'03, pages 400–417, Berlin, Heidelberg, 2003. Springer-Verlag. Citado na página 17.

[162] A. Shrake and J. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351 – 371, 1973. Citado na página 3.

[163] E. Silla and J. L. Pascual-Ahuir. Gepol: An improved description of molecular surfaces. i. building the spherical surface set. *J. Comp. Chem.*, 11:1047, 1990. Citado na página 31.

[164] I. R. Silva, L. M. Dos Reis, and A. Caliri. Topology-dependent protein folding rates analyzed by a stereochemical model. *Journal of Chemical Physics*, 123(154906), 2005. Citado nas páginas 11 e 19.

[165] K. T. Simons, R. Bonneau, I. Ruczinski, , and D. Baker. Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins Suppl 3*, pages 171–176, 1999. Citado nas páginas 13, 15 e 16.

[166] M. S. Skaf and F. da Silva. *Explorando as propriedades moleculares de solventes*. Química Nova, 17 edition, 1994. Citado na página 38.

[167] N. Srinivas and K. Deb. Multiobjective function optimization using non-dominated sorting genetic algorithm evolutionary computation. 2(3):221–248, 1994. Citado nas páginas 47, 52 e 53.

[168] P. J. Steinbach and B. R. Brooks. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.*, 15(7):667–683, July 1994. Citado na página 27.

[169] G. Sywerda. Uniform crossover in genetic algorithms. In *Proceedings of the third international conference on Genetic algorithms*, pages 2–9, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc. Citado na página 124.

[170] B. M. N. A. J. Talbi, E.-G. and E. Alba. Multiobjective optimization using metaheuristics: non-standard algorithms. *International Transactions in Operational Research*, 19:283–305, 2012. Citado na página 49.

[171] S. Tanizaki and M. Feig. A new generalized born formalism for heterogeneous dielectric environments: Application to the implicit modeling of biological membranes. *Journal of Physical Chemistry*, 2005. Citado na página 30.

[172] S. Tanizaki and M. Feig. Molecular dynamics simulations of large integral membrane proteins with an implicit membrane model. *Journal of Physical Chemistry*, 2006. Citado na página 30.

[173] O. Tapia and O. Goscinski. Self-consistent reaction field theory of solvent effects. *Molecular Physics*, 29:1653–1661, 1975. Citado na página 30.

[174] M. M. Teeter. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. *Proceedings of the National Academy of Sciences of the United States of America*, 81:6014–6018, 1984. Citado na página 86.

[175] G. B. Thomas and L. Finney. *Calculus and Analytic Geometry*. Addison Wesley, 9 edition, 1996. Citado na página 22.

[176] W. Ticona. Algoritmos evolutivos para otimização multiobjetivo. Technical report, Universidade de São Paulo (ICMC), 2008. Citado nas páginas 48 e 50.

[177] D. Tolkatchev, A. Ng, W. Vranken, and F. Ni. Citado na página 86.

[178] J. Tomasi, R. Cammi, and B. Mennuci. Medium effects on the properties of chemical systems: Electric and magnetic response of donor-acceptor systems within the polarizable continuum model. *J. Quantum Chem.*, 75:783, 1999. Citado na página 30.

[179] E. K. Tuominen, J. M. Holopainen, J. Chen, G. D. Prestwich, P. R. Bachiller, P. K. Kinnunen, and P. A. Janmey. Fluorescent phosphoinositide derivatives reveal specific binding of gelsolin and other actin regulatory proteins to mixed lipid bilayers. *The Federation of*

European Biochemical Societies Journal, 263:85–92, 1999. Citado nas páginas 68 e 86.

[180] S. Tyukhtenko, E. K. Tiburu, L. Deshmukh, O. Vinogradova, D. R. Janero, and A. Makriyannis. Nmr solution structure of human cannabinoid receptor-1 helix 7/8 peptide: candidate electrostatic interactions and microdomain formation. *Biochemical and Biophysical Research Communications*, 390:441–446, 2009. Citado nas páginas 69 e 86.

[181] R. Unger, D. Harel, S. Wherland, and J. Sussman. A 3-d building blocks approach to analyzing and predicting structure of proteins. *Proteins: Struct. Funct. Genet.*, 5:355–373, 1989. Citado nas páginas 13 e 15.

[182] R. Unger and J. Moult. Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*, 231:75–81, 1993. Citado nas páginas 11, 13 e 17.

[183] S. C. Valvani, S. H. Yalkowsky, and G. L. Amidon. Solubility of nonelectrolytes in polar solvents: V. estimation of the solubility of aliphatic monofunctional compounds in water using the molecular surface area approach. *J. Phys. Chem.*, page 829, 1976. Citado nas páginas 31 e 34.

[184] A. Vullo. On the role of machine learning in protein structure determination. *AIIA*, 2002. Citado nas páginas 13 e 15.

[185] A. Wallqvist and R. D. Mountain. Molecular models of water: Derivation and description. *Reviews in Computational Chemistry*, 13:183–247, 1999. Citado na página 3.

[186] L. Wesson and D. Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, 1:227–35, 1992. Citado na página 35.

[187] M. Wittlich, B. W. Koenig, M. Stoldt, H. Schmidt, and D. Willbold. Nmr structural characterization of hiv-1 virus protein u cytoplasmic domain in the presence of dodecylphosphatidylcholine micelles. *The FEBS journal*, 276:6560–6575, 2009. Citado nas páginas 69 e 86.

[188] K.-C. Wong, K.-S. Leung, and M.-H. Wong. Protein structure prediction on a lattice model via multimodal optimization techniques. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, GECCO '10, pages 155–162, New York, NY, USA, 2010. ACM. Citado nas páginas 15 e 16.

[189] D. Xu, J. Zhang, A. Roy, and Y. Zhang. Automated protein structure modeling in casp9 by i-tasser pipeline combined with quark-based ab initio folding and fg-md-based structure refinement. *Proteins: Structure, Function, and Bioinformatics*, 79(S10):147–160, 2011. Citado nas páginas 1, 2, 16 e 91.

[190] D. Xu and Y. Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, Mar. 2012. Citado nas páginas 2, 16 e 28.

[191] A. Zemla. Lga: A method for finding 3d similarities in protein structures. *Nucleic Acids Res*, 31(13):3370–3374, July 2003. Citado nas páginas 71, 72 e 73.

[192] Q. Zhang and H. Li. Moea/d: A multiobjetive evolutionary algorithm based ond decomposition. *IEEE Transation on Evolutionary Computation*, 11 (6):712–731, 2007. Citado nas páginas 5, 47, 52, 55, 57, 64 e 66.

[193] Y. Zhang. Template-based modeling and free modeling by i-tasser in casp7. *Proteins*, 69:108–117, 2007. Citado nas páginas 1, 2, 16 e 91.

[194] Y. Zhang and J. Skolnick. Citado na página 72.

[195] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. In K. C. Giannakoglou, D. T. Tsahalis, J. Périaux, K. D. Papailiou, and T. Fogarty, editors, *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems*, pages 95–100, Athens, Greece, 2001. International Center for Numerical Methods in Engineering. Citado nas páginas 47, 52, 55, 56, 57, 64 e 66.

[196] E. Zitzler and L. Thiele. An evolutionary algorithm for multiobjective optimization: The strength pareto approach. Technical Report 43, Computer Engineering and Communication Networks Lab (TIK), Swiss Federal Institute of Technology (ETH), 1998. Citado nas páginas 47, 52 e 56.

[197] X. Zou, Y. Chen, M. Liu, and L. Kang. A new evolutionary algorithm for solving many-objective optimization problems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 38(5):1402–1412, 2008. Citado na página 59.

APÊNDICE

A

Aminoácidos

As estruturas dos vinte aminoácidos apresentados no código genético universal, e seus respectivos nomes, símbolos e abreviações são os seguintes [133]:

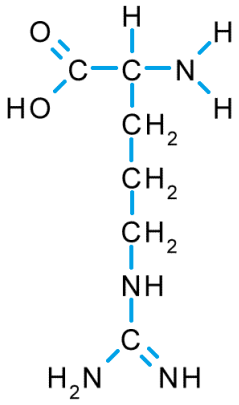

Figura A.1: Arginina - Arg - R.

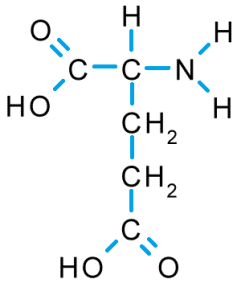

Figura A.2: Ácido Glutâmico - Glu - E.

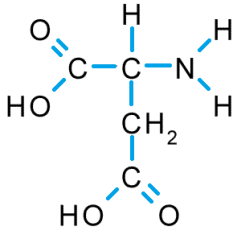

Figura A.3: Ácido Aspártico - Asp - D.

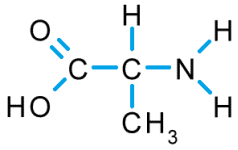

Figura A.4: Alanina - Ala - A.

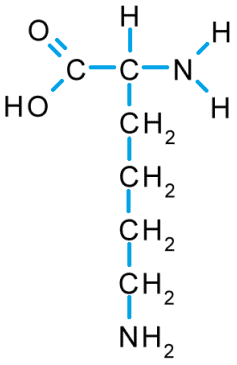

Figura A.5: Lisina - Lys, Lis - K.

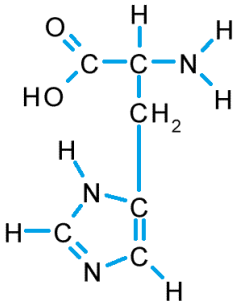

Figura A.6: Histidina - His - H.

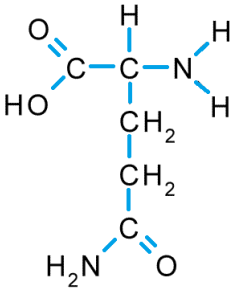

Figura A.7: Glutamina - Gln - Q.

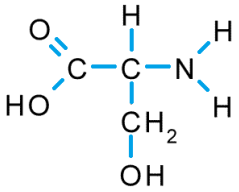

Figura A.8: Serina - Ser - S.

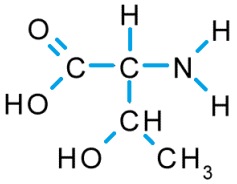

Figura A.9: Treonina - Thr, The - T.

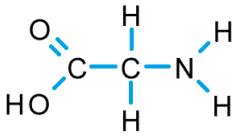

Figura A.10: Glicina - Gly, Gli - G.

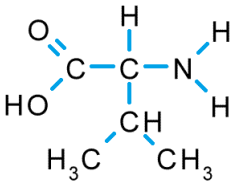

Figura A.11: Valina - Val - V.

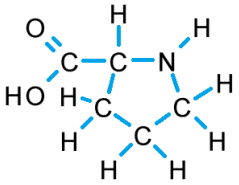

Figura A.12: Prolina - Pro - P.

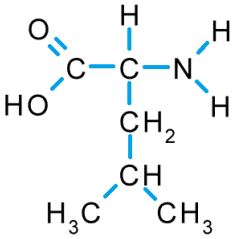

Figura A.13: Leucina - Leu - L.

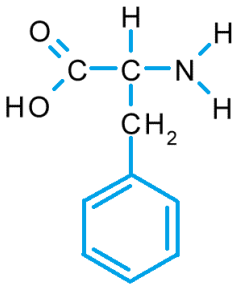

Figura A.14: Fenilalanina - Phe, Fen - F.

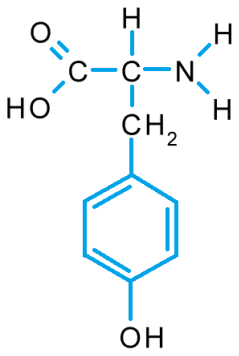

Figura A.15: Tirosina - Tyr, Tir - Y.

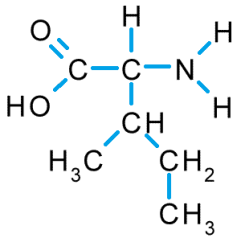

Figura A.16: Isoleucina - Ile - I.

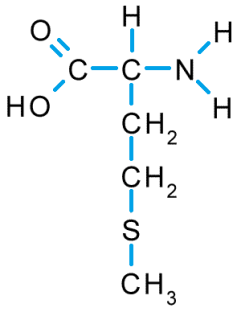

Figura A.17: Metionina - Met - M.

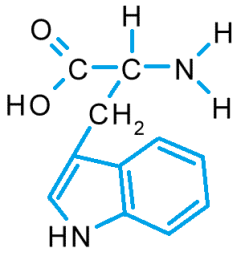

Figura A.18: Triptofano - Trp, Tri - W.

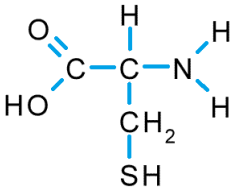

Figura A.19: Cisteína - Cys, Cis - C.

APÊNDICE

B

Algoritmos Evolutivos

Algoritmos evolutivos (AEs) são métodos de busca e otimização [75, 92, 38], baseados na teoria da seleção natural de Darwin, que prevê que os indivíduos mais aptos sobrevivem. Essa técnica não garante soluções ótimas globais, podendo encontrar ótimos locais, no entando pode gerar soluções adequadas para os problemas considerados.

Para implementar um AE é necessário enfatizar dois aspectos importantes: a codificação da solução em cromossomos e uma função de aptidão (*fitness*). Um cromossomo representa uma possível solução, ou indíviduo. Deste modo, a codificação dos cromossomos podem usar cadeias de *bits*, base binária, números reais etc, enquanto que a função de aptidão é responsável por avaliar as possíveis soluções para o problema. Essa função avalia um cromossomo, com um critério que depende do problema considerado, devolvendo um número real, o qual representa seu grau de adaptabilidade, informando quão distante esse cromossomo está da solução ótima.

No processo da evolução, os cromossomos (indivíduos) mais aptos são identificados e mantidos na população, enquanto que os mais fracos são excluídos. Para isso, é necessário que se utilize de métodos de seleção de indivíduos, que podem ser seleção por *ranking* e por torneio [75], por exemplo.

Selecionados os indivíduos para reprodução, por meio de operadores de reprodução, são gerados novos indivíduos. Tais operadores genéticos são: *crossover* e mutação.

O operador *crossover*, também chamado cruzamento ou recombinação, permite a troca de material genético entre dois indivíduos denominados pais, combinando informações de maneira que exista uma probabilidade razoável dos novos indivíduos serem melhores [46, 129]. O operador de

cruzamento mais usado é o *crossover* de um ponto, mas também existem outras variações, tais como de dois pontos, uniforme, aritmético, BLX-α, de simulação binária, de recombinação *Fuzzy*, de distribuição uniforme, simplex e outros [46, 63, 128, 129, 169].

A mutação altera o valor genético de um indivíduo, substituindo o valor de um gene por outro valor aleatório [128, 129]. No caso do indivíduo ser representado por uma cadeia binária, ela consiste em escolher, aleatoriamente, um gene do cromossomo e inverter seu valor de 1 para 0 ou vice-versa [92]. Em [42] podem ser encontrados três tipos de mutação aplicadas ao problema de PSP, as quais foram utilizadas no desenvolvimento deste trabalho. O objetivo da mutação é manter a diversidade da população, buscando que soluções não fiquem retidas em regiões de sub-ótimos, garantindo que se alcance uma parte suficientemente grande do espaço de busca. Geralmente, a mutação é aplicada em baixas frequências, para que não se torne uma busca aleatória.

Neste trabalho de Doutorado, foram usados três tipos de *crossover* e três tipos de mutação [120] aplicados ao problema de PSP, que são descritos a seguir:

Crossover 1 é o operador de reprodução que usa BLX-α [63], cruzando os ângulos Φ, Ψ e os ângulos laterais para cada resíduo da proteína. A Equação B.1 mostra a maneira que se efetua esse operador sobre o ângulo Φ, por exemplo. O mesmo cálculo é feito para o ângulo Ψ.

$$filho.res.phi = pai_1.res.phi + \alpha(pai_1.res.phi - pai_2.res.phi) \qquad \text{(B.1)}$$

tal que $filho.res.phi$ é o ângulo Φ de cada resíduo do novo indivíduo gerado pelo cruzamento, $pai_1.res.phi$, $pai_2.res.phi$ são os ângulos Φ de cada resíduo dos indivíduos a serem cruzados, e $\alpha = 0.5$. De acordo com [120], o melhor valor para α é 0.5.

Crossover 2 é o cruzamento de um ponto, que escolhe aleatoriamente a posição P de um resíduo no cromossomo, e copia os genes do pai_1 até a posição P, e da posição $P + 1$ copia os genes do pai_2 gerando o $filho_1$, em que pai_1 e pai_2 são os indivíduos selecionados para serem cruzados, e $filho_1$ é o novo indivíduo gerado.

Crossover 3 é o cruzamento de dois pontos, que escolhe aleatoriamente duas posições diferentes P_1 e P_2 de dois resíduos no cromossomo, e gera dois filhos diferentes. Até o P_1, copia-se o pai_1 no $filho_1$, e o pai_2 no $filho_2$. Da posição P_1 até P_2, copia-se pai_2 no $filho_1$, e pai_1 no $filho_2$. Da posição P_2 até o numero total de genes, copia-se pai_1 no $filho_1$, e pai_2 no $filho_2$.

Mutação 1 é o operador de reprodução escolhe um gene (resíduo) aleatoriamente, e muta os ângulos Φ, Ψ, e laterais com um pequena taxa de perturbação sobre os mesmos. A Equação B.2 mostra como ocorre essa mutação nos ângulos Phi. O mesmo ocorre para os outros ângulos da cadeia principal e lateral. Esse operador é efetuado M vezes na mesma proteína, em que $M = 1+(nresiduos/4)exp(-(ingeneration+1)/Neff)$ (aplica-se essa função a fim de evitar a convergência prematura [120]).

$$r = random(); \quad \text{(B.2)}$$

$$\delta = (180 - |prot.res.phi|)); \quad \text{(B.3)}$$

$$prot.res.phi = prot.res.phi + (r - 0.5)\delta. \quad \text{(B.4)}$$

tal que r é calculado com distribuição uniforme no intervalo [0,1], δ é o ângulo suplementar de Phi, $prot.res.phi$ representa os ângulos Phi do resíduo a ser mutacionado da proteína, $nresiduos$ é o número de resíduos da proteína, $ingeneration$ é a geração atual da evolução, $Neff = 150$ (em [120] tem-se que 150 é o melhor valor).

Mutação 2 é o operador de reprodução que funciona como a *mutação 1*, alterando a taxa de perturbação nos ângulos. A Equação B.5 mostra o cálculo dessa mutação. Esse operador é efetuado M vezes na mesma proteína, em que $M = 1 + (nresiduos/4)exp(-(ingeneration + 1)/Neff)$.

$$r = random(); \quad \text{(B.5)}$$

$$\delta = (180 - |prot.phi|); \quad \text{(B.6)}$$

$$prot.phi = prot.phi + (r - 0.05)\delta. \quad \text{(B.7)}$$

Mutação 3 é o operador de reprodução que muta todos os ângulos de um gene (resíduo), selecionando os novos ângulos a partir das regiões restritas, definidas pelo diagrama de Ramachandram (Seção 2.3). Esse operador é efetuado M vezes na mesma proteína uniformemente, em que $M = 1 + (inPopSize)exp(-(ingeneration + 1)/Neff)$ e $inPopSize$ é o tamanho da população.

Os métodos desenvolvidos neste trabalho utilizam 30% de probabilidade de efetuar o *crossover 1*, e 20% de probabilidade de realizar o *crossover 2*, e 50% o *crossover 3*. Para todos os indivíduos da população, realiza-se a mutação, com 25% de ser *mutação 1*, 25% de ser *mutação 2* e 50% de ser *mutação 3*.