UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

WILLIAM WOLMANN GONÇALVES

Um Estudo da Aplicação de Algoritmos Genéticos na Predição da Estrutura 3-D Aproximada de Proteínas

Profª. Drª. Luciana Salete Buriol
Orientador

Me. Márcio Dorn
Co-orientador

Porto Alegre, Julho de 2011

“The beginning is
the most important part of the work.”
— PLATO

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
AG	Algoritmo Genético
AMBER	*Assisted Model Building with Energy Refinement*, pacote de parâmetros de mecânica molecular para simulação de biomoléculas
CHARMM	*Chemistry at HARvard Molecular Mechanics*, pacote de parâmetros de mecânica molecular para simulação de biomoléculas
DSSP	*Dictionary Secondary Structure of Proteins*, algoritmo utilizado na identificação das estruturas secundárias de uma proteína
PDB	*Protein Data Bank*, repositório de dados estruturais tridimensionais de proteínas
PF	*Protein Folding prediction problem*, Problema da Predição do Enovelamento de Proteínas
PSP	*Protein Structure Prediction problem*, Problema da Predição da Estrutura de Proteínas
RMSD	*Root-Mean-Square Deviation*, Raiz do Desvio Quadrático Médio

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O Problema da Predição da Estrutura Tridimensional de Proteínas (3D-PSP, sigla em inglês) é um dos mais importantes problemas em Bioinformática Estrutural. Diversos algoritmos têm sido propostos ao longo dos últimos anos. Entretanto, o problema continua desafiante por causa da complexidade e dimensionalidade do espaço de busca conformacional das proteínas. A estrutura nativa de uma proteína dita sua função bioquímica em um organismo. Conhecer sua estrutura 3D implica em também conhecer a sua função. Assim, conhecendo a sua estrutura é possível interferir ativando ou inibindo a sua função, como em doenças onde os alvos dos fármacos são as proteínas. Neste trabalho de diplomação é apresentado um estudo da aplicação de algoritmos genéticos com população estruturada no 3D-PSP. Diferentes estratégias de seleção e organização populacional são utilizadas com a finalidade de garantir diversidade de indivíduos e convergência das conformações preditas. A eficácia dos métodos estudados é conferida com a execução de seis estudos de caso.

Palavras-chave: Algoritmos genéticos, predição da estrutura de polipeptídeos, 3D-PSP, bioinformática, random-key, método *ab initio*.

A Study of Application of Genetic Algorithms in Predicting Approximate 3-D Structure of Proteins

ABSTRACT

The Protein 3D Structure Prediction problem (3D-PSP) is one of the most important problems in Structural Bioinformatics. Several algorithms have been proposed over the last years. However, the problem still remains challenging because the complexity and dimensionality of the protein conformational search space. The native structure of a protein dictates its biochemical function in an organism. Knowing its 3D structure also implies knowing its function. Hence, knowledge of a protein structure allows one to interfere with it, either by enhancing or inhibiting its function, such as in diseases in which the drug targets are proteins. This graduation work presents a study of the application of genetic algorithms with structured population to the 3D-PSP. Different selection and population organization strategies are used with the purpose of ensuring diversity of individuals and convergence of the predicted conformations. The efficacy of these methods is conferred by the execution of six case studies.

Keywords: genetic algorithms, structure prediction of polypeptides, 3D-PSP, bioinformatics, random-key, *ab initio* method.

1 INTRODUÇÃO

1.1 Motivação

As proteínas são macromoléculas sintetizadas pelos organismo vivos e responsáveis por diversos processos complexos nos mesmos. Constituídas de uma ou mais cadeias polipeptídicas – sequências de aminoácidos unidos por meio de ligações peptídicas – as proteínas possuem funções fundamentais no organismo dos seres vivos, tais como hormonal, de defesa, de transporte, energética e enzimática e, portanto, consideradas os componentes químicos mais importantes do ponto de vista estrutural (WIKIPEDIA, 2011a).

Em condições fisiológicas uma proteína adota uma estrutura 3D funcional estável e única que define sua função bioquímica no organismo (GIBAS; JAMBECK, 2001; DORN, 2008). Técnicas experimentais existentes para obtenção dessa estrutura demandam altos custos e podem consumir muito tempo (em alguns casos, semanas). Essas dificuldades na determinação das estruturas tridimensionais, aliadas a projetos genoma, deram origem a grande volume de dados e proteínas cujas estruturas 3D ainda não são conhecidas.

Atualizando os levantamentos realizados em (DORN, 2008), em Abril de 2011 havia aproximadamente 135 milhões de sequências de proteínas no GenBank[1]. Desse conjunto, aproximadamente sete milhões são consideradas únicas, para ser exato, $6,53$ milhões em Outubro de 2008 (PRICE; DEHAL; ARKIN, 2008). Até 14 de Junho de 2011, apenas 68.344 estruturas 3D de proteínas[2] estavam disponíveis no *Protein Data Bank* (PDB) (BERMAN et al., 2000). Eliminando-se estruturas tridimensionais redundantes, considerando apenas estruturas com topologias únicas, tem-se apenas 1.393 topologias ou tipos de enovelamentos distintos[3]. Ainda de acordo com as pesquisas realizadas em (DORN, 2008) há três anos, isso significa que, ainda, somente cerca de $0,02\%$ das estruturas funcionais de sequências únicas de proteínas são conhecidas.

Essa disparidade, aliada com a estimativa de que apenas, aproximadamente, 2.000 tipos de enovelamentos existam entre proteínas de ocorrência natural (GOVINDARAJAN; RECABARREN; GOLDSTEIN, 1999), motivou pesquisadores a desenvolverem metodologias computacionais de busca pela predição de estruturas funcionais de proteínas, bem como a criarem grandes repositórios de dados de enovelamentos prováveis na natureza. Métodos de predição computacionais (*in silico*) da estrutura nativa de proteínas têm sido largamente estudados (CRESCENZI et al., 1998; FLOUDAS et al., 2006; JONES, 1997; ESWAR et al., 2006; OSGUTHORPE, 2000; ROHL et al., 2004).

Dado o grande volume de dados disponíveis, o problema da predição computacional

[1] ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

[2] http://www.pdb.org/pdb/statistics/holdings.do

[3] http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=fold-scop

da estrutura de proteínas (PSP, sigla em inglês) é atualmente um dos mais importantes problemas em Bioinformática Estrutural (FLOUDAS et al., 2006), paralelo ao problema da predição e estudo do processo do enovelamento de proteínas (PF, sigla em inglês). Visto que a estrutura 3D de uma proteína dita a função da mesma em uma célula, o PSP consiste em determinar a estrutura tridimensional de uma proteína dado, em alguns casos, apenas sua sequência de aminoácidos (FLOUDAS et al., 2006)

Recentemente, muitos métodos têm sido aplicados ao PSP. Essas metodologias podem ser classificadas em quatro grupos (FLOUDAS et al., 2006): ao primeiro grupo pertencem os métodos de modelagem comparativa por homologia (ESWAR et al., 2006); ao segundo grupo pertencem os métodos baseados no reconhecimento de enovelamentos (JONES, 1997); ao terceiro grupo pertencem os métodos de primeiros-princípios sem informações de bancos de dados (*ab initio*) (OSGUTHORPE, 2000) e, por fim, ao quarto grupo pertencem os métodos de primeiros-princípios com informações de repositórios de estruturas tridimensionais de proteínas (*de novo*) (ROHL et al., 2004). Os dois últimos grupos comportam as técnicas que possuem a capacidade de predizer novas formas de enovelamento cujas homologias, sequenciais e estruturais, não podem ser determinadas através de modelos oriundos de repositórios de estruturas (DORN, 2008).

Métodos *ab initio* não utilizam informações de bases de dados estruturais de proteínas e são fundamentados na observação que a estrutura nativa de uma proteína corresponde a uma estrutura cuja energia livre é mínima. O objetivo do método é encontrar o valor mínimo global da energia livre de uma proteína que corresponderia, ou à estrutura nativa, ou a uma conformação funcional da mesma (OSGUTHORPE, 2000). A principal vantagem dos métodos desse grupo é que eles são capazes de prever novos enovelamentos de proteínas, pois não estão limitados ao uso de estruturas já conhecidas.

Em métodos *de novo*, padrões estruturais são extraídos do PDB e utilizados na construção de estruturas tridimensionais de proteínas. Métodos pertencentes a esta classe, frequentemente, não comparam inteiramente uma sequência alvo contra uma sequência de uma proteína modelo com estrutura tridimensional conhecida, mas comparam pequenos fragmentos de resíduos de aminoácidos e os combinam (DORN, 2008) com o objetivo de encontrar a estrutura 3D da proteína cuja energia potencial seja a menor, logo, possibilitando, também, a predição de novos enovelamentos.

Através do alinhamento da sequência alvo em uma sequência modelo, de estrutura 3D conhecida no PDB (ESWAR et al., 2006), o objetivo da modelagem comparativa é predizer a estrutura de determinada proteína tridimensional com uma precisão comparável aos melhores resultados alcançados experimentalmente. Após o alinhamento e detecção de estruturas homólogas – limiar estabelecido, geralmente, em 30%, inclusive para proteínas longas –, o procedimento de modelagem é realizado através das cópias de coordenadas, distâncias interatômicas e ângulos de torção da proteína modelo. Os métodos desse grupo são os que apresentam maior acurácia nos resultados finais (JONES; TAYLOR; THORNTON, 1992; MARTÌ-RENOM M.A.; SALI, 2000; TRAMONTANO, 2006; DORN, 2008), embora esta precisão seja dependente do grau de identidade entre proteína alvo e modelos. São os métodos, comumente, empregados quando a estrutura de uma proteína não pode ser determinada por ressonância magnética nuclear, nem por cristalografia de difração de raios-X.

Métodos baseados no reconhecimento de enovelamentos organizam os resíduos da proteína alvo espacialmente conforme os moldes utilizados de estruturas conhecidas e para cada arranjo é calculada sua aptidão através de funções de "pontuação". Habitualmente, o objetivo também está relacionado a minimizar a energia livre de determinados

arranjos, inclusive com o refinamento posterior das estruturas da molécula predita. O método mais comum de reconhecimento é o alinhavamento de proteínas (*protein threading*) (JONES; TAYLOR; THORNTON, 1992; TRAMONTANO, 2006; DORN, 2008). Este método trabalha utilizando conhecimento estatístico da relação entre estruturas conhecidas em repositórios e da sequência da proteína alvo (BUJNICKI, 2006; WIKIPEDIA, 2011b). Os métodos desse grupo permitem tratar casos em que somente resultados de baixa homologia são obtidos via modelagem comparativa – abaixo de 15%, por exemplo –, mas é identificado que determinada proteína alvo possui um estado nativo muito próximo ao de estruturas já conhecidas (PENG; XU, 2010).

Métodos *ab initio* podem prever novas conformações de proteínas, mas precisam tratar a complexidade e alta dimensionalidade do espaço de busca conformacional (NGO; MARKS; KARPLUS, 1997). Métodos *de novo* também têm de tratar com a magnitude do espaço de busca e possuem como pré-requisito a necessidade de encontrar uma maneira inteligente de manipular os dados estruturais das bases de dados, como o PDB, e usá-los para prever estruturas tridimensionais próximas às nativas.

Devido ao grande número de conformações que determinada sequência linear de aminoácidos pode assumir, dada a alta dimensionalidade do espaço conformacional, Algoritmos Genéticos (AG) ganharam reconhecimento como metodologia computacional de busca na otimização de problemas relacionados a estruturas de proteínas e, particularmente, em métodos *ab initio* ao problema PSP (PEDERSEN; MOULT, 1997; CUTELLO; NARZISI; NICOSIA, 2006). Alguns estudos sobre a utilização de AGs ao PSP foram realizados nos últimos vinte anos (SCHULZE-KREMER, 1993; PEDERSEN; MOULT, 1996; UNGER, 2004). Por serem inspirados em processos naturais, serem uma boa metodologia de busca e largamente estudados em Otimização Combinatória, dentre outras heurísticas, suas características tornam os AGs atrativos no desenvolvimento de soluções que lidam com vastos espaços de busca, tais como os problemas biológicos, embora não garantam convergência a soluções ótimas. Sua eficiência concentra-se nas estratégias adotadas que garantam diversidade populacional e elevada exploração do espaço de soluções. Por conta da explosão combinatória do número de conformações factíveis definido pelo problema (LEVINTHAL, 1969), em Otimização Combinatória e Teoria da Complexidade, o PSP é classificado como um problema NP-Completo (CRESCENZI et al., 1998).

1.2 Objetivo e Organização do Trabalho

O objetivo deste trabalho é estudar a aplicação de diferentes estratégias de seleção e organização populacionais sobre algoritmos genéticos para o problema da predição da estrutura tridimensional aproximada de proteínas e realizar a análise da qualidade dos resultados da melhor estratégia. O estudo da aplicação de populações estruturadas será realizado com o intuito de investigar métodos que garantam a diversidade durante a fase de simulação e a convergência para soluções com conformações preditas próximas às conformações nativas em um curto espaço de tempo. Essas conformações aproximadas podem servir como entrada para algoritmos *ab initio* que empregam técnicas de refinamento mais complexas e custosas, tais os que utilizam simulação por dinâmica molecular. A importância em predizer estruturas 3D aproximadas e, sobre as mesmas, utilizar técnicas de refinamento para obter conformações altamente precisas está, principalmente, relacionada à análise de proteínas patogênicas e ao estudo e desenvolvimento de drogas racionalmente desenhadas que possam inibir a ação dessas proteínas ao se fixarem em

seus sítio de ligação. Não é objetivo deste trabalho realizar um estudo da complexidade e análise de desempenho dos métodos apresentados – embora sejam informações relevantes ao problema.

Este trabalho está organizado em sete capítulos. No Capítulo 2, é realizada uma breve revisão de assuntos relacionados à proteômica, tais como polipeptídeos, ligações peptídicas, níveis hierárquicos estruturais de proteínas e repositórios de informações estruturais. No Capítulo 3, é apresentada uma classificação em quatro grupos dos métodos computacionais que buscam a solução do PSP. No Capítulo 4, são apresentadas decisões de projeto relacionadas com o estudo e desenvolvimento de métodos *ab initio* para o PSP, são, então, apontadas as escolhas realizadas e são descritos os algoritmos genéticos com população estruturada. Esses algoritmos foram estudados e desenvolvidos durante o trabalho como metodologia de busca de conformações e como uma abordagem à predição *ab initio*. No Capítulo 5, resultados computacionais e análises qualitativas sobre as melhores estruturas tridimensionais preditas são apresentados. No Capítulo 6, são apresentadas as considerações finais sobre o trabalho e suas contribuições.

2 PROTEÍNAS

Proteínas são macromoléculas responsáveis por diversos processos complexos nos seres vivos e são suas principais constituintes. São formadas por um ou mais cadeias polipeptídicas – sequências de aminoácidos ligados através de ligações peptídicas – que em condições fisiológicas adotam estruturas funcionais estáveis, únicas e invariáveis (DORN, 2008) através de um processo chamado *enovelamento*, ou *dobramento* (BRANDEN; TOOZE, 1998).

A estrutura funcional, ou conformação nativa de determinada proteína, estabelece sua função bioquímica no organismo. Dentre as funções desempenhadas pelas proteínas estão (BRANDEN; TOOZE, 1998):

- Função enzimática (ou de catálise): desempenhada pelas enzimas – proteínas capazes de catalisar reações bioquímicas específicas, mas não limitadas a catalisar, cada uma, apenas uma reação química, desde que respeitado o sítio ativo;
- Função estrutural: função relacionada à estrutura dos tecidos;
- Função de defesa: função desempenhada pelos denominados *anticorpos* que, ao combinar-se, quimicamente, com um antígeno específico, neutralizam ações de proteínas "estranhas" no organismo;
- Função hormonal: exercida sobre algum órgão de um organismo vivo; e
- Função de transporte: realizada, por exemplo, no transporte dos gases oxigênio e carbônico realizado por proteínas como a hemoglobina e hemocianina.

É possível conhecer a função específica de determinada proteína a partir do conhecimento de sua estrutura tridimensional. Esse conhecimento é fundamental no desenvolvimento de drogas racionalmente desenhadas, as quais são substâncias capazes de se ligarem quimicamente a determinada proteína e, deste modo, estimular, restringir e até mesmo suspender a ação biológica da proteína.

Este capítulo fará uma breve revisão de assuntos diretamente ligados à proteômica e que precisam ser introduzidos para um completo entendimento do objetivo deste trabalho.

2.1 Níveis de Organização Estrutural

Proteínas podem ser representadas e estudadas em até quatro níveis distintos de organização estrutural (LEHNINGER; NELSON; COX, 2005): primário, secundário, terciário e quartenário. Essas estruturas são apresentadas nas subseções seguintes.

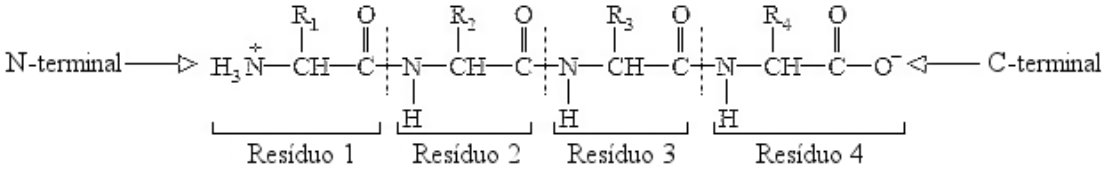

Figura 2.1: Representação esquemática da estrutura primária de um polipeptídeo. Indicações das regiões de *C-terminal* e *N-terminal*, ligações peptídicas representadas pelas linhas verticais tracejadas e indicação de cada resíduo de aminoácido R_i que compõe o polipeptídeo.

2.1.1 Estrutura Primária

É o nível estrutural mais simples, representado pela sequência de resíduos de aminoácidos ao longo da cadeia polipeptídica em ordem linear (LEHNINGER; NELSON; COX, 2005), sem preocupação com orientação espacial da molécula. É deste nível que todo arranjo espacial da molécula é derivado, portanto, possui grande importância na proteômica (WIKIPEDIA, 2011a). Cada resíduo é ligado a outro resíduo de aminoácido através de uma ligação peptídica (Figura 2.1). Esta longa cadeia resultante é determinada pela duas extremidades "amino terminal", ou *N-terminal* e "carboxi terminal", ou *C-terminal* (Figura 2.1).

2.1.2 Estrutura Secundária

Embora as proteínas sejam são polímeros lineares e, portanto, estarem aptas a assumir diversas conformações tridimensionais, na maioria dos casos essas conformações apresentam certa regularidade. Arranjos estáveis de resíduos de aminoácidos formam padrões estruturais, ou regulares que representam o nível secundário de organização estrutural de uma proteína (DORN, 2008). Essa regularidade é mantida devido a presença de interações intermoleculares, tais como *pontes de hidrogênio* entre os átomos de hidrogênio dos grupos amino ($R - NH-$) e os átomos de oxigênio dos grupos carboxilos ($R - CO-$), na cadeia polipeptídica.

Neste nível estrutural, duas estruturas regulares são as que mais se destacam pela ocorrência: as α-*hélices* e as *folhas*-β (PAULING; COREY; BRANSON, 1951). Há também outras estruturas "periódicas" e irregulares tais como *voltas* e *alças* que são responsáveis pela união das estruturas secundárias regulares. Abaixo, são descritas as estruturas secundárias comumente encontradas.

- *Hélices* α são estruturas regulares cuja principal força de estabilização são as pontes de hidrogênio entre os grupos amino e carboxilo do mesmo segmento (LESK, 2000; PAULING; COREY; BRANSON, 1951). Alguns resíduos possuem maior propicialidade em formar as α-hélices (BRANDEN; TOOZE, 1998) cujas ligações de hidrogênio entre cada volta sucessiva e voltas adjacentes são as interações responsáveis em assegurar a estabilidade da estrutura helicoidal. Os ângulos diedros (ϕ e ψ) dos resíduos de aminoácidos com estrutura regular α-hélice variam no mapa de Ramachandran (RAMACHANDRAN; SASISEKHARAN, 1968) em torno de -30° a -120° para ϕ e -60° a 20° para ψ (DORN, 2008).

- *Folhas* β são estruturas regulares formadas quando as estruturas polipeptídicas estão dispostas lado a lado (PAULING; COREY; BRANSON, 1951). As folhas-β consistem em cadeias polipeptídicas estendidas que possuem outras cadeias poli-

peptídicas vizinhas adjacentes e também são estabilizadas por pontes de hidrogênio que são formadas entre os grupos amino e carboxilo das duas cadeias. As cadeias adjacentes em uma folha-β podem ser ou paralelas, ou antiparalelas, cada uma possuindo padrões de ligações de hidrogênio distintos (PAULING; COREY, 1951). Os ângulos diedros destas estruturas secundárias assumem valores que variam de $-180°$ a $-45°$ para ϕ e $45°$ a $225°$ para ψ (DORN, 2008).

- *Alças e voltas* são os nomes denominados às estruturas secundárias irregulares que são formadas em regiões onde o polipeptídeo muda sua direção, após segmentos de estruturas secundárias regulares. Devido sua irregularidade não possuem uma região específica no mapa de Ramachandran, logo, seus ângulos diedros podem ocupar quaisquer regiões incluindo regiões de folhas-β e de α-hélices (DORN, 2008). Visto o vasto espaço conformacional no qual podem ser encontradas, essas dobras dificultam a predição computacional das estruturas funcionais dos polipeptídeos.

2.1.3 Estrutura Terciária

A estrutura terciária é resultante do enrolamento e distribuição espacial das estruturas secundárias. A forma tridimensional assumida pela proteína é também chamada de estrutura nativa da proteína ou estrutura funcional (DORN, 2008). A estrutura nativa da proteína é determinada por interações moleculares de longa distância – diferentemente das estruturas secundárias – tais como interações hidrofóbicas, eletrostáticas, pontes de hidrogênio, pontes de sulfeto e forças de van der Waals (GIBAS; JAMBECK, 2001). Além disso, as cadeias laterais, definidas por cada resíduo de aminoácido, também possuem um papel principal na conformação funcional final de um polipeptídeo (SCHEEF; FINK, 2003).

A estrutura terciária confere a atividade biológica às proteínas, através do seu conhecimento é possível analisar e predizer a função de determinada proteína em uma célula. Com isso, é possível identificar o sítio ativo, ou de ligação de uma proteína (LEHNINGER; NELSON; COX, 2005) possibilitando o desenvolvimento de fármacos que interfiram em seu funcionamento. Por exemplo, no caso de uma proteína patogênica de estrutura conhecida, drogas podem ser quimicamente desenhadas para se adaptarem em sítios específicos da mesma e, assim, bloquearem sua ação no organismo.

2.1.4 Estrutura Quaternária

Algumas proteínas podem apresentar duas ou mais cadeias polipeptídicas, cada uma denominada de "subunidade", exibindo um nível de organização estrutural a mais. O arranjo espacial dessas subunidades em suas formas terciárias e suas interações formam a estrutura quaternária. Esta estrutura é mantida pelas mesmas forças que determinam os níveis estruturais anteriores (DORN, 2008).

2.2 Métodos Experimentais para Determinação da Estrutura 3D das Proteínas

Experimentalmente, através de técnicas de cristalografia por difração de raios-X, ou por ressonância magnética nuclear de proteínas – comumente, complementares – é possível obter a estrutura tridimensional de uma proteína.

Técnicas cristalográficas de difração de raios-X são as técnicas mais antigas, porém precisas, utilizadas na determinação das estruturas nativas de proteínas. As técnicas per-

mitem determinar os arranjos tridimensionais das proteínas, não limitando o tamanho das moléculas em estudo, porém, causam danos às amostras devido a radiação aplicada, impossibilitando que a dinâmica das interações entre as proteínas e substratos seja analisada (BAXEVANIS; QUELLETTE, 1990).

A ressonância magnética nuclear é uma técnica mais nova cuja vantagem é permitir estudar a estrutura e dinâmica de determinada proteína em substratos líquidos ou em ambiente fisiológico (DORN, 2008). No caso de proteínas que só podem ser estudadas em solução aquosa, a ressonância nuclear magnética é ferramenta fundamental.

A desvantagem dos métodos experimentais – altos custos e grau de complexidade dos experimentos – e grande volume de estruturas ainda não solucionadas motivou pesquisadores a desenvolverem metodologias computacionais que buscam a predição correta, ou aproximada de estruturas funcionais de proteínas, apenas a partir de suas estruturas primárias e, em alguns casos, informações estruturais semelhantes. No capítulo 3, uma revisão dessas metodologias *in silico* é apresentada.

2.3 Repositórios de Informações Estruturais de Proteínas

Entre os repositórios de informações estruturais de proteínas mais conhecidos, está o *Protein Data Bank* (PDB) (BERMAN et al., 2000), localizado nos Estados Unidos, cujo propósito principal é o armazenamento e a catalogação das informações estruturais de macromoléculas. Outros bancos de dados estruturais tais como PDBe (localizado na Europa) (VELANKAR et al., 2004) e PDBj (localizado no Japão) também existem com o mesmo propósito de organização e distribuição de dados no formato PDB.

Projetos que visam uma base de dados única de estruturas tridimensionais estão em desenvolvimento. A iniciativa, que é composta pelos pesquisadores das organizações mencionadas e pesquisadores do BMRB (*Biological Magnetic Resonance Data Bank*), também localizado nos Estados Unidos, trata-se do wwPDB (*Worldwide Protein Data Bank*) (BERMAN et al., 2007). Os esforços desses pesquisadores está na missão de manter um único repositório PDB de informações estruturais de macromoléculas gratuito e disponível a comunidade global[1].

Esses repositórios são fundamentais para o estudo da Bioinformática Estrutural, visto que fornecem conteúdo para o estudo estrutural de biomoléculas e para construção e validação de estruturas preditas por metodologias computacionais.

[1] http://www.wwpdb.org/

3 MÉTODOS DE PREDIÇÃO *IN SILICO* DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS

Muitos métodos e algoritmos têm sido propostos e analisados, ao longo dos anos, como soluções para o complexo problema da predição da estrutura nativa de proteínas (ZHANG, 2008; WU; SKOLNICK; ZHANG, 2007; XU; PENG; ZHAO, 2009; HILDEBRAND et al., 2009; KRIEGER et al., 2009; ZHANG, 2007; MOULT, 2005; ZHOU; SKOLNICK, 2009; OSGUTHORPE, 2000; CUTELLO; NARZISI; NICOSIA, 2006; TRAMONTANO, 2006; BUJNICKI, 2006; ROHL et al., 2004; SIMONS et al., 1999; JONES; TAYLOR; THORNTON, 1992; JONES, 2001; SRINIVASAN; ROSE, 2002, 1995).

Na literatura, encontra-se diversas classificações de métodos de predição de estruturas 3D de proteínas. Neste estudo, adotou-se uma classificação similar a descrita em (FLOUDAS et al., 2006) a qual classifica os métodos computacionais em quatro grupos:

- Métodos de Modelagem Comparativa,
- Métodos de Reconhecimento de Enovelamentos,
- Métodos de primeiros princípios sem informação de base de dados (*ab initio*) e
- Métodos de primeiros princípios com informação de base de dados (*de novo*).

Neste capítulo, uma breve revisão da metodologia utilizada por cada grupo é realizada nas seções seguintes.

3.1 Método de Modelagem Comparativa

A modelagem comparativa é baseada na observação que a estrutura terciária de proteínas é mais bem conservada que suas sequências de aminoácidos (MARTÌ-RENOM M.A.; SALI, 2000) para proteínas homólogas; entretanto, identidade entre sequências de aminoácidos inferior a 20% pode indicar estruturas funcionais muito distintas (CHOTHIA; LESK, 1986). Através do alinhamento da sequência alvo em uma sequência modelo, de estrutura 3D conhecida no PDB (ESWAR et al., 2006), o objetivo da modelagem comparativa é predizer a estrutura de determinada proteína tridimensional com uma precisão comparável aos melhores resultados alcançados experimentalmente. Este método é empregado, principalmente, quando determinada proteína é muito longa para sua estrutura ser predita via ressonância magnética e não pode ser predita por cristalografia de difração de raios-X, mas há o conhecimento da estrutura de proteínas da mesma família. Após o alinhamento e detecção de estruturas homólogas – limiar estabelecido, geralmente, em

30%, inclusive para proteínas longas –, o procedimento de modelagem é realizado através das cópias de coordenadas, distâncias interatômicas e ângulos de torção da proteína modelo.

A modelagem comparativa por homologia é o método de predição com maior acurácia nos resultados finais (JONES; TAYLOR; THORNTON, 1992; MARTÌ-RENOM M.A.; SALI, 2000; TRAMONTANO, 2006; DORN, 2008), embora esta precisão seja dependente do grau de identidade entre proteína alvo e modelos utilizados. Geralmente, uma identidade de sequencia acima de 50% produz estruturas bem confiáveis e apresenta erros típicos aos obtidos por ressonância magnética nuclear. Abaixo de um limiar de 30%, muitos erros podem ocorrer levando a predição de estruturas mal formadas e não confiáveis. O método possui como desvantagem a impossibilidade da descoberta de novos enovelamentos e do estudo do processo de enovelamento de proteínas, ou seja, o processo iterativo de dobra e estabilização das moléculas até seu estado nativo (DORN, 2008).

3.2 Método de Reconhecimento de Enovelamentos

Métodos baseados nos reconhecimento de enovelamentos são motivados pela observação de que poucos enovelamentos distintos devem existir na natureza – aproximadamente, 2.000 (GOVINDARAJAN; RECABARREN; GOLDSTEIN, 1999) – e que muitas das estruturas submetidas ao PDB nos últimos anos possuíam estruturas funcionais muito similares às estruturas já existentes no repositório. Esses métodos permitem tratar casos em que somente resultados de baixa homologia sequencial são obtidos via modelagem comparativa – abaixo de 15%, por exemplo –, mas é identificado que determinada proteína alvo possui um estado nativo muito próximo ao de estruturas já conhecidas (PENG; XU, 2010).

Dentro dessa metodologia, o método mais comum de reconhecimento é o alinhavamento de proteínas (*protein threading*) (JONES; TAYLOR; THORNTON, 1992; TRAMONTANO, 2006; DORN, 2008). Este método trabalha utilizando conhecimento estatístico da relação entre estruturas conhecidas em repositórios e da sequência da proteína alvo (BUJNICKI, 2006; WIKIPEDIA, 2011b). Os resíduos da proteína alvo são arranjados espacialmente conforme os moldes utilizados de estruturas conhecidas e para cada arranjo é calculada sua aptidão através de funções de "pontuação". Habitualmente, o objetivo também está relacionado a minimizar a energia livre de determinados arranjos, inclusive com o refinamento posterior das estruturas da molécula predita.

3.3 Métodos *ab initio*

Métodos *ab initio* não utilizam informações de bases de dados estruturais de proteínas e são fundamentados na observação que a estrutura nativa de uma proteína corresponde a uma estrutura cuja energia livre é mínima. O objetivo do método é encontrar o valor mínimo global da energia livre de uma proteína que corresponderia, ou à estrutura nativa, ou a uma conformação funcional da mesma (OSGUTHORPE, 2000). Pode-se dividir a predição aproximada por métodos *ab initio* em dois subproblemas: 1) calcular a energia de uma determinada conformação e 2) adotar uma estratégia de busca, no espaço conformacional, por conformações factíveis (TRAMONTANO, 2006). Essa estratégia de busca corresponde ao método utilizado na exploração do espaço conformacional através de alterações das conformações correntes a cada iteração, na tentativa de encontrar uma estrutura tridimensional que apresente a mais baixa energia potencial (OSGUTHORPE,

2000; DORN, 2008).

A principal vantagem dos métodos pertencentes a esta classe é que eles são capazes de prever novos enovelamentos de proteínas, visto que não são limitados a modelos encontrados, por exemplo, no PDB.

3.4 Métodos *de novo*

Em métodos *de novo*, padrões estruturais são extraídos do PDB e utilizados na construção de estruturas tridimensionais de proteínas (TRAMONTANO, 2006); entretanto, a comparação com estas estruturas extraídas de repositórios não é a sua essência. Métodos pertencentes a esta classe, frequentemente, não comparam inteiramente uma sequência alvo contra uma sequência de uma proteína modelo com estrutura tridimensional conhecida, mas comparam pequenos fragmentos de resíduos de aminoácidos e os combinam com o objetivo de encontrar a estrutura 3D da proteína cuja energia potencial seja a menor (estratégia derivada de métodos *ab initio*), logo, possibilitando, também, a predição de novos enovelamentos. As estratégias utilizadas na classificação e exploração de todas as possíveis combinações de fragmentos é o aspecto fundamental das predições *de novo* (DORN, 2008). A classificação dos fragmentos tem o objetivo de organizar os fragmentos mais aptos a ocupar determinadas regiões da cadeia principal de uma proteína-alvo, através das análises de similaridade e interações entre outros fragmentos da proteína que podem influenciar o enovelamento da mesma (ROHL et al., 2004; DAS et al., 2007; DORN, 2008).

4 ALGORITMOS GENÉTICOS PARA A PREDIÇÃO APROXIMADA DA ESTRUTURA 3-D DE PROTEÍNAS

Neste capítulo, as diferentes estratégias de seleção e organização populacional sobre os algoritmos genéticos desenvolvidos durante o estudo são apresentadas. A primeira seção apresenta as decisões preliminares do desenvolvimento de algoritmos de predição *ab initio*. A segunda seção do capítulo descreve os dados utilizados como entrada dos algoritmos estudados e as estratégias em comum entre eles. As demais seções apresentam as estratégias particulares de cada algoritmo genético aplicado ao PSP durante o estudo. Os pseudocódigos desses métodos também são apresentados e foram escritos utilizando o arquivo de estilo disponibilizado por (KREHER; STINSON, 1999).

4.1 Preliminares

Nesta seção, é dada uma visão geral de certas decisões de projeto relacionadas com o estudo e desenvolvimento de métodos *ab initio* para o PSP e apontadas as escolhas realizadas para os métodos estudados neste trabalho.

4.1.1 Representação da Cadeia Polipeptídica

Um polipeptídeo é uma molécula composta de dois ou mais aminoácidos ligados, em cadeia, por uma ligação peptídica, ver tracejado vertical da Figura 4.1. Uma ligação peptídica é um enlace amida, uma ligação química covalente – onde há o compartilhamento de um ou mais pares de elétrons entre os átomos – formada entre duas moléculas quando o grupo carboxilo de uma molécula reage com o grupo amino de outra molécula, desde modo liberando uma molécula de água (H_2O). Os resíduos de aminoácidos diferem um dos outros através do grupo R ligado ao C_α. A cadeia principal de um peptídeo possui três ângulos de torção chamados phi ϕ, psi ψ e ômega ω. No modelo apresentado na Figura 4.1 as ligações entre N e C_α e entre C_α e C possuem rotação livre. Estas rotações são descritas pelos ângulos de rotação ϕ e ψ respectivamente. Os ângulos de torção da cadeia principal ϕ e ψ são os maiores responsáveis por determinar a conformação de um polipeptídeo (BRANDEN; TOOZE, 1998).

De acordo com (CUTELLO; NARZISI; NICOSIA, 2006), existem cinco representações de cadeias polipeptídicas comumente utilizadas no estudo de suas conformações:

1. coordenadas tridimensionais de todos os átomos;

2. coordenadas de todos *átomos pesados*;

3. coordenadas dos átomos da cadeia principal e centroides das cadeias laterais;

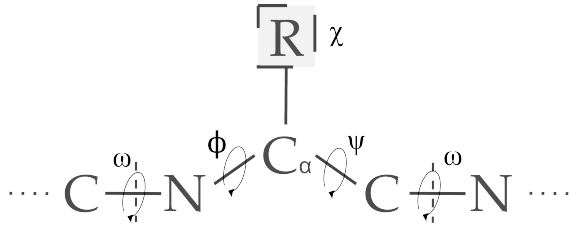

Figura 4.1: Representação esquemática de um modelo de peptídeo. N é nitrogênio, C e C_α são carbonos e R é uma cadeia lateral arbitrária. As linhas verticais tracejadas identificam a ligação peptídica.

4. coordenadas do C_α; e

5. ângulos de torção das cadeias principal e laterais.

Neste trabalho, a estrutura de um polipeptídeo é representada baseando-se apenas nos ângulos de torção da cadeia principal (ϕ, ψ, ω) e da cadeia lateral (χ_i). A Figura 4.1 representa, esquematicamente, toda informação acerca do modelo de peptídeo adotado. Todos os ângulos de torção das ligações peptídicas (ω) têm seus valores fixados em $180°$.

4.1.2 Função de Energia Potencial

Embora funções de avaliação mais precisas, tais as utilizadas em mecânica quântica, estejam disponíveis, estas são, computacionalmente, muito complexas e pouco práticas para uso em grandes sistemas computacionais de modelagem dos quais breve "rendimento" é esperado. Deste modo, funções de energia mais "baratas" e menos precisas, por usarem física clássica, são utilizadas como funções de avaliação de conformações de polipeptídeos.

A energia interna da proteína e suas interações com o ambiente no qual está inserida é descrita por uma função de energia. Funções de energia são utilizadas em métodos *ab initio* para avaliar a conformação no decorrer da simulação e, desta forma, encontrar a conformação com o valor global mínimo de energia livre. Uma função de energia potencial – como é chamada – retorna o valor aproximado da energia potencial de uma determinada conformação da proteína e incorpora dois tipos de termos: ligados e não-ligados (ver Eq. 4.1). Os termos ligados (`lig`, `ang` e `die`) referem-se as interações covalentemente ligadas e correspondem, respectivamente, a energia potencial liberada pela ligação de dois átomos, o potencial angular e o potencial dos ângulos de torção diedros. Os dois primeiros termos auxiliam em restringir os comprimentos das ligações e ângulos para valores próximos aos de equilíbrio. O terceiro termo modela as barreiras energéticas periódicas encontradas durante a rotação de uma ligação. Os termos não-ligados representam ligações iônicas, interações hidrofóbicas, pontes de hidrogênio, interações de *van der Waals* e interações *Dipolo-Dipolo*. São forças mais fracas que as forças de ligações covalentes; entretanto, elas contribuem de maneira expressiva para a estabilidade da molécula. Nos algoritmos genéticos do estudo, a função de energia potencial AMBER94 (CORNELL et al., 1995) é utilizada como função objetivo. AMBER possui um bom desempenho é uma das mais largamente utilizadas famílias de campos de força – usados para descrever

a energia potencial – cuja forma funcional é apresentada por 4.1.

$$Energia = \sum_{\texttt{lig}} \frac{1}{2} K_b (b - b_0)^2 + \sum_{\texttt{ang}} \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \sum_{\texttt{die}} \frac{1}{2} K_\eta (1 + \cos(\eta_\omega - \gamma)) + \sum_{\text{j=1}}^{N-1} \sum_{\text{i=j+1}}^{N-1} \left\{ \epsilon_{i,j} \left[\left(\frac{R_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \tag{4.1}$$

onde:

- O termo `lig` representa a energia entre átomos covalentemente ligados;
- O segundo termo `ang` representa a energia devida a geometria dos orbitais de elétrons envolvida na ligação covalente;
- O termo `die` representa a energia despendida na torção de uma ligação e é relacionada à ordem de ligação e ligações vizinhas ou pares solitários de elétrons;
- O último termo representa a energia das forças fracas entre todos pares de átomos e pode ser decomposta em *van der Waals* (primeiro termo do somatório duplo) e energias eletrostáticas (segundo termo do somatório duplo).

4.1.3 Avaliação de Similaridade

Quando a conformação nativa, isto é, obtida por meios experimentais, de um polipeptídeo está disponível, é possível avaliar o quão similar é uma conformação obtida *in silico* de uma experimental. Para isso, empregamos o cálculo do desvio quadrático médio (RMSD) dos átomos "carbono-alfa" das cadeias principais dos polipeptídeos envolvidos. O RMSD é dado pela fórmula

$$RMSD(a, b) = \sqrt{\frac{\sum_{i=1}^{n} (r_{ai} - r_{bi})^2}{n}}, \tag{4.2}$$

onde r_{ai} e r_{bi} são as posições do átomo i da estrutura a e estrutura b, respectivamente, estando estas estruturas sobrepostas no átomo de carbono central de cada uma.

4.2 Entrada e Estratégias Padrão

4.2.1 Dados Pré-Processados

Como entrada, os algoritmos recebem um modelo do polipeptídeo a ser analisado. Este modelo é um arquivo contendo os dois primeiros níveis estruturais do polipeptídeo: a sequência de aminoácidos da cadeia polipeptídica (estrutura primária) e as subestruturas regulares definidas por padrões de pontes de hidrogênio entre os grupos peptídicos (estrutura secundária) (PAULING; COREY; BRANSON, 1951).

Visto que os resultados obtidos são comparados às estruturas determinadas experimentalmente do *Protein Data Bank*, o programa PROMOTIF (HUTCHINSON; THORNTON, 1996) é utilizado na análise e determinação das estruturas secundárias das proteínas experimentais. Entretanto, seria perfeitamente possível determinar estruturas secundárias de polipeptídeos sem o uso desse software, através de métodos padronizados de atribuição de estrutura secundária, tais como o algoritmo DSSP (KABSCH; SANDER, 1983). Utilizando dados de predição de estrutura secundária, diminui-se o espaço conformacional de

Estrutura Secundária	ϕ	ψ
H (α-hélice)	$[-67°, -47°]$	$[-57°, -37°]$
B (folha-β)	$[-130°, -110°]$	$[110°, 130°]$
E (folha-β)	$[-130°, -110°]$	$[110°, 130°]$
G	$[-59°, -39°]$	$[-36°, -16°]$
I	$[-67°, -47°]$	$[-80°, -60°]$
T	$[-180°, 180°]$	$[-180°, 180°]$
S	$[-180°, 180°]$	$[-180°, 180°]$
indefinida	$[-180°, 180°]$	$[-180°, 180°]$

Tabela 4.1: Regiões correspondentes às restrições definidas pelas estruturas secundárias regulares e irregulares.

Resíduo	Número de ângulos χ
GLY, ALA, PRO	nenhum
SER, CYS, THR, VAL	χ_1
ILE, LEU, ASP, ASN, HIS, PHE, TYR, TRP	χ_1, χ_2
MET, GLU, GLN	χ_1, χ_2, χ_3
LYS, ARG	$\chi_1, \chi_2, \chi_3, \chi_4$

Tabela 4.2: Número de ângulos χ necessários para estabelecer os ângulos de torção das cadeias laterais para cada tipo de resíduo.

busca, pois a cada subestrutura está associado um intervalo no espaço. A Tabela 4.1 descreve as regiões factíveis do espaço conformacional para os ângulos de torção da cadeia principal, cuja estrutura secundária – regular, ou irregular – foi atribuída.

Como apresentado anteriormente, os métodos de predição também consideram os ângulos de torção das cadeias laterais. Para prover tal informação, como número de ângulos χ por tipo de resíduo e suas regiões restritas, os algoritmos utilizam dados de predição geométrica de cadeias laterais fornecidos pela biblioteca de rotâmeros de cadeia lateral SCWRL4 (KRIVOV; SHAPOVALOV; DUNBRACK, 2009). A Tabela 4.2 mostra o número de ângulos χ necessários para fixar a posição dos átomos de cadeia lateral em cada tipo de resíduo.

4.2.2 Representação e População Inicial

Visto que todos os algoritmos estudados trabalham sobre valores de ângulos de torção e uma simples sequência de aminoácidos define cada polipeptídeo, foi utilizada uma codificação trivial baseada em vetores como representação interna dos dados de uma população. Cada cromossomo A de um indivíduo é representado por um vetor de estruturas de aminoácidos – os genes (ver Figura 4.2). Cada estrutura de aminoácido consiste de informações básicas, tais como tipo do aminoácido, estrutura secundária predita, número de ângulos da cadeia lateral e valores correntes para os ângulos ϕ, ψ e χ_i, quando apropriado.

Levando em consideração que os métodos estudados centram-se em manter a diversidade populacional enquanto aprimoram a melhor solução disponível, foi assumido, como melhor alternativa, produzir a primeira geração aleatoriamente. A população inicial é ge-

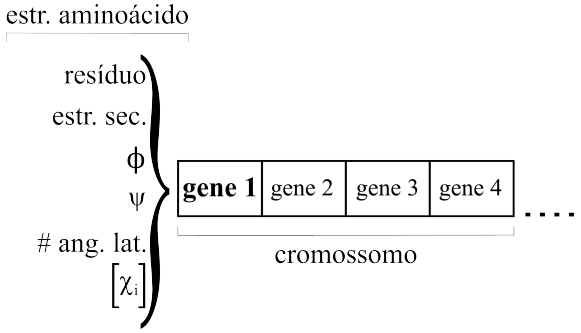

Figura 4.2: Representação esquemática do cromossomo de um indivíduo.

rada de maneira aleatória escolhendo valores factíveis do espaço conformacional de busca para cada ângulo de torção de cada estrutura de aminoácido de um cromossomo.

4.2.3 Avaliação de Indivíduos

A fim de avaliar a conformação de uma determinada solução, a função de energia potencial AMBER (Eq. 4.1), descrita anteriormente, foi utilizada como função objetivo. Para associar cada solução ao seu valor de aptidão e converter a representação baseada em ângulos de torção na representação de coordenadas tridimensionais atômicas, foram usadas as rotinas externas `protein` e `analyze` do pacote de mecânica molecular TINKER (PONDER; RICHARDS, 1987; PONDER, 1998) e o conjunto de parâmetros de mecânica molecular para o campo de força AMBER. A função de avaliação retorna um valor para a energia baseada na conformação molecular provida pela solução. Conforme os métodos *ab initio*, considerando que a busca pela estrutura nativa de um polipeptídeo é baseada somente na sequência primária de aminoácidos e estrutura secundária predita, os métodos estudados buscam pelo valor global mínimo de energia, logo, quanto menor a energia, mais apta a solução.

4.2.4 Política de Substituição de Indivíduos

Foi implementada uma política de substituição para evitar a convergência prematura dos métodos. O algoritmo não permite a coexistência de duas ou mais soluções de mesma energia. Portanto, cada solução avaliada em um mesmo valor de uma outra é marcada, substituída por uma solução factível aleatória e realocada na população. Consequentemente, esse procedimento contribui garantindo a diversidade em cada nova população estabelecida.

4.2.5 Estratégia de Crossover

O procedimento de recombinação utilizado sobre os pares de pais selecionados por cada método é chamado *random keys* (BEAN, 1994). Com a finalidade de combinar-se dois pais p_1 (mais apto) e p_2 (menos apto), primeiro o algoritmo gera um vetor R de número reais entre 0 e 1, escolhidos ao acaso, de tamanho $|A|$. Um número de corte K, no intervalo $(0,5\ ;\ 1]$, determinará se um gene – uma estrutura de aminoácido – é herdado ou de p_1, ou de p_2. Se um valor na posição i no vetor R for menor que K, um gene da mesma posição é selecionado do pai p_1, caso contrário, é selecionado de p_2. Essa abordagem, especificamente, é conhecida como *biased random-key* (GONÇALVES; RESENDE, 2010), pois dá ao pai de maior aptidão maior probabilidade de passar suas características a gerações futuras.

Na implementação do método, adotamos um corte $K = 0,7$ o qual, experimentalmente, produziu bons resultados relacionados a convergência, embora não seja incomum o algoritmo genético experienciar convergências prematuras para mínimos locais.

4.3 Método Utilizando Seleção Aleatória

4.3.1 Organização Populacional

A população é organizada de maneira trivial, em um vetor não ordenado.

4.3.2 Seleção de Pais

A seleção dos pais para recombinação é baseada no algoritmo de Seleção por Torneio (BAECK, 1996), de maneira aleatória não considerando a aptidão dos indivíduos.

4.3.3 Mutação

Após a seleção e recombinação dos pais, mutações sobre os indivíduos são aplicadas com uma determinada probabilidade. O procedimento de mutação aplicado foi algoritmo de mutação Gaussiana (SCHWEFEL, 1981). A probabilidade adotada, com fins de aumentar a diversidade e exploração do espaço conformacional, foi de 70%.

4.3.4 Pseudocódigo

Um pseudocódigo das estratégias fundamentais do método é apresentado pelo Algoritmo 4.3.1. No pseudocódigo, uma população P, no tempo t, é representada por $P[t]$. O código faz uso da variável temporária P', a qual armazena os indivíduos selecionados por torneio.

Algoritmo 4.3.1: METODOALEATORIO(P)

$t \leftarrow 0$
INICIALIZA($P[t]$)
AVALIA($P[t]$)
APLICAPOLITICASUBSTITUICAO($P[t]$)
se $houveSubstituicao$
  então AVALIA($P[t]$)
enquanto $criterioParadaNaoAtendido$
faça
- $t \leftarrow t + 1$
- $P' \leftarrow$ SELECAOPORTORNEIO($P[t-1]$)
- $P[t] \leftarrow$ REALIZACROSSOVER(P')
- $P[t] \leftarrow$ APLICAMUTACAOGAUSSIANA($P[t]$)
- AVALIA($P[t]$)
- APLICAPOLITICASUBSTITUICAO($P[t]$)
- **se** $houveSubstituicao$
  - **então** AVALIA($P[t]$)

4.4 Método Utilizando Seleção Probabilística

4.4.1 Organização Populacional

A população é organizada de maneira trivial, em um vetor não ordenado.

4.4.2 Seleção de Pais

A seleção dos pais para recombinação é dada através do algoritmo de seleção probabilística da Roleta (DE JONG, 1975; GOLDBERG; DEB, 1991; BAECK, 1996). Os pais são selecionados de acordo com seus valores de aptidão. Quanto mais apto um indivíduo for, maiores suas chances de ser selecionado e transmitir seus genes para gerações futuras.

4.4.3 Mutação

Para o método, também foi adotado o algoritmo de mutação Gaussiana com uma probabilidade de 70% de chance de um indivíduo sofrer mutações em genes aleatórios.

4.4.4 Pseudocódigo

O pseudocódigo do método é apresentado através do Algoritmo 4.4.1. Como no caso do pseudocódigo apresentado na Seção 4.3, uma população P, em um determinado tempo t, é representada por $P[t]$ e a variável temporária P' armazena os pais selecionados, neste caso, pela técnica probabilística da roleta.

Algoritmo 4.4.1: METODOPROBABILISTICO(P)

```
t ← 0
INICIALIZA(P[t])
AVALIA(P[t])
APLICAPOLITICASUBSTITUICAO(P[t])
se houveSubstituicao
  então AVALIA(P[t])
enquanto criterioParadaNaoAtendido
  faça { t ← t + 1
         P' ← SELECAOPORROLETA(P[t − 1])
         P[t] ← REALIZACROSSOVER(P')
         P[t] ← APLICAMUTACAOGAUSSIANA(P[t])
         AVALIA(P[t])
         APLICAPOLITICASUBSTITUICAO(P[t])
         se houveSubstituicao
           então AVALIA(P[t])
```

4.5 Método Utilizando Árvores Ternárias

4.5.1 Organização Populacional

Neste método a população é estruturada, hierarquicamente, em uma árvore ternária, seguindo o proposto em (MOSCATO; TINETTI, 1994; BURIOL; FRANCA; MOSCATO, 2004). Como ilustrado na Figura 4.3, a estrutura em árvore ternária escolhida possui 4 níveis, altura 3, totalizando 40 nodos e 80 soluções factíveis. Cada nodo da árvore

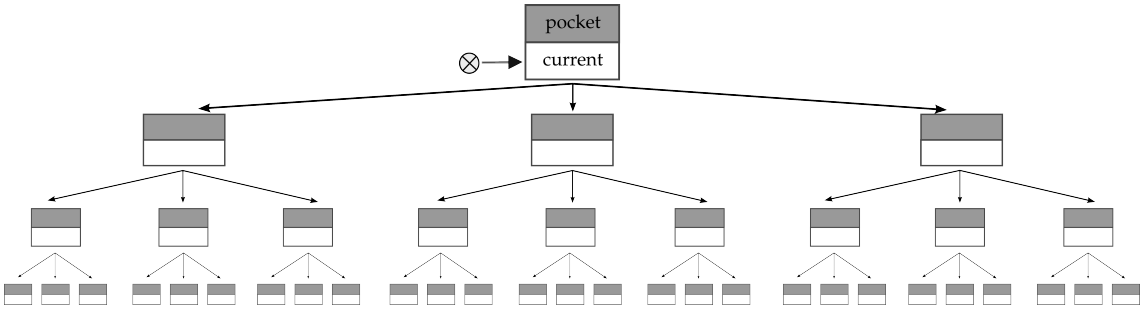

Figura 4.3: Representação esquemática da estrutura de população em árvore ternária estudada. O símbolo de crossover '⊗' indica que a substituição de uma solução factível *current* por um novo indivíduo proveniente de uma recombinação.

armazena duas soluções factíveis: *pocket* e *current*. A solução *pocket* atua como uma memória de longo prazo, mantendo sempre o controle do melhor indivíduo, de acordo com a função objetivo. A solução *current*, por sua vez, representa a prole resultante de um processo de recombinação – este podendo envolver soluções *pocket* e *current* de outro nodos –, mas que não figura o quadro de melhor indivíduo de seu nível na estrutura.

4.5.2 Seleção de Pais

A estratégia padrão adotada para garantir a diversidade populacional foi selecionar sempre um *pocket* e um *current* de níveis distintos. Esta estratégia só é quebrada no nível zero, onde dois *pockets* são selecionados como pais, um do próprio nível e outro, aleatoriamente, do nível um. Logo, as regras adotadas para seleção foram as seguintes:

- Caso o indivíduo resultante da recombinação for substituir a solução *current* do nível zero, selecionar a solução *pocket* raiz e uma solução *pocket* de um nó filho como pais do cruzamento.

- Caso contrário, na substituição de outra solução *current*, selecionar a solução *pocket* do nó pai e uma solução *current* do próprio nó como pais no crossing over.

4.5.3 Mutação

O procedimento de mutação é aplicado logo após a seleção e recombinação dos pais. A única solução factível exposta a mutação é a solução *current*. Para o método, também foi adotado o algoritmo de mutação Gaussiana com a probabilidade de 70%.

4.5.4 Políticas de Troca e Organização da Estrutura

Por fim, após a seleção, recombinação, mutação e avaliação das soluções, dois algoritmos são executadas com o objetivo de manter a estrutura organizada e hierárquica em relação a melhor solução. O método `updatePocket()`, de custo linear, efetua a troca entre soluções *pocket* e *current* de um mesmo nodo, caso a segunda seja melhor que a primeira. O método recursivo `pocketPropagation()` re-estrutura a árvore para que soluções *pocket* melhores estejam em níveis superiores ao de soluções *pocket* com menor aptidão. O procedimento recursivo explora a estrutura até os nós-folha efetuando a troca entre soluções *pocket* de um nó filho e de um nó pai caso a primeira seja melhor que a segunda. Esses dois mecanismos de re-estruturação garantem o fluxo das melhores soluções em direção ao topo da hierarquia (BURIOL; FRANCA; MOSCATO, 2004).

4.5.5 Pseudocódigo

O pseudocódigo indicado pelo Algoritmo 4.5.1 apresenta uma visão geral do método. A variável temporária P' armazena os pais selecionados conforme as regras descritas anteriormente.

Algoritmo 4.5.1: METODOARVORESTERNARIAS(P)

$t \leftarrow 0$
INICIALIZA($P[t]$)
AVALIA($P[t]$)
APLICAPOLITICASUBSTITUICAO($P[t]$)
se $houveSubstituicao$
  então AVALIA($P[t]$)
$P[t] \leftarrow$ UPDATEPOCKET($P[t]$)
$P[t] \leftarrow$ POCKETPROPAGATION($P[t]$)
enquanto $criterioParadaNaoAtendido$
faça
- $t \leftarrow t + 1$
- $P' \leftarrow$ SELECAOPAIS($P[t-1]$)
- $P[t] \leftarrow$ REALIZACROSSOVER(P')
- $P[t] \leftarrow$ APLICAMUTACAOGAUSSIANA(SOLUCOESCURRENT($P[t]$))
- AVALIA($P[t]$)
- APLICAPOLITICASUBSTITUICAO($P[t]$)
- **se** $houveSubstituicao$
  - **então** AVALIA($P[t]$)
- $P[t] \leftarrow$ UPDATEPOCKET($P[t]$)
- $P[t] \leftarrow$ POCKETPROPAGATION($P[t]$)

4.6 Método Utilizando Sistema de Castas

4.6.1 Organização Populacional

Esse método emprega um sistema de Castas, originalmente proposto em (ERICSSON; RESENDE; PARDALOS, 2002), na estruturação de populações. Usando parâmetros de controle de dimensão das castas do sistema, α e β, após os indivíduos serem ordenados de acordo com seus valores de aptidão, a população é divida em três castas. A classe alta, casta A, de proporção populacional α, é denominada casta elite. A próxima, casta B, de proporção β, é chamada de casta média. Os indivíduos da última e mais baixa casta, casta C, são o restante da população. Utilizando os parâmetros α e β é possível controlar o fator de elitismo e um balanço entre convergência e diversidade. No estudo, o valor adotado para ambos parâmetros α e β foi de 40% por ter apresentado os melhores resultados durante fases experimentais.

4.6.2 Seleção de Pais

O algoritmo combina seleção familiar e seleção individual por escolher, ao acaso, um pai da casta elite e outro de uma casta não-elite (ou casta B, ou C) (DORN; BURIOL; LAMB, 2011). Como ilustrado na Figura 4.4, para produzir uma nova geração o algoritmo: promove, inteiramente, a casta A sem qualquer alteração para a próxima geração;

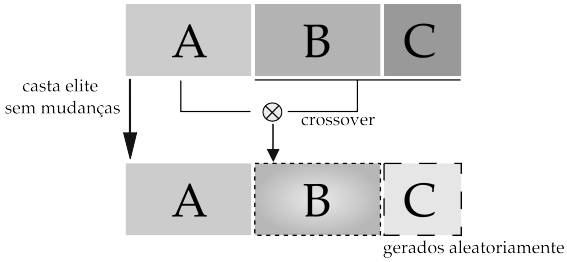

Figura 4.4: Representação esquemática da estrutura e reprodução de uma nova geração utilizando o método de castas.

seleciona uma proporção β de pares de pais, como estabelecido, para originar novos indivíduos que substituirão, em proporção, a casta média corrente; e substitui toda casta C, 20% da população, por soluções factíveis geradas aleatoriamente.

4.6.3 Garantia de Diversidade

Durante a concepção do método, não foi idealizada a implementação de uma rotina de mutação. Deste modo, uma prole final diversificada a cada geração é garantida pelo procedimento padrão de substituição de indivíduos de mesma energia.

4.6.4 Pseudocódigo

Um simples pseudocódigo do método é apresentado pelo Algoritmo 4.6.1. O código faz uso de variáveis temporárias: P' armazena indivíduos da casta elite e de castas não-elite selecionados para recombinação, conforme descrito acima; A' é uma cópia dos indivíduos da classe elite da geração anterior; B' detém os novos indivíduos produzidos pelos pais selecionados; e C' armazena os indivíduos factíveis gerados aleatoriamente.

Algoritmo 4.6.1: METODOSISTEMADECASTAS(P)

$t \leftarrow 0$
INICIALIZA($P[t]$)
AVALIA($P[t]$)
APLICAPOLITICASUBSTITUICAO($P[t]$)
se $houveSubstituicao$
  então AVALIA($P[t]$)
APLICADIVISAOEMCASTAS($P[t]$)
enquanto $criterioParadaNaoAtendido$
faça
- $t \leftarrow t + 1$
- $A' \leftarrow$ COPIACASTAELITE($P[t-1]$)
- $P' \leftarrow$ SELECIONAPAIS($P[t-1]$)
- $B' \leftarrow$ REALIZACROSSOVER(P')
- $C' \leftarrow$ INDIVIDUOSFACTIVEISAOACASO()
- $P[t] \leftarrow A' + B' + C'$
- AVALIA($P[t]$)
- APLICAPOLITICASUBSTITUICAO($P[t]$)
- **se** $houveSubstituicao$
  então AVALIA($P[t]$)
- APLICADIVISAOEMCASTAS($P[t]$)

5 TESTES COMPUTACIONAIS

Neste capítulo, são apresentados os resultados obtidos pelos métodos de predição *ab initio* desenvolvidos e estudados. Utilizando os algoritmos genéticos do estudo, foram preditas as estruturas 3D aproximadas de seis proteínas. Estas proteínas possuem estrutura 3D conhecida, determinada experimentalmente, e armazenada no PDB. Foram escolhidos os polipeptídeos com os seguintes códigos PDB: 1A11 (OPELLA et al., 1999), 1CRN (TEETER, 1984), 1PLW (MARCOTTE et al., 2004), 1ROP (BANNER; KOKKINIDIS; TSERNOGLOU, 1987), 1ZDD (STAROVASNIK; BRAISTED; WELLS, 1997), 2JR8 (DAI et al., 2008). Foram escolhidos esses polipeptídeos com o objetivo de comparar os resultados obtidos, analisando suas qualidades e conformações preditas, com as informações experimentais dos mesmos encontradas no PDB. Para proteínas maiores seria inviável, em curto espaço de tempo, obter dados suficientes para análise utilizando a metodologia *ab initio*.

5.1 Materiais e Métodos

Foram utilizados os algoritmos descritos, no Capítulo 4, para a predição da estrutura tridimensional aproximada das seis proteínas experimentais em estudo. Para todos métodos estudados, cada instância foi executada quatro vezes. Adotou-se dois critérios de parada para os algoritmos genéticos desenvolvidos: 1) execução atingiu 2000 gerações, ou 2) há 200 gerações um indivíduo mais apto do que o atual não foi estabelecido. Atendendo a um desses critérios, o algoritmo grava a solução corrente e termina. Para os testes dos métodos aleatório, probabilístico e baseado em castas uma população de 100 indivíduos foi utilizada. Nos testes do método utilizando árvores ternárias, a taxa de crossover adotada na criação de novos indivíduos *current* foi de 100%; para os demais métodos, a probabilidade de crossover adotada foi de 70% em todas as execuções. A probabilidade de mutação adotada foi de 70% nas execuções dos métodos que a aplicam, conforme descrito no Capítulo 4. A alta taxa de mutação foi aplicada com o propósito de permitir que os algoritmos escapem de convergências prematuras a mínimos locais e de acelerar o processo de diversificação dos indivíduos como descrito em (ESHELMAN, 1990).

As estruturas tridimensionais preditas das execuções de melhor resultado – menor valor de energia da conformação predita final de uma proteína – foram analisadas (ver Seção 5.3). A qualidade estereoquímica das estruturas 3D preditas de menor energia foram analisadas através do programa PROCHECK (LASKOWSKI et al., 1993); os mapas Ramachandran da distribuição dos resíduos também foram obtidos pelo mesmo programa através do PDBsum[1]. As representações gráficas das estruturas 3D foram geradas atra-

[1] http://www.ebi.ac.uk/pdbsum/

vés do programa PYMOL[2]. Os cálculos de RMSD, efetuados entre os átomos C_α da cadeia principal das conformações iniciais e finais preditas pelo melhor método e os átomos da cadeia principal das conformações experimentais foram realizados pelo programa Swiss-PDBViewer[3]. Nos cálculos de RMSD foram desconsiderados os dois resíduos de aminoácidos iniciais (região N-terminal) e os dois resíduos de aminoácidos finais (região C-terminal), pois tratam-se de estruturas secundárias irregulares e indefinidas. Todos os testes foram executados em uma máquina PC Intel Core i7 2.8GHZ 8MB Cache e 8GB de RAM sobre o sistema operacional Ubuntu.

5.2 Resultados dos Experimentos

A seguir, são apresentados os resultados das execuções dos algoritmos revisados. Para cada estudo de caso escolhido, é apresentada uma tabela com os valores de energia das conformações iniciais e finais de cada execução.

5.2.1 Experimento 1: 1A11

No primeiro estudo de caso, os métodos estudados foram testados na predição da estrutura 3D aproximada da proteína cujo código no PDB é 1A11 (OPELLA et al., 1999). A estrutura experimental (ver Figura 5.6-A, em vermelho) é composta por 25 aminoácidos e conhecida pelo seu arranjo de uma estrutura secundária regular única em forma de α-hélice. Os resultados, para cada umas das execuções da instância, são mostrados na Tabela 5.1.

5.2.2 Experimento 2: 1CRN

No segundo estudo de caso, os métodos estudados foram testados na predição da estrutura 3D aproximada da proteína cujo código no PDB é 1CRN (TEETER, 1984). A estrutura experimental (ver Figura 5.6-F, em vermelho) é composta por uma cadeia de 46 aminoácidos e conhecida pelo seu arranjo de estruturas regulares em α-hélices e folhas-β. Os resultados, para cada umas das execuções da instância, são mostrados na Tabela 5.2.

5.2.3 Experimento 3: 1PLW

No terceiro estudo de caso, os métodos estudados foram testados na predição da estrutura 3D aproximada da mini-proteína cujo código no PDB é 1PLW (MARCOTTE et al., 2004). A estrutura experimental (ver Figura 5.6-D, em vermelho) é composta por somente 5 aminoácidos e conhecida pela sua estrutura quase indefinida parecendo do com a de uma folha-β. Os resultados, para cada umas das execuções da instância, são mostrados na Tabela 5.3.

5.2.4 Experimento 4: 1ROP

No quarto estudo de caso, os métodos estudados foram testados na predição da estrutura 3D aproximada da proteína cujo código no PDB é 1ROP (BANNER; KOKKINIDIS; TSERNOGLOU, 1987). A estrutura experimental (ver Figura 5.6-C, em vermelho) é composta por uma cadeia de 56 aminoácidos e conhecida pelo seu arranjo de estruturas regulares em α-hélice. Os resultados, para todas execuções, são mostrados na Tabela 5.4.

[2]http://www.pymol.org
[3]http://spdbv.vital-it.ch

Instância		Energia ($kcal/mol$)	
		Inicial	Final
(R)	A	189898	$-455,932$
	B⋆	$1,46 \cdot 10^6$	$-473,802$
	C	$3,89 \cdot 10^6$	$-403,143$
	D	$5,62 \cdot 10^6$	$-470,816$
(P)	A	836085	$-3,9586$
	B	$2,10 \cdot 10^6$	$51107,9$
	C	$1,79 \cdot 10^6$	$8971,56$
	D⋆	242054	$-344,901$
(T)	A	$3,05 \cdot 10^6$	$-329,329$
	B	$1,54 \cdot 10^6$	$-17,943$
	C	232375	$-313,26$
	D⋆	$8,58 \cdot 10^6$	$-351,88$
(C)	A	$2,47 \cdot 10^6$	$-475,74$
	B	$9,18 \cdot 10^6$	$-477,13$
	C	$1,48 \cdot 10^6$	$-488,46$
	D⋆	407980	$-491,33$

Tabela 5.1: Resultado das soluções geradas para a proteína 1A11. Os valores da energia potencial da solução encontrada na população inicial (*Inicial*) e da melhor solução final (*Final*) são apresentados. Os códigos, entre parênteses, R, P, T e C indicam, respectivamente, execuções dos métodos *Aleatório*, *Probabilístico*, *Árvore Ternária* e *Castas*. Os sufixos A, B, C e D indicam, respectivamente, cada uma das quatro execuções do estudo de caso. Um asterisco ('*') assinala a menor energia potencial dentre as execuções de um método.

Instância		Energia ($kcal/mol$)	
		Inicial	Final
(R)	A	$4,34 \cdot 10^7$	$-113,187$
	B*	$1,03 \cdot 10^7$	$-358,526$
	C	$5,01 \cdot 10^6$	$-277,254$
	D	$1,33 \cdot 10^8$	$-143,747$
(P)	A	$3,06 \cdot 10^7$	598179
	B*	$3,89 \cdot 10^7$	468168
	C	$2,13 \cdot 10^7$	473096
	D	$4,26 \cdot 10^8$	$5,95464 \cdot 10^6$
(T)	A	$9,09 \cdot 10^6$	194,413
	B	$2,18 \cdot 10^7$	3,239
	C	$1,59 \cdot 10^6$	563,386
	D*	$1,14 \cdot 10^7$	$-72,2543$
(C)	A	$4,81 \cdot 10^7$	$-84,45$
	B	$1,77 \cdot 10^6$	$-256,59$
	C*	$1,04 \cdot 10^7$	$-374,32$
	D	$1,62 \cdot 10^8$	$-297,81$

Tabela 5.2: Resultado das soluções geradas para a proteína 1CRN. Os valores da energia potencial da solução encontrada na população inicial (*Inicial*) e da melhor solução final (*Final*) são apresentados. Os códigos, entre parênteses, R, P, T e C indicam, respectivamente, execuções dos métodos *Aleatório*, *Probabilístico*, *Árvore Ternária* e *Castas*. Os sufixos A, B, C e D indicam, respectivamente, cada uma das quatro execuções do estudo de caso. Um asterisco ('*') assinala a menor energia potencial dentre as execuções de um método.

Instância		Energia ($kcal/mol$)	
		Inicial	Final
(R)	A	44,933	−98,323
	B	71,863	−97,345
	C*	27,4127	−100,565
	D	82,0778	−94,748
(P)	A	69,7421	−95,07
	B	78,6533	27,735
	C*	66,5811	−100,218
	D	61,7718	−91,343
(T)	A	83,082	−98,958
	B*	471,258	−102,627
	C	77,973	−94,311
	D	32,77	−91,714
(C)	A	102,79	−92,96
	B*	94,17	−102,84
	C	35,96	−102,19
	D	68,29	−93,03

Tabela 5.3: Resultado das soluções geradas para a mini-proteína 1PLW. Os valores da energia potencial da solução encontrada na população inicial (*Inicial*) e da melhor solução final (*Final*) são apresentados. Os códigos, entre parênteses, R, P, T e C indicam, respectivamente, execuções dos métodos *Aleatório*, *Probabilístico*, *Árvore Ternária* e *Castas*. Os sufixos A, B, C e D indicam, respectivamente, cada uma das quatro execuções do estudo de caso. Um asterisco ('*') assinala a menor energia potencial dentre as execuções de um método.

Instância		Energia ($kcal/mol$)	
		Inicial	Final
(R)	A	$4,06 \cdot 10^8$	$-365,823$
	B	$1,77 \cdot 10^9$	$-257,843$
	C*	$5,67 \cdot 10^8$	$-415,162$
	D	$4,89 \cdot 10^8$	$-354,17$
(P)	A*	$2,02 \cdot 10^9$	$581,227$
	B	$3,36 \cdot 10^9$	225039
	C	$7,86 \cdot 10^8$	$92077,3$
	D	$3,24 \cdot 10^8$	$51832,6$
(T)	A	$1,13 \cdot 10^9$	$36939,9$
	B*	$3,59 \cdot 10^8$	$1475,01$
	C	$6 \cdot 10^8$	$3489,75$
	D	$7,89 \cdot 10^7$	$1615,19$
(C)	A	$5,18 \cdot 10^8$	$-658,97$
	B*	$9,53 \cdot 10^8$	$-710,74$
	C	$2,77 \cdot 10^9$	$-707,01$
	D	$5,18 \cdot 10^8$	$-680,28$

Tabela 5.4: Resultado das soluções geradas para a proteína 1ROP. Os valores da energia potencial da solução encontrada na população inicial (*Inicial*) e da melhor solução final (*Final*) são apresentados. Os códigos, entre parênteses, R, P, T e C indicam, respectivamente, execuções dos métodos *Aleatório*, *Probabilístico*, *Árvore Ternária* e *Castas*. Os sufixos A, B, C e D indicam, respectivamente, cada uma das quatro execuções do estudo de caso. Um asterisco ('*') assinala a menor energia potencial dentre as execuções de um método.

Instância		Energia ($kcal/mol$)	
		Inicial	Final
(R)	A	$1,76 \cdot 10^8$	$-872,034$
	B*	$3,70 \cdot 10^8$	$-955,218$
	C	$1,57 \cdot 10^8$	$-523,03$
	D	$5,89 \cdot 10^7$	$-812,593$
(P)	A	$2,73 \cdot 10^8$	$24793,7$
	B	$6,07 \cdot 10^8$	-6.984
	C*	$1,77 \cdot 10^8$	$-615,403$
	D	$4,72 \cdot 10^7$	$268,366$
(T)	A	$1,85 \cdot 10^8$	$9829,99$
	B*	$2,51 \cdot 10^8$	$-822,272$
	C	$9 \cdot 10^7$	$-318,024$
	D	$5,26 \cdot 10^8$	$-691,296$
(C)	A	$2,25 \cdot 10^8$	$-1022,56$
	B	$4,53 \cdot 10^7$	$-997,85$
	C	$1,59 \cdot 10^8$	$-824,22$
	D*	$5,85 \cdot 10^8$	$-1049,17$

Tabela 5.5: Resultado das soluções geradas para a mini-proteína 1ZDD. Os valores da energia potencial da solução encontrada na população inicial (*Inicial*) e da melhor solução final (*Final*) são apresentados. Os códigos, entre parênteses, R, P, T e C indicam, respectivamente, execuções dos métodos *Aleatório*, *Probabilístico*, *Árvore Ternária* e *Castas*. Os sufixos A, B, C e D indicam, respectivamente, cada uma das quatro execuções do estudo de caso. Um asterisco ('*') assinala a menor energia potencial dentre as execuções de um método.

5.2.5 Experimento 5: 1ZDD

No quinto estudo de caso, os métodos estudados foram testados na predição da estrutura 3D aproximada da proteína cujo código no PDB é 1ZDD (STAROVASNIK; BRAISTED; WELLS, 1997). A estrutura experimental (ver Figura 5.6-B, em vermelho) é composta por uma cadeia de 34 aminoácidos e, similar à proteína 1ROP, conhecida pelo seu arranjo de duas estruturas regulares em forma de α-hélices conectadas por uma volta. Os resultados, para cada umas das execuções da instância, são mostrados na Tabela 5.5.

5.2.6 Experimento 6: 2JR8

No sexto estudo de caso, os métodos estudados foram testados na predição da estrutura 3D aproximada da proteína cujo código no PDB é 2JR8 (DAI et al., 2008). A estrutura experimental (ver Figura 5.6-E, em vermelho) é composta por uma cadeia de 42 aminoácidos e conhecida pela sua forma curva em α-hélice. Os resultados, para cada umas das execuções da instância, são mostrados na Tabela 5.6.

Instância		Energia ($kcal/mol$)	
		Inicial	Final
(R)	A	$4,06 \cdot 10^8$	1299,2
	B	$1,77 \cdot 10^9$	1310,01
	C	$5,67 \cdot 10^8$	1434,1
	D⋆	$4,9 \cdot 10^8$	1140,43
(P)	A⋆	$8,44 \cdot 10^7$	1440,33
	B	$1,47 \cdot 10^8$	2100,77
	C	$1,04 \cdot 10^9$	15313
	D	$9,41 \cdot 10^7$	53481
(T)	A	$4,01 \cdot 10^8$	468482
	B	$8,08 \cdot 10^8$	5341,31
	C	$3,97 \cdot 10^8$	3623,01
	D⋆	$2,76 \cdot 10^8$	2040,66
(C)	A	$2,52 \cdot 10^7$	1153,76
	B	$5,20 \cdot 10^8$	1081,93
	C	$1,15 \cdot 10^8$	1137,59
	D⋆	$4,42 \cdot 10^7$	1017,37

Tabela 5.6: Resultado das soluções geradas para a proteína 2JR8. Os valores da energia potencial da solução encontrada na população inicial (*Inicial*) e da melhor solução final (*Final*) são apresentados. Os códigos, entre parênteses, R, P, T e C indicam, respectivamente, execuções dos métodos *Aleatório*, *Probabilístico*, *Árvore Ternária* e *Castas*. Os sufixos A, B, C e D indicam, respectivamente, cada uma das quatro execuções do estudo de caso. Um asterisco ('*') assinala a menor energia potencial dentre as execuções de um método.

5.3 Análise dos Resultados

De acordo com os resultados obtidos e apresentados na Seção 5.2, o algoritmo genético que aplicou estratégias baseadas em um sistema de Castas (apresentado na Seção 4.6) foi o destaque do estudo, visto que seus melhores resultados foram os melhores em todos os experimentos. Para todas as instâncias, sua predição aproximada do estado conformacional nativo foi a que obteve os menores valores de energia potencial. As estratégias de seleção e organização populacional distintas, aplicadas pela metodologia baseada em Castas, contribuíram para que o método obtivesse os melhores resultados. Acredita-se que a seleção e a recombinação entre indivíduos *elite* e *não-elite* da população colaboraram para que o espaço de busca conformacional fosse melhor explorado, visto que a recombinação de dois indivíduos bastante distintos auxiliariam sempre a manter uma boa diversidade populacional. Além do procedimento padrão de substituição de indivíduos de mesma energia, relatado na Seção 4.2.4, a substituição de 20% da população – indivíduos da casta mais baixa – a cada geração também contribuiu no aumento da diversidade populacional e na exploração do espaço de busca, visto que cada novo indivíduo era uma solução factível gerada aleatoriamente.

Para medir a qualidade das estruturas 3D aproximadas preditas pelo melhor método, foram calculados os valores de RMSD de todos os seus resultados. A Tabela 5.7 apresenta o C_α RMSD (desvio quadrático médio entre átomos C_α) das conformações iniciais e finais preditas e aproximadas, em relação às estruturas experimentais disponíveis no PDB, pelo AG baseado em Castas. As conformações das estruturas 3D aproximadas finais de menor energia, preditas pelo método, podem ser vistas na Figura 5.6-A (1A11), B (1ZDD), C (1ROP), D (1PLW), E (2JR8) e F (1CRN), em azul.

Foi observado que para os estudos de caso 1A11, 1ZDD, 1PLW, e 2JR8 foram obtidos resultados precisos (1,09Å, 4,48Å, 0,25Å e 6,14Å, respectivamente). Os experimentos com as proteínas 1CRN e 1ROP apresentaram altos valores de RMSD (10,80Å e 8,27Å, respectivamente). Isso era esperado, visto que a proteína 1CRN possui um dos mais complexos padrões de enovelamento quando comparado aos outros estudos de caso e a proteína 1ROP possui uma conformação nativa na qual suas duas α-hélices ficam próximas. Como o algoritmo trabalha na minimização da energia potencial, ele busca uma conformação que reduza o número de choques estereoquímicos – os quais elevariam a energia interna de uma molécula – e, portanto, acaba afastando as duas estruturas secundárias regulares. Atualmente, os melhores métodos *in silico* apresentam desvios de 2Å a 6Å na avaliação de similaridade entres estruturas preditas e experimentais (DORN, 2008). Em (CUTELLO; NARZISI; NICOSIA, 2006) experimentos foram realizados com os polipeptídeos 1PLW, 1ZDD, 1ROP e 1CRN – os valores de RMSD obtidos foram 0,49Å, 2,27Å, 3,70Å e 4,43Å respectivamente. Em (Ó, 2009) experimentos foram realizados com os polipeptídeos 1PLW e 1CRN cujos valores de RMSD obtidos foram 6Å e 17,51Å respectivamente. Entretanto, não é viável realizar comparações fieis entre os trabalhos, visto que os métodos e parâmetros adotados pelos algoritmos evolucionários diferem.

A Tabela 5.8 apresenta a percentagem de resíduos que ocorrem em um dos quatro estados conformacionais das proteínas experimentais: folhas-β, α-hélices, 3_{10}-hélices e outras (regiões de alças e voltas). Foi observado que a formação das estruturas secundárias das estruturas preditas estão muito próximas das presentes nas estruturas experimentais. A mini-proteína 1PLW não foi considerada nesta análise.

A distribuição dos resíduos no mapa de Ramachandran foi analisada para cada estrutura predita de menor energia pelo método baseado em Castas. A análise foi realizada com o software PROCHECK. Os resultados estão resumidos na Tabela 5.9. Novamente

Instância	C_α RMSD (Å)	
	Inicial	Final
1A11_A	1,09	1,00
1A11_B	0,83	0,80
1A11_C	1,82	0,99
1A11_D *	1,51	1,09
1CRN_A	8,53	9,66
1CRN_B	13,43	9,13
1CRN_C *	13,14	10,80
1CRN_D	9,84	10,87
1PLW_A	0,51	0,15
1PLW_B *	0,09	0,25
1PLW_C	0,17	0,13
1PLW_D	0,47	0,13
1ROP_A	13,34	12,70
1ROP_B *	14,56	8,27
1ROP_C	12,77	8,19
1ROP_D	12,43	7,42
1ZDD_A	5,32	5,92
1ZDD_B	8,99	5,83
1ZDD_C	8,09	4,91
1ZDD_D *	5,51	4,48
2JR8_A	3,77	6,43
2JR8_B	6,12	5,86
2JR8_C	4,49	5,14
2JR8_D *	4,68	6,14

Tabela 5.7: Valores de C_α RMSD das soluções geradas pelo AG baseado no sistema de Castas. Os valores de RMSD das soluções encontradas na população inicial (*Inicial*) e da melhor solução final (*Final*) são apresentados. Os sufixos A, B, C e D indicam, respectivamente, cada uma das quatro execuções dos estudos de caso. Um asterisco ('*') assinala a execução com menor energia potencial final dentre as execuções para um estudo de caso.

Cód. PDB	Folha-β	α-hélice	3_{10}-hélice	Outras	# Resíduos
1A11	0,0	92,0	0,0	8,0	25
1A11-P	0,0	92,0	0,0	8,0	25
1CRN	8,7	41,3	6,5	43,5	46
1CRN-P	0,0	43,5	6,5	50,0	46
1ROP	0,0	89,3	0,0	10,7	56
1ROP-P	0,0	89,3	0,0	10,7	56
1ZDD	0,0	73,5	0,0	26,5	34
1ZDD-P	0,0	76,5	0,0	23,5	34
2JR8	0,0	73,8	0,0	26,2	42
2JR8-P	0,0	73,8	0,0	26,2	42

Tabela 5.8: Análise da disposição das estruturas secundárias das conformações aproximadas preditas e experimentais. O sufixo '-P' indica as estruturas preditas de menor energia final dentre as quatro execuções. Os valores da tabela estão expressos em %.

Cód. PDB	Região Mais Favorável	Região Permitida	Região Ainda Aceitável	Região Não Permitida
1A11	91,3	4,3	4,3	0,0
1A11-P	100,0	0,0	0,0	0,0
1ZDD	87,1	12,9	0,0	0,0
1ZDD-P	87,1	12,9	0,0	0,0
1ROP	98,1	1,9	0,0	0,0
1ROP-P	96,3	1,9	0,0	1,9
2JR8	85,3	8,8	2,9	2,9
2JR8-P	85,3	14,7	0,0	0,0
1CRN	94,3	5,7	0,0	0,0
1CRN-P	82,9	17,1	0,0	0,0

Tabela 5.9: Valores numéricos no mapa de Ramachandran das estruturas secundárias das conformações aproximadas preditas e experimentais. O sufixo '-P' indica as estruturas preditas de menor energia final dentre as quatro execuções. Os valores da tabela estão expressos em %.

a mini-proteína 1PLW não foi considerada nesta análise, pois o software possui uma limitação quanto ao número mínimo de aminoácidos. Os mapas de Ramachandran, em comparação, das estruturas preditas e experimentais são apresentados nas Figuras 5.1, 5.2, 5.3, 5.4 e 5.5.

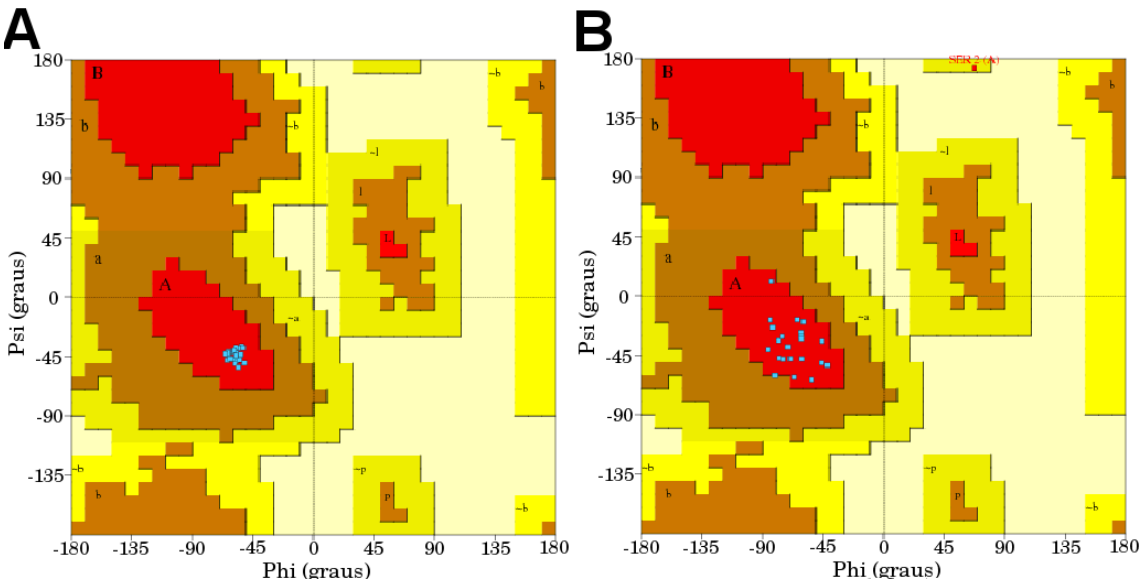

Figura 5.1: Mapas de Ramachandran das estruturas secundárias da conformação predita de menor energia final pelo AG baseado em castas (A) e da estrutura experimental (B), ambas de código PDB 1A11.

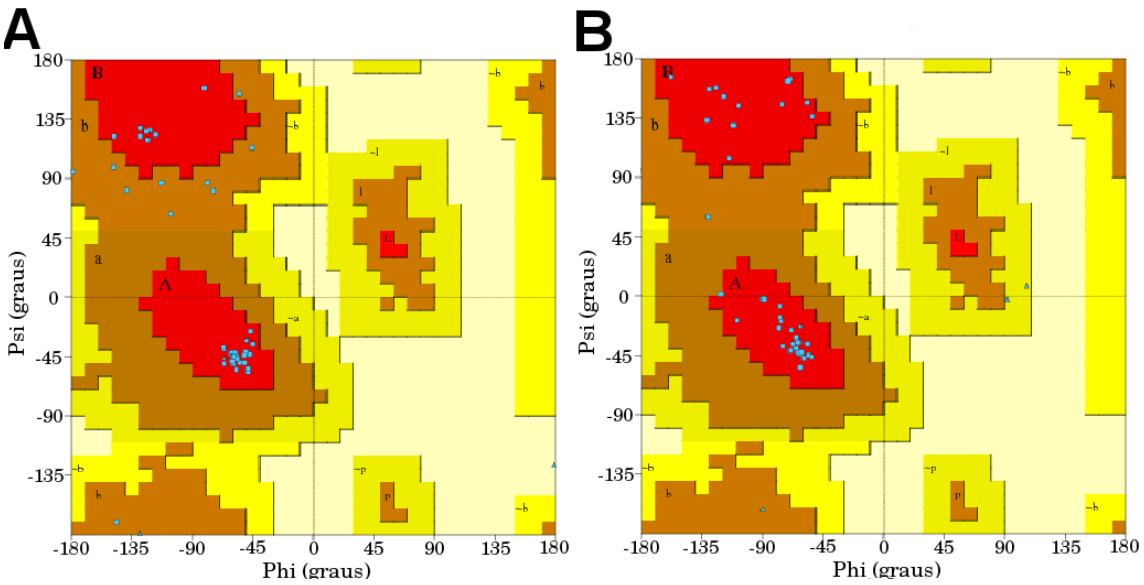

Figura 5.2: Mapas de Ramachandran das estruturas secundárias da conformação predita de menor energia final pelo AG baseado em castas (A) e da estrutura experimental (B), ambas de código PDB 1CRN.

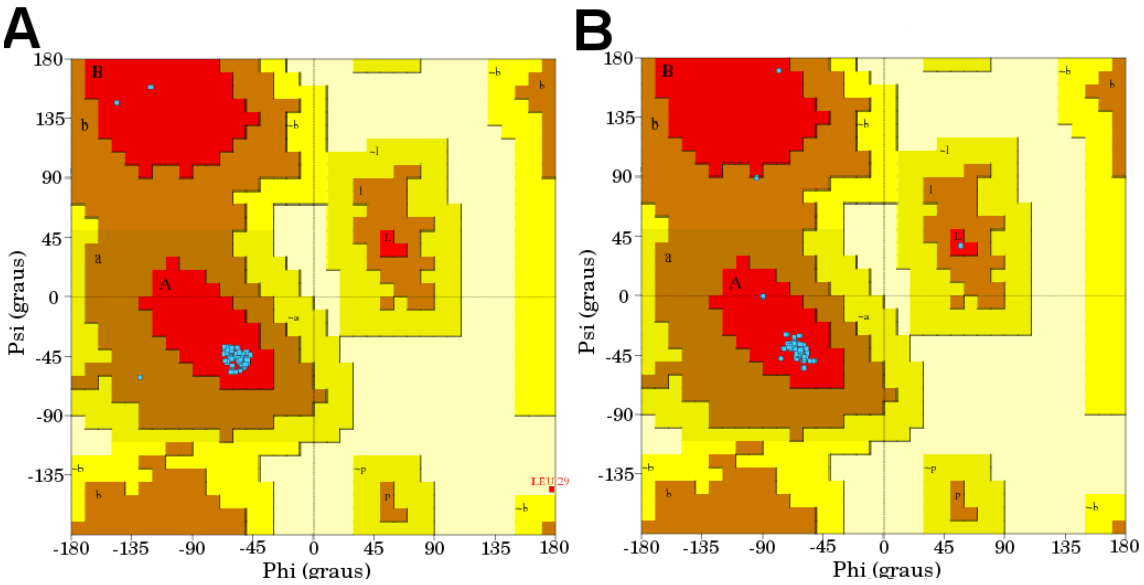

Figura 5.3: Mapas de Ramachandran das estruturas secundárias da conformação predita de menor energia final pelo AG baseado em castas (A) e da estrutura experimental (B), ambas de código PDB 1ROP.

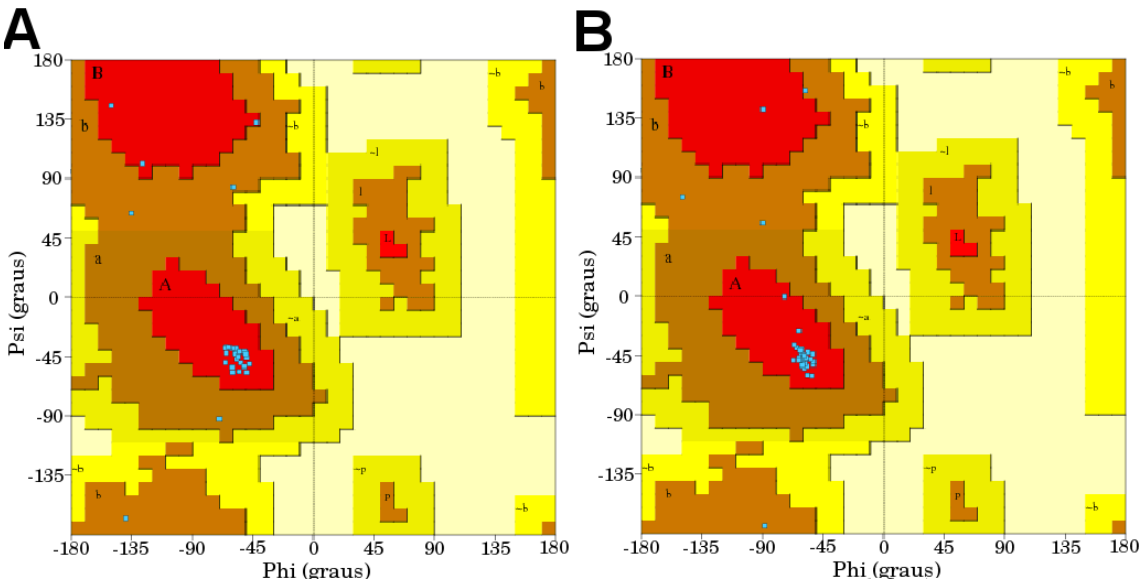

Figura 5.4: Mapas de Ramachandran das estruturas secundárias da conformação predita de menor energia final pelo AG baseado em castas (A) e da estrutura experimental (B), ambas de código PDB 1ZDD.

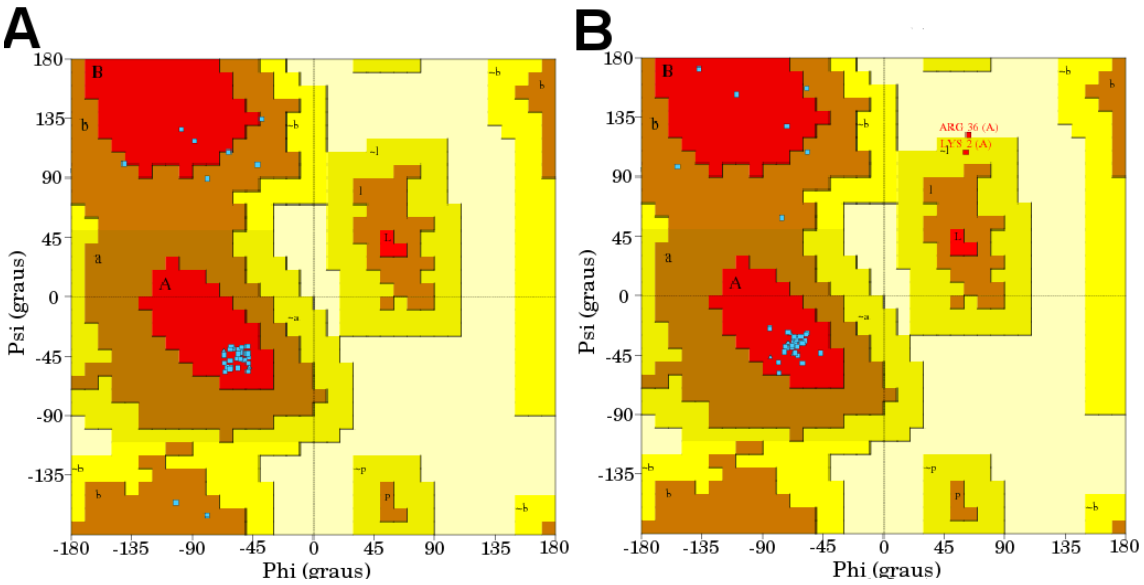

Figura 5.5: Mapas de Ramachandran das estruturas secundárias da conformação predita de menor energia final pelo AG baseado em castas (A) e da estrutura experimental (B), ambas de código PDB 2JR8.

Observou-se que, nas estruturas preditas, $\geq 95\%$ dos resíduos de aminoácidos estão localizados nas regiões mais favoráveis do mapa de Ramachandran (regiões em vermelho). Isso indica que essas estruturas possuem um baixo número de contatos que causam choques estereoquímicos e não mantêm a estrutura estável e que suas estruturas secundárias estão bem formadas. Ao comparar os resultados obtidos pelas estruturas preditas contra os obtidos pelas estruturas experimentais do PDB, observou-se que ambos estão muito similares. Esta semelhança pode ser verificada pela distribuição dos resíduos nos mapas de Ramachandran e é um indicador positivo da qualidades das estruturas regulares preditas. Na Figura 5.6 as estruturas 3D preditas e experimentais são apresentadas para fins de comparação dos enovelamentos.

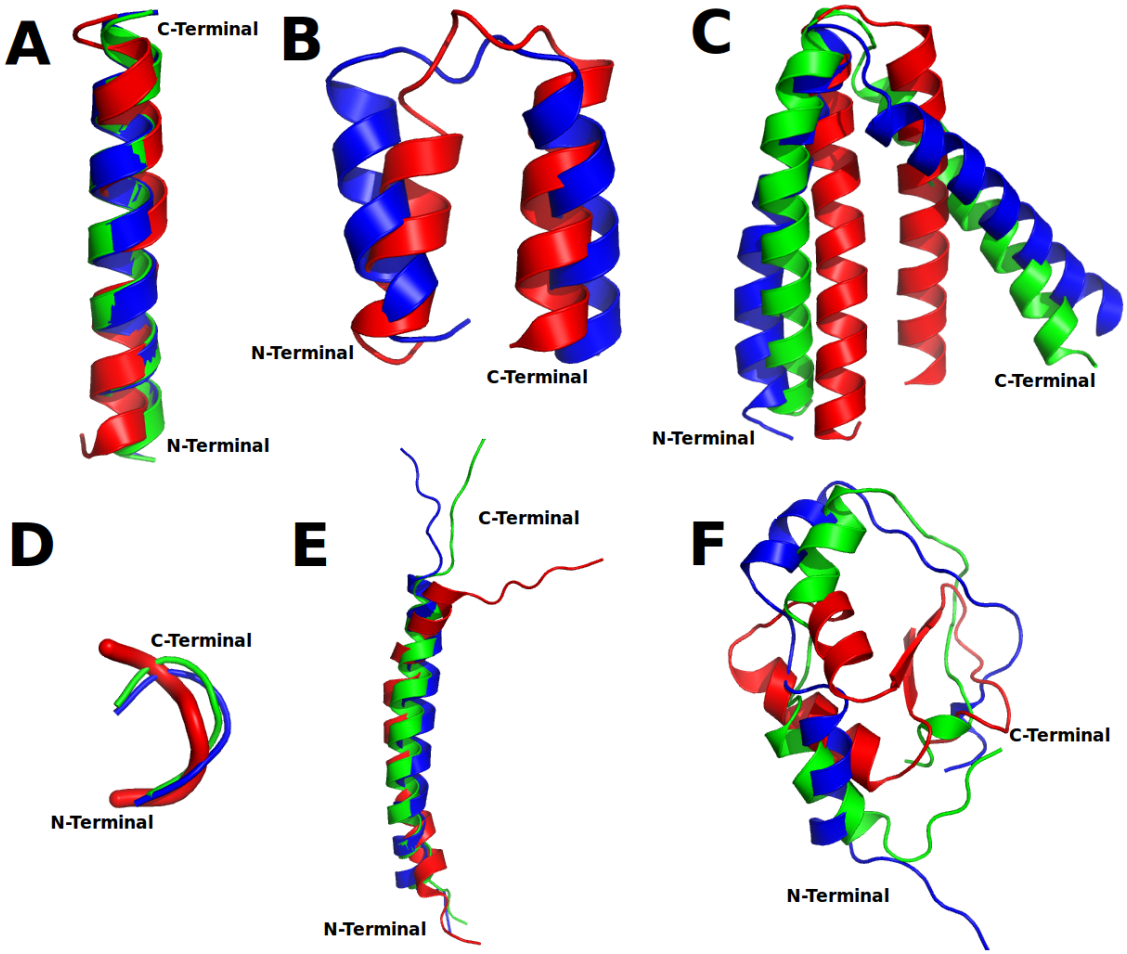

Figura 5.6: Representação em fita das estruturas experimentais (vermelho), das estruturas preditas (azul) e das estruturas com menor valor de RMSD encontradas pelo AG baseado em castas para cada instância (verde). O C_α central das estruturas experimentais e preditas estão fixos numa mesma coordenada. A, B, C, D, E e F apresentam, respectivamente, as estruturas 3D preditas e experimentais de código PDB: 1A11 (A), 1ZDD (B), 1ROP (C), 1PLW (D), 2JR8 (E) e 1CRN (F). As cadeias de aminoácidos não são mostradas para maior clareza.

6 CONSIDERAÇÕES FINAIS

Neste trabalho, foram estudados e apresentados alternativas na aplicação de algoritmos genéticos ao problema da predição aproximada de estruturas 3D de proteínas. Diferentes estratégias de seleção e organização populacionais sobre algoritmos genéticos foram testadas e pode-se constatar que simples decisões na concepção de um algoritmo genético podem ser responsáveis ou pelo fracasso, ou pelo sucesso de um método. O melhor método do estudo demonstrou que, de fato, é possível realizar predições de enovelamentos muito próximas às experimentais utilizando apenas informações de nível estrutural primário de proteínas e procedimentos evolutivos como algoritmo genéticos. Poderiam, ainda, ser utilizados métodos de dinâmica molecular no refinamento das conformações preditas, com a finalidade de prover soluções mais próximas das experimentais. Embora as outras estratégias apresentadas tenham mostrados resultados tímidos, elas poderiam ser aprimoradas pelo afinamento dos parâmetros dos algoritmos, ou desenvolvimento de novas rotinas que aperfeiçoassem as conformações preditas a cada iteração do algoritmo.

É um ponto a ser destacado que os métodos de predição desenvolvidos e estudados conseguiram predizer estruturas 3D próximas das estrutura 3D experimentais em um relativo curto espaço de tempo – em média, quatro horas de execução por instância –, embora este não tenha sido o objetivo do trabalho. Outros métodos de predição existentes demandam muito tempo de execução e processamento e, portanto, necessitam de uma plataforma computacional de custo elevado, tal qual o método ROSETTA (DAS et al., 2007) cujas predições despendem semanas de processamento dedicado e são avaliadas em um desvio entre 2Å e 6Å através de cálculos de RMSD (DORN, 2008). Devido ao uso de sub-rotinas externas, muito do tempo demandado pelos métodos estudados é efeito do grande volume de acessos a disco durante as execuções. Como trabalho futuro, o desenvolvimento de bibliotecas de rotâmeros que trabalham em memória primária contribuiria no aumento de desempenho e rápida obtenção de predições de estruturas pequenas por esses algoritmos.

Pode-se concluir que os AGs estudados utilizando seleção aleatória, árvores ternárias e sistemas de castas – apresentados nas seções 4.3, 4.5 e 4.6 respectivamente – forneceram bons resultados embora terem utilizado estratégias diferentes de organização e controle populacional. O método baseado no sistema de castas obteve os melhores resultados. Embora ainda não exista técnica *in silico* capaz de predizer novos enovelamentos com extrema acurácia (DORN, 2008), a análise desses resultados, através dos cálculos de RMSD e distribuição dos resíduos no mapa de Ramachandran, demonstrou que predições *ab initio* de qualidade podem ser desenvolvidas utilizando técnicas de busca evolucionárias. Quando não há convergências prematuras, tem-se boa diversidade e um fluxo constante de melhores indivíduos é atingido, essas técnicas podem prover conformações aproximadas precisas em curto espaço de tempo e que apresentam boas propriedades físicas e estereoquímicas.

6.1 Contribuições

As principais contribuições deste trabalho que podem ser destacadas são:

- A utilização de técnicas computacionais efetivas aplicadas a um problema biológico e de grande importância;
- A compilação e revisão dos principais conceitos relacionados ao PSP;
- O desenvolvimento de algoritmos genéticos com o emprego de diferentes estratégias de seleção e organização populacionais, as quais podendo abrir diversas possibilidades na pesquisa de aplicações de uso potencial na biologia computacional, como, por exemplo, a simples aplicação dos algoritmos apresentados a outras classes de proteínas;
- A investigação de estratégias para redução do espaço conformacional de busca do PSP sem utilizar informações de bancos de dados estruturais;
- A escrita e submissão de um artigo científico em evento de Bioinformática:

  Título: A Structured-Population Genetic Algorithm for the 3-D Protein Structure Prediction Problem.

  Conferência: Proceedings of the 2011 Brazilian Symposium on Bioinformatics.

  Ano: 2011 (aceito).

  Tipo: artigo completo.

  Qualificação: Qualis B4.

  Citação: GONÇALVES, W. W. ; DORN, M. ; BURIOL, L. S. ; LAMB, L. C. A Structured-Population Genetic Algorithm for the 3-D Protein Structure Prediction Problem. In: Proceedings of the 2011 Brazilian Symposium on Bioinformatics, 2011, Brasília.

REFERÊNCIAS

BAECK, T. **Evolutionary algorithms in theory and practice**: evolution strategies, evolutionary programming, genetic algorithms. Oxford, UK: Oxford University Press, 1996.

BANNER, D. W.; KOKKINIDIS, M.; TSERNOGLOU, D. Structure of the ColE1 rop protein at 1.7Å resolution. **J Mol Biol**, [S.l.], v.196, n.3, p.657–75, 1987.

BAXEVANIS, A.; QUELLETTE, B. **Bioinformatics**: a practical guide to the analysis of genes and proteins. 2.ed. New York, USA: John Wiley and Sons, Inc., 1990. 488p.

BEAN, J. Genetic algorithms and random keys for sequencing and optimization. **ORSA J. on Comp.**, [S.l.], v.6, p.154–160, 1994.

BERMAN, H. M.; HENRICK, K.; NAKAMURA, H.; MARKLEY, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. **Nucleic Acids Research**, [S.l.], p.301–303, 2007.

BERMAN, H.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BATH, T.; WEISSIG, H.; SHINDYALOV, I.; BOURNE, P. The protein data bank. **Nucleic Acids Res.**, [S.l.], v.28, n.1, p.235–242, 2000.

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

BUJNICKI, J. Protein structure prediction by recombination of fragments. **ChemBioChem**, [S.l.], v.7, n.1, p.19–27, 2006.

BURIOL, L.; FRANCA, P. M.; MOSCATO, P. A New Memetic Algorithm for the Asymmetric Traveling Salesman Problem. **Journal of Heuristics**, [S.l.], v.10, p.483–506, September 2004.

CHOTHIA, C.; LESK, A. M. The relation between the divergence of sequence and structure in proteins. **The European Molecular Biology Organization Journal**, [S.l.], 1986.

CORNELL, W.; CIEPLAK, P.; BAYLY, C.; GOULD, I.; MERZ JR., K.; FERGUSON, D.; SPELLMEYER, D.; FOX, T.; CALDWELL, J.; KOLLMAN, P. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. **J. Am. Chem. Soc.**, [S.l.], v.117, n.19, p.5179–5197, 1995.

CRESCENZI, P.; GOLDMAN, D.; PAPADIMITRIOU, C.; PICCOLBONI, A.; YANNAKAKIS, M. On the complexity of protein folding. **J. Comput. Biol.**, [S.l.], v.5, n.3, p.423–466, 1998.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. **Journal of the Royal Society Interface, Royal Society Publications London**, [S.l.], v.3, n.6, p.139–151, 2006.

DAI, H.; RAYAPROLU, S.; GONG, Y.; HUANG, R.; PRAKASH, O.; JIANG, H. Solution structure, antibacterial activity, and expression profile of Manduca sexta moricin. **J Pept Sci**, [S.l.], v.14, n.7, p.855–63, 2008.

DAS, R.; QIAN, B.; RAMAN, S.; VERNON, R.; THOMPSON, J.; BRADLEY, P.; KHARE, S.; TYKA, M.; BHAT, D.; CHIVIAN, D.; KIM, D.; SHEFFLER, W.; MALMSTROEM, L.; WOLLACOTT, A.; WANG, C.; ANDRE, I.; BAKER, D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. **Proteins**, [S.l.], v.68, n.S8, p.118–128, 2007.

DE JONG, K. A. **An analysis of the behavior of a class of genetic adaptive systems.** 1975. Tese (Doutorado em Ciência da Computação) — , Ann Arbor, MI, USA. AAI7609381.

DORN, M. **Uma Proposta para a Predição Computacional da Estrutura 3D Aproximada de Polipeptídeos com Redução do Espaco Conformacional Utilizando Análise de Intervalos**. 2008. Dissertação (Mestrado em Ciência da Computação) — Pontifícia Universidade do Rio Grande do Sul, Porto Alegre, RS, Brasil.

DORN, M.; BURIOL, L. S.; LAMB, L. C. A hybrid genetic algorithm for the 3-D protein structure prediction problem using a path-relinking strategy. **IEEE Congress on Evolutionary Computation, 2011, New Orleans**, [S.l.], p.1–8, 2011.

ERICSSON, M.; RESENDE, M.; PARDALOS, P. A genetic algorithm for the weight setting problem in OSPF routing. **J. Comb. Optim.**, [S.l.], v.6, p.299–2002, 2002.

ESHELMAN, L. J. The CHC Adaptive Search Algorithm: how to have safe search when engaging in nontraditional genetic recombination. In: FOUNDATIONS OF GENETIC ALGORITHMS, 1990. **Anais...** [S.l.: s.n.], 1990. p.265–283.

ESWAR, N.; MARTÍ-RENOM, M.; WEBB, B.; MADHUSUDHAN, M. S.; ERAMIAN, D.; SHEN, M.; PIEPER, U.; SALI, A. Comparative protein structure modeling with MODELLER. **Curr. Protoc. Bioinf.**, [S.l.], v.15, p.5.6.1–5.6.30, 2006.

FLOUDAS, C. A.; FUNG, H. K.; MCALLISTER, S. R.; MNNIGMANN, M.; RAJGARIA, R. Advances in protein structure prediction and de novo protein design: a review. **Chem. Eng. Sci.**, [S.l.], v.61, n.3, p.966–988, 2006.

GIBAS, C.; JAMBECK, P. **Developing bioinformatics computer skills**. 1.ed. USA: O'Reilly, 2001. 448p.

GOLDBERG, D. E.; DEB, K. A comparative analysis of selection schemes used in genetic algorithms. In: FOUNDATIONS OF GENETIC ALGORITHMS, 1991. **Anais...** San Francisco: CA: Morgan Kaufmann, 1991. p.69–93.

GONÇALVES, J.; RESENDE, M. Biased random-key genetic algorithms for combinatorial optimization. **Journal of Heuristics**, [S.l.], p.1–39, 2010. 10.1007/s10732-010-9143-1.

GOVINDARAJAN, S.; RECABARREN, R.; GOLDSTEIN, R. A. Estimating the total number of protein folds. **Proteins: Structure, Function, and Bioinformatics**, [S.l.], v.35, n.4, p.408–414, 1999.

HILDEBRAND, A.; REMMERT, M.; BIEGERT, A.; SÖDING, J. Fast and accurate automatic structure prediction with HHpred. **Proteins**, [S.l.], v.77, n.S9, p.128–132, 2009.

HUTCHINSON, E.; THORNTON, J. PROMOTIF - A program to identify and analyze structural motifs in proteins. **Protein Sci.**, [S.l.], v.5, n.2, p.212–220, 1996.

JONES, D. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. **Proteins**, [S.l.], v.S1, p.185–191, 1997.

JONES, D. Predicting novel protein folds by using Fragfold. **Proteins**, [S.l.], v.45, n.S5, p.127–132, 2001.

JONES, D.; TAYLOR, W.; THORNTON, J. A new approach to protein fold recognition. **Nature**, [S.l.], v.358, n.6381, p.86–89, 1992.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, [S.l.], v.22, n.12, p.2577–2637, 1983.

KREHER, D. L.; STINSON, D. R. **Combinatorial Algorithms**: generation, enumeration, and search. [S.l.]: CRC Press, 1999. 329p.

KRIEGER, E.; JOO, K.; LEE, J.; LEE, J.; RAMAN, S.; THOMPSON, J.; TYKA, M.; BAKER, D.; KARPLUS, K. Improving physical realism, stereo-chemistry, and side-chain accuracy in homology modeling: four approaches that performed well in casp8. **Proteins**, [S.l.], v.77, n.S9, p.114–122, 2009.

KRIVOV, G.; SHAPOVALOV, M. V.; DUNBRACK, R. L. Improved prediction of protein side-chain conformations with SCWRL4. **Proteins: Structure, Function, and Bioinformatics**, [S.l.], v.77, n.4, p.778–795, 2009.

LASKOWSKI, R. A.; MACARTHUR, M. W.; MOSS, D. S.; THORNTON, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. **J. Appl. Crystallogr.**, [S.l.], v.26, n.2, p.283–291, 1993.

LEHNINGER, A.; NELSON, D.; COX, M. **Principles of Biochemistry**. 4.ed. New York, USA: W.H. Freeman, 2005. 1100p.

LESK, A. **Introduction to protein architecture**: the structural biology of proteins. 1.ed. Cambridge, UK: Oxford University Press, 2000.

LEVINTHAL, C. How to fold graciously. **Mossbauer Spectroscopy in Biological Systems**, [S.l.], v.Proceedings of a meeting held at Allerton House, Monticello, IL, p.22–24, 1969.

MARCOTTE, I.; SEPAROVIC, F.; AUGER, M.; GAGNE, S. M. A multidimensional 1H NMR investigation of the conformation of methionine-enkephalin in fast-tumbling bicelles. **Biophys J**, [S.l.], v.86, n.3, p.1587–600, 2004.

MARTÌ-RENOM M.A., S. A. F. A. S. R. M. F.; SALI, A. Comparative protein structure modelling of genes and genomes. **Annu. Rev. Biophys. Biomol. Struct.**, [S.l.], v.29, n.16, p.291–235, 2000.

MOSCATO, P.; TINETTI, F. Blending Heuristics with a Population-Based Approach: a memetic algorithm for the traveling salesman problem. In: REPORT 92-12, UNIVERSIDAD NACIONAL DE LA PLATA, C.C. 75, 1900 LA PLATA, 1994. **Anais...** [S.l.: s.n.], 1994.

MOULT, J. A. Decade of CASP: progress, bottlenecks an prognosis in protein structure prediction. **Curr. Opin. Struct. Biol.**, [S.l.], v.15, n.3, p.285–289, 2005.

MERZ JR, K.; GRAND, S. (Ed.). **The Protein Folding Problem and Tertiary Structure Prediction**. Boston, MA, USA: [s.n.], 1997.

OPELLA, S. J.; MARASSI, F. M.; GESELL, J. J.; VALENTE, A. P.; KIM, Y.; OBLATT-MONTAL, M.; MONTAL, M. Structures of the M2 channel-lining segments from nicotinic acetylcholine and NMDA receptors by NMR spectroscopy. **Nat Struct Biol**, [S.l.], v.6, n.4, p.374–9, 1999.

OSGUTHORPE, D. Ab initio protein folding. **Curr. Opin. Struct. Biol.**, [S.l.], v.10, n.2, p.146–152, 2000.

PAULING, L.; COREY, R. The pleated sheet, a new layer configuration of polypeptide chains. **Proc. Natl. Acad. Sci. U. S. A.**, [S.l.], v.37, n.5, p.251–256, 1951.

PAULING, L.; COREY, R.; BRANSON, H. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. **PNAS**, [S.l.], v.37, n.4, p.205–234, 1951.

PEDERSEN, J.; MOULT, J. Genetic algorithms for protein structure prediction. **Current Opinion in Structural Biology**, [S.l.], v.6, n.2, p.227–231, 1996.

PEDERSEN, J.; MOULT, J. Protein folding simulations with genetic algorithms and a detailed molecular description. **Journal of Molecular Biology**, [S.l.], v.269, n.2, p.240–259, 1997.

PENG, J.; XU, J. Low-homology protein threading. **Bioinformatics**, [S.l.], v.26, n.12, p.i294–i300, June 2010.

PONDER, J. TINKER: software tools for molecular design. , [S.l.], 1998.

PONDER, J.; RICHARDS, F. An efficient newton-like method for molecular mechanics energy minimization of large molecules. **Journal of Computational Chemistry**, [S.l.], v.8, n.7, p.1016–1024, 1987.

PRICE, M. N.; DEHAL, P. S.; ARKIN, A. P. FastBLAST: homology relationships for millions of proteins. , [S.l.], v.3, n.10, p.e3589+, Oct. 2008.

RAMACHANDRAN, G.; SASISEKHARAN, V. Conformation of polypeptides and proteins. **Adv. Protein Chem.**, [S.l.], v.23, p.238–438, 1968.

ROHL, C.; STRAUSS, C.; MISURA, K.; BAKER, D. Protein structure prediction using rosetta. **Methods Enzymol.**, [S.l.], v.383, p.66–93, 2004.

BOURNE, P.; WEISSIG, H. (Ed.). **Fundamentals of protein structure**: structural bioinformatics. [S.l.: s.n.], 2003. p.15.

SCHULZE-KREMER, S. Genetic algorithms for protein tertiary structure prediction. In: BRAZDIL, P. (Ed.). **Machine Learning**. [S.l.]: Springer, 1993. p.262–279. (Lect. Notes Comp. Scien., v.667).

SCHWEFEL, H. P. **Numerical optimization of computer models**. [S.l.]: Wiley, Chichester ; New York :, 1981. 389p.

SIMONS, K.; RUCZINKI, I.; KOOPERBERG, C.; FOX, B.; BYSTROFF, C.; BAKER, D. Improved recognition of native-like structures using a combination of sequence-dependent and sequence-independent features of proteins. **Proteins: Structure, Function, and Bioinformatics**, [S.l.], v.34, n.1, p.82–95, 1999.

SRINIVASAN, R.; ROSE, G. LINUS - A hierarchic procedure to predict the fold of a protein. **Proteins**, [S.l.], v.22, n.2, p.81–99, 1995.

SRINIVASAN, R.; ROSE, G. Ab initio prediction of protein structure using LINUS. **Proteins**, [S.l.], v.47, n.4, p.489–495, 2002.

STAROVASNIK, M. A.; BRAISTED, A. C.; WELLS, J. A. Structural mimicry of a native protein by a minimized binding domain. **Proc Natl Acad Sci U S A**, [S.l.], v.94, n.19, p.10080–5, 1997.

TEETER, M. M. Water structure of a hydrophobic protein at atomic resolution: pentagon rings of water molecules in crystals of crambin. **Proc Natl Acad Sci U S A**, [S.l.], v.81, n.19, p.6014–8, 1984.

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

UNGER, R. The Genetic Algorithm Approach to Protein Structure Prediction. In: **Structure**. [S.l.: s.n.], 2004. v.110, p.153–175.

VELANKAR, S.; MCNEIL, P.; MITTARD-RUNTE, V.; SUAREZ, A.; BARREL, D.; APWEILER, R.; HENRICK, K. E-MSD: an integrated data resource for bioinformatics. **Nucleic Acids Res**, [S.l.], v.32, p.211–216, 2004.

WIKIPEDIA. **Proteína**. Online; acessado em 29 de Junho de 2011, `http://pt.wikipedia.org/wiki/Prote%C3%ADna`.

WIKIPEDIA. **Threading (protein sequence)**. Online; acessado em 29 de Junho de 2011, `http://en.wikipedia.org/wiki/Threading_(protein_sequence)`.

WU, S.; SKOLNICK, J.; ZHANG, Y. Ab initio modeling of small proteins by iterative TASSER simulations. **BMC Biol.**, [S.l.], v.5, n.17, p.1–10, 2007.

XU, J.; PENG, J.; ZHAO, F. Template-based and free modeling by RAPTOR11 in CASP8. **Proteins**, [S.l.], v.77, n.S9, p.133–137, 2009.

ZHANG, Y. Template-based modeling and free modeling by I-TASSER in CASP7. **Proteins**, [S.l.], v.8, p.108–117, 2007.

ZHANG, Y. I-TASSER server for protein 3D structure prediction. **BMC Bioinf.**, [S.l.], v.9, n.40, p.1–8, 2008.

ZHOU, H.; SKOLNICK, J. Protein structure prediction by Pro-Sp3-TASSER. **Biophys. J.**, [S.l.], v.96, n.6, p.2119–2127, 2009.

Ó, V. T. **Técnicas de Controle da Diversidade de Populações em Algoritmos Genéticos para Determinação de Estruturas de Proteínas**. 2009. Dissertação (Mestrado em Ciência da Computação) — Universidade de São Paulo, São Paulo, SP, Brasil.