UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
DEPARTAMENTO DE BIOLOGIA MOLECULAR E BIOTECNOLOGIA
CURSO DE BIOTECNOLOGIA

MARIANA DOS SANTOS OLIVEIRA

Identificação de padrões conformacionais em estruturas experimentais e projeto de metaheurísticas para a predição da estrutura tridimensional de proteínas

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Biotecnologia

Orientador: Prof. Dr. Márcio Dorn

Porto Alegre
2016

AGRADECIMENTOS

Primeiramente, gostaria de agradecer ao meu orientador, Professor Doutor Márcio Dorn, por todo apoio, incentivo e disposição cedidos desde minha entrada no laboratório e principalmente nesta etapa final de conclusão de curso. Apesar de sermos de áreas diferentes, sempre se mostrou acessível e compreensível. Espero que este seja apenas o início de trabalhos e pesquisas que virão.

Aos colegas de laboratório gostaria de agradecer por toda ajuda e disponibilidade provindos os momentos em que foram necessários. Principalmente ao Bruno Borguesan, sem o qual este trabalho não seria possível. Aos colegas de curso de Biotecnologia que me acompanharam ao longo dessa trajetória.

Em especial a minha família, principalmente minha mãe Maria Clara Oliveira,g que me apoiou desde o dia em que resolvi mudar o rumo da minha graduação e por todo auxílio e amparo, sem os quais não seria capaz de chegar neste momento, a conclusão do curso de graduação em Biotecnologia.

RESUMO

O conhecimento da estrutura tridimensional de proteínas permite a inferência e estudo de sua função. Predizer essa estrutura 3D a partir da sequência linear de aminoácidos é um dos problemas mais difíceis da Bioinformática Estrutural. Fragmentos de *loops* da cadeia polipeptídica são ditos estruturas flexíveis e de alta variabilidade sendo consideradas mais difíceis de serem preditas e manipuladas. Devido à alta seletividade do processo de enovelamento proteico, o espaço de busca tridimensional é extenso, podendo ser manipulado como um problema de otimização. Assim, este trabalho tem como proposta a criação de uma Biblioteca de Padrões Estruturais (BPE) para regiões de *loops*, criada a partir de dados experimentais extraídos do *Protein Data Bank* (PDB) a fim de diminuir o espaço de busca conformacional. Esta biblioteca foi empregada com conhecimento na metaheurística *Self-Adapting Differential Evolution* (SADE) aplicada ao problema de predição de estruturas 3D de proteínas. Para validação da BPE foram implementadas duas versões do algoritmo SADE, em que a primeira utilizava os dados da BPE e APL-1 como conhecimento para geração e evolução da população e a segunda utilizava apenas a APL-1, considerando a não existência dos padrões estruturais. Para análise dos resultados foi considerada a função de aptidão e calculados os valores de RMSD e GDT_TS para as estruturas preditas em relaçao as estruturas experimentais. Os valores de RMSD médio mostraram-se 93% melhores para as execuções utilizando a BPE, além disso, 92% das estruturas com menor valor de RMSD foram geradas utilizando a BPE. Os valores de GDT_TS médio revelaram-se 93% melhores em estruturas preditas com a utilização da BPE, além disso, 61,5% das estruturas geradas com a utilização da BPE apresentaram os maiores valores de GDT_TS. Os valores da função de energia não indicaram diferenças entre as diferentes variações do método SADE. Em suma, a partir dos resultados apresentados é possível afirmar que a utilização da BPE como conhecimento na metaheurística SADE é capaz de gerar estruturas 3D de proteínas mais próximas às estruturas experimentais do que quando comparado à implementação do mesmo método sem a utilização da BPE.

Palavras-chave: Bioinformática Estrutural. Predição de Estruturas 3D de Proteínas. Metaheurísticas. Métodos Baseados em Conhecimento. Regiões de *Loops*.

Identification of conformational patterns in experimental structures and metaheuristic applied to 3D Protein Structural Prediction.

ABSTRACT

The knowledge of three-dimensional (3D) protein structure allows the determination and study of biological function. One of the most difficult problems in Structural Bioinformatics is the prediction of 3D protein structure from amino acids sequence. Loops fragments are flexible and highly variable structures in the polypeptide chain, so loops are more difficult to predict. Due to the high selectivity of protein folding process, the 3D search space is extensive. Therefore, this work proposes a loop Structure Pattern Library (BPE) which was created using experimental information extracted from Protein Data Bank (PDB) aiming to constrain the conformational search space. Self-Adapting Differential Evolution (SADE) metaheuristic was implemented for the 3D protein structure prediction problem using BPE as knowledge. SADE algorithm was tested in two versions. First, BPE and APL-1 were used as information to population generation and evolution. Second, only APL-1 was employ. Average RMSD results were better in 93% of cases using BPE, besides 92% of structures with minimum RMSD were predicted using BPE. Average GDT_TS were higher in 93% of cases using BPE and maximum GDT_TS were higher in 61,5% of BPE cases. Fitness function did not show different results between SADE versions. Thereby, our results allows us to state that BPE application as knowledge in SADE metaheuristic is capable to create 3D protein structures closer to experimental structures than SADE application without BPE.

Keywords: Structural Bioinformatics, 3D-Protein Structure Prediction, Metaheuristics, Knowledge-Based Search Methods, Loops Regions.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
APL	*Angle Probability List*
BNL	*Brookhaven National Laboratory*
BPE	Biblioteca de Padrões Estruturais
CASP	*Critical Assessment of Structure Prediction*
DE	*Differential Evolution*
EA	Algoritmos Evolutivos
GDT_TS	*Global Distance Total Score Test*
PDB	*Protein Data Bank*
RefSeq	*Reference Sequence Database*
RMN	Ressonância Magnética Nuclear
RMSD	*Root Mean Square Deviation*
SADE	*Self-Adapting Differential Evolution*
SBCB	*Structural Bioinformatics and Computational Biology Lab*
SI	*Swarm Intelligence*
STRIDE	*STRuctural IDEntification*

SUMÁRIO

1 INTRODUÇÃO

O sucesso do projeto Genoma e o crescimento de técnicas experimentais - como sequenciamento, espectrometria de massas e análises de expressão por microarranjo - possibilitou a geração de dados de maneira explosiva. Porém, essa quantidade de dados é altamente complexa para ser analisada manualmente, sendo indispensável o uso de técnicas computacionais. Surge, assim, a **Bioinformática**, área de estudo que une a biologia molecular com a ciência da computação.

Bioinformática é uma área interdisciplinar, a confluência de um conjunto de tecnologias em que técnicas computacionais e métodos estatísticos são aplicados para a manipulação e organização de dados de biologia molecular (LUSCOMBE; GREENBAUM; GERSTEIN, 2001; CHEN, 2005). Os dois maiores desafios (CHEN, 2005) para área de Bioinformática são: (i) manipulação de dados e (ii) descoberta de conhecimento. Com o crescimento de dados tridimensionais de estruturas (3D) de macromoléculas se tornou necessária a criação de métodos para armazenamento e organização dos mesmos, surgindo, assim, o *Protein Data Bank*[1] (PDB). A representação, armazenamento, análise e distribuição de informação estrutural são focos da subárea **Bioinformática Estrutural**.

Proteínas são sequências de aminoácidos os quais adotam uma estrutura 3D única quando em meio fisiológico. A estrutura 3D está intrinsecamente conectada a sua função, logo o conhecimento dessa estrutura permite o estudo e determinação da atividade biológica de proteínas. A determinação da estrutura 3D de proteínas ocorre, principalmente, de maneira experimental, entretanto as técnicas aplicadas são altamente custosas e levam demasiado tempo (KOSÍNSKI et al., 2009; DORN et al., 2014).

Atualmente, o número de sequências de proteínas armazenadas no RefSeq (N.A. et al., 2016) já ultrapassa 70 milhões[2], enquanto que o número de estruturas 3D de proteínas determinadas e armazenadas no PDB é aproximadamente 115 mil[1]. Essa discrepância no volume de dados assim como a importância em determinar novas estruturas 3D de proteínas implica na necessidade de aplicar métodos computacionais no **problema de predição da estrutura 3D de proteínas**.

Determinar a estrutura 3D de proteínas a partir da sequência linear de aminoácidos é um dos problemas mais difíceis da Bioinformática Estrutural. Devido à alta seletividade do processo de enovelamento, em que a sequência de aminoácido adota uma estrutura única dentre as inúmeras possibilidades conformacionais, o espaço de busca tri-

[1] http://www.rcsb.org/pdb/home/home.do
[2] http://www.ncbi.nlm.nih.gov/refseq/

dimensional é extenso (LEVINTHAL, 1978), podendo ser modelado como um problema de otimização.

Problemas de otimização, geralmente, não apresentam soluções ótimas capazes de serem encontradas em tempo razoável por métodos exatos. Uma alternativa é utilizar métodos aproximativos (TABLI, 2009; BOUSSAÏD; LEPAGNOT; SIARRY, 2013), os quais geram soluções de qualidade em tempo praticável, entretanto sem garantia de encontrar a solução ótima. Metaheurística é uma classe de algoritmos aproximativos aplicada em problemas de otimização, como o problema de predição da estrutura 3D de proteínas (TABLI, 2009; BOUSSAÏD; LEPAGNOT; SIARRY, 2013). Metaheurísticas comumente implementadas no problema de predição de estruturas são classificadas como algoritmos evolutivos, os quais simulam a evolução de indivíduos através dos processos de seleção, reprodução e recombinação buscando de maneira iterativa produzir a melhor solução.

A eficiência dos métodos é avaliada pelos experimentos *Critical Assessment of Structure Prediction* (CASP)[3], os quais demonstraram que os melhores resultados têm sido gerados por aqueles que utilizam conhecimento para diminuir o espaço de busca conformacional (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014). Estruturas secundárias regulares, como hélices e folhas, apresentam padrões estruturais comuns e, assim, possíveis combinações de valores de ângulos diedros dentro de um limite menor. Já estruturas de *loops* são ditas regiões flexíveis, visto que apresentam ausência de padrões, variadas combinações de valores de ângulos permitidas e são mais suscetíveis a mutações. Em suma, a predição estrutural de regiões de *loops* é mais difícil e complexa em virtude da alta variabilidade encontrada nessas estruturas (OFFMANN; TYAGI; BREVERN, 2007; SHEHU; KAVRAKI, 2012; LI, 2013).

1.1 Objetivos

O conhecimento da estrutura 3D de proteínas permite que sua função seja determinada e estudada. Entretanto a determinação da mesma de maneira experimental é altamente custosa e leva demasiado tempo; além disso, o desenvolvimento de técnicas em biologia molecular levou a uma explosão de dados. Assim, há uma enorme lacuna no volume de dados, o que motiva a aplicação e desenvolvimento de métodos computacionais para predizer estruturas 3D de proteínas.

Os melhores resultados de estruturas preditas computacionalmente têm sido gera-

[3] http://predictioncenter.org/

dos por métodos que utilizam conhecimento experimental a fim de diminuir o espaço de busca conformacional. Este conhecimento é extraído de banco de dados, como o PDB, e pode ser aplicado de diferentes formas, como bibliotecas de fragmentos. Entretanto, um ponto crucial para gerar soluções satisfatórias é a aplicação do conhecimento de maneira inteligente. Assim este trabalho de conclusão de curso em Biotecnologia tem como objetivos:

1. Desenvolver uma **biblioteca de fragmentos de padrões estruturais para regiões de *loops*** a partir de dados extraídos do PDB.
2. Aplicar o conhecimento gerado através da biblioteca de fragmentos em uma **metaheurística** de maneira a diminuir o espaço de busca para o problema de predição da estrutura 3D de proteínas.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte forma: **Capítulo 2**, em que são apresentados conceitos e tópicos sobre proteínas e seu estudo estrutural; **Capítulo 3**, em que são revisados os métodos computacionais aplicados no problema de predição da estrutura 3D de proteínas; **Capítulo 4**, em que metaheurísticas são introduzidas e descritas; **Capítulo 5**, em que estruturas de voltas e regiões irregulares são apresentadas, além da descrição da biblioteca de fragmentos (BPE) criada e descrição da metaheurística proposta; **Capítulo 6**, em que são mostrados e discutidos os resultados; e, finalizando, **Capítulo 7**, em que as conclusões e trabalhos futuros são apresentados.

2 PROTEÍNAS

Proteínas são as macromoléculas biológicas mais abundantes, mediando praticamente todas as funções e estruturas essenciais das células (LEHNINGER; NELSON; COX, 2005; LESK, 2010). Proteínas são polímeros lineares formados por diferentes combinações de aminoácidos, os quais estão conectados por interações covalentes chamadas ligação peptídica.

Todas as proteínas são formadas pelo mesmo conjunto de 20 aminoácidos básicos e possuem a mesma cadeia principal, constituída pelo átomo de Carbono, os grupos carboxila e amino. A variedade entre proteínas é dependente da cadeia lateral, já que o grupo R de cada aminoácido possui propriedades químicas distintas (LEHNINGER; NELSON; COX, 2005; LESK, 2010; SCHEEFF; FINK, 2003).

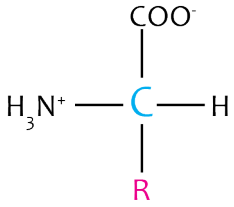
Para cada sequência de aminoácidos é atribuída uma estrutura 3D única, a qual é essencial para determinar a função exercida pela proteína. As diversas possíveis estruturas 3D permitem que proteínas desempenham variadas funções bioquímicas; sendo elas: receptores celulares, proteínas enzimáticas (catálise), ou componentes estruturais da célula. Em suma, outros tipos de moléculas não são versáteis o suficiente para assumir todas as funções atribuídas a proteínas (LEHNINGER; NELSON; COX, 2005; LESK, 2010; SCHEEFF; FINK, 2003).

2.1 Aminoácidos

Proteínas são polímeros formados por aminoácidos, as variadas propriedades e atividades são consequência das diferentes combinações e sequências de um conjunto de 20 aminoácidos básicos. A estrutura básica de um aminoácido é composta por um Carbono central (Carbono α), ao qual está ligado um Hidrogênio, um grupo amino (-NH_2), um grupo carboxila (-COO^-) e um grupo R (Figura 2.1). Cada aminoácido se diferencia pelo grupo R, também chamado de cadeia lateral, o qual varia em estrutura, tamanho e carga elétrica; com exceção da glicina, cuja cadeia lateral é um átomo de Hidrogênio, todos os aminoácidos possuem 4 grupos diferentes ligados ao Carbono α, o que o torna um centro quiral (LEHNINGER; NELSON; COX, 2005; LESK, 2010; SCHEEFF; FINK, 2003).

A propriedade de quiralidade do Carbono faz com que haja apenas dois possíveis arranjos espaciais entre os 4 grupos ligados, desta maneira aminoácidos possuem dois

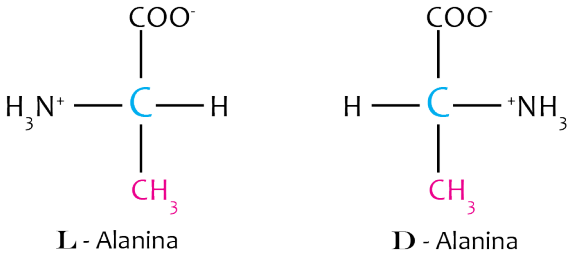
Figura 2.1: Estrutura Geral Molécula de Aminoácido



Fonte: adaptado de Lehninger et al. (2005)

possíveis estereoisômeros (Figura 2.2). Naturalmente, compostos biológicos ocorrem em apenas umas das duas formas, ou L-estereoisômero ou D-estereoisômero; no caso, aminoácidos presentes em proteínas são exclusivamente L-estereoisômero (LEHNINGER; NELSON; COX, 2005; SCHEEFF; FINK, 2003).

Figura 2.2: Estereoisômeros de Aminoácidos



Fonte: adaptado de Lehninger et al. (2005)

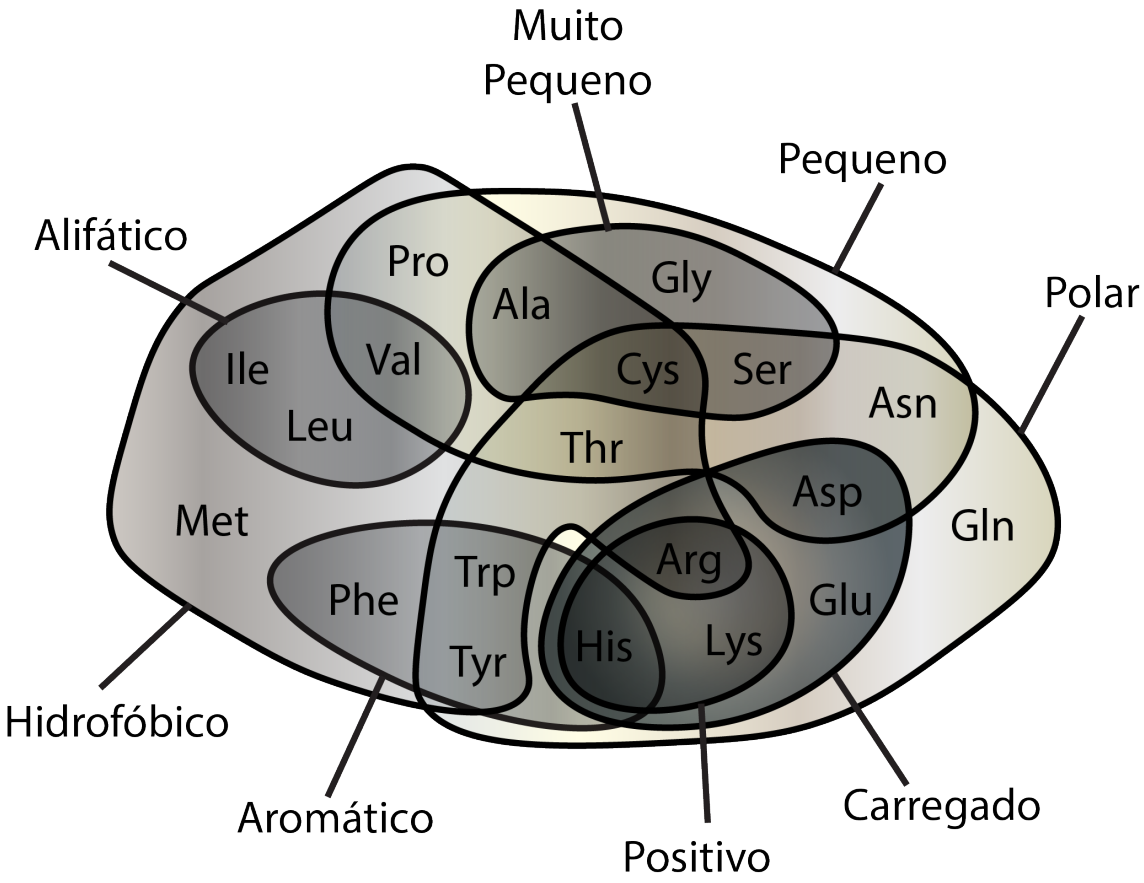
Aminoácidos podem ser classificados conforme características da cadeia lateral, como tamanho, forma e polaridade (propriedades físico-químicas) (TAYLOR, 1986). Esta classificação é representada através do Diagrama de *Venn*, o qual iclui subgrupos de aminoácidos, apresentado na Figura 2.3.

Proteínas podem sofrer modificações químicas na cadeia lateral e/ou interagir com moléculas não polipeptídicas, como íons, pequenos ligantes orgânicos e moléculas de água. Em alguns casos, são modificações cruciais para a formação da correta estrutura 3D; enquanto em outros casos irá diversificar a capacidade funcional da proteína (LEHNINGER; NELSON; COX, 2005; LESK, 2010; SCHEEFF; FINK, 2003).

2.2 Ligação Peptídica

Polipeptídios podem variar em tamanho, podendo conter entre 2 e 3 aminoácidos até milhares de aminoácidos conectados. Moléculas de aminoácidos são covalentemente conectadas por uma ligação amida, chamada ligação peptídica, a qual é formada por uma

Figura 2.3: Classificação de Aminoácidos



Ala - Alanina	Gln - Glutamina	Leu - Leucina	Ser - Serina
Arg - Arginina	Glu- Ác. Glutâmico	Lys - Lisina	Thr - Treonina
Asn - Asparagina	Gly - Glicina	Met - Mationina	Trp - Triptofano
Asp - Ác. Aspártico	His - Histidina	Phe - Fenilalanina	Tyr - Tirosina
Cys - Cisteína	Ile - Isoleucina	Pro - Prolina	Val - Valina

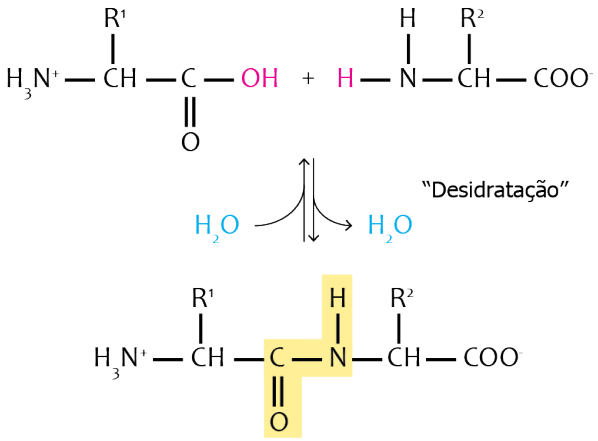
Fonte: adaptado de Taylor (1986)

reação de condensação (LEHNINGER; NELSON; COX, 2005). A ligação é formada pela união do grupo α-carboxila de um aminoácido com o grupo α-amino de um segundo aminoácido, havendo a formação e liberação de uma molécula de água (Figura 2.4), caracterizando um processo de desidratação (LEHNINGER; NELSON; COX, 2005).

A ligação peptídica (átomos C-N) apresenta um caráter de ligação dupla parcial, ou seja, sendo limitada a livre rotação ao redor da ligação; assim, o posicionamento dos grupos R é restrito, sendo a ligação peptídica capaz de adotar duas configurações: *cis* ou *trans* (LEHNINGER; NELSON; COX, 2005; LESK, 2010). Devido ao volume das cadeias laterais, a configuração *trans* é mais estável para maior parte dos aminoácidos (LEHNINGER; NELSON; COX, 2005).

Peptídeos sempre iniciam com um resíduo de aminoácido α-amino livre, N-terminal, e terminam com um resíduo de aminoácido α-carboxila, C-terminal; visto que, durante

Figura 2.4: Reação de Formação da Ligação Peptídica



Fonte: adaptado de Lehninger et al. (2005)

síntese de proteínas nas células, a adição de aminoácidos ocorre nesta ordem.

2.3 Conformação Proteica

A maneira com que os átomos de uma proteína se organizam no espaço tridimensional representa sua conformação. Teoricamente, existem inúmeras possíveis conformações as quais uma proteína pode assumir, ou seja, qualquer conformação em que não haja quebra de ligações covalentes é uma possibilidade. Entretanto, em condições biológicas, existem conformações preferenciais, chamadas estruturas nativas.

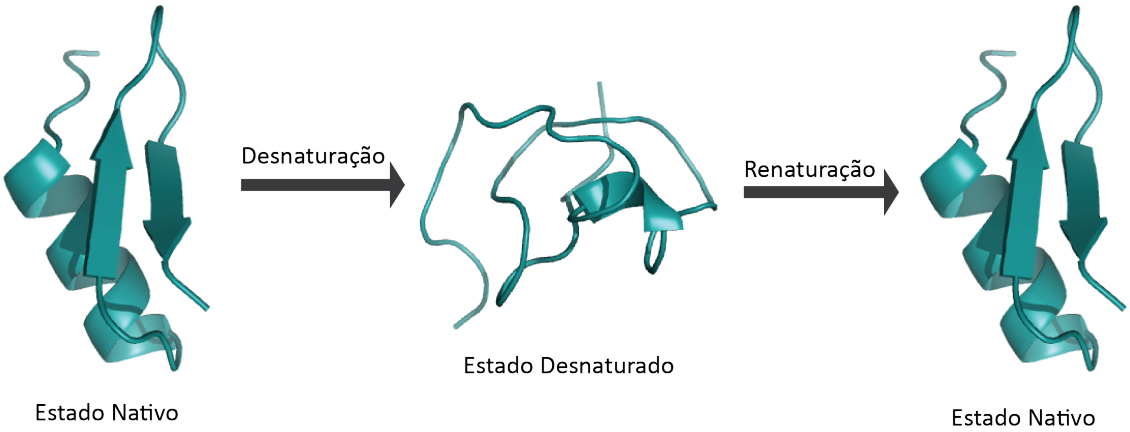
2.3.1 Estado Nativo

A estrutura 3D de proteínas está implícita na sequência de aminoácidos; em condições biológicas, proteínas enovelam espontaneamente em sua estrutura nativa (LEHNINGER; NELSON; COX, 2005; LESK, 2010). Proteínas em seu estado nativo são ditas funcionais, ou seja, estão biologicamente ativas; além de apresentarem uma estrutura compacta e marginalmente estável. A estrutura nativa de uma proteína é estabilizada por ligações dissulfetos e interações não-covalentes fracas, como ligações de Hidrogênio e interações de *van der Walls*. Em suma, conformações com o número máximo de interações fracas apresentam valores de energia livre mais baixos, ou seja, são mais estáveis (LEHNINGER; NELSON; COX, 2005).

Alterações no ambiente fisiológico, como calor, pH extremos, presença de com-

postos desnaturantes, resultam em modificações estruturais, as quais, quando suficientes, provocam a perda da função biológica (desnaturação). Proteínas desnaturadas apresentam sua estrutura 3D nativa desestabilizada; além de não exibirem uma estrutura definitiva, ou seja, podem assumir múltiplas conformações. Contudo, proteínas em estado desordenado são capazes de enovelar em sua estrutura nativa e funcional quando restabelecidas as condições fisiológicas (renaturação), como visto na Figura 2.5 (LEHNINGER; NELSON; COX, 2005).

Figura 2.5: Reação de Desnaturação-Renaturação Proteica



Fonte: do autor (2016)

2.3.2 Processo de Enovelamento Proteico

O processo de enovelamento se inicia durante a síntese de proteínas no ribossomo e requer $\sim 10^{13}$s; entretanto, descrever o processo de enovelamento logicamente ainda não é possível (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

Proteínas desnaturadas podem assumir diversas conformações; porém, para todas as moléculas com a mesma sequência de aminoácidos em mesmas condições a estrutura 3D nativa adotada será a mesma. Portanto, pode-se assumir que existem várias trajetórias "estado desnaturado $\rightarrow$ estado nativo" (LESK, 2010). Se o processo de enovelamento ocorresse de maneira randômica, em que fossem testadas todas possíveis conformações até encontrar a estrutura nativa, levaria aproximadamente 10^{10} anos para explorar o espaço conformacional de pequenos polipeptídios (LESK, 2010; LEHNINGER; NELSON; COX, 2005). Em outras palavras, proteínas não podem enovelar utilizando um processo de tentativa e erro, de fato, ocorre em escalas de segundos; contradição conhecida como Paradoxo de Levinthal (LEVINTHAL, 1978).

Termodinamicamente, a estabilidade do estado nativo corresponde à variação de energia livre entre os estados desnaturado (D) e nativo (N) (Equação 2.1). A conformação nativa estável terá um valor negativo de energia livre de *Gibbs* (ΔG^{Θ}), caracterizado por valores altamente negativos de entropia (ΔS^{Θ}) devido à redução de liberdade do estado nativo compactado. Geralmente, proteínas estabilizam entre -20 e -60 kJ mol^{-1} (LESK, 2010).

$$\Delta G_{D \rightarrow N} = G(N) - G(D) \tag{2.1}$$

Durante o processo de enovelamento, proteínas assumem conformações parciais, estados intermediários, também chamados de "*molten globule*". São estruturas relativamente compactas, com uma quantidade significativa de estruturas secundárias, relativamente flexíveis, com o interior pouco denso, e apresentando valores de energia livre intermediários entre os estados desnaturados e nativos.

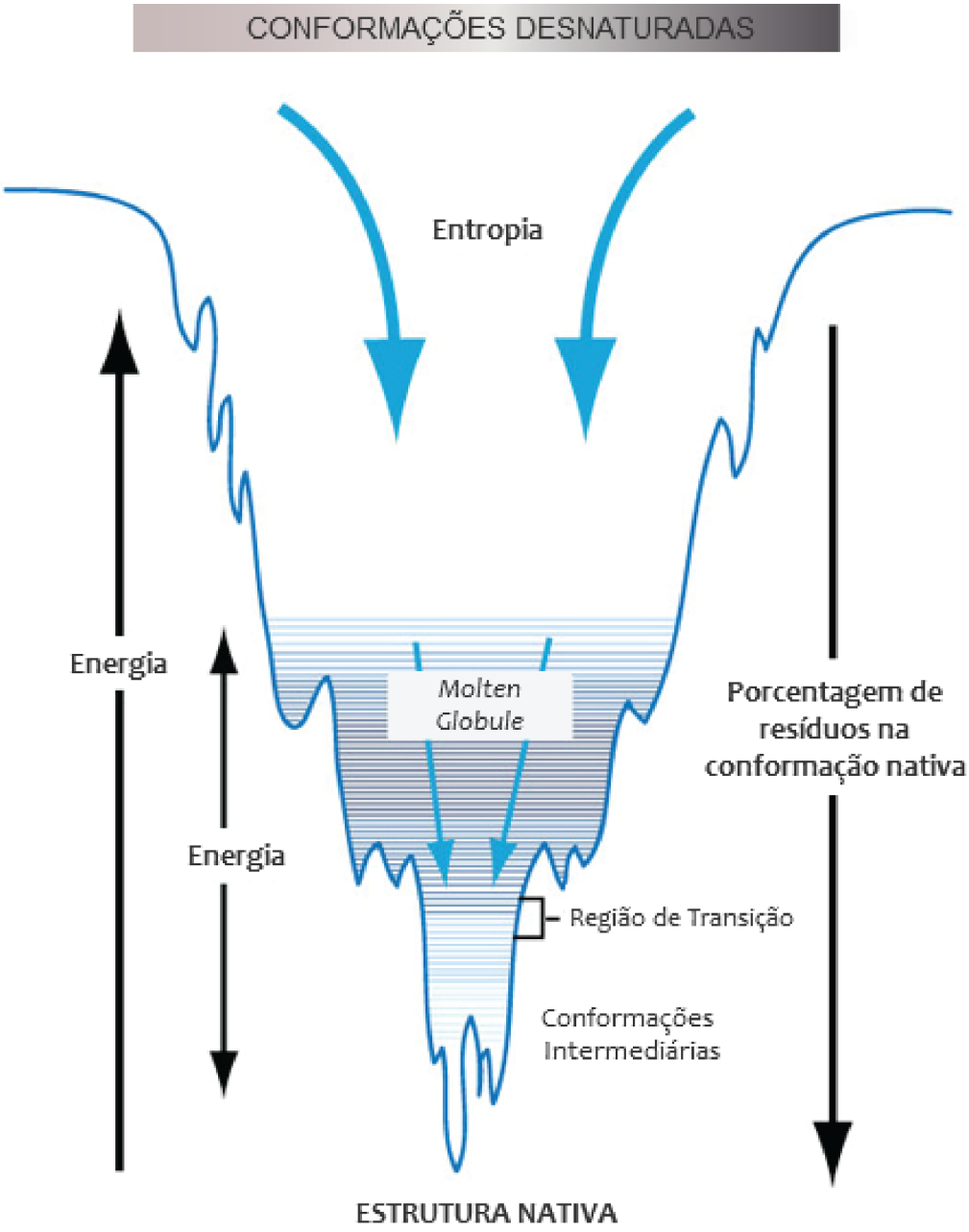
Assim, o processo de enovelamento proteico é descrito como um funil de energia livre (Figura 2.6), demonstrando que o alto número de conformações no estado desnaturado (altos valores de entropia) converge para a pequena fração de conformações no estado nativo (decréscimo dos valores de entropia). Pequenas depressões representam os mínimos locais, ou seja, estados intermediários semi-estáveis (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

2.3.3 Preferências Conformacionais

A conformação de uma proteína pode ser descrita quantitativamente como ângulos internos ao redor das ligações da cadeia principal. Devido ao caráter parcial de ligação dupla, rotações ao redor da ligação peptídica não são permitidas; sendo permitidas apenas rotações entre os átomos das ligações N-Cα e Cα-C. Assim, a cadeia polipeptídica é representada como consecutivos planos rígidos compartilhando um ponto de rotação no Cα, como apresentado na Figura 2.7.

Grupos de quatro átomos sucessivos definem um ângulo diedral. Para um resíduo *i*, o ângulo definido pelos átomos $C_{(i-1)}$-N-Cα-C representa o ângulo *phi* (ϕ) e o ângulo definido pelos átomos N-Cα-C-$N_{(i+1)}$ representa o ângulo *psi* (ψ). Os valores do par de ângulos são definidos como 180,0° para uma cadeia polipeptídica estendida ao longo de um único plano, podendo assumir qualquer valor entre o intervalo de -180,0° a +180,0°.

Figura 2.6: *Protein Folding Funnel Tunnel*

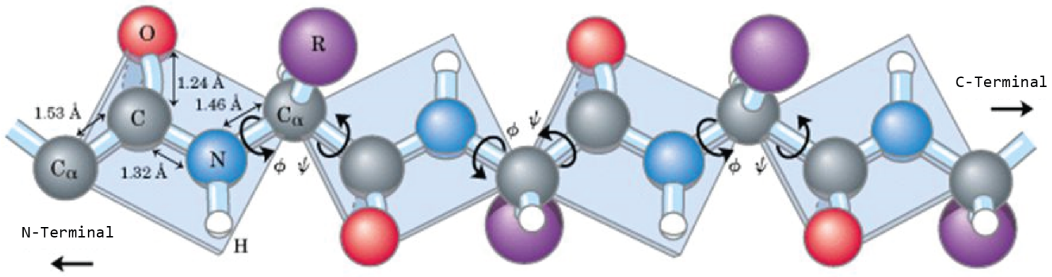


Fonte: adaptado de Lehninger et al. (2005)

No entanto, inúmeros valores não são permitidos devido os choques estereoquímicos entre os átomos da cadeia principal e cadeia lateral (LEHNINGER; NELSON; COX, 2005; LESK, 2010; SCHEEFF; FINK, 2003).

As combinações de valores permitidos para ângulos *phi* e *psi* são mostradas graficamente pelo diagrama de Ramachandran (RAMACHANDRAN; SASISEKHARAN,

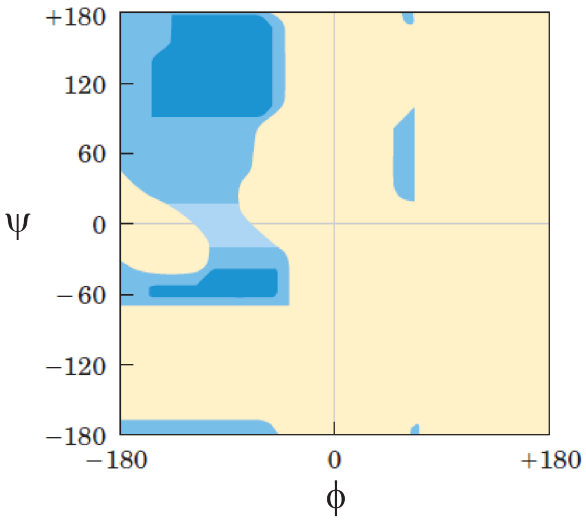
Figura 2.7: Representação Planar da Cadeia Polipeptídica



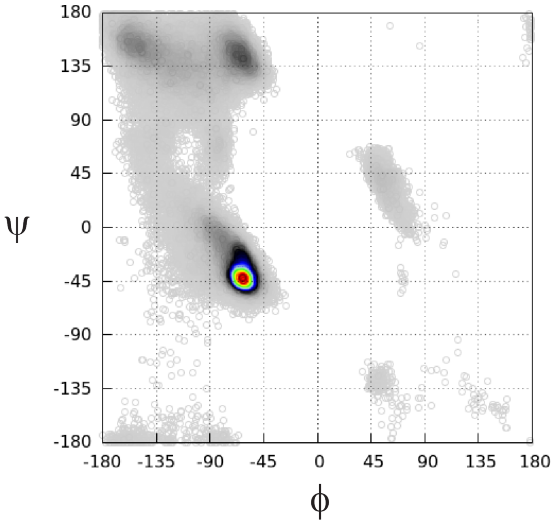
Fonte: adaptado de Lehninger et al. (2005)

1968). O gráfico é colorido utilizando uma escala de intensidade; áreas mais escuras representam regiões permitidas visto que não resultam em interferências estéricas, áreas intermediárias representam conformações permitidas em determinadas condições, áreas claras ou em branco representam regiões proibidas. A Figura 2.8 apresenta dois exemplos de representação para o **Mapa de Ramachandran** do aminoácido Alanina.

Figura 2.8: Exemplos Mapa de Ramachandran para o Aminoácido Alanina



Fonte: Lehninger et al. (2005)

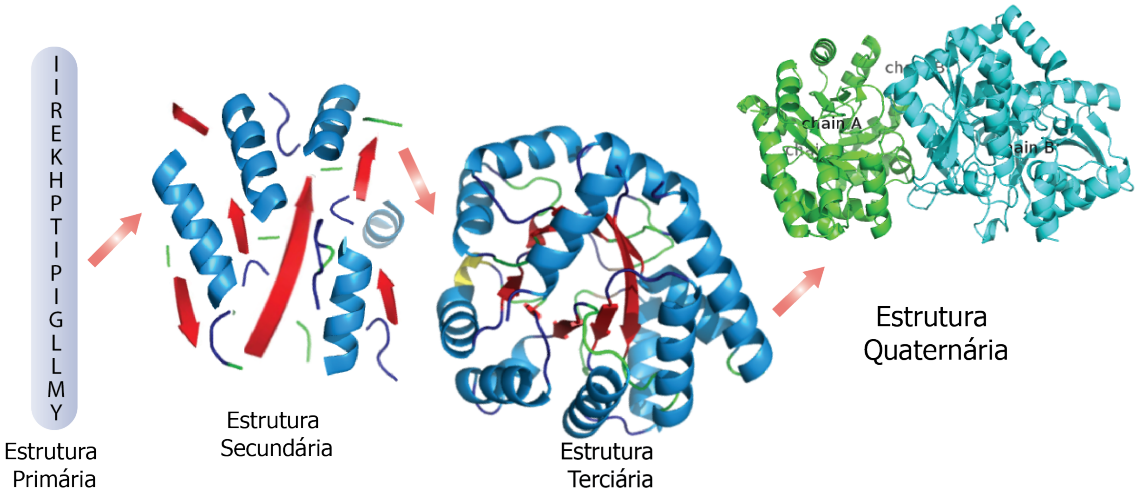


Fonte: Borguesan et al. (2015)

2.4 Organização Estrutural

Proteínas são macromoléculas claramente distintas pela sua estrutura, o entendimento e descrição de estruturas proteicas são feitos através de níveis, os quais são organizados de maneira hierárquica (Figura 2.9).

Figura 2.9: Níveis Estruturais de Proteínas



Fonte: do autor (2016)

2.4.1 Estrutura Primária

A estrutura primária de uma proteína consiste em sua sequência linear de aminoácidos. Cada proteína possui um tamanho e sequência única de aminoácidos, a qual define sua estrutura 3D única, e a mesma está relacionada a função desta proteína. Assim, alterações na estrutura primária podem acarretar modificações na função da proteína (LEHNINGER; NELSON; COX, 2005; LESK, 2010; SCHEEFF; FINK, 2003).

2.4.2 Estrutura Secundária

A estrutura secundária de uma proteína consiste em conformações locais que ocorrem ao longo da sequência de aminoácidos, ou seja, padrões estruturais presentes na cadeia principal, os quais são estabilizados por ligações de Hidrogênio entre os grupos N-H e C=O das ligações peptídicas. Alguns tipos de estruturas secundárias são mais estáveis e ocorrem mais comumente em proteínas, como hélices e folhas (LEHNINGER; NELSON; COX, 2005; LESK, 2010; SCHEEFF; FINK, 2003).

2.4.2.1 Hélices

α-hélices foram descritas por PAULING, COREY and BRANSON (1951) como uma estrutura em que a cadeia polipeptídica se projetava ao redor de um eixo imaginário que passava longitudinalmente pelo centro da hélice. α-hélices (Figura 2.10.a) possuem

3,6 resíduos por volta, com uma ligação de Hidrogênio entre os átomos CO do resíduo *n* e os átomos NH do resíduo *n*+4; cada unidade de volta se estende por 5,4Å. Aminoácidos presentes em estruturas de α-hélices com torção anti-horária, a qual é encontrada em proteínas, apresentam ângulos conformacionais de aproximadamente ϕ = -60,0° e ψ = -45,0° a -65,0° (Figura 2.10.a) (RICHARDSON, 1981).

Devido às inúmeras ligações de Hidrogênio formadas entre as ligações peptídicas, α-hélices são as estruturas secundárias mais estáveis e que ocorrem mais comumente em proteínas. São estruturas locais que podem ter uma influência sobre a estabilidade e organização do estado nativo superior a qualquer outro elemento estrutural. Entretanto, nem todos polipeptídios podem formar hélices estáveis, devido as interações entre a cadeia lateral dos diferentes aminoácidos.

A diferença de cargas entre os átomos da ligação peptídica gera um pequeno dipolo elétrico. Em estruturas de hélices os dipolos elétricos são conectados por ligações de Hidrogênio formando um macro dipolo, em que a extremidade N-terminal da hélice está carregada positivamente e a extremidade C-terminal negativamente. Além do dipolo elétrico ao longo de hélices, a repulsão eletrostática entre sucessivos aminoácidos com carga, a sobreposição espacial entre átomos da cadeia lateral, a presença de aminoácidos Prolina e Glicina são fatores que interferem na estabilidade de estruturas secundárias de hélices (LEHNINGER; NELSON; COX, 2005).

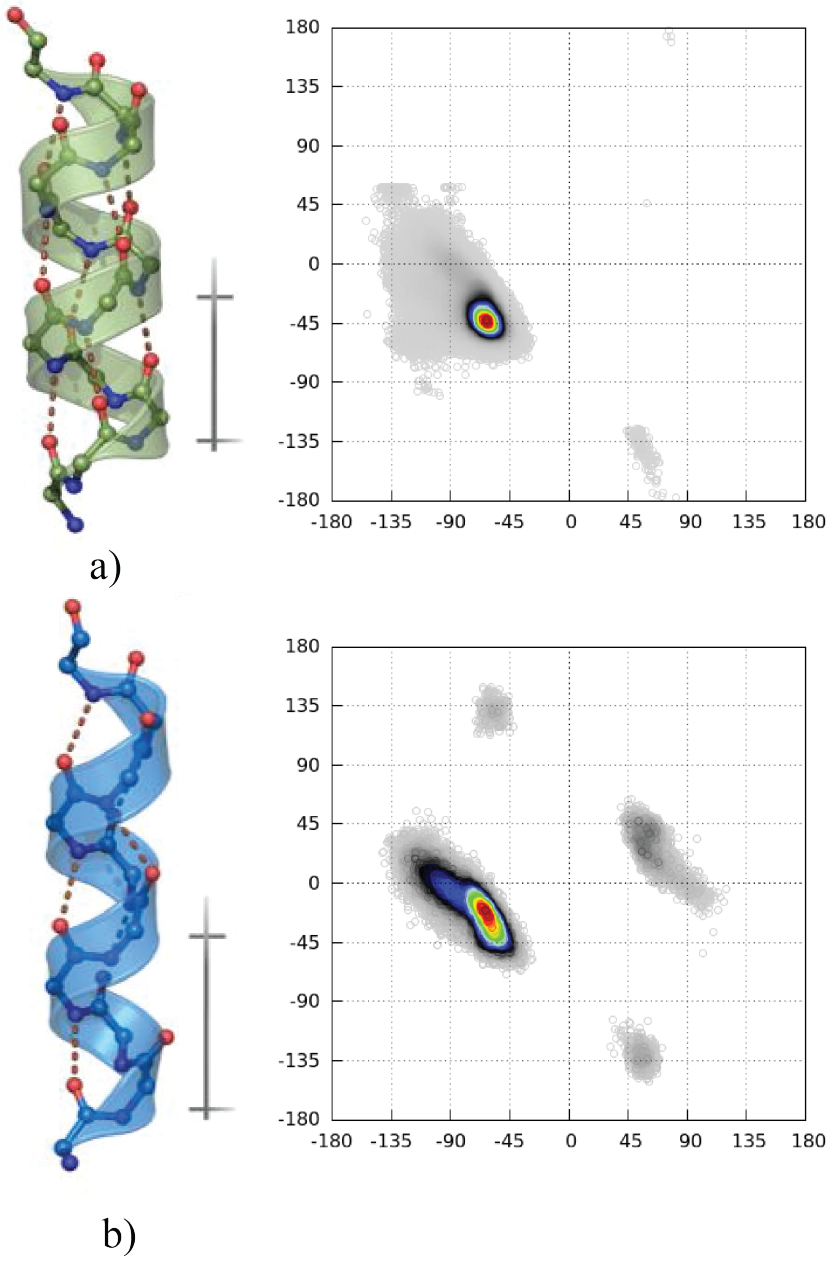
Menos favoráveis energeticamente e, consequentemente, menos comuns, 3_{10}-hélices (Figura 2.10.b) são estruturas similares a α-hélices. Todavia, cada unidade apresenta apenas 3 resíduos de aminoácidos com uma ligação de Hidrogênio entre os átomos CO do resíduo *n* e os átomos NH do resíduo *n*+3. Os ângulos conformacionais da cadeia principal apresentam valores de aproximadamente ϕ = -50,0° e ψ = -25,0° (Figura 2.10.b) (RICHARDSON, 1981; SCHEEFF; FINK, 2003).

2.4.2.2 Folhas Beta

Estruturas β foram descritas por PAULING and COREY (1951) como estruturas de conformação estendida e pregueada ao longo da cadeia polipeptídica, ou seja, uma conformação em "*zig-zag*". Folhas-β interagem formando ligações de Hidrogênio entre conjuntos de resíduos independentes e adjacentes, podendo aproximar regiões distantes na sequência de aminoácidos. As cadeias laterais ao longo de uma mesma estrutura de folha se direcionam de maneira a criar um padrão alternado.

Folhas-β podem interagir de maneira paralela ou antiparalela (Figura 2.11.a e b,

Figura 2.10: Representação das estruturas a) α-hélice e b) 3_{10}-hélice, com respectivos Mapas de Ramachandran

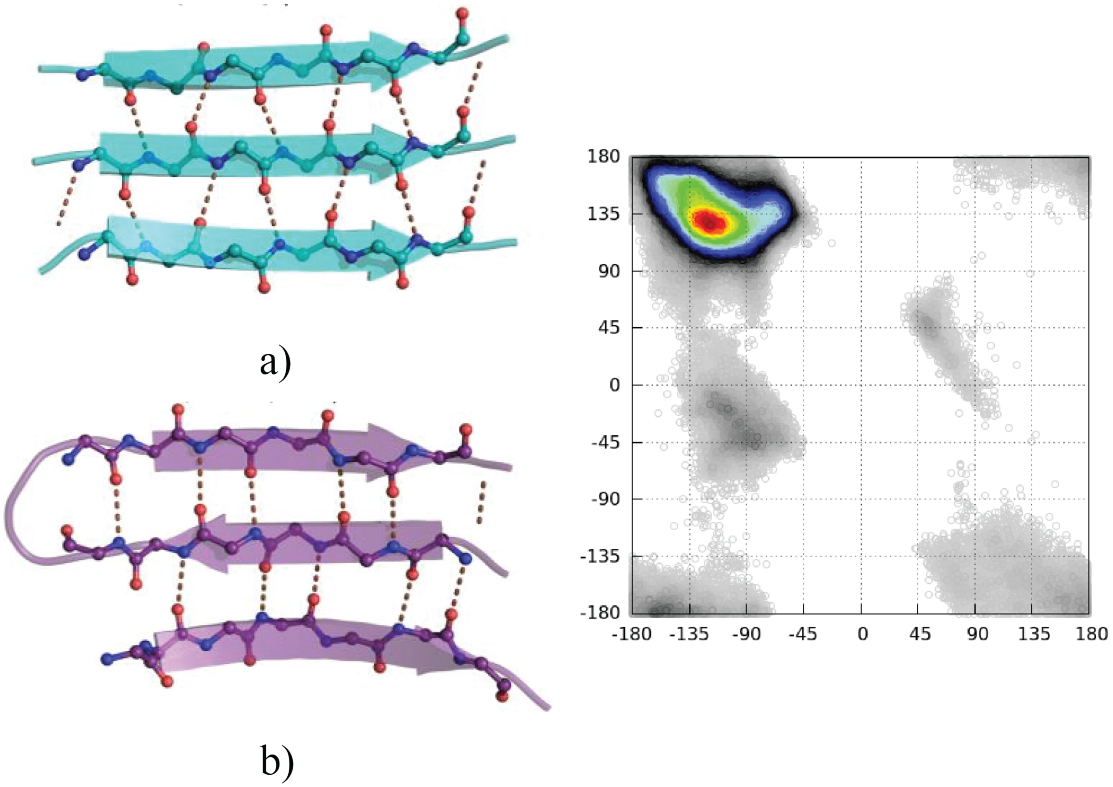


Fonte: adaptado de Verli (2014); Borguesan et al. (2015)

respectivamente), ou seja, tendo a mesma orientação amino-carboxila ou oposta, respectivamente. Apesar de serem estruturas similares, apresentam diferentes comprimentos e padrões de ligações de Hidrogênio. A unidade de repetição em folhas-β antiparalelas apresenta 7,0Å e ligações de Hidrogênio perpendiculares à folha; enquanto que em folhas-β paralelas são mais curtas com 6,5Å e ligações de Hidrogênio espaçadas alternativamente . Em geral, folhas-β paralelas são mais regulares e ocorrem em estruturas com pelo menos 5 folhas, enquanto que folhas-β antiparalelas podem se formar em estruturas de apenas 2 folhas (RICHARDSON, 1981). A preferência conformacional para folhas-β

é apresentada no Mapa de Ramachandran da Figura 2.11.

Figura 2.11: Representação estruturas folhas-β a) paralelas e b) antiparalelas com respectivo Mapas de Ramachandran



Fonte: adaptado de Verli (2014); Borguesan et al. (2015)

Conexões entre folhas-β devem se encaixar entre dois possíveis padrões: *hairpin* ou *crossover* (RICHARDSON, 1981). Em conexões *hairpin* a região N-terminal de uma folha está adjacente a região C-terminal da folha seguinte, gerando um *loop* entre ambas. Em conexões *crossover*, o *loop* entre as folhas ocorre em terminais oposto.

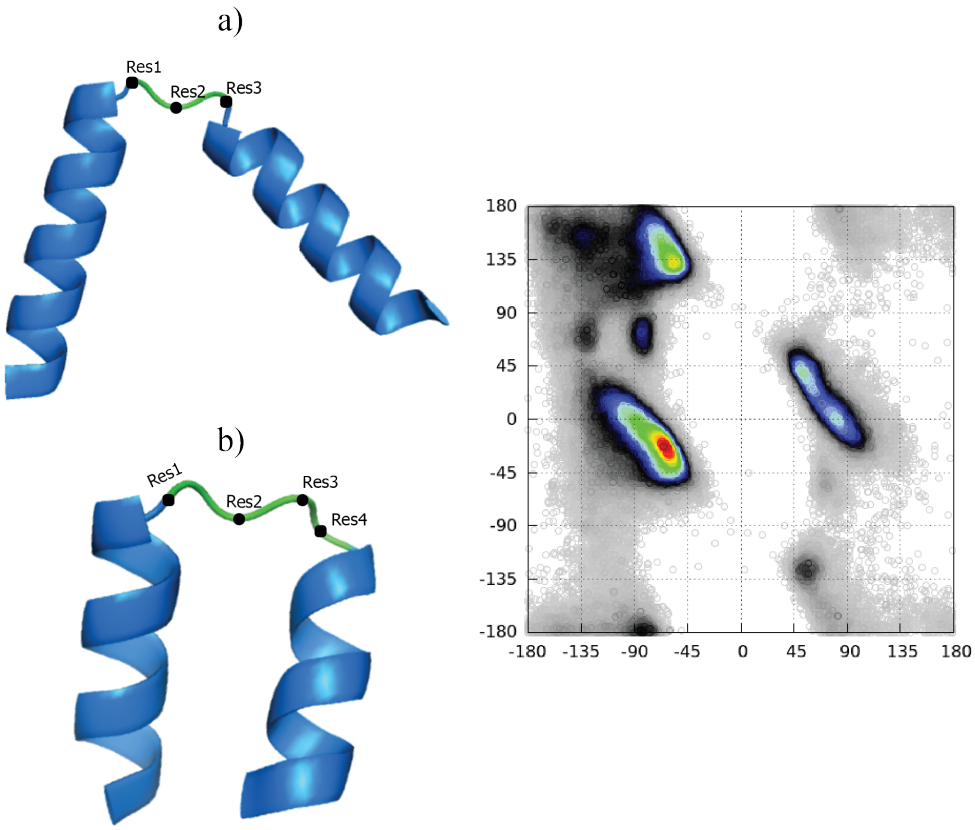
2.4.2.3 Voltas e Regiões Desordenadas

Estruturas secundárias de voltas e regiões desordenadas são descritas como estruturas flexíveis e irregulares as quais não apresentam padrões repetitivos comuns, logo, são difíceis de serem classificadas.

Voltas são regiões de pequenas estruturas secundárias conectando estruturas repetitivas de hélices e folhas. São constituídas de n aminoácidos consecutivos, em que a distância entre o C do resíduo i e $i+n$ não deve ultrapassar 7,0Å. Voltas são compostas por voltas-γ (n = 3), voltas-β (n = 4), voltas-α (n = 5) e voltas-π (n = 6). A classificação de voltas em tipos é feita a partir dos valores de ângulos diedros (*phi* e *psi*) dos aminoácidos

centrais (OFFMANN; TYAGI; BREVERN, 2007). Visto que estruturas de voltas orientam estruturas de hélices e folhas, seu principal papel está na topologia final da proteína (OFFMANN; TYAGI; BREVERN, 2007). Exemplos de estruturas de voltas são apresentados na Figura 2.12, assim como sua preferência conformacional apresentada no Mapa de Ramachandran.

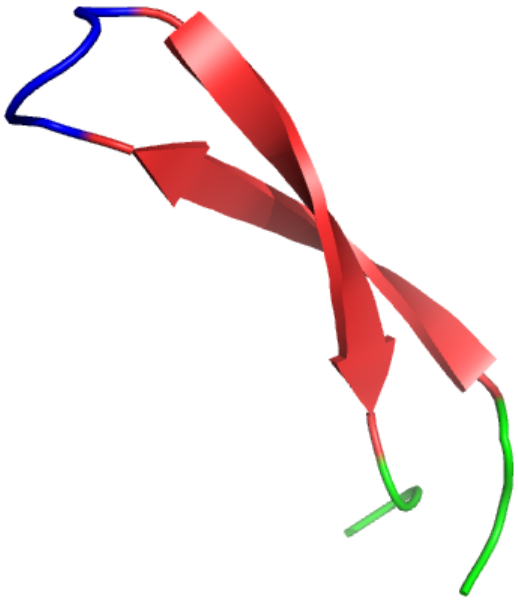
Figura 2.12: Representação das estruturas de voltas a) γ e b) β com respectivo Mapas de Ramachandran



Fonte: do autor (2016); Borguesan et al. (2015)

Regiões Desordenadas são flexíveis e conectam estruturas rígidas além de, geralmente, não possuírem estruturas secundárias. Devido sua flexibilidade e diversidade, são importantes no processo de enovelamento e estabilidade proteica, além de exercerem papel crítico na funcionalidade da proteína. Regiões Desordenadas são classificadas conformacionalmente baseadas nas estruturas secundárias. β-*hairpin* (Figura 2.13) é um exemplo em que duas fitas antiparalelas adjacentes são conectadas por regiões desordenadas (OFFMANN; TYAGI; BREVERN, 2007) .

Figura 2.13: Representação Estrutura β-*hairpin*



Fonte: do autor (2016)

2.4.3 Estrutura Terciária

A estrutura terciária de uma proteína consiste no arranjo espacial entre as estruturas de folhas, hélices, voltas e regiões desordenadas, formando um padrão conformacional global. Aminoácidos que se encontram distantes ao longo da sequência polipeptídica e que estão em diferentes estruturas secundárias podem interagir na estrutura terciária, também chamada tridimensional.

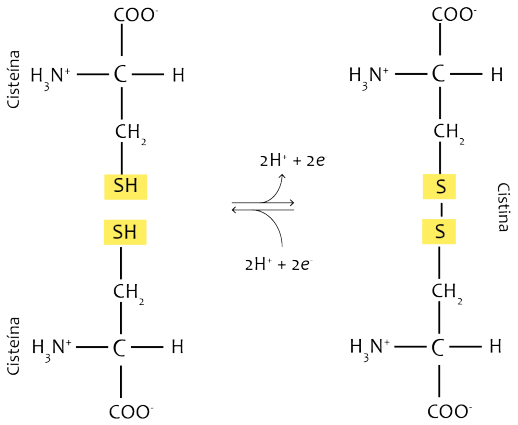
Estruturas terciárias são estabilizadas por interações não-covalentes fracas. Interações eletrostáticas entre aminoácidos polares são dependentes do meio, sendo as ligações de Hidrogênio e as interações de *van der Walls* as principais contribuintes. Além disso, pode haver ligações covalentes entre as cadeias laterais de aminoácidos Cisteína na forma de ligações dissulfeto (S-S) (Figura 2.14), as quais contribuem para estabilizar a estrutura 3D (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

Proteínas podem ser classificadas em dois grandes grupos em função das propriedades bioquímicas de sua estrutura terciária: proteínas fibrosas e proteínas globulares.

2.4.3.1 Proteínas Fibrosas

A cadeia polipeptídica de proteínas fibrosas é empacotada em formato de folhas ou fios. Geralmente apresentam sequências repetidas, as quais adotam apenas um tipo de estrutura secundária regular ou estruturas secundárias atípicas, como regiões desordena-

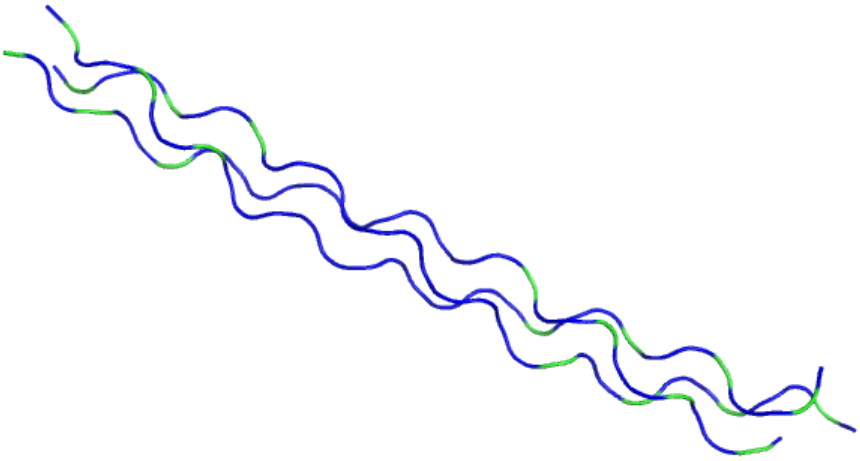
Figura 2.14: Reação de formação das Ligações Dissulfeto



Fonte: adaptado de Lehninger et al. (2005)

das (LEHNINGER; NELSON; COX, 2005; SCHEEFF; FINK, 2003). Proteínas fibrosas possuem propriedades que concedem força e flexibilidade às estruturas em que ocorrem; como a propriedade de insolubilidade em água, consequência da alta concentração de aminoácidos hidrofóbicos no interior e superfície da proteína. Alguns exemplos de proteínas fibrosas são: α-queratinas, colágeno e fibroína (LEHNINGER; NELSON; COX, 2005; LESK, 2010). A Figura 2.15 mostra um exemplo de proteína fibrosa, em que três "fios"compostos de estruturas de voltas (azul) e regiões desordenadas (verde) se entrelaçam formando uma estrutura helical.

Figura 2.15: Representação de Proteínas Fibrosas (Colágeno)

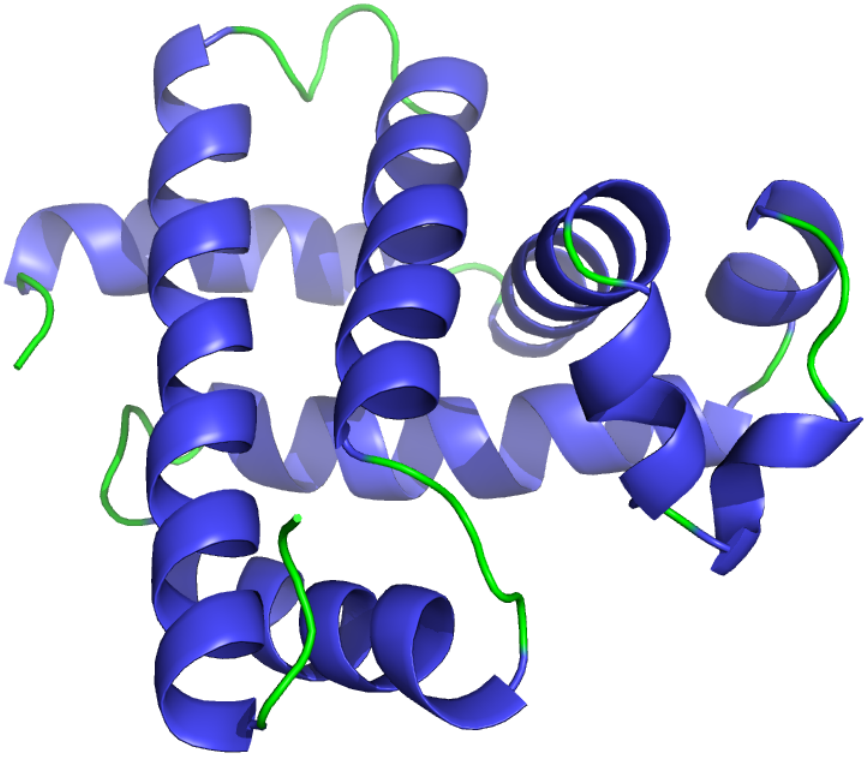


Fonte: *Protein Data Bank* (PDB_ID: 1BKV)

2.4.3.2 Proteínas Globulares

Em proteínas globulares a cadeia polipeptídica é empacotada em estruturas de formato globular solúvel, e podem apresentar variados tipos de estruturas secundárias. Essa diversidade é fundamental para a determinação da função exercida pela proteína; proteínas globulares incluem enzimas, proteínas transportadoras, motoras, reguladoras, imunoglobulinas e inúmeras outras funções (LEHNINGER; NELSON; COX, 2005; LESK, 2010). Proteínas globulares, normalmente, são encontradas em ambientes aquosos, assim são estruturas compactas com um núcleo hidrofóbico e uma superfície hidrofílica (SCHEEFF; FINK, 2003). Exemplos de proteínas globulares são: mioglobinas, citocromo *c*, lisoenzima e ribonuclease (LEHNINGER; NELSON; COX, 2005; LESK, 2010). A Figura 2.16 mostra um exemplo de proteínas globulares, onde estruturas de hélices estão representadas em azul e estruturas de voltas em verde.

Figura 2.16: Representação de Proteínas Globulares (Mioglobina)



Fonte: *Protein Data Bank* (PDB_ID: 1MBN)

2.4.4 Estrutura Quaternária

A estrutura quaternária de uma proteína descreve como duas ou mais cadeias polipeptídicas, ou subunidades, se conectam para formar uma única proteína funcional. Assim como estruturas terciárias, a conformação é determinada por interações iônicas e hidrofóbicas entre os grupos R de aminoácidos.

2.5 Métodos Experimentais para Determinação de Estrutura Tridimensional

A determinação da estrutura 3D de macromoléculas inclui métodos experimentais e computacionais, além da combinação de ambos. As estruturas 3D da mioglobina e hemoglobina foram as primeiras a serem determinadas em meados de 1950 através da técnica de cristalografia por difração de raios-X (LESK, 2010). Por anos a técnica foi a única fonte de estruturas de macromoléculas, até que, em 1980, surgiu a técnica de ressonância nuclear magnética (RMN). Atualmente, existem mais de 115 mil[1] estruturas 3D de proteína determinadas e armazenadas em banco de dados, como o PDB. Teoricamente, métodos computacionais seriam capazes de deduzir a estrutura 3D de uma proteína a partir de sua sequência de aminoácidos, entretanto, até o momento, métodos experimentais tem sido a principal fonte de dados.

2.5.1 Cristalografia por Difração de Raio-X

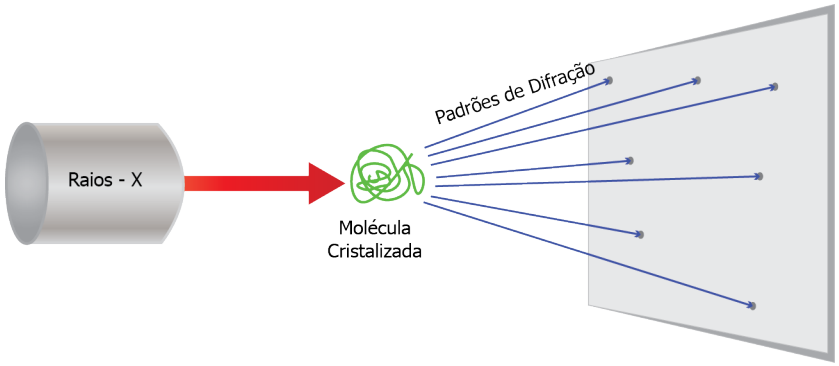
A determinação da estrutura 3D de uma macromolécula inicia com o isolamento, purificação e cristalização da mesma, buscando um cristal de ordem completa e tamanho adequado. O cristal é bombardeado com feixes de raios-X, ondas de comprimento entre 0,07 e 0,15 *nm*, sendo observados padrões de difração gerados pelos elétrons dos átomos (Figura 2.17). Todos os átomos do cristal agem como fonte secundária de radiação, emitindo uma fração da energia absorvida, entretanto, a maior parte dos raios dispersos por diferentes átomos se cancelam, ou seja, apenas uma pequena parcela é responsável pelos padrões de difração (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

Com base nos dados de difração é gerado um mapa de densidade eletrônica para a proteína aplicando o método matemático transformação de Fourier. Regiões de picos na

[1] http://www.rcsb.org/pdb/home/home.do

densidade eletrônica correspondem a posição de átomos; esta informação é computada e utilizada na construção do modelo estrutural final (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

Figura 2.17: Técnica de Cristalografia por Difração de Raio-X



Fonte: adaptado de Lesk (2010)

2.5.2 RMN

A técnica de Ressonância Magnética Nuclear (RMN) permitiu a determinação de estruturas em solução, podendo ser aplicada àquelas moléculas que não cristalizam. Espectros de RMN medem o momento angular de *spin* nuclear, uma propriedade mecânica quântica do núcleo atômico; isto é, na presença de um campo magnético externo, são medidos os valores de transição entre os níveis de energia do núcleo magnético dos átomos. Os níveis de energia de um núcleo magnético são dependentes do momento magnético nuclear, das características do átomo e da força do campo magnético externo aplicado; portanto, cada grupo químico irá aparecer em regiões diferentes do espectro de RMN (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

Aminoácidos de proteínas desnaturadas do mesmo tipo apresentam energias de transição similares, entretanto apresentam dispersão de espectro quando encontrados em estruturas conformacionais. Espectros de RMN podem determinar valores de ângulos conformacionais (*phi* e *psi*) da cadeia principal através da transição química de átomos de H, Cα, Cβ, carbonil-C e N. Além disso, é possível identificar interações entre átomos não ligados, mas com distância menor que 5,0Å, provendo informação sobre o padrão conformacional global. Para gerar estruturas 3D, os dados de distâncias e ângulos conformacionais são enriquecidos com dados geométricos conhecidos, como quiralidade, raio de *van der Walls*, e comprimento de ligações e ângulos. Computacionalmente é gerado

um grupo de estruturas relacionadas, o qual representa a variação de conformações para determinada molécula (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

2.6 Banco de Dados Estruturais

Banco de dados são responsáveis por armazenar e distribuir informação, além de organizarem o conhecimento impondo uma estrutura lógica de maneira que o acesso seja rápido e fácil. Banco de dados estruturais armazenam e distribuem dados de coordenadas atômicas, sendo o *wwPDB* (*world-wide Protein Data Bank*)[2] o maior banco de dados de estruturas de macromoléculas biológicas.

2.6.1 *Protein Data Bank*

Em 1971, W. Hamilton criou o PDB[3] no Laboratório Nacional Brookhaven como um depósito para estruturas cristalizadas de macromoléculas biológicas (BERMAN et al., 2000). O avanço nas técnicas envolvidas no processo de cristalização, assim como o desenvolvimento do método de determinação estrutural por RMN levou ao aumento no número de estruturas depositadas com o passar dos anos.

Uma estrutura depositada no PDB recebe um identificador de quatro caracteres. Cada determinação experimental de uma proteína aparece como uma única entrada, cada entrada irá conter: proteína e espécie correspondente; autor e referências bibliográficas; informações do experimento; sequência de aminoácidos; coordenadas atômicas; informação de moléculas adicionais, caso haja; estrutura secundária; pontes dissulfetos. Os dados são distribuídos no formato PDB ou mmCIF (*macromolecule crystalographic information file*).

2.7 Classes de Proteínas

Ao longo dos anos, o número de proteínas com sua estrutura 3D determinada tem aumentado. Essas proteínas apresentam similaridades estruturais e muitas vezes uma his-

[2] http://www.wwpdb.org/
[3] http://www.rcsb.org/pdb/home/home.do

tória evolutiva em comum; para facilitar o acesso a essa informação estrutural e evolutiva, proteínas são classificadas e organizadas em níveis hierárquicos estruturais.

2.7.1 SCOP

O SCOP[4] (*Structural Classification of Proteins*) organiza proteínas hierarquicamente de acordo com sua origem evolutiva e similaridades estruturais (MURZIN et al., 1995). Em seu nível mais baixo, são classificados os **domínios** individuais de cada proteína, a partir de dados extraídos do PDB. Proteínas que apresentam significante similaridade em sua sequência de aminoácidos e demonstram similar função e estrutura conformacional são agrupadas em **famílias** de homólogos. Famílias de proteínas geralmente apresentam forte relação evolutiva, ou seja, possuem uma origem evolutiva comum. Quando duas ou mais famílias apresentam estrutura conformacional comum, ou, então, estrutura e função comum, mas a sequência de aminoácidos possui baixa similaridade, essas famílias são agrupadas em **superfamílias**. Superfamílias possivelmente possuem relação evolutiva.

A classificação entre os dois níveis mais altos é puramente estrutural, logo, superfamílias de proteínas que compartilham conformações em comum são agrupadas em ***folds***. Em seu nível mais alto, o SCOP classifica proteínas entre as principais **classes** conforme a predominância dos padrões estruturais: α, β, α+β (segmentos distribuídos em diferentes partes da molécula), α/β (segmentos intercalados ou alternados), e "pequenas proteínas".

Tabela 2.1: Classificação SCOP para proteína Uteroglobin de Coelhos (1UTG)

Classe	Proteínas α
Fold	Uteroglobin-like (multihelical)
Superfamília	Uteroglobin-like (dímeros dissulfetos de duas cadeias idênticas com 4 hélices cada)
Família	Uteroglobin-like
Proteína	Uteroglobin
Espécie	Coelho (*Oryctolagus cuniculus*)

Fonte: SCOP *website*

[4]http://scop.mrc-lmb.cam.ac.uk/scop/index.html

2.7.2 CATH

CATH[5] (*Class, Architecture, Topology, and Homologous superfamily*) apresenta um sistema de classificação similar ao SCOP, as letras do nome são referentes aos níveis de hierarquia: classe, arquitetura, topologia e superfamília de homólogos (SILLITOE et al., 2015). Proteínas que apresentam estrutura, função e sequência de aminoácidos similares são agrupadas em **superfamílias de sequências**. Proteínas classificadas em **superfamílias homólogas** apresentam evidência de ancestral comum, a qual é baseada em similaridade de sequência e estrutura. Superfamílias de homólogas que compartilham arranjo espacial e conectividade de hélices e folhas são classificadas na mesma **topologia**. Em uma mesma **arquitetura**, são classificadas proteínas com arranjos similares de hélices e folhas, mas que possuem diferente conformação espacial. Em seu nível mais alto, proteínas são classificadas nas principais **classes** de arquiteturas: α, β, α-β, e domínios com poucas estruturas secundárias.

Tabela 2.2: Classificação CATH para proteína Uteroglobin de Coelhos (1UTG)

Classe	Proteínas α
Arquitetura	Ortogonal
Topologia	Uteroglobin
Superfamília de Homologos	Uteroglobin

Fonte: gCATH *website*

2.8 Resumo do Capítulo

Proteínas são as macromoléculas mais abundantes, estando envolvidas em praticamente todas as funções e estruturas das células. São polímeros lineares formados por diferentes combinações de aminoácidos conectados pela ligação peptídica. Aminoácidos possuem uma estrutura básica e comum, assim, todas as moléculas de proteínas apresentam uma mesma cadeia principal, sendo a cadeia lateral responsável pela variabilidade. A sequência de aminoácidos irá determinar a estrutura 3D adotada pela proteína, a qual está relacionada a sua função biológica.

[5]http://www.cathdb.info/

Teoricamente proteínas podem se organizar no espaço tridimensional de inúmeras maneiras possíveis, entretanto, em condições biológicas, assumem uma estrutura 3D única, chamada estrutura nativa. O estado nativo de uma proteína corresponde ao seu estado funcional. Qualitativamente, conformações estruturais podem ser descritas por ângulos de torção, *phi* (ϕ) e *psi* (ψ). Devido às interações estereoquímicas entre átomos da cadeia principal e cadeia lateral, certos valores de ângulos não são permitidos, o que pode ser representado graficamente pelo Mapa de Ramachandran.

Proteínas apresentam 4 níveis estruturais: estrutura primária, correspondente a sequência linear de aminoácidos; estrutura secundária, correspondente a conformações locais, como hélices e folhas; estrutura terciaria, correspondente a conformação global; e estrutura quaternária, correspondente ao arranjo de mais de uma subunidade ou cadeia polipeptídica.

Conhecer a estrutura terciária de uma proteína permite a determinação e estudo de sua função. A determinação de estruturas terciárias pode ser feita através de técnicas experimentais (cristalografia e RMN) e métodos computacionais. Os dados conformacionais de proteínas são armazenados e organizados em banco de dados, como PDB; além disso, proteínas são classificadas estruturalmente e evolutivamente de maneira hierárquica, exemplos são SCOP e CATH.

3 PREDIÇÃO DE ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS: MÉTODOS COMPUTACIONAIS

A estrutura 3D de proteínas está intrinsecamente conectada à sua função. O conhecimento da conformação ajuda a determinar e entender o processo biológico de proteínas, ou seja, qual a sua atividade, qual substância é processada, qual ligante deve se ligar. Além disso, a informação tridimensional permite o estudo de motivos estruturais, conformações (*folds*), processo de enovelamento, relações evolutivas e estrutural-funcional.

A determinação de estruturas 3D é feita de maneira experimental pelas técnicas de cristalografia por difração de raios-X e RMN; entretanto são técnicas limitadas e hoje poucas proteínas possuem sua estrutura 3D conhecida. Por outro lado, houve uma explosão de dados de sequências ao longo dos anos, gerando uma lacuna entre o volume de dados obtidos pelas diferentes técnicas. Desta forma, o investimento em pesquisas em métodos computacionais para a predição de estruturas 3D de proteínas é essencial.

3.1 Contextualização do Problema

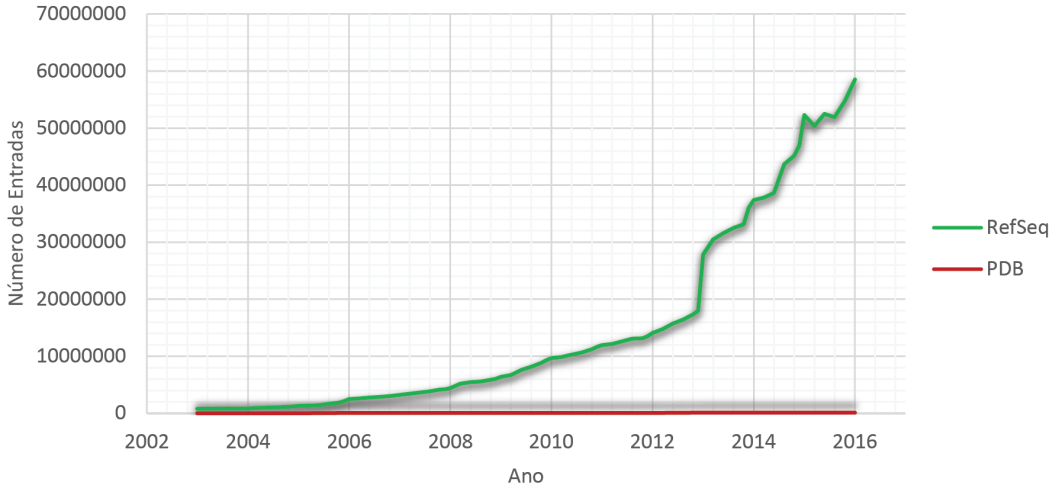
O Projeto Genoma e o desenvolvimento das técnicas de biologia molecular acarretaram numa explosão de dados, e, hoje, o número de sequências proteicas conhecidas e armazenadas no RefSeq é 70.427.238[1], enquanto que o número de estruturas 3D de proteínas no PDB é 115.119 [2] (Figura 3.1). Essa discrepância no volume de dados ilustra a dificuldade para determinar estruturas 3D através das técnicas experimentais.

A determinação de estruturas 3D pelo método de cristalografia por difração de raios-X é limitada pela necessidade de altos níveis de expressão, eficiente purificação e cristalização das proteínas-alvo. Já a técnica de RMN é limitada pelo tamanho das amostras que podem ser analisadas. Além disso, ambas as estratégias são extremamente custosas, levam muito tempo e necessitam de equipamentos altamente especializados. Logo, a necessidade do desenvolvimento e utilização de métodos computacionais para a predição de estruturas 3D de proteínas é demonstrada e motivada (DORN et al., 2014).

A predição estrutural de proteínas a partir da sequência de aminoácidos é um problema interdisciplinar e computacionalmente classificado como NP-Completo (CRESCENZI et al., 1998). A complexidade é devido à alta seletividade do processo de enove-

[1] Valores obtidos em novembro de 2016. http://www.ncbi.nlm.nih.gov/refseq/
[2] Valores obtidos em novembro de 2016. http://www.rcsb.org/

Figura 3.1: Crescimento do número de entradas do RefSeq e PDB.



Fonte: do autor (2016)

lamento proteico, em que uma sequência de aminoácidos irá adotar uma estrutura conformacional única entre as inúmeras possibilidades.

Ao longo dos anos, variadas estratégias e algoritmos foram desenvolvidos a fim de solucionar o problema de predição estrutural, as quais são divididas em 4 classes (DORN et al., 2014) e descritos:

i. Métodos de Primeiros Princípios sem Informação da Base Experimental;
ii. Métodos de Primeiros Princípios com Informação da Base Experimental;
iii. Métodos de *Fold Recognition* e *Threading*;
iv. Métodos de Modelagem Comparativa e Alinhamento de sequências.

3.2 Métodos de Primeiros Princípios sem Informação da Base Experimental

Métodos *ab initio* são fundamentados nas leis da termodinâmica e baseados no fato que a estrutura nativa de proteínas apresenta a menor energia livre global. Considerando que a sequência de aminoácidos contém toda a informação necessária para o processo de enovelamento proteico (ANFINSEN, 1973), são métodos que utilizam apenas a sequência de aminoácidos para a predição da estrutura 3D. Logo, não é permitido o uso de dados estruturais provindos de banco de dados e, assim, não havendo limitações dos modelos já existentes permitindo a predição de novos *folds*.

Métodos *ab initio* conduzem uma busca conformacional guiada por uma função de energia, em outras palavras, para todas conformações possíveis, calculam sua ener-

gia livre e selecionam a estrutura correspondente a energia livre global mínima (estrutura nativa) (DORN et al., 2014). Entretanto, os métodos ficam limitados pelo tamanho do espaço de busca (Paradoxo de Levinthal) (LEVINTHAL, 1978), tornando-o custoso computacionalmente.

São considerados três fatores (LEE et al., 2009) como base para a criação de métodos *ab initio*: (i) representação geométrica da cadeia polipeptídica; (ii) função de energia; e (iii) metodologia de busca. Representações geométricas mais detalhadas incluem todos os átomos da cadeia polipeptídica e as moléculas de solvente (*all-atom*), porém são computacionalmente custosas. Assim, inúmeras simplificações podem e são utilizadas, como representação *united-atom*, solvente implícito, rotâmeros.

A função de energia deve ser capaz de encontrar a energia livre global mínima, e pode ser classificada em dois grupos (LEE et al., 2009): (i) funções de energia baseadas em parâmetros físicos; e (ii) funções de energia baseadas em estruturas conhecidas. Como mencionado anteriormente, métodos *ab initio* não permitem a utilização de dados experimentais, entretanto estruturas já conhecidas são utilizadas para a parametrização de campos de força. Alguns exemplos de funções de energia são: AMBER (CORNELL et al., 1995), CHARMM (BROOKS et al.,) e GROMOS (CHRISTEN et al., 2005).

Por fim, os métodos devem ser capazes de encontrar a conformação de menor energia dentre as possibilidades de maneira eficiente. Simulações de dinâmica molecular (MD), simulações de Monte Carlo (MC) e algoritmo genético (GA) são exemplos de técnicas utilizadas. Atualmente, métodos *ab initio* geram modelos estruturais aceitáveis apenas para pequenos modelos (< 100 resíduos) (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014).

3.3 Métodos de Primeiros Princípios com Informação da Base Experimental

Novas conformações de proteínas, quando determinadas, apresentam em sua composição motivos estruturais comuns ou fragmentos de estruturas super-secundárias de proteínas cuja estrutura 3D é conhecida (TRAMONTANO, 2006). Com base nesta observação, métodos de primeiros princípios que utilizam informação da base experimental comparam fragmentos (sequências curtas) de uma sequência alvo contra fragmentos de sequências de proteínas com estrutura 3D determinada. Sendo chamados de métodos de fragmentos.

Métodos de fragmentos utilizam estruturas locais para predizer a estrutura 3D da proteína alvo, entretanto a relação entre sequências de aminoácidos locais e estruturas locais é degenerativa, ou seja, sequências locais similares podem apresentar estruturas locais diferentes (KOSÍNSKI et al., 2009). Isto ocorre porque estruturas 3D apresentam inúmeras interações físico-químicas. Desta forma, a sequência de aminoácidos alvo não pode ser fragmentada randomicamente, sendo necessária a aplicação de critérios, como funções de escore (DORN et al., 2014). Funções de ranqueamento são derivadas de estatísticas conformacionais de estruturas cuja estrutura 3D é conhecida e armazenada em banco de dados, podendo ser utilizada informações adicionais, como dados de estrutura secundária.

Métodos de primeiros princípios que utilizam informação da base experimental ocorrem em 5 etapas (Figura 3.2) (DORN et al., 2014):

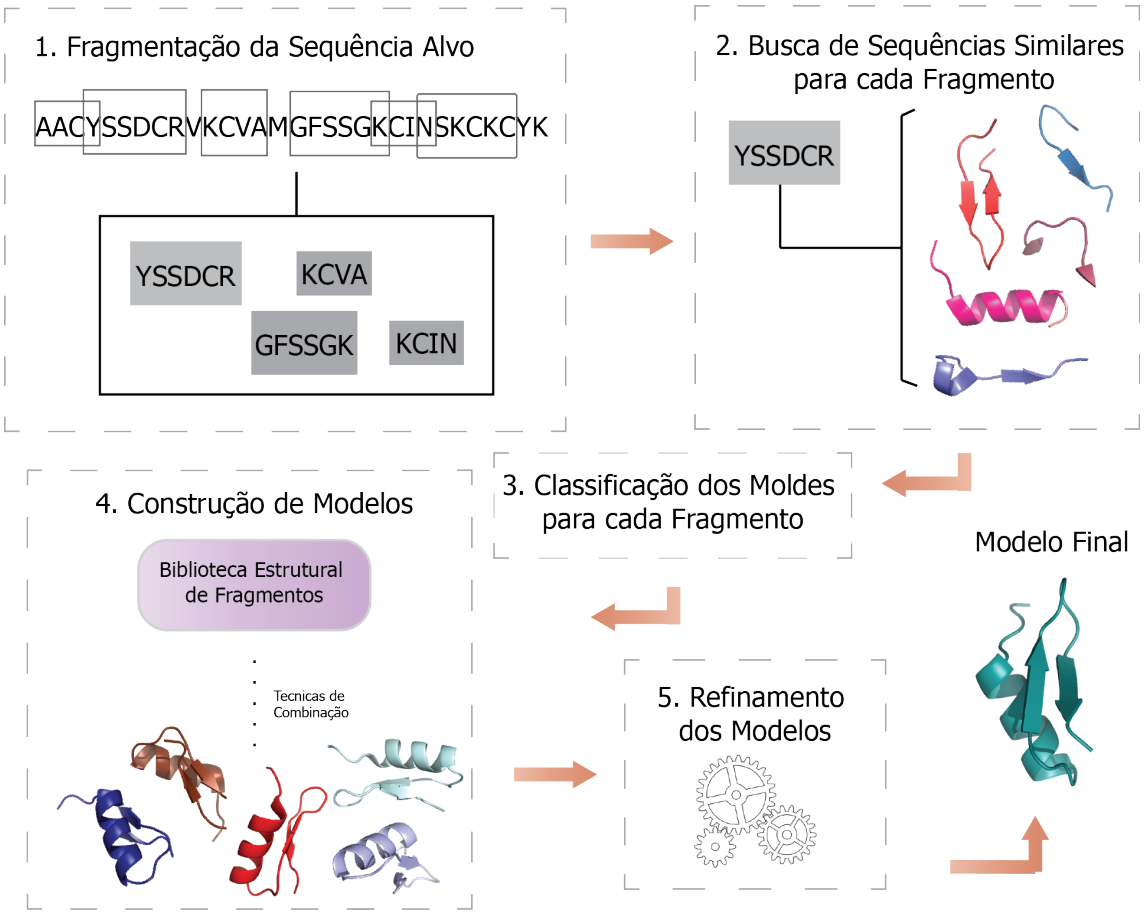
1) Fragmentação da sequência alvo;
2) Busca de sequências similares cuja estrutura 3D é conhecida para cada fragmento;
3) Classificação dos fragmentos;
4) Construção do modelo através de técnicas de combinação;
5) Refinamento conformacional.

A combinação dos fragmentos é feita visando encontrar o modelo conformacional de menor energia potencial, de maneira similar aos métodos *ab initio*, entretanto, com redução do espaço de busca quando comparados. Geralmente o processo de combinação de fragmentos é feito por metaheurísticas (DORN et al., 2014). Além disso, assim como os métodos *ab initio*, métodos de fragmentos são capazes de gerar novas conformações. Contudo, são limitados pelo enorme espaço de busca conformacional devido às inúmeras possibilidades de combinações entre os fragmentos, sendo computacionalmente custosos (DORN et al., 2014). Exemplos de métodos de fragmentos são: I-TASSER (ZHANG, 2008), ROSETTA (ROHL et al., 2004) e FRAGFOLD (JONES, 2001).

3.4 Métodos de *Fold Recognition* e *Threading*

Métodos de *Fold Recognition* (FR) e *Threading* são embasados no fato que estruturas são evolutivamente mais conservadas do que sequências, isto é, proteínas não relacionadas evolutivamente (ausência de sequências similares) podem apresentar padrões

Figura 3.2: Fluxograma Método de Fragmentos



Fonte: adaptado de Dorn et al. (2014)

conformacionais similares (DORN et al., 2014). Proteínas, cuja estrutura 3D é conhecida, são classificadas entre poucas distintas conformações (*folds*), além disso, o número de conformações novas descobertas cresce de maneira muito lenta, indicando que esta quantidade é finita (MARSDEN et al., 2006).

O objetivo desses métodos é encontrar o encaixe correto entre a sequência de aminoácidos da proteína-alvo (desconhecida) e o modelo estrutural conhecido. Pode se afirmar que são problemas de classificação, os quais envolvem dois processos: (a) selecionar um modelo apropriado e (b) encontrar o melhor alinhamento entre sequência-alvo e modelo estrutural conhecido.

Métodos FR e *Threading* podem ser classificados em dois grupos (RUSSELL; COPLEY; BARTON, 1996): (i) baseados em perfis e (ii) baseados em pares de potenciais. Basicamente, métodos do grupo i armazenam as informações dos modelos de estruturas conhecidas de maneira linear (matrizes, vetores), enquanto que métodos do grupo ii utilizam pares de potenciais que medem a probabilidade de contato resíduo-resíduo entre

modelo e alvo.

O processo de *Threading* e *Fold Recognition* ocorre em 3 etapas (DORN et al., 2014) (Figura 3.3):

1) Construção da biblioteca de padrões conformacionais a partir de um banco de dados (PDB);
2) Avaliação do alinhamento "modelo estrutural x sequência proteína-alvo";
3) Identificação de possíveis padrões de substituições entre sequência-estrutura que melhorem o alinhamento global.

Para avaliação do alinhamento são utilizadas referências que pontuam as interações entre resíduos da sequência-alvo e do modelo estrutural e a soma das interações resulta na pontuação global, ou seja, o quão favorável é determinado alinhamento. Em suma, busca-se o alinhamento que irá minimizar o potencial energético. A fim de diminuir o espaço de busca entre as possíveis substituições, os métodos utilizam informações estruturais como padrões de contato resíduo-resíduo, estruturas secundárias e acessibilidade ao solvente (DORN et al., 2014; KOSÍNSKI et al., 2009).

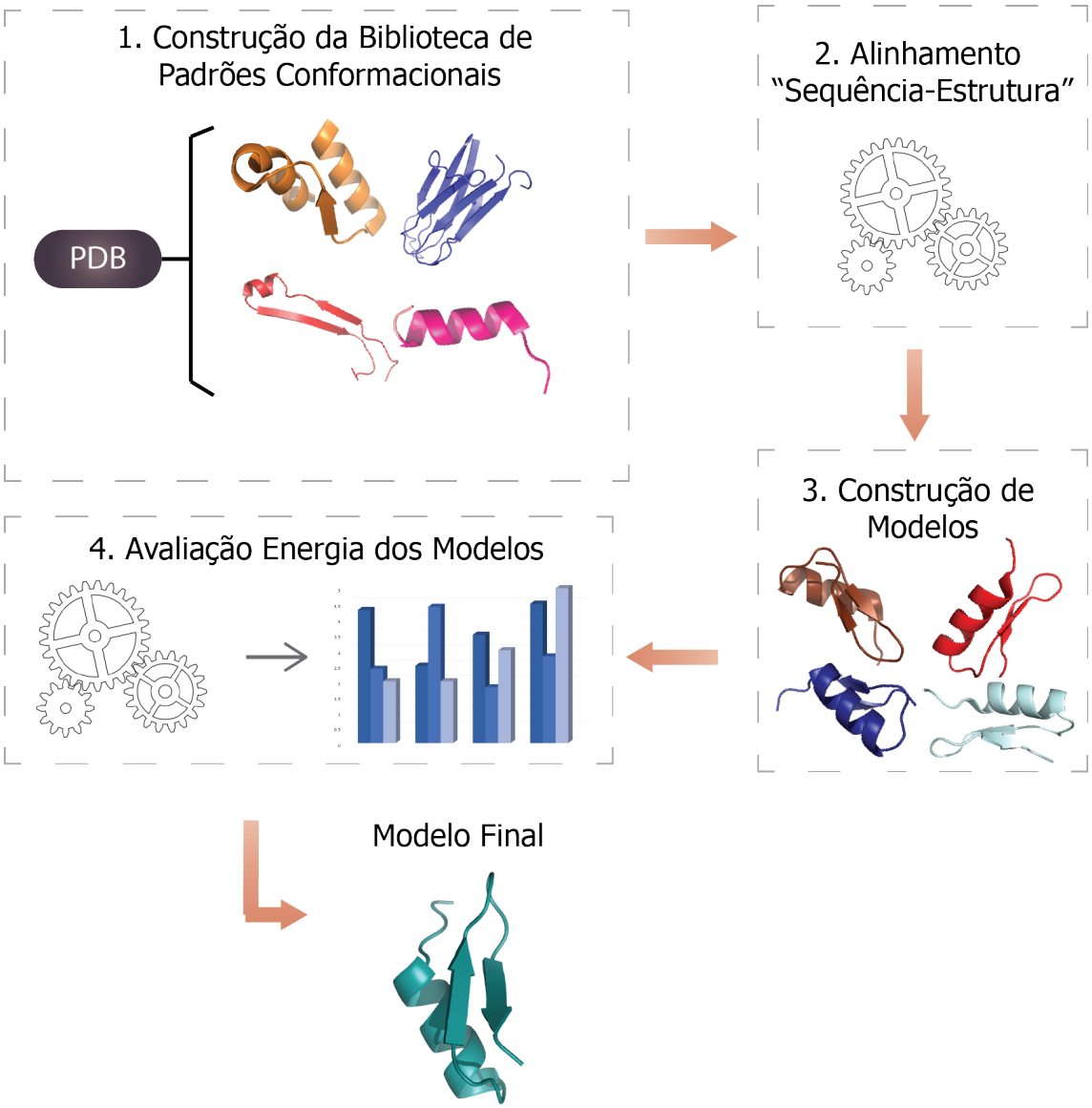
Exemplos de métodos FR e *Threading* são: GENTHREADER (JONES, 1999), Bioshell-Threading (GNIEWEK et al., 2014), RaptorX server (KALLBERG et al., 2014) e Phyre server (KELLEY; STERNBERG, 2009). Entretanto, os melhores resultados têm se mostrados para métodos *meta-servers*, os quais combinam diferentes áreas (DORN et al., 2014).

3.5 Métodos de Modelagem Comparativa e Alinhamento de sequências

Proteínas relacionadas evolutivamente (sequências de aminoácidos similares) tendem em adotar estruturas similares, isso porque pequenas alterações na sequência de aminoácidos geralmente resultam em pequenas mudanças na estrutura 3D. Além disso, estruturas de proteínas são evolutivamente mais conservadas e estáveis, havendo perpetuação de alterações de modo muito mais lento do que em sequências de aminoácidos (CHOTHIA; LESK, 1986).

Métodos de Modelagem Comparativa (MC) buscam construir um modelo estrutural de resolução atômica para sequências de aminoácidos (proteína-alvo) a partir da estrutura 3D conhecida de proteínas relacionadas evolutivamente a sequência-alvo.

Figura 3.3: Fluxograma Método *Fold Recognition* e *Threading*



Fonte: adaptado de Dorn et al. (2014)

A precisão de métodos MC depende do número de estruturas conhecidas para proteínas cujas sequências são similares, o que acaba sendo uma limitação, visto que o número de proteínas com estrutura 3D determinada é pequeno (DORN et al., 2014; FISER, 2009). Entretanto, o projeto *Protein Structure Initiative* (PSI) tem o objetivo de determinar pelo menos uma estrutura para cada família de proteína (BURLEY et al., 2008); e viu-se que aproximadamente 70% de todas as sequências conhecidas apresentam pelo menos um domínio relacionado a proteínas cuja estrutura 3D é conhecida (PIEPER et al., 2006). Desta forma, a aplicabilidade de métodos de modelagem comparativa cresce com o aumento do número de estruturas determinadas.

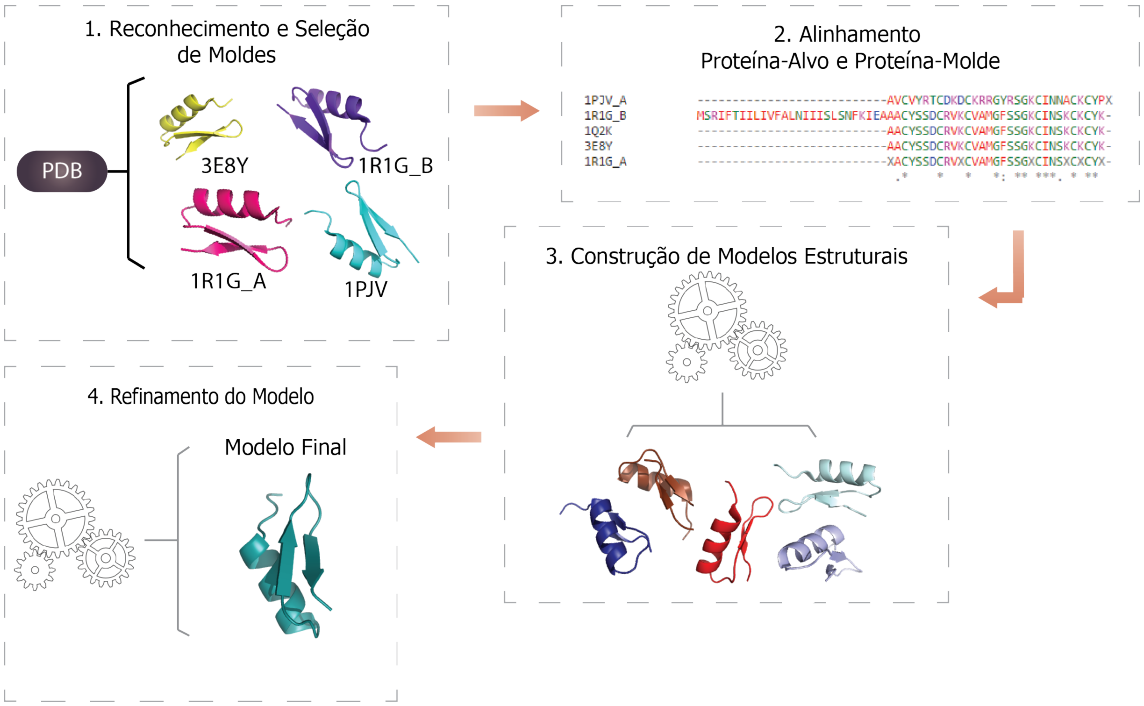
A qualidade de estruturas geradas por modelagem comparativa é dependente da

qualidade do alinhamento entre a sequência alvo e molde (estrutura 3D conhecida). Existem duas classes de métodos de alinhamento (DORN et al., 2014; FISER, 2009): (i) comparação par a par e (ii) comparação múltipla. BLAST (ALTSCHUL, 1990) e FASTA*server* (PEARSON; LIPMAN, 1988) são exemplos do grupo i, em que comparam a sequência alvo contra cada modelo do banco de dados de maneira independente. CLUSTAL-W (THOMPSON; HIGGINS; GIBSON, 1994), PSI-BLAST (ALTSCHUL, 1997) e T-COFFEE (NOTREDAME; HIGGINS; HERINGA, 2000) são exemplos do grupo ii, em que se compara a sequência alvo contra múltiplas sequências, incorporando buscas por informações evolutivas, como regiões conservadas e motivos.

Processos de modelagem comparativa contêm 4 etapas básicas (Figura 3.4):

1) Reconhecimento e seleção de moldes;
2) Alinhamento "proteína-alvo e proteína-molde";
3) Construção do modelo estrutural;
4) Refinamento do modelo estrutural.

Figura 3.4: Fluxograma Método de Modelagem Comparativa



Fonte: adaptado de Dorn et al. (2014)

Na etapa 1 se busca, em banco de dados, estruturas de proteínas que sejam relacionadas à sequência alvo; para a seleção do modelo mais adequado são utilizados critérios de similaridade entre sequências ou de relação evolutiva. A partir do alinhamento entre sequência alvo e molde (etapa 2) o modelo estrutural é construído (etapa 3). A cadeia

principal em regiões homólogas é construída a partir da estrutura molde podendo ser feito através de 3 métodos: junção de corpos rígidos; reconstrução de coordenadas; e satisfação de restrições espaciais. Em seguida são traçadas as regiões não homólogas, regiões de voltas e, por fim, cadeias laterais. Em suma, para gerar modelos estruturais adequados por métodos de modelagem comparativa são necessárias duas condições (FISER, 2009): a) detectar similaridades entre sequência alvo e possíveis sequências moldes e b) alinhamento adequado entre as sequências.

Métodos de modelagem comparativa são capazes de gerar modelos estruturais de alta acurácia quando há alta similaridade entre sequência alvo e molde, além disso, a qualidade dos modelos gerados pode ser estimada a priori. Entretanto, não é possível gerar novas conformações nem estudar o processo de enovelamento proteico a partir dos mesmos (DORN et al., 2014). Exemplos de métodos de modelagem comparativa são: SWISS-MODEL (ARNOLD et al., 2006), MODELLER (SALI; BLUNDELL, 1993) e PyMOD (JANSON et al., 2016).

3.6 CASP

Os experimentos CASP foram propostos em 1994 por John Moult com o objetivo de avaliar o "estado da arte" em predição de estruturas 3D de proteínas. Desta maneira, identificando os progressos e direcionando os futuros esforços para a área.

Os experimentos são testes cegos para métodos de predição de estruturas de proteínas. Proteínas cuja estrutura 3D foi determinada experimentalmente, pelas técnicas de cristalografia por raios-X ou RMN, mas cujos dados não foram divulgados, têm sua sequência de aminoácidos disponibilizada para membros da comunidade científica de predição. Os modelos produzidos computacionalmente para essas proteínas são depositados e analisados por avaliadores do CASP (conduzidos por John Moult e Krzystof Fidelis) a fim de gerar conclusões gerais sobre o estado em que os métodos se encontram.

Atualmente, os métodos são classificados em duas categorias: a) *Template-Based Modelling* e b) *Free-Modelling* (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014). Métodos classificados em a utilizam informações de estruturas determinadas experimentalmente, as quais a proteína alvo está relacionada. Métodos classificados em b não utilizam conhecimento experimental, visto que não há relações utéis ou detectáveis para a proteína alvo.

Inicialmente, haviam poucas proteínas com estrutura 3D determinada e armazenada em banco de dados, assim os modelos eram praticamente gerados por métodos *Free-Modelling*. O aumento na quantidade de dados experimentais disponíveis permitiu o crescimento de métodos de predição que utilizassem conhecimento (*Template-Based Modelling*). Ao longo dos experimentos, métodos das 4 classes mostraram progressos, sendo os resultados mais importantes provindos de métodos híbridos, ou seja, aqueles baseados em primeiros princípios, mas que utilizam, de alguma maneira, dados experimentais (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014).

3.7 Resumo do Capítulo

Conhecer a estrutura 3D de proteínas permite um estudo mais direto e profundo de sua função biológica. Atualmente, métodos experimentais são as maiores fontes de dados estruturais, entretanto são técnicas de alto custo e que levam demasiado tempo. Assim, existem poucas proteínas com sua estrutura 3D determinada; além de que há uma discrepância entre o número de sequências proteicas conhecidas e estrutura 3D.

A utilização de métodos computacionais para predição de estruturas 3D de proteínas surgiu como uma alternativa. Inúmeras estratégias e algoritmos foram desenvolvidos e aplicados no problema de predição, os quais podem ser classificados em 4 grupos: (i) métodos de primeiros princípios sem informação experimental (ii) métodos de primeiros princípios com informação da base experimental (iii) métodos de *Fold Recognition* e *Threading* (iv) métodos de modelagem comparativa e alinhamento de sequências. A eficiência dos métodos é avaliada através dos experimentos CASP, sendo apresentados importantes resultados para métodos classificados no grupo ii.

4 METAHEURÍSTICAS

A predição de estruturas 3D de proteínas é um dos maiores problemas em biologia computacional. A utilização de métodos computacionais de primeiros princípios (baseados em leis físicas) fica limitada ao enorme espaço de busca gerado pelas diversas conformações possíveis que uma proteína pode adotar. Métodos computacionais que utilizam informações de banco de dados experimentais ficam restritos a qualidade e/ ou existência de conhecimento. Em suma, predizer a estrutura 3D de proteínas é um problema de otimização classificado computacionalmente como NP-Completo (CRESCENZI et al., 1998).

Como visto anteriormente (Seção 2.3.1), proteínas adotam sua estrutura nativa quando em condições fisiológicas; entretanto, proteínas são altamente dinâmicas e, assim, capazes de modificar sua conformação na presença de moléculas/ligantes ou em resposta a alterações do meio (TEILUM; OLSEN; KRAGELUND, 2009). Isto é, proteínas são flexíveis, logo, determinar uma única estrutura 3D não tem fundamento, além de ser computacionalmente inviável e custoso. Assim, utilizar algoritmos aproximativos é uma alternativa capaz de gerar soluções satisfatórias. Metaheurísticas são algoritmos aproximativos, os quais se mostram bem-sucedidos em problemas de otimização (BOUSSAÏD; LEPAGNOT; SIARRY, 2013; TABLI, 2009).

4.1 Metaheurísticas e Problemas de Otimização

Problemas que não podem ser solucionados de maneira ótima por métodos exatos (determinísticos) em um intervalo de tempo razoável são ditos problemas de otimização complexos. Esses problemas são classificados como NP-Completo, os quais requerem um intervalo de tempo exponencial e só permitem soluções aproximativas. Métodos aproximativos são utilizados com a intensão de gerar soluções de qualidade em tempo razoável, entretanto sem garantia de encontrar a solução ótima global (BOUSSAÏD; LEPAGNOT; SIARRY, 2013; TABLI, 2009).

Metaheurísticas constituem uma família de algoritmos que resolvem de forma aproximativa problemas de otimização difíceis, gerando soluções satisfatórias em tempo razoável e praticável. Entretanto não garantem o ótimo global nem soluções limites. A palavra "*heurística*" significa descobrir novas estratégias para solucionar problemas e o sufixo "*meta*" significa metodologias de alto nível. De fato, metaheurísticas encontram

soluções aceitáveis em problemas de grande porte de maneira generalizada (BOUSSAÏD; LEPAGNOT; SIARRY, 2013; TABLI, 2009).

Praticamente todas metaheurísticas são estratégicas baseadas na natureza (princípios da física, biologia e etiologia), que utilizam componentes estocásticos (regras e variáveis randômicas) e apresentam parâmetros ajustados ao problema. Ao desenvolver uma metaheurística dois critérios devem ser considerados: diversificação e intensificação. Em diversificação há maior exploração do espaço de busca, não ficando restrito a pequenas áreas. Em intensificação, regiões com soluções de maior qualidade são identificadas e exploradas de maneira a encontrar as melhores soluções (BOUSSAÏD; LEPAGNOT; SIARRY, 2013; TABLI, 2009).

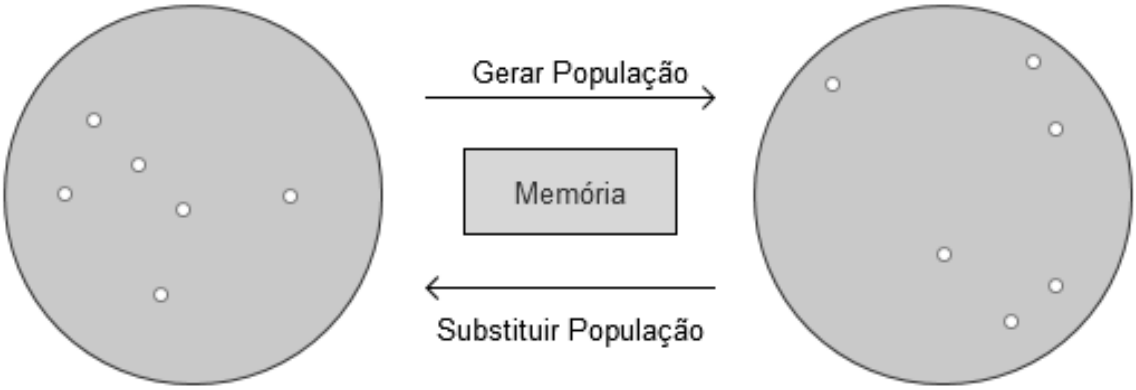
Metaheurísticas podem ser classificadas a partir de diversos critérios. Baseando-se em critério de busca são classificadas em baseadas em População (*Population-based search*) e baseadas em Solução Única (*Single-Solution based search*). Algoritmos baseados em Solução Única, chamados métodos de trajetória, manipulam e descrevem uma solução única inicial durante o processo de busca, logo são estratégias orientadas por intensificação. Algoritmos baseados em População manipulam um conjunto de soluções ao longo do processo, logo, são estratégias orientadas por diversificação (BOUSSAÏD; LEPAGNOT; SIARRY, 2013; TABLI, 2009).

4.2 Métodos Baseados em População

Metaheurísticas baseadas em população são algoritmos que iterativamente melhoram um conjunto (população) de soluções (BOUSSAÏD; LEPAGNOT; SIARRY, 2013). Basicamente, começam com uma população de possíveis soluções, repetidamente é gerado um novo conjunto de soluções e a população atual evolui conforme critérios de seleção (Figura 4.1). O processo é feito até que seja satisfeita a condição de parada.

Métodos baseados em população comumente usados são Algoritmos Evolutivos (EA) e *Swarm Intelligence* (SI) (BOUSSAÏD; LEPAGNOT; SIARRY, 2013; TABLI, 2009). EA são inspirados nos princípios evolutivos darwinianos, simulando a evolução de espécies. Algoritmos de SI simulam interações sociais segundo o comportamento coletivo entre colônias de insetos e outros animais (BOUSSAÏD; LEPAGNOT; SIARRY, 2013).

Figura 4.1: Princípios de Metaheurísticas baseadas em População

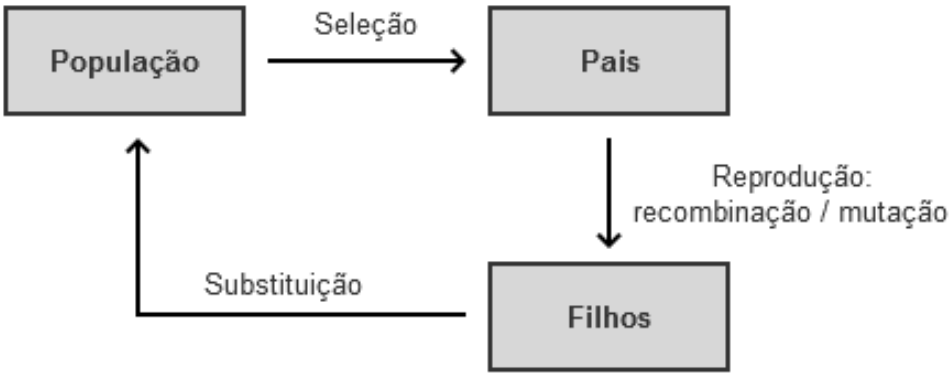


Fonte: adaptado de Tabli (2009)

4.2.1 Algoritmos Evolutivos

EA são baseados na competição entre espécies e inspirados na teoria evolutiva de Darwin, em que os indivíduos mais adaptados ao ambiente são mais capazes de sobreviver e gerar mais sobreviventes. São algoritmos que simulam a evolução de indivíduos através dos processos de seleção, reprodução (operadores de recombinação e mutação) e substituição buscando produzir a melhor solução (Figura 4.2).

Figura 4.2: Princípios de Algoritmos Evolutivos



Fonte: adaptado de Tabli (2009)

O Algoritmo 1 apresenta o pseudocódigo genérico para Algoritmos Evolutivos, em que a população inicial de soluções é criada de maneira randômica, a cada iteração indivíduos são selecionados para gerar novos candidatos conforme o paradigma: "*quanto melhor o indivíduo, maiores as chances de ser selecionado*". Novas soluções são criadas aplicando operadores de reprodução; mutação, em que ocorrem pequenas mudanças em um indivíduo, e recombinação (*crossover*), em que características de 2 ou mais indivíduos são combinadas. Os novos indivíduos são avaliados calculando o valor de *fitness* (função

objetiva) e, então, são determinados quais irão ser mantidos ou descartados (tamanho populacional constante). A função objetiva representa a habilidade do indivíduo sobreviver ao ambiente. O processo é repetido até a condição de parada ser satisfeita, a qual pode ser o número de gerações ou o número de avaliações da função objetiva (TABLI, 2009).

Algoritmo 1 Pseudocódigo Algoritmos Evolutivos

Criação População Inicial (P_0)
Avaliação cada indivíduo
enquanto *não atingir critério de parada* **faça**
    Seleção indivíduos-pais
    Mutação e *Crossover*
    Avaliação novos indivíduos
    Substituição indivíduos em P_{atual}
fim
Saída: Melhor indivíduo ou melhor população

4.3 *Differential Evolution*

Differential evolution é uma poderosa técnica para resolver problemas de otimização contínuos, sendo bem-sucedida em inúmeras aplicações. Desenvolvido por STORN and PRICE (1995) , o método apresenta uma estratégia diferenciada para gerar novos vetores (indivíduos), em que utiliza diferentes vetores para realizar perturbações nos vetores da população atual.

O Algortimo 2 a seguir apresenta o pseudocódigo de um DE genérico. A cada geração G, o método evolui uma população de NP (tamanho da população) vetores $X_{i,G}$ de dimensão D, chamados indivíduos. O valor de NP se mantem constante durante o processo de busca. A população inicial (P_G) é gerada de maneira randômica quando não se tem informações do sistema e deve cobrir o máximo do espaço de busca possível. Para gerar novos indivíduos, é adicionada a diferença ponderada entre dois vetores da população a um terceiro vetor (mutação); o novo vetor $U_{i,G}$ (*mutant vector*) é combinado com outro vetor (*target vector*) da população (*crossover*). Para o vetor resultante (*trial vector*) é calculado a função objetiva $f(U_{i,G})$ e o valor comparado com seu respectivo vetor na população corrente (*target vector*), se o valor for menor o indivíduo da população atual é substituído por este novo indivíduo gerado (seleção).

Algoritmo 2 Pseudocódigo DE

Criação da População Inicial $P_G = \{X_{1,G}, ..., X_{NP,G}\}$ randomicamente
em que $X_{i,G} = \{x^1_{i,G}, ..., x^D_{i,G}\}$
Avaliação cada indivíduo $F_{i,G} = \{f_{1,G}, ..., f_{NP,G}\}$
enquanto *não atingir critério de parada* **faça**
    Gerar um $U_{i,G} = \{u^1_{i,G}, ..., u^D_{i,G}\}$ para cada $X_{i,G}$
    Crossover
    Avaliação novos indivíduos
    se $f(U_{i,G}) \leq f(X_{i,G})$ **então**
        $X_{i,G+1} = U_{i,G}$ e $f(X_{i,G+1}) = f(U_{i,G})$
    fim
    Substituição indivíduos em P_{atual}
fim
Saída: Melhor indivíduo ou melhor População

A maior vantagem de algoritmos DE é que são constituídos de apenas 3 parâmetros que controlam o processo de busca: **NP**, tamanho da população; **F**, constante de diferenciação, a qual controla a amplitude da diferença entre os vetores; e **CR**, controle de *crossover*. Durante o processo de otimização os parâmetros são mantidos fixos, entretanto diferentes problemas requerem diferentes atribuições de valores.

Algoritmos DE possuem a conotação **DE/x/y/z**, onde x corresponde ao vetor que será perturbado (rand ou best), y corresponde ao número de vetores usado para a perturbação, e z corresponde a estratégia de *crossover* (bin ou exp). Variadas estratégias foram desenvolvidas para algoritmos DE, sendo DE/rand/1/bin sua estratégia básica.
(STORN; PRICE, 1995; STORN; PRICE, 1997)

4.4 *Self-Adapting Differential Evolution*

Em algoritmos de DE, existem diversas estratégias para gerar novos indivíduos que podem ser aplicadas, além disso, há 3 parâmetros cruciais envolvidos no processo de busca (STORN; PRICE, 1997). Essas estratégias e parâmetros são dependentes do problema, logo o sucesso do algoritmo está relacionado à escolha adequada dos mesmos. A escolha dos parâmetros e estratégias é feita através do processo de busca de tentativa

e erro, o qual é computacionalmente custoso. Além disso, durante diferentes etapas do processo evolutivo, a população pode transitar por diferentes regiões do espaço de busca, logo aplicar diferentes estratégias e parâmetros ao longo das etapas poderia ser mais eficiente (QIN; SUGANTHAN, 2005; QIN; HUANG; SUGANTHAN, 2009).

A utilização de algoritmos adaptativos evita altos custos computacionais. SADE é uma variação de algoritmos DE em que a estratégia de mutação e parâmetros são gradualmente adaptados ao longo do processo evolutivo conforme aprendizado prévio em que soluções bem-sucedidas foram geradas (QIN; SUGANTHAN, 2005; QIN; HUANG; SUGANTHAN, 2009).

Durante o processo evolutivo é mantida um *pool* de candidatos de estratégias, a escolha é de acordo a probabilidade calculada a partir da taxa de sucesso. Utilizar diferentes estratégias permite que distintas estratégias lidem com etapas específicas do problema. Os dados de sucesso e falha são armazenados e utilizados para atualizar a probabilidade de cada estratégia a cada geração. O parâmetro NP é mantido constante ao longo do processo, sendo este valor definido pelo usuário. O parâmetro F é randomicamente escolhido dentro de um intervalo de distribuição normal, com média igual 0,5 e desvio padrão igual 0,3. O parâmetro CR é gradualmente alterado ao longo do processo evolutivo devido sua maior sensibilidade ao erro; é randomicamente selecionado dentro de um intervalo de distribuição normal, com média inicialmente igual a 0,5 e desvio padrão igual 0,1. O valor do CR médio é atualizado a cada geração conforme prévios valores capazes de gerar soluções bem-sucedidas (QIN; HUANG; SUGANTHAN, 2009).

O Algoritmo 3 apresenta o pseudocódigo para um SADE genérico. Semelhante aos algoritmos DE, é criada uma população inicial com NP vetores $X_{i,G}$ de dimensão D, para os quais é calculado a função objetiva $f(X_{i,G})$. Durante o período de aprendizagem são testados os valores para os parâmetros F e CR e testadas as estratégias de mutação, para então cálculo de probabilidade $Prob_{mut}$. O processo evolutivo evolui a população até que seja atingido o critério de parada. A cada geração G, é gerado um novo vetor $U_{i,G}$ para seu respectivo vetor da população atual utilizando a estratégia de mutação selecionada conforme $Prob_{mut}$. A função objetiva $f(U_{i,G})$ é calculada e comparada com $f(X_{i,G})$, sendo selecionado para a população atual o vetor correspondente ao menor valor.

Algoritmo 3 Pseudocódigo SADE

Criação da População Inicial $P_G = \{X_{1,G}, ..., X_{NP,G}\}$ randomicamente em que $X_{i,G} = \{x^1_{i,G}, ..., x^D_{i,G}\}$

Avaliação cada indivíduo $F_{i,G} = \{f_{1,G}, ..., f_{NP,G}\}$

Estabelecer valores para CRmédio, estratégias de mutação (k) e periodo de aprendizagem (LP)

Periodo de Aprendizagem

Testar parâmetros F e CR

Testar estratégias de mutação

Calcular $Prob_{mut}$

Processo Evolutivo

enquanto *não atingir critério de parada* **faça**

    Selecionar estratégia de mutação

    $F = Normrnd(0.5, 0.3)$

    $CR = Normrnd(CRm, 0.1)$

    Gerar um $U_{i,G} = \{u^1_{i,G}, ..., u^D_{i,G}\}$ para cada $X_{i,G}$

    Avaliação novos indivíduos

    se $f(U_{i,G}) \leq f(X_{i,G})$ **então**

        $X_{i,G+1} = U_{i,G}$ e $f(X_{i,G+1}) = f(U_{i,G})$;

        Atualização $Prob_{muta}$

    fim

fim

Saída: Melhor indivíduo ou melhor população

4.5 Resumo do Capítulo

Devido ao enorme espaço de busca, a predição de estruturas 3D de proteínas é considerada um problema de otimização, em que encontrar soluções exatas em tempo prático é inviável. Metaheurísticas são algoritmos aproximativos capazes de encontrar soluções satisfatórias para problemas difíceis e complexos em um intervalo de tempo razoável. Utilizar metaheurísticas em estratégias para a predição estrutural de proteínas tem sido uma alternativa.

Metaheurísticas podem ser classificadas como baseadas em solução única (i) e ba-

seadas em população (ii). Métodos do grupo i manipulam uma solução única inicial e são orientados por intensificação, ou seja, exploram profundamente regiões promissoras buscando a melhor solução. Métodos do grupo ii evoluem um conjunto de soluções iniciais e são orientados por diversificação, em que exploram maior parte possível do espaço de busca.

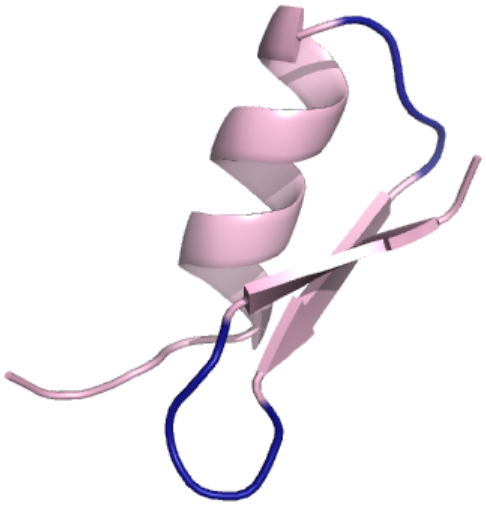
EA são uma classe de metaheurísticas baseadas em população. São algoritmos inspirados na teoria evolutiva de Darwin, simulando a evolução de espécies através dos processos de seleção, reprodução e recombinação. Os indivíduos melhor adaptados ao ambiente, no caso aqueles com melhores valores de *fitness*, são selecionados buscando a população de melhores soluções. Um exemplo de EA é DE, em que vetores são utilizados para representar os indivíduos de uma população. A maior vantagem do método é a existência de apenas 3 parâmetros envolvidos no processo de busca: NP, F e CR.

Em algoritmos DE, os valores atribuídos aos parâmetros e a escolha da estratégia para geração de novos indivíduos é diretamente dependente do problema. A escolha adequada dos mesmos é computacionalmente custosa devido ao processo de tentativa e erro, assim, a implementação de algoritmos adaptativos é uma estratégia utilizada. No algoritmo SADE os parâmetros e a estratégia aplicada são atualizadas ao longo do processo evolutivo conforme dados gerados do processo de aprendizagem.

5 PREDIÇÃO ESTRUTURAL DE VOLTAS E REGIÕES DESORDENADAS

Loops são fragmentos flexíveis de uma cadeia polipeptídica, os quais conectam dois elementos de estrutura secundária (hélices e folhas) em uma proteína, como representado na Figura 5.1 em azul escuro. Regiões de *loops* desempenham papeis importantes no processo de enovelamento e estabilidade proteica, além de, frequentemente, contribuírem para a especificidade funcional de proteínas. *Loops* são encontrados em sítios ativos e de ligação em proteínas, mediando interações entre antígenos e imunoglobulinas; toxinas e receptores; íons metálicos e proteínas; moléculas de DNA e proteínas-ligantes; e substratos proteicos e proteases (SHEHU; KAVRAKI, 2012). Diferenças funcionais em membros de uma mesma família de proteínas geralmente são consequências das variações estruturais entre *loops*.

Figura 5.1: Representação Estrutura de *Loops*.



Fonte: *Protein Data Bank* (PDB_ID: 1Q2K)

Constantemente, segmentos de *loops* são encontrados em superfícies proteicas e sujeitos a exposição a solventes. Logo, são mais suscetíveis a sofrerem inserções, deleções e substituições; e, assim, são ditas regiões variáveis. Consequentemente, regiões de *loops* são mais difíceis de serem modeladas e analisadas (SHEHU; KAVRAKI, 2012).

Muitos métodos de predição de regiões de *loops* consideram a estrutura da proteína rígida, ou seja, sua conformação é conhecida, sendo necessária predição apenas das regiões flexíveis. Em geral, os métodos podem ser classificadas em dois grupos (SHEHU; KAVRAKI, 2012; LI, 2013): (i) métodos *ab initio* e (ii) métodos que utilizam banco de

dados. Métodos do grupo i são fundamentados em explorar diferentes conformações no espaço de busca, guiado pela minimização de uma função de energia. São métodos capazes de modelarem conformações desconhecidas, entretanto são computacionalmente custosos (LI, 2013). Exemplos de métodos *ab initio* são: LEAP (LIANG; ZHANG; ZHOU, 2014), RCD+ (LÓPEZ-BLACO et al., 2016), ModLoop (FISER; SALI, 2003), FALC-Loop (KO et al., 2011), DiSGro (TANG; ZHANG; LIANG, 2014). Métodos do grupo ii buscam moldes estruturais em banco de dados, como PDB, que se encaixam no *loop*-alvo. São dependentes da qualidade e número de proteínas com estrutura determinada, e limitados a regiões de *loops* curtas. Exemplos de métodos que utilizam banco de dados são: ArchPRED (FERNANDEZ-FUENTES; OLIVA; FISER, 2006), LoopIng (MESSIH M.A.; TRAMONTANO, 2015), SuperLooper2 (ISMER et al., 2016).

5.1 Métodos que utilizam Banco de Dados

Métodos que utilizam banco de dados buscam modelos conformacionais de *loops* em proteínas que satisfaçam uma dada restrição. Estruturas são consideradas apropriadas e selecionadas quando satisfazem critérios de comprimento (número de aminoácidos) e geometria. São métodos rápidos, entretanto limitados pela disponibilidade de conhecimento e a *loops* curtos, visto que o número de possibilidades conformacionais cresce exponencialmente com o comprimento (SHEHU; KAVRAKI, 2012).

Exemplos de métodos que utilizam banco de dados e de bibliotecas de *loops* são descritos a seguir:

5.1.1 ArchDB

ArchDB é um banco de dados em que *loops* extraídos de proteínas, cuja estrutura 3D é conhecida, são classificados com base em sua geometria e estruturas secundárias limitadoras (BONET et al., 2014). As estruturas secundárias consideradas são: folhas-β (E), α-hélices (H) e 3_{10}-hélices (G). A geometria de um *loop* é composta por 4 variáveis internas: distância entre extremos e 3 ângulos. A classificação de um *loop* é definida pela sua estrutura secundária limite, comprimento e geometria.

ArchDB possui 2 classificações independentes baseadas na estratégia de cluste-

rização: *Density Search* (DS) e *Markov Clustering Algorithm* (MCL) (BONET et al., 2014).

5.1.2 ArchPRED

ArchPRED é uma ferramenta para predição de regiões de *loops* em proteínas baseado em um método de busca de fragmentos (FERNANDEZ-FUENTES; OLIVA; FISER, 2006). O método utiliza uma biblioteca de fragmentos multidimensional, chamada *Search Space*, a qual é organizada em 3 níveis hierárquicos: (i) tipo de estrutura secundária limitadoras: $\alpha\alpha$-*loops*, $\beta\alpha$-*loops*, $\alpha\beta$-*loops* e $\beta\beta$-*loops*; (ii) comprimento ou número de resíduos; e (iii) geometria de estrutura secundária limitadoras, definida pela distância e 3 ângulos internos.

O algoritmo é composto de 3 etapas: Seleção, Filtragem e *Ranking*. Os *loops* candidatos são selecionados do *Search Space* a partir da distância entre extremidades, geometria e posição inicial do *loop* na proteína alvo; logo, são selecionados aqueles com maiores chances de apresentarem configurações similares. Na etapa de busca, candidatos não favoráveis são descartados. Por fim, na etapa de Ranking, os candidatos são ordenados a partir do Z-Score, o qual combina pontuações de similaridade de sequência e probabilidades de ângulos diedros da cadeia principal (FERNANDEZ-FUENTES; OLIVA; FISER, 2006).

5.1.3 BriX

BriX é um banco de dados de fragmentos de proteínas contendo entre 4 e 14 aminoácidos de comprimento (VANHEE et al., 2011). A biblioteca é composta por 7290 fragmentos de proteínas provindos do Astral40 (CHANDONIA et al., 2004), os quais são clusterizados de maneira hierárquica pela similaridade da cadeia principal. Além da biblioteca de fragmentos, há o **Loop BriX** (VANHEE et al., 2011), um banco de dados composto de 14525 estruturas de *loops* provindo do Astral95. O Loop BriX é clusterizado conforme a similaridade entre *loops*, a qual é medida pela distância entre extremidades e pela sobreposição de aminoácidos das estruturas secundárias limitadoras.

5.1.4 LoopIng

LoopIng é uma ferramenta para selecionar modelos estruturais de *loops* para proteínas a partir de um banco de dados utilizando o modelo de regressão linear *Random Forest* (MESSIH M.A.; TRAMONTANO, 2015). A partir de informações de sequência e geometria (sequência do *loop*, similaridade de sequências, distância entre extremidades, estrutura secundária das extremidades, geometria da estrutura secundária limitadora) o modelo é treinado para selecionar os moldes de *loops* que apresentam os menores valores de distância entre *loops* predito e proteína alvo. Por fim as estruturas de *loops* candidatas são ordenadas conforme o nível de confiança de predição, podendo ser selecionado pelo usuário o número de modelos desejados.

5.1.5 SuperLooper 2

SuperLooper2 é uma ferramenta para a predição, visualização e seleção de segmentos de *loops* em moléculas de proteínas (ISMER et al., 2016). Os possíveis *loops* são selecionados de um banco de dados contendo aproximadamente 700 milhões de fragmentos de proteínas extraídas do PDB. Todos os fragmentos contêm entre 3 e 35 aminoácidos de comprimento, e, para cada fragmento, são armazenadas as informações de sequência de aminoácido, PDB ID, localização na proteína e identidade geométrica. O número de fragmentos disponíveis diminui conforme aumenta o número de resíduos.

A identidade geométrica é definida por 4 variáveis (distância e 3 ângulos internos) e utilizada para avaliar a combinação entre átomos das terminações C- e N- de cada fragmento da base com átomos do *gap* (localização do *loop*) existente na proteína-alvo. Os 100 candidatos mais aptos são disponibilizados ao usuário (ISMER et al., 2016).

5.2 Biblioteca de Padrões Estruturais

Como apresentado em capítulo anterior, trabalhos que obtiveram resultados satisfatórios em edições do CASP utilizavam conhecimento de banco de dados (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014), o qual pode ser aplicado de variadas maneiras, como biblioteca de fragmentos. Apesar de regiões de *loops* não apresentarem padrões estruturais, a utilização de conhecimento em métodos de predição para estas regiões diminui

o espaço de busca conformacional, o qual, como visto, é enorme. ArchPRED e BriX são exemplos citados de bibliotecas de fragmentos aplicadas na predição de regiões de *loops*. Assim, nesta monografia é proposta uma Biblioteca de Padrões Estruturais (BPE) que irá apresentar informação de preferência conformacional de padrões estruturais encontrados em regiões de *loops*, ou seja, regiões unindo estruturas regulares. Esta biblioteca será utilizada como conhecimento em metaheurísticas aplicadas ao problema de predição de estruturas 3D de proteínas.

5.2.1 Base de Dados

Para a criação da BPE foram selecionadas estruturas de proteínas experimentalmente determinadas pela técnica de cristalização por difração de raios-X com resolução $\leq$ 2,5Å e armazenadas no PDB até dezembro de 2014. O valor de resolução indica o nível de detalhamento de determinação. Foram consideradas apenas estruturas com fator-R < 0,2 e ocupância igual a 1. O fator-R indica a qualidade do modelo atômico obtido, ou seja, a concordância entre modelo construído e dados experimentais. O valor de ocupância reflete a probabilidade do correto posicionamento espacial do átomo. No caso de proteínas homólogas foram aplicados filtros de similaridade, selecionando apenas aquelas com identidade máxima de 30$\%$. Desta forma, a base de dados é composta de 11.130 estruturas 3D de proteínas (BORGUESAN; INOSTROZA-PONTA; DORN, 2016). Outros trabalhos utilizaram parâmetros similares para filtrar a base de dados (HOVMÖLLER; ZHOU; OHLSON, 2002; BORGUESAN et al., 2015; BORGUESAN; INOSTROZA-PONTA; DORN, 2016)

5.2.2 Extração de Dados

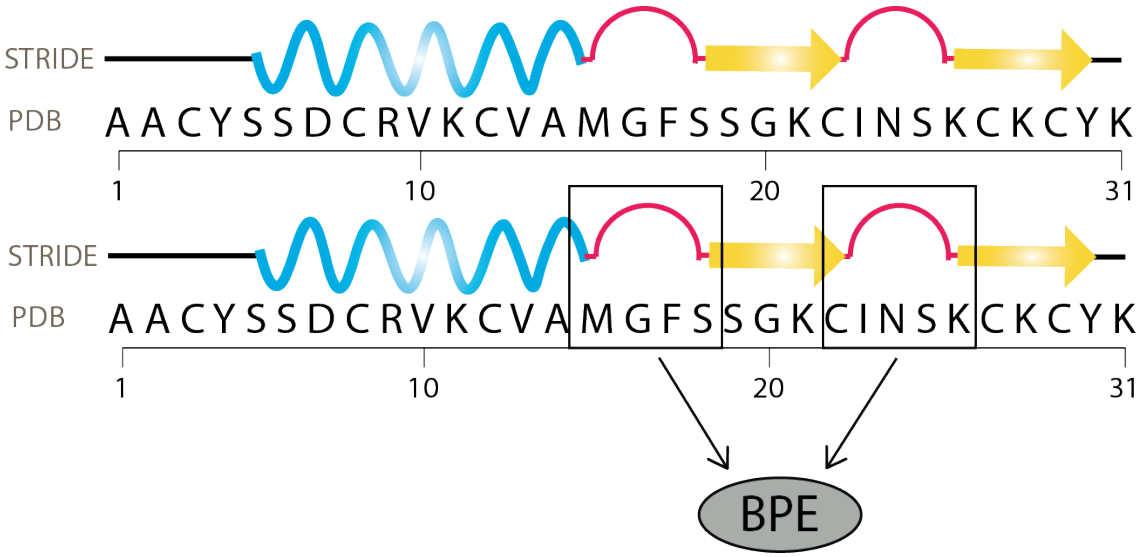
O conjunto final de estruturas representa um total de 5.255.768 aminoácidos (BORGUESAN; INOSTROZA-PONTA; DORN, 2016), para os quais foram extraídos os dados de ângulos diedros (*phi* e *psi*). A estrutura secundária foi atribuída a cada aminoácido através do software STRIDE (HEINIG; FRISHMAN, 2004), em que foram consideradas 6 entre 8 das diferentes estruturas: α-hélice (H); 3_{10}-hélice (G); folhas-β (E); ponte isolada (B); volta (T); alças ou região desordenada (C).

Assim como estudos internos do grupo *Structural Bioinformatics and Computati-*

onal Biology[1] (SBCB), os fragmentos foram gerados pela quebra de estrutura secundária. Durante o processo de fragmentação (Figura 5.2) a sequência de estruturas secundárias das 11.130 estruturas 3D que compunham a base de dados foram percorridas em busca de regiões de *loops*. Quando fragmentos compostos por estruturas de voltas e regiões desordenadas (rosa) unindo estruturas secundárias regulares, como hélices (azul) e folhas (amarelo), eram encontrados havia a quebra da estrutura. Ao final foram encontrados um total de aproximadamente 5000 padrões estruturais para regiões de *loops*.

Para cada padrão estrutural encontrado foi gerado um conjunto de combinações de valores de ângulos diedros possíveis a partir da informação conformacional das ocorrências. A partir do número de ocorrências na base de dados, os padrões foram ranqueados e selecionados para compor a BPE. Assim, a BPE é uma biblioteca de padrões estruturais de regiões de *loops*, os quais aparecem frequentemente em estruturas do PDB (Tabela 5.1).

Figura 5.2: Processo de Fragmentação



Fonte: do autor (2016)

5.3 Algoritmo Proposto

Conforme capítulo anterior, metaheurísticas são utilizadas em grande número de problemas de Bioinformática Estrutural, sendo a predição de estrutura 3D de proteínas um exemplo. A classe de EA é a mais comumente aplicada. Desta forma, é proposta a implementação do SADE, uma variação de algoritmo evolutivo, utilizando a base de

[1] http://sbcb.inf.ufrgs.br/home

Tabela 5.1: Padrões da BPE

Padrão	N° Ocorrências
ECCCE	3002
ECCCH	2631
ETTE	15284
ETTTE	6518
ETTTTE	5655
GCG	743
HCCCE	5686
HCCCH	2966
HCCCCH	1221
HCCCCCH	754
HCCE	5343
HCCH	3114
HTTTE	1606
HTTTH	811

Fonte: do autor (2016)

dados BPE como conhecimento a fim de diminuir o espaço de busca do problema.

Visto que o algoritmo proposto é aplicado ao problema de predição de estrutura 3D de proteínas, os dados de entrada são a sequência linear de aminoácidos e a sequência de estrutura secundária. Os indivíduos, possíveis conformações, são representados na forma de ângulos diedros combinados com a função de aptidão. Por apresentar bons resultados nos últimos CASP (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014), foi utilizada a plataforma PyRosetta (CHAUDHURY; LYSKOV; GRAY, 2010) para as rotinas da função objetiva.

O Algoritmo 4 apresenta o pseudocódigo do algoritmo SADE proposto. Para cada proteína-alvo são informadas a sequência de aminoácidos e de estrutura secundária como dados de entrada. Durante o Processo de Inicialização é criada a população inicial (Seção 5.3.2), além de serem estabelecidos os valores iniciais para os parâmetros F e $CR_{médio}$ e calculada a função objetiva (Seção 5.3.1) para cada indivíduo. Na etapa seguinte, Processo de Aprendizagem (Seção 5.3.3), são realizadas LP iterações a fim de armazenar os resultados obtidos para cada estratégia de mutação e valores de parâmetros aplicados. Resultados bem-sucedidos são computados na Memória de Sucesso e, consequentemente, resultados não bem-sucedidos na Memória de Falha. Ao final da etapa, os valores são utilizados para calcular a Probabilidade de cada estratégia de mutação. Durante o Processo Evolutivo (Seção 5.3.4) é gerado um novo vetor para cada vetor da população corrente utilizando a estratégia de mutação selecionada conforme os valores de Probabilidade e a

troca de padrões (descritas a seguir). A função objetiva (*fitness*) é calculada para cada vetor novo e, assim, comparada com o valor do respectivo vetor da população corrente. O vetor que apresentar o menor valor de *fitness* será selecionado para a próxima geração. O processo evolutivo ocorre repetidamente até atingir o critério de parada, sendo selecionado o melhor indivíduo da população final para dado de saída.

Algoritmo 4 Pseudocódigo SADE Proposto

Entrada: Sequência de AA e SS
Inicialização
Gerar população inicial utilizando base de dados;
Estabelecer valores iniciais dos parâmetros;
Avaliação indivíduos;
Processo de Aprendizagem
Atualização Memoria Sucesso e Falha;
Cálculo probabilidade;
Processo Evolutivo
enquanto *não atingir critério de parada* **faça**
    Selecionar estratégia de mutação k;
    Aplicar estratégia de mutação;
    Aplicar Troca de padrão;
    Calcular *fitness* vetores mutados;
    se *vetor mutado com fitness melhor* **então**
        Vetor mutado substitui vetor testado na população corrente
    fim
fim
Saída: Melhor indivíduo ou melhor População

5.3.1 Avaliação de Indivíduos

A avaliação do indivíduo refere-se a encontrar o quanto este é apto a resolver o problema em questão. No algoritmo proposto é utilizada a função de energia (Equação 5.1) implementada pelo PyRosetta como função de aptidão, sendo aplicado, como auxílio, o reforço da formação da estrutura secundária, os quais serão descritos a seguir. Desta forma o objetivo é minimizar o somatório E_N + E_{SS} a fim de encontrar o melhor possível indivíduo.

$$\textbf{Energia} = E_{PyRosetta} + E_{SS} \tag{5.1}$$

5.3.1.1 Função de Energia

A função de energia do PyRosetta utilizada foi a SCORE3 (ROHL et al., 2004), a qual é utilizada no quarto estágio do protocolo clássico *ab initio* do Rosetta, além de ser comumente aplicada e incluir os termos comuns de funções centroides. Funções de modo centroide consideram todos os átomos da cadeia principal, entretanto as cadeias laterais são representadas como um átomo único. A função score3 é composta por 10 termos e um *cutoff* para distância de sequência (*STRAND_STRAND_WEIGHTS*).

Baseado no efeito hidrofóbico, o termo *env* descreve o processo de solvatação para determinado resíduo. Para interações entre pares de resíduos (ligações eletrostáticas e pontes dissulfetos) é utilizado o termo *pair*. O termo *cbeta* é uma correção de solvatação. Repulsões estéricas entre pares de átomos são representadas pelo termo *vdw*, para atrações estéricas são utilizadas outros termos. O RMSD entre centroides de resíduos é indicado pelo termo *rg*. O empacotamento entre hélices e folhas é descrito pelo termo *hs_pair*. O termo *ss_pair* descreve ligações de Hidrogênio entre folhas-β. A probabilidade de um número de folhas-beta se arranjarem é descrita pelo termo *sheet* (Rohl et al. 2004). A distância entre folhas é calculada pelo termo *rsigma* (SHMYGELSK; LEVITT, 2008). O termo *cenpack* reforça o enovelamento ao estado nativo da proteína (ROHL et al., 2004).

5.3.1.2 Reforço da Formação da Estrutura Secundária

O reforço da formação da estrutura secundária auxilia o método a formar a mesma estrutura secundária que a passada como dado de entrada. O termo é baseado no *Framework model*, o qual considera que a a formação da estrutura secundária da proteína ocorre antes do processo de enovelamento. O PyRosetta implementa uma versão simplificada do DSSP (KABSCH; SANDER, 1983a), sendo identificadas apenas 3 estruturas secundárias: hélices, folhas e voltas. Assim, a formação correta de estrutura secundária reflete um reforço positivo ao valor de aptidão, enquanto que a formação incorreta acarreta um reforço negativo.

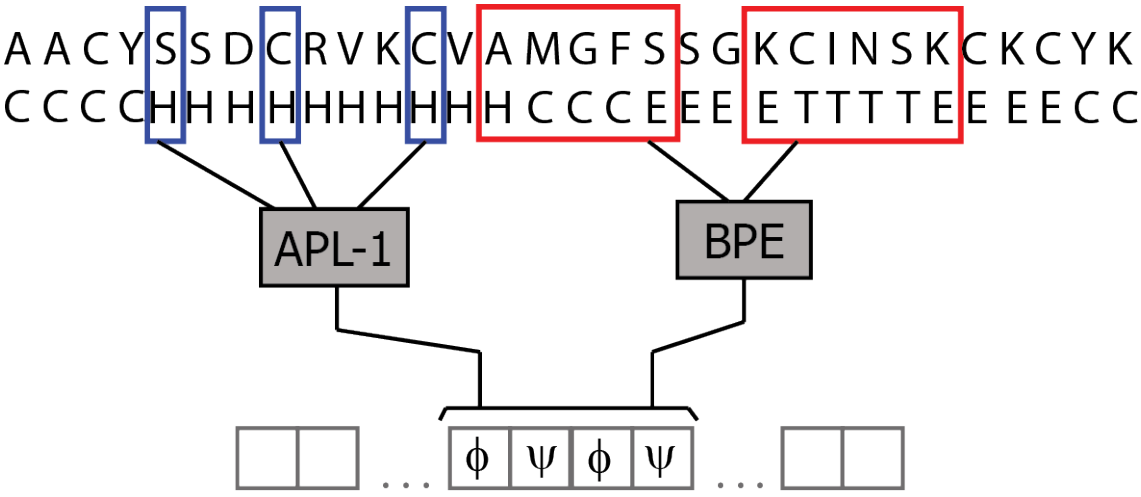
5.3.2 População Inicial

Um indivíduo da população é representado por um vetor de ângulos diedros, em que para cada aminoácido são computados dois valores (ϕ e ψ). A atribuição desses valores é feita utilizando base de dados: APL-1 e BPE, como mostra a Figura 5.3. APL-1

é uma base de dados que representa a frequência observada entre pares de aminoácidos e estrutura secundária existente no PDB, ou seja, os dados de preferência conformacional de cada aminoácido são combinados com a informação de estrutura secundária (BORGUESAN et al., 2015). Para cada aminoácido e estrutura secundária da proteína-alvo são selecionados valores de ângulos diedros da APL-1 de maneira randômica; porém, há maior probabilidade de selecionar valores que apresentam maior frequência, desta maneira é possível percorrer a base por inteiro.

Como visto na Figura 2.12, o Mapa de Ramachandran para estruturas de *loops* apresenta variadas regiões de valores de ângulos permitidos, assim encontrar a combinação correta é custoso. Contendo a informação conformacional para padrões estruturais de regiões de *loops*, a BPE é capaz de limitar o espaço de busca conformacional, ou seja, restringir as regiões para valores de ângulos diedros permitidas. Assim, para regiões da sequência de estrutura secundária da proteína-alvo que apresentam algum dos padrões existentes na BPE é selecionado randomicamente um conjunto de valores de ângulos, o qual é aplicado as respectivas posições no vetor-indivíduo.

Figura 5.3: Representação da geração do Vetor Indivíduo



Fonte: do autor (2016)

Visto que para cada aminoácido são atribuídos dois valores de ângulos diedros, um indivíduo será um vetor de dimensão igual a 2 vezes o tamanho da sequência de aminoácidos. O tamanho da população se mantem constante ao longo de todos processos, sendo igual a 10 vezes a dimensão do vetor. Para cada indivíduo da população é calculado seu valor de *fitness* pela função de aptidão.

5.3.3 Processo de Aprendizagem

Como visto anteriormente, algoritmos DE tem a vantagem de conter apenas 3 parâmetros (NP, F e CR) que controlam o processo de busca. Entretanto esses parâmetros são altamente dependentes do problema em questão, e sua escolha adequada é computacionalmente custosa devido ao processo de tentativa e erro. Algoritmos *self-adapting* evitam esse alto custo por gradualmente adaptarem parâmetros e estratégias ao longo do processo evolutivo conforme aprendizado prévio. Assim o processo de aprendizagem é utilizado para que o algoritmo SADE "aprenda" quais valores de parâmetros e estratégia de mutação geram soluções bem-sucedidas.

O processo de aprendizagem (LP) corresponde a 40 iterações (QIN; HUANG; SUGANTHAN, 2009) em que são testadas as diferentes estratégias de mutação e valores para os parâmetros F e CR, com seus resultados armazenados na Memória de Sucesso (n_s) e Falha (n_f). Como já dito, o parâmetro NP é mantido constante ao longo dos processos. Foram utilizadas 4 diferentes estratégias de mutação (QIN; HUANG; SUGANTHAN, 2009):

- DE/rand/1/bin:
$$u_{i,j} = \begin{cases} x_{r_1,j} + F \cdot (x_{r_2,j} - x_{r_3,j}), & \text{se } rand \leq CR \\ x_{i,j} \end{cases}$$
- DE/rand-to-best/2/bin:
$$u_{i,j} = \begin{cases} x_{i,j} + F \cdot (x_{best,j} - x_{i,j}) + F \cdot (x_{r_1,j} - x_{r_2,j}) + F \cdot (x_{r_3,j} - x_{r_4,j}) \\ x_{i,j} \end{cases}$$
- DE/rand/2/bin:
$$u_{i,j} = \begin{cases} x_{r_1,j} + F \cdot (x_{r_2,j} - x_{r_3,j}) + F \cdot (x_{r_4,j} - x_{r_5,j}), & \text{se } rand \leq CR \\ x_{i,j} \end{cases}$$
- DE/current-to-rand/1:
$$U_{i,G} = X_{i,G} + K \cdot (X_{r_1,G} - X_{i,G}) + F \cdot (X_{r_2,G} - X_{r_3,G})$$

onde $u_{i,j}$ corresponde ao novo vetor-indivíduo mutado, $x_{i,j}$ corresponde a vetores-indivíduo da população atual, F é a constante de diferenciação e G corresponde a geração corrente.

Para cada nova possível solução gerada é calculada sua função de aptidão. Estratégias que geram vetores bem-sucedidos, ou seja, em que há minimização da função energética são armazenadas em n_s; enquanto que o contrário é armazenado em n_f. Ao

final do período de aprendizagem, esses dados são utilizados para calcular a probabilidade de cada estratégia de mutação gerar vetores bem-sucedidos.

$$p_{k,G} = \frac{S_{k,G}}{\sum_1^k S_{k,G}}, \text{ onde } S_{k,G} = \frac{n_{s_{k,G}}}{n_{s_{k,G}} + n_{f_{k,G}}} * 0,01 \tag{5.2}$$

O parâmetro F é definido a cada iteração, em que é randomicamente selecionado num intervalo de distribuição normal cujo valor médio é 0,5 e o desvio padrão 0,3. Desta maneira F pode assumir valores entre [-0,4, 1,4]. Visto que o parâmetro CR é mais sensível ao problema, este é gradualmente ajustado ao longo do processo de busca. Inicialmente CR é randomicamente selecionado num intervalo de distribuição normal cujo valor médio é 0,5 e o desvio padrão 0,1; entretanto, o valor médio é alterado conforme são gerados resultados. Valores de CR que geram vetores bem-sucedidos são armazenados na *Memória*$_{CR}$, a cada iteração, um novo valor médio para CR é calculado e aplicado na distribuição normal (QIN; HUANG; SUGANTHAN, 2009).

5.3.4 Processo Evolutivo

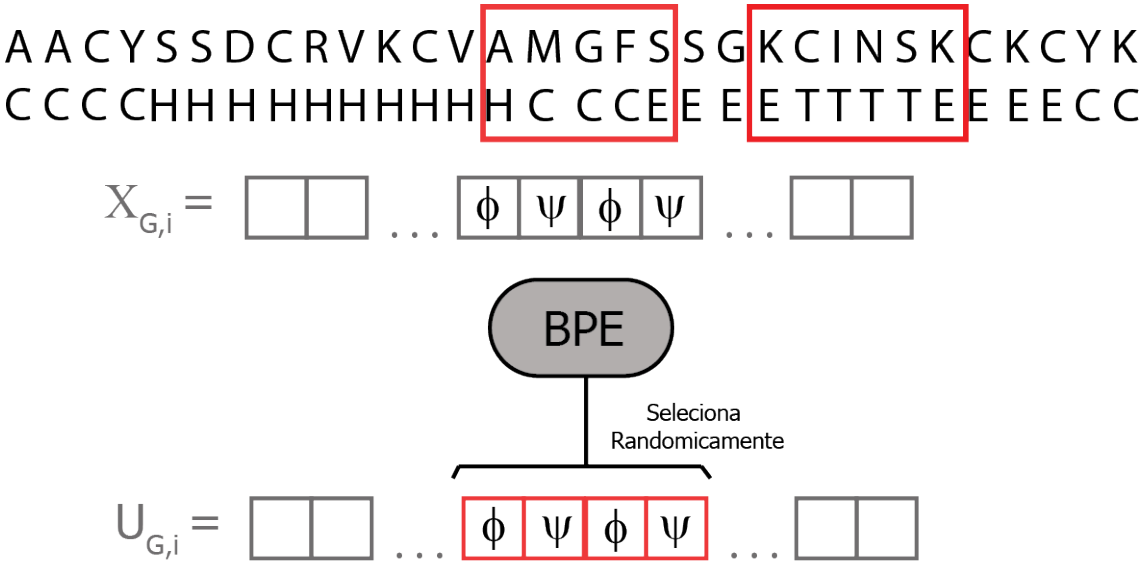
O processo evolutivo se refere ao momento em que o método aplica operadores de variação, como mutação e *crossover*, afim de gerar indivíduos melhores. No problema em questão, encontrar melhores indivíduos significa encontrar indivíduos cuja a função de energia foi minimizada, logo esses serão os indivíduos selecionados e mantidos na população corrente.

Para cada vetor indivíduo da população corrente é gerado um vetor teste, o qual contém as alterações. Para que não houvesse quebra dos padrões de voltas, optou-se por não realizar operações de mutação e *crossover* nas regiões em que há padrões estruturais da BPE. Entretanto, aplicou-se uma operação de troca de padrões nessas regiões. A fim de manter a diversidade, metade da população, um total de NP/2 indivíduos, sofreu as operações de mutação e *crossover* nos resíduos não pertencentes aos padrões estruturais.

- **Troca de Padrões**: para cada indivíduo é escolhido randomicamente o número de padrões e quais serão trocados. A operação consiste em escolher randomicamente um novo conjunto de valores de ângulos diedros na BPE, atribuindo esses novos valores ao vetor teste (Figura 5.4).
- **Mutação e *Crossover***: a cada iteração é escolhida uma estratégia de mutação conforme os dados de probabilidades, sendo a mesma aplicada e gerando o vetor teste.

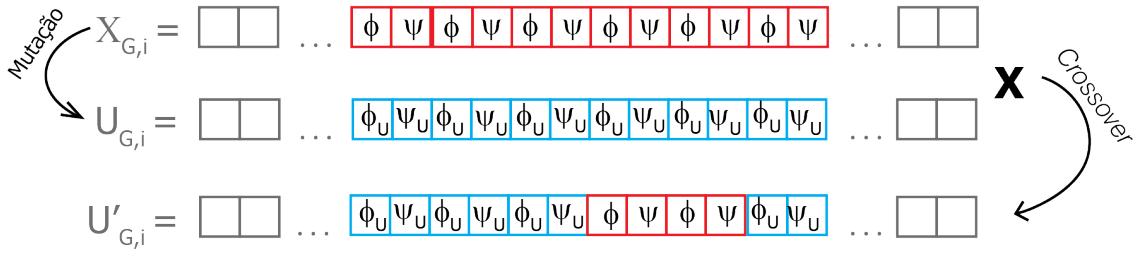
A operação de *crossover* ocorre sucessivamente, onde o vetor teste e vetor indivíduo são combinados (Figura 5.5).

Figura 5.4: Estratégia de Troca de Padrões



Fonte: do autor (2016)

Figura 5.5: Estratégia de Mutação e *Crossover*



Fonte: do autor (2016)

Para cada vetor teste gerado é calculada sua função de aptidão; vetores teste que apresentam minimização do seu valor energético passam a substituir os respectivos vetores indivíduos na população corrente. Todo esse processo evolutivo ocorre de maneira repetida até que seja atingido um critério de parada.

- **Critério de Parada**: corresponde ao momento que o algoritmo deve terminar. Condições de parada podem ser previamente estabelecidas a fim de manter um controle, como um número fixo de iterações, um limite nos recursos da CPU, ou um número máximo de avaliações da função objetiva (TABLI, 2009). No problema em questão, se estabeleceu o critério de 1.000.000 avaliações de função de aptidão.

Ao ser atingido o critério de parada, o melhor indivíduo (indivíduo de menor valor de função de energia) da população final é selecionado.

5.4 Resumo do Capítulo

Loops são regiões flexíveis encontradas ao longo da cadeia polipeptídica, as quais conectam estruturas secundárias, como hélices e folhas. São regiões mais suscetíveis a sofrerem alterações como inserções e deleções e por isso são ditas regiões variáveis. Além disso, *loops* são importantes na funcionalidade proteica, estão presentes em sítios ativos e de ligação de proteínas. Devido à sua flexibilidade e variabilidade, *loops* são regiões de difícil determinação.

A predição de *loops* pode ser feita por métodos *ab initio* ou métodos que utilizam conhecimento experimental, exemplos citados. Nos últimos experimentos CASP, os melhores resultados gerados utilizavam informação de banco de dados, a qual pode ser aplicada na forma de biblioteca de fragmentos. Assim, foi proposta uma nova biblioteca de fragmentos para padrões de regiões de *loops* (BPE) a ser utilizada como conhecimento em uma metaheurística aplicada ao problema de predição de estruturas 3D de proteínas.

Para a criação da BPE se fragmentou um conjunto de estruturas selecionadas do PDB. Buscaram-se fragmentos que fossem a união de regiões de estruturas secundárias regulares. Para os padrões encontrados, haviam armazenados os dados de estrutura secundária e os valores de ângulos diedros (*phi* e *psi*) de cada aminoácido.

Foi proposta a utilização da metaheurística **SADE** utilizando a **BPE** como informação para diminuir o espaço de busca. Os indivíduos da população são vetores contendo os valores de ângulos diedros para cada aminoácido, assim como sua função de aptidão. A função de aptidão mede o quão um indivíduo é apto a solucionar o problema, neste caso foi utilizada a função de energia do PyRosetta. De maneira iterativa, a metaheurística realiza um processo de evolução nos indivíduos da população buscando encontrar a melhor solução.

6 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados e discutidos os resultados gerados na validação da biblioteca de fragmentos desenvolvida (BPE) como conhecimento em uma metaheurística (SADE) aplicada no problema de predição da estrutura 3D de proteínas. São apresentados e descritos os critérios utilizados para avaliação dos resultados. Para validação da BPE foi feita a comparação dos resultados gerados pela implementação da metaheurística SADE proposta utilizando a BPE como conhecimento (SADE-BPE) com a implementação sem a BPE. Para a base de teste foram selecionadas 14 diferentes proteínas a partir de estudo realizado pelo grupo anteriormente BORGUESAN et al. (2015).

6.1 Critérios de Avaliação dos Resultados

Como visto em capítulo anterior, as possíveis soluções são representadas por vetores-indivíduos de ângulos diedros (*phi* ϕ e *psi* ψ) combinados com sua função de avaliação. A cada implementação, o melhor indivíduo da população final, o qual apresenta o menor valor para função de avaliação, é selecionado e seus valores de ângulos diedros são utilizados para gerar um arquivo de formato .pdb contendo a informação de coordenadas cartesianas da estrutura. Para avaliação dos resultados foram considerados os critérios: Função de Avaliação, RMSD e GDT_TS. Os valores de RMSD e GDT_TS foram calculados em relação a estrutura experimental da proteína armazenada no PDB.

6.1.1 Função de Avaliação

A função de avaliação foi apresentada e descrita em capítulo anterior (Seção 5.3.1), sendo o objetivo do método a minimização da função de energia (Equação 6.1), de maneira que o melhor indivíduo da população seja o que apresenta o menor valor.

$$\textbf{Energia} = E_{PyRosetta} + E_{SS} \tag{6.1}$$

6.1.2 RMSD

O *Root Mean Square Deviation* é utilizado para avaliar a similaridade entre duas estruturas 3D a partir do cálculo da distância entre átomos equivalentes (Equação 6.2) (ZHANG; SKOLNICK, 2004). Neste trabalho, foram selecionados apenas os átomos de Cα de cada estrutura para o cálculo, o qual foi implementado pelo PyRosetta (CHAUDHURY; LYSKOV; GRAY, 2010). O programa realiza uma série de rotinas; inicialmente é feito o alinhamento dos centros de massa das estruturas, seguido pela rotação de uma das estruturas até que seja encontrada a melhor superposição (KABSCH; SANDER, 1983b).

Para validação da BPE, as estruturas preditas pelo método SADE (ambas implementações) foram comparadas com as estruturas experimentais e armazenadas no PDB. Devido à alta flexibilidade das regiões terminais, os dois resíduos terminais foram desconsiderados para o cálculo.

$$RMSD = \sqrt{\left(\sum_{i=2}^{n-1} \left\| dp_{C\alpha i} - de_{C\alpha i} \right\|^2\right) / n}, \tag{6.2}$$

onde $dp_{C\alpha i}$ e $de_{C\alpha i}$ representam a posição (i) do C_α da estrutura predita (p) e experimental (e) ambas de tamanho n.

Valores de RMSD igual a 0 correspondem a estruturas idênticas, com aumento do valor conforme estruturas se diferenciam.

6.1.3 GDT_TS

O *Global Distance Total Score Test* de uma estrutura é uma porcentagem que indica o quanto uma estrutura predita convergiu para a estrutura nativa. Resíduos da estrutura alvo e modelo são utilizados para encontrar a superposição ótima em 4 diferentes distâncias (Equação 6.3) (ZHANG; SKOLNICK, 2004). É um método de avaliação constantemente utilizado nos experimentos CASP (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014).O cálculo de GDT_TS é executado por uma rotina disponibilizada pelo Zhang Lab (ZHANG; SKOLNICK, 2004).

$$GDT\_TS = \left(GDT_{P1} + GDT_{P2} + GDT_{P4} + GDT_{P8}\right) / 4 \tag{6.3}$$

onde P1, P2, P4 e P8 são as porcentagens do número de resíduos alinhados com distância menor que 1Å, 2Å, 4Åe 8Å, respectivamente.

Valores de GDT_TS igual a 100% correspondem a resíduos das estruturas predita e experimental alinhados com distância menor que 1Å. Conforme os resíduos são afastados, ocorrem descontos no valor de GDT_TS.

6.2 Base de Teste

Para validação da BPE foram selecionadas 14 diferentes proteínas que apresentam 1 ou mais dos padrões presentes na biblioteca. As proteínas foram escolhidas baseadas em trabalho anterior do grupo SBCB (BORGUESAN et al., 2015). Na Tabela 6.1 são apresentadas as diferentes estruturas.

Tabela 6.1: Proteínas da Base de Teste

PDB_ID	*Referência*	*Tamanho* (aa)	*Topologia*	*Padrões de Loops*
1AB1	Yamano et al. (1997)	46	2 hélices 1 folha	ECCCH, HCCCCH, HCCE
1ACW	Blanc et al. (1996)	29	1 hélice, 1 folha	ETTE, HTTTE
1CRN	Teeter (1984)	46	3 hélices, 1 folha	ECCCH, HCCCCH, HCCE
1D5Q	Vita et al. (1999)	27	1 hélice, 1 folha	ETTE, HCCCE
1ENH	Clarke et al. (1994)	54	3 hélices	HCCCCCH, HCCCH
1K43	Pastor et al. (2002)	14	1 folha	ETTE
1Q2K	Cai et al. (2004)	31	1 hélice, 1 folha	ETTTTE, HCCCE
1UTG	Morize et al. (1987)	70	5 hélices	HCCCCH, HCCCH, HCCH
1WQC	Chagot et al. (2005)	26	2 hélices	HCCCCH
2MR9	Nowicka et al. (2015)	44	3 hélices	HCCCH, HTTTH
2P5K	Garnett et al. (2007)	63	3 hélices, 1 folha	ETTTE, HCCCCCH, HCCE
2P6J	Shah et al. (2007)	52	3 hélices	HCCCH
2P81	Religa et al. (2007)	44	2 hélices	HCCCH
2PMR	n/a	76	3 hélices	HCCH

Fonte: adaptado de Borguesan et al. (2015)

6.3 Determinação de Parâmetros

Como visto em capítulo anterior, alguns parâmetros utilizados na metaheurística SADE proposta são determinados dependentemente do problema. Os valores atribuídos foram escolhidos com base em QIN, HUANG and SUGANTHAN (2009) e adaptados ao problema de predição de estruturas 3D de proteínas, os quais são apresentados na Tabela 6.2. Visto que proteínas são estruturalmente representadas por vetores

de ângulos diedros, cada indivíduo da população é um vetor de dimensão (***dim***) igual a $2 * tamanho(sequência\_de\_AA)$. O tamanho da população (**NP**) é igual a $10 * dim$. Os parâmetros envolvidos no processo de busca (**F** e **CR**) são gradualmente alterados ao longo da implementação como já descrito (Seção 5.3.3). O número de iterações do processo de aprendizagem (**LP**) é definido como 40. Por fim, o critério de parada aplicado corresponde ao número de avaliações de função (**num_faval**) realizadas, sendo esse igual a 1.000.000.

Tabela 6.2: Parâmetros utilizados na metaheurística SADE.

Sigla	*Correspondência*	*Valor*
AA	Sequência de Aminoácidos	Proteína
SS	Estrutura Secundária	Proteína
dim	Dimensão vetor-indivíduo	$2 * tamanho(AA)$
NP	Tamanho População	$10 * dim$
F	Fator de Mutação	$N(0.5, 0.3)$
CR	Taxa de *Crossover*	$N(CR_{médio}, 0.1)$
$CR_{médio}$	Valor mdio de CR	0.5 (inicial)
LP	Gerações do Período de Aprendizagem	40
num_faval	Avaliações da Função de Energia	1000000

Fonte: do autor (2016)

6.4 Validação da BPE

Teoricamente proteínas podem assumir inúmeras possíveis conformações estruturais; entretanto, a alta seletividade do processo de enovelamento proteico (Seção 2.3.2) faz com que uma estrutura única e dinâmica, seja adotada. As inúmeras possíveis estruturas correspondem ao espaço de busca conformacional em que as possíveis soluções serão exploradas pela metaheurística SADE. A biblioteca de fragmentos BPE foi proposta com o objetivo de diminuir este espaço de busca conformacional para o problema de predição da estrutura 3D de proteínas.

Para validação da informação, foi implementada a metaheurística SADE (seção 5.3) para comprovar que sua aplicação produz resultados melhores. Desta forma, foram realizadas 32 execuções do SADE utilizando a BPE e a base de dados APL-1 como conhecimento e 32 execuções do SADE sem a BPE, apenas aplicando a APL-1 e desconsiderando a presença de padrões estruturais de voltas. Para as implementações em que

a BPE não foi considerada, foi empregada a mesma estratégia de mutação (seção 5.3.4 - "Mutação e *Crossover*") para todos indivíduos da população NP.

Cada implementação resultou em um vetor-indivíduo final combinado com seu valor de função energética. Como visto (Seção 6.1.2), o RMSD e GDT_TS foram calculados para cada estrutura predita em relação sua estrutura experimental proveniente do PDB. Para cada proteína foram calculados os valores de RMSD médio e desvio padrão (Tabela 6.3 e Figura 6.1), GDT_TS médio e desvio padrão (Tabela 6.4 e Figura 6.2) e Função de Energia média e desvio padrão (Tabela 6.5). Os valores foram empregados na comparação entre implementações com uso da BPE (SADE-BPE) e implementações sem uso.

A Figura 6.3 mostra a sobreposição das estruturas preditas com SADE aplicando a BPE (azul), estruturas preditas com SADE sem BPE (verde) e estruturas extraídas do PDB determinadas de maneira experimental (vermelho). Para cada proteína foram selecionadas as estruturas preditas de menor RMSD, para ambas implementações, em relação a estrutura experimental. As estruturas foram sobrepostas pelo alinhamento dos átomos Cα através do programa PyMol.

Nas Figuras 6.1 e 6.2 são apresentadas as variabilidades dos valores de RMSD e GDT_TS para as estruturas geradas pelas 2 versões do método SADE. Os retângulos em cinza representam 50% dos valores, a linha em vermelho corresponde a mediana de cada conjunto de valores, as barras em azul marcam os valores máximos e mínimos e as cruzes em preto correspondem aos valores atípicos encontrados (*outliers*).

A análise visual da sobreposição das estruturas preditas em comparação com a estrutura experimental (Figura 6.3) mostra que as estruturas geradas pelas duas versões do SADE foram capazes de gerar as estruturas secundárias regulares (hélices e folhas) e capazes de formar estruturas empacotadas. Entretanto, uma análise mais detalhada revela que na maior parte dos casos a melhor sobreposição ocorre entre as estruturas preditas pelo SADE-BPE (azul) e as estruturas experimentais (vermelho). Além disso, a sobreposição das regiões de *looping* mostra-se melhor para estruturas geradas com a utilização da BPE (azul). Porém, está análise visual não é suficiente para garantir que empregar a BPE como conhecimento no método SADE gera resultados melhores, ou seja, mais próximos do experimental. Para melhor conclusão, os dados presentes nas Tabelas 6.3 e 6.4 e Figuras 6.1 e 6.2 foram considerados.

A análise da Tabela 6.3 demonstra que 93% das estruturas preditas com método SADE-BPE apresentam RMSD médio melhor do que estruturas preditas sem a BPE.

Tabela 6.3: Valores de RMSD para implementações de SADE considerando a utilização ou não da BPE.

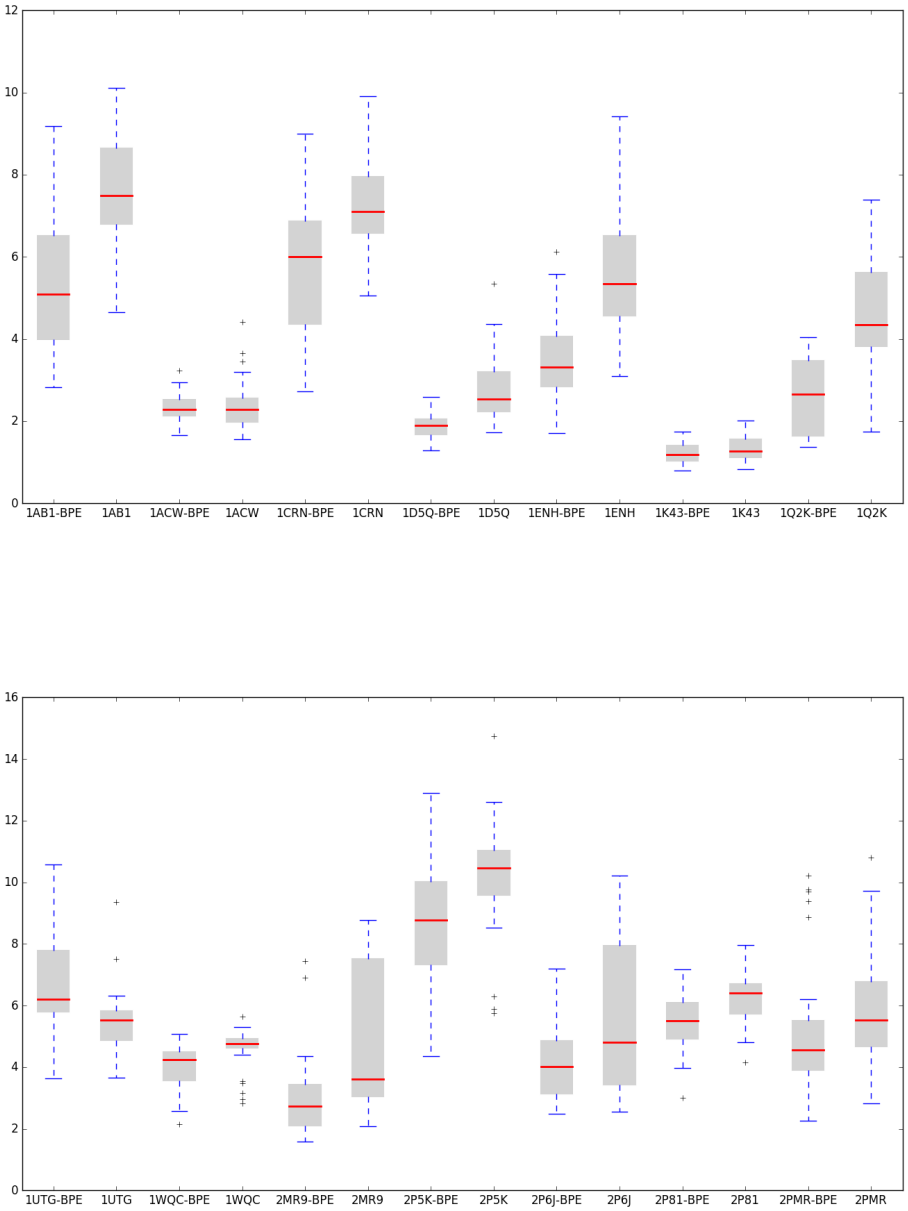
PDB_ID	Menor RMSD (Å)	$\text{RMSD}_{Médio}$
1AB1-BPE	2,83	5,28 ± 1,61
1AB1	4,65	7,49 ± 1,61
1ACW-BPE	1,67	2,33 ± 0,37
1ACW	1,56	2,37 ± 0,62
1CRN-BPE	2,72	5,76 ± 1,67
1CRN	5,06	7,20 ± 1,18
1D5Q-BPE	1,29	1,85 ± 0,29
1D5Q	1,72	2,78 ± 0,83
1ENH-BPE	1,70	3,54 ± 0,94
1ENH	3,10	5,63 ± 1,48
1K43-BPE	0,80	1,22 ± 0,23
1K43	0,83	1,34 ± 0,31
1Q2K-BPE	1,37	2,61 ± 0,93
1Q2K	1,75	4,65 ± 1,34
1UTG-BPE	3,62	6,63 ± 1,56
1UTG	3,65	5,49 ± 1,01
1WQC-BPE	2,14	3,96 ± 0,76
1WQC	2,83	4,61 ± 0,66
2MR9-BPE	1,58	3,02 ± 1,29
2MR9	2,08	4,68 ± 2,19
2P5K-BPE	4,36	8,75 ± 2,09
2P5K	5,76	10,2 1± 1,86
2P6J-BPE	2,48	4,14 ± 1,31
2P6J	2,55	5,53 ± 2,38
2P81-BPE	3,01	5,47 ± 0,94
2P81	4,14	6,28 ± 0,87
2PMR-BPE	2,27	5,12 ± 2,12
2PMR	2,82	5,94 ± 1,78

Fonte: do autor (2016)

Além disso, o método SADE-BPE gerou estruturas com o menor valor de RMSD em 92% dos casos testados. A análise da Figura 6.1 mostra resultados semelhantes à Tabela 6.3, em que são observados valores mínimos de RMSD e valores da mediana melhores em 86% das estruturas geradas pelo SADE-BPE. A distribuição dos valores de RMSD se revela mais uniforme em 57% das estruturas geradas pelo SADE-BPE. Além disso, em 86% dos casos os valores máximos de RMSD são encontrados em estruturas geradas pelo método sem a utilização da BPE.

Em relação aos dados de GDT_TS, a análise da Tabela 6.4 revela melhor resultado para GDT_TS médio em 93% das estruturas preditas pelo método SADE-BPE. Os dados

Figura 6.1: Distribuição dos valores de RMSD para implementação de SADE considerando a utilização ou não da BPE.



Fonte: do autor (2016)

de maior GDT_TS são melhores para 61,5% das estruturas preditas com método SADE-BPE. A Figura 6.2 mostra que 65% das estruturas geradas pelo SADE-BPE apresentam valores máximo para GDT_TS, enquanto que 86% dos casos apresentam melhores valores de mediana. A distribuição dos valores de GDT_TS se mostra mais uniforme em 71,5% das estruturas geradas pelo SADE-BPE, sendo encontrados os menores valores mínimos em 76% das estruturas geradas sem a utilização da BPE.

Tabela 6.4: Valores de GDT_TS para implementações de SADE considerando a utilização ou não da BPE.

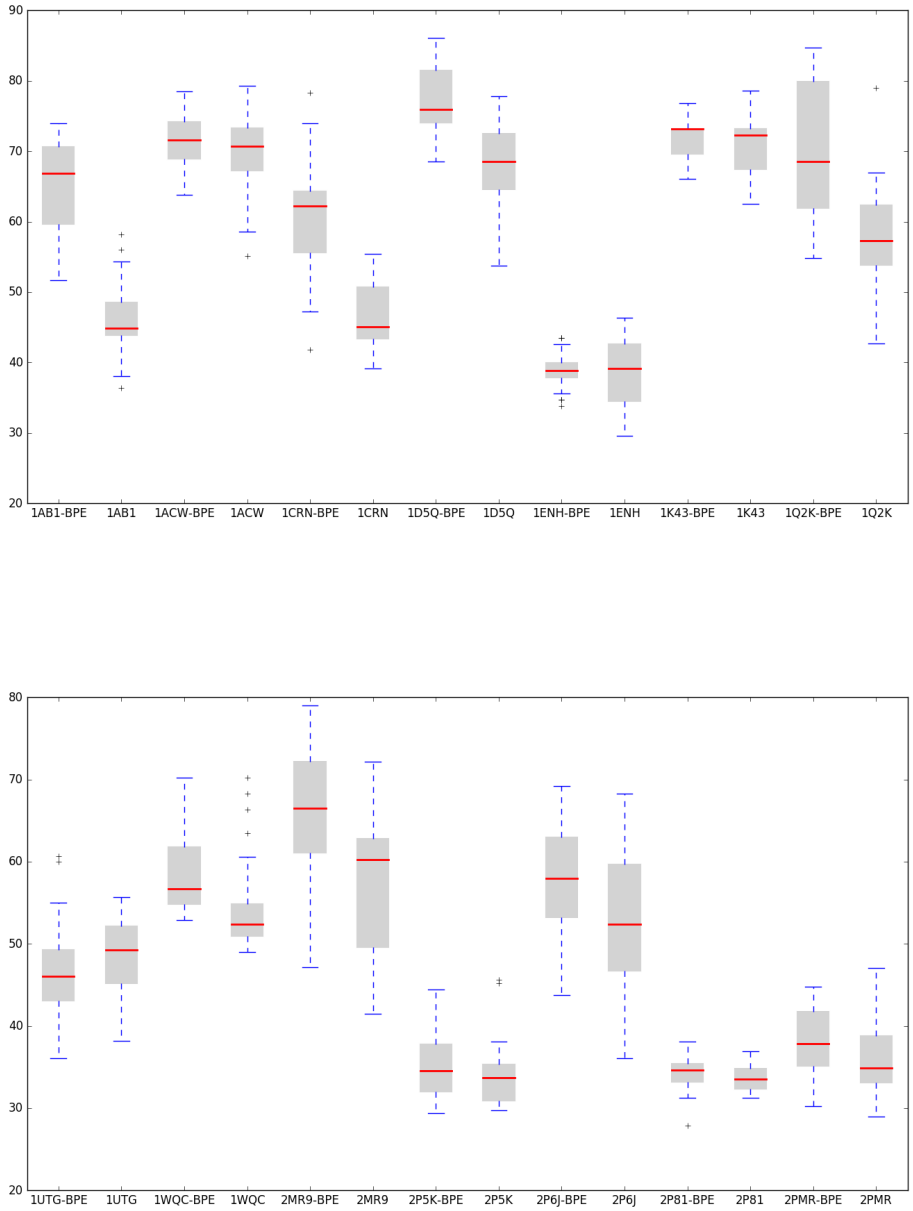
PDB_ID	Maior GST_TS (%)	$\text{GST\_TS}_{Médio}$
1AB1-BPE	73,91	65,81 ± 6,33
1AB1	58,15	46,16 ± 4,60
1ACW-BPE	78,45	71,53 ± 3,20
1ACW	79,31	69,88 ± 5,05
1CRN-BPE	78,26	60,58 ± 7,72
1CRN	55,43	46,35 ± 4,71
1D5Q-BPE	86,11	77,66 ± 4,69
1D5Q	77,78	68,09 ± 6,49
1ENH-BPE	43,52	38,95 ± 2,53
1ENH	46,30	38,53 ± 4,66
1K43-BPE	76,79	71,99 ± 2,58
1K43	78,57	70,70 ± 4,37
1Q2K-BPE	84,68	69,96 ± 9,00
1Q2K	79,03	58,01 ± 6,68
1UTG-BPE	60,71	46,32 ± 5,73
1UTG	55,71	48,67 ± 4,41
1WQC-BPE	70,19	59,01 ± 5,29
1WQC	70,19	54,36 ± 5,43
2MR9-BPE	78,98	66,14 ± 7,71
2MR9	72,16	57,12 ± 8,61
2P5K-BPE	44,44	35,00 ± 3,78
2P5K	45,63	33,99 ± 3,76
2P6J-BPE	69,23	57,93 ± 6,51
2P6J	68,27	53,11 ± 8,17
2P81-BPE	38,07	34,18 ± 2,03
2P81	36,93	33,81 ± 1,63
2PMR-BPE	44,74	38,28 ± 3,71
2PMR	47,04	35,64 ± 4,14

Fonte: do autor (2016)

Na Tabela 6.5 são apresentados os resultados para valores de energia média, assim como o menor valor atingido por determinada estrutura. A análise desses dados não indica diferença entre ambas as implementações do método SADE visto que os valores são muito próximos. Em outras palavras, a utilização da BPE não necessariamente gera estruturas de menor energia em relação ao não uso.

A Tabela 6.6 apresenta a comparação dos menores valores de RMSD para estruturas geradas pelo método SADE com a utilização da BPE e para estruturas geradas pelo método *Genetic Algorithm* (GA) com utilização da APL-1 (BORGUESAN et al., 2015). Os resultados se mostram melhores em todos os casos para estruturas geradas pelo

Figura 6.2: Distribuição dos valores de GDT_TS para implementação de SADE considerando a utilização ou não da BPE.



Fonte: do autor (2016)

SADE-BPE.

Em suma, os resultados apresentados mostram que a implementação da metaheurística SADE utilizando a BPE como conhecimento é capaz de gerar estruturas com menor valor de RMSD, menor valor de RMSD médio, maior valor de GDT_TS e maior valor de GDT_TS médio quando comparadas ao mesmo método sem utilização da BPE. Além disso, é capaz de gerar estruturas cujos valores de RMSD e GDT_TS se distribuem de ma-

Tabela 6.5: Valores de Energia para implementações de SADE considerando a utilização ou não da BPE.

PDB_ID	Menor Energia	$\text{Energia}_{Médio}$
1AB1-BPE	-21935,87	-21913,50 $\pm$ 7,32
1AB1	-21903,67	-21850,08 $\pm$ 23,04
1ACW-BPE	-18980,22	-18974,01 $\pm$ 2,88
1ACW	-18972,27	-18959,51 $\pm$ 6,01
1CRN-BPE	-21926,44	-21912,38 $\pm$ 6,15
1CRN	-21905,91	-21847,14 $\pm$ 25,24
1D5Q-BPE	-18972,51	-18969,39 $\pm$ 1,76
1D5Q	-18972,87	-18964,75 $\pm$ 3,48
1ENH-BPE	-37947,90	-37940,42 $\pm$ 2,91
1ENH	-37939,08	-37933,43 $\pm$ 2,10
1K43-BPE	-5978,47	-5977,37 $\pm$ 0,62
1K43	-5979,44	-5977,92 $\pm$ 0,70
1Q2K-BPE	-18966,99	-18963,39 $\pm$ 1,80
1Q2K	-18946,57	-18928,93 $\pm$ 11,11
1UTG-BPE	-55897,52	-55890,78 $\pm$ 3,58
1UTG	-55907,88	-55901,21 $\pm$ 2,85
1WQC-BPE	-15977,66	-15974,72 $\pm$ 1,46
1WQC	-15981,78	-15977,23 $\pm$ 1,54
2MR9-BPE	-29949,81	-29946,04 $\pm$ 1,89
2MR9	-29950,84	-29947,19 $\pm$ 1,86
2P5K-BPE	-45918,93	-45910,28 $\pm$ 4,40
2P5K	-45899,91	-45888,30 $\pm$ 7,21
2P6J-BPE	-32963,39	-32958,31 $\pm$ 2,50
2P6J	-32961,41	-32956,36 $\pm$ 2,16
2P81-BPE	-26951,79	-26948,84 $\pm$ 1,31
2P81	-26952,61	-26949,15 $\pm$ 1,59
2PMR-BPE	-60880,91	-60872,91 $\pm$ 3,20
2PMR	-60878,31	-60870,52 $\pm$ 2,82

Fonte: do autor (2016)

neira mais uniforme, ou seja, não havendo grandes flutuações e valores atípicos. Por fim, os menores valores de RMSD encontrados para estruturas geradas pelo método SADE-PE revelaram-se melhores em 100% dos casos quando comparados com estruturas geradas pelo método GA-APL. O método GA-APL (BORGUESAN et al., 2015) é uma estratégia similar ao SADE, ou seja, ambas são metaheurísticas baseadas em população e de caráter evolutivo. Além disso, ambas fazem utilização de conhecimento experimental visando a diminuição do espaço de busca conformacional. Assim, é possível afirmar que a utilização da BPE na metaheurística SADE é capaz de gerar estruturas 3D de proteínas mais próximas das estruturas experimentais, quando comparado com estruturas preditas pelo

Tabela 6.6: Comparação dos menores valores de RMSD entre estruturas geradas com SADE-BPE e com GA-APL.

PDB_ID	SADE-BPE (Å)	GA-APL (Å)
1AB1	2,83	5,53
1ACW	1,67	7,99
1CRN	2,72	5,80
1D5Q	1,29	4,08
1ENH	1,70	10,92
1K43	0,80	1,39
1Q2K	1,37	5,64
1UTG	3,62	8,90
1WQC	2,14	3,49
2MR9	1,58	7,74
2P5K	4,36	8,77
2P6J	2,48	11,16
2P81	3,01	3,90
2PMR	2,27	19,22

Fonte: do autor (2016) e Borguesan et al. (2015)

mesmo método sem o uso da BPE ou por método semelhante sem uso da BPE.

6.5 Resumo do Capítulo

Para validação da BPE criada (Seção 5.2) foram implementadas duas versões da metaheurística SADE. A primeira versão (SADE-BPE) utilizou a BPE e APL-1 como conhecimento, enquanto que a segunda versão utilizou apenas a APL-1, desconsiderando padrões de voltas na estrutura. O objetivo das implementações foi garantir que com a utilização da BPE se consegue gerar estruturas melhores, ou seja, mais próximas da estrutura experimental.

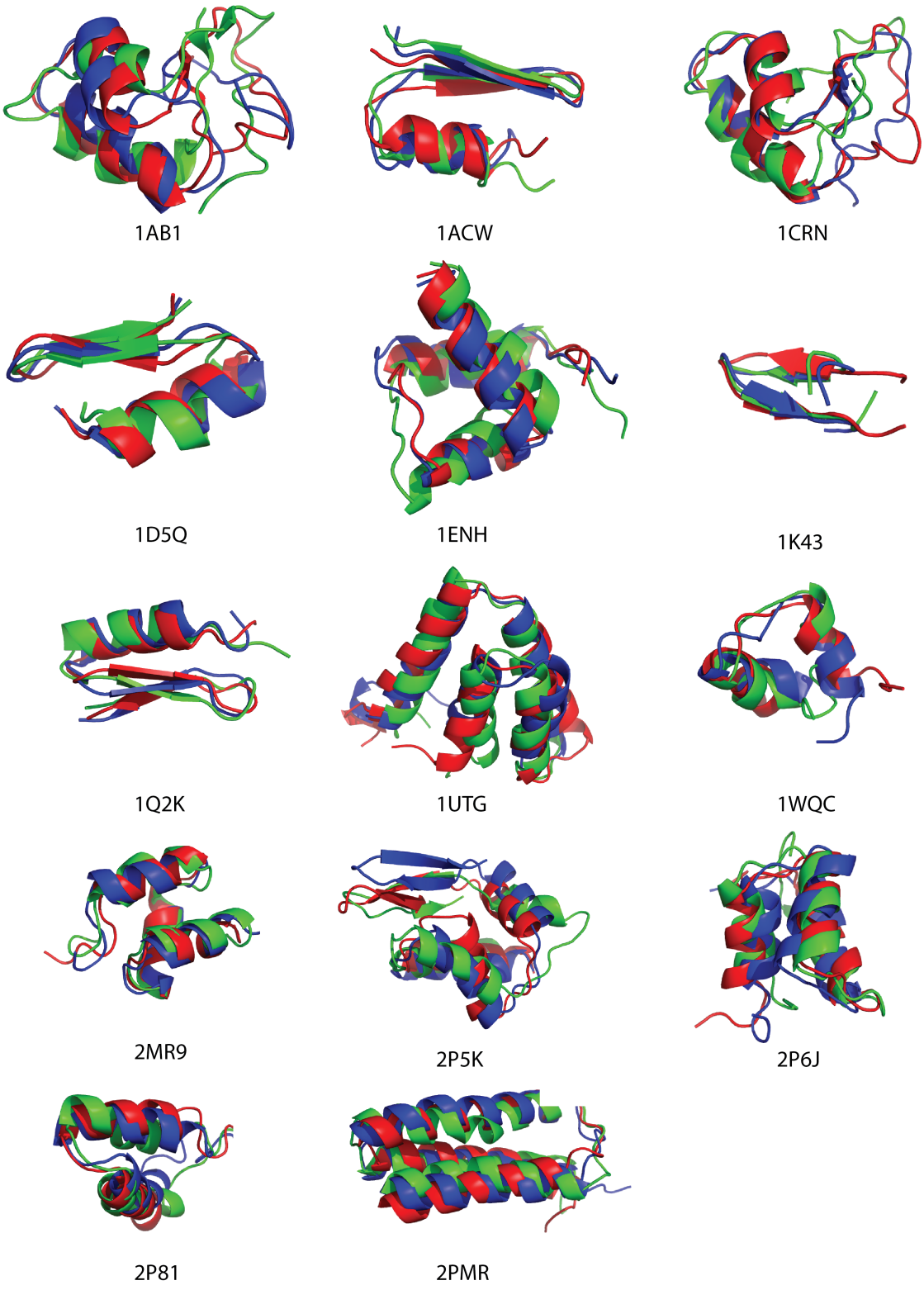
Para comparação das estruturas preditas, foram considerados os critérios de RMSD, GDT_TS e Função de Energia. Os cálculos de RMSD e GDT_TS foram feitos pela comparação da estrutura predita e estrutura experimental do PDB. Para análises, foram selecionadas 14 proteínas que compunham a base de teste.

A comparação dos resultados de RMSD médio e menor valor de RMSD demonstraram que 93% e 92%, respectivamente, das estruturas preditas pelo SADE-BPE foram melhores. Os valores de GDT_TS médio e maior valor de GDT_TS foram 93% e 61,5%, respectivamente, melhores nos casos em que o método SADE-BPE foi aplicado. A distri-

buição dos valores mostrou-se mais uniforme em 57% dos casos para valores de RMSD e em 71,5% dos casos para GDT_TS. Os valores de Função de Energia média e menor Função de Energia não mostraram diferenças entre as implementações.

A comparação dos valores de menor RMSD para estruturas geradas pelo SADE-BPE com estruturas geradas pelo GA-APL mostrou-se melhor em 100% dos casos. Logo, a partir das análises efetuadas é possível afirmar que a utilização da BPE na metaheurística SADE é capaz de produzir estruturas mais próximas das estruturas experimentais quando comparada com a implementação do mesmo algoritmo sem o uso da BPE.

Figura 6.3: Representação gráfica da sobreposição das estruturas 3D preditas e experimental.



Em **vermelho** estão representadas as estruturas experimentais extraídas do PDB. Em **azul** e **verde** estão representadas as estruturas geradas pelos métodos SADE-BPE e SADE respectivamente.

Fonte: do autor (2016)

7 CONCLUSÃO

Proteínas são polímeros lineares formados por diferentes combinações de aminoácidos, os quais adotam uma estrutura 3D única quando em condições fisiológicas. A estrutura 3D de proteínas está intrinsecamente relacionada àa sua atividade biológica, logo, variadas estruturas 3D possibilitam que proteínas desempenhem inúmeras funções. O conhecimento da estrutura 3D de proteínas permite, então, a determinação, estudo e análise direta da sua função (LEHNINGER; NELSON; COX, 2005; LESK, 2010).

A maior fonte de estruturas 3D de macromoléculas são técnicas experimentais, como cristalografia por raio-x e RMN; entretanto, são metodologias custosas e que levam demasiado tempo, além de necessitarem de equipamentos específicos. Devido às limitações, hoje poucas proteínas possuem sua estrutura 3D determinada. Por outro lado, houve uma explosão de dados biológicos com o sucesso do Projeto Genoma e desenvolvimento das técnicas de biologia molecular; acarretando em uma discrepância de volume de dados. Assim, a utilização de métodos computacionais no problema de predição da estrutura 3D de proteínas é motivada e necessária (DORN et al., 2014).

Predizer a estrutura 3D de proteínas a partir da sequência linear de aminoácidos é um problema interdisciplinar e um dos mais difíceis da Bioinformática Estrutural. Visto que o processo de enovelamento é altamente seletivo, a sequência de aminoácidos irá adotar uma conformação única dentre as inúmeras possibilidades. Encontrar esta conformação única é computacionalmente custoso devido ao enorme espaço de busca (Paradoxo de Levinthal) (LEVINTHAL, 1978), representando um problema de Otimização Complexo, classificado computacionalmente como NP-Completo (CRESCENZI et al., 1998).

Em proteínas, *loops* são fragmentos flexíveis os quais estão conectando dois elementos de estrutura secundária regulares, como hélices e folhas. Desempenham importantes papéis no processo de enovelamento e estabilidade proteicas, além de serem encontrados em sítios ativos e de ligações. *Loops* são regiões variáveis, os quais não apresentam padrões comuns, além de serem mais suscetíveis a sofrerem mutações. Assim, *loops* podem adotar inúmeras possíveis conformações, sendo regiões mais difíceis e complexas de serem manipuladas (OFFMANN; TYAGI; BREVERN, 2007; SHEHU; KAVRAKI, 2012).

Encontrar a solução exata para o problema de predição de estrutura 3D de proteínas não tem fundamento devido à flexibilidade das moléculas de proteínas, além de ser inviável em tempo razoável. A utilização de algoritmos aproximativos, como Metaheurís-

ticas (TABLI, 2009; BOUSSAÏD; LEPAGNOT; SIARRY, 2013), é uma alternativa capaz de gerar um conjunto de soluções satisfatórias em tempo razoável.

Os últimos resultados dos experimentos CASP revelaram que métodos de predição utilizando conhecimento experimental obtiveram resultados satisfatórios (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014). A aplicação do conhecimento pode ser feita de variadas maneiras, sendo a abordagem em fragmentos uma das mais utilizadas. Desta forma, esta monografia propôs uma nova biblioteca de padrões estruturais (BPE), a qual é composta de fragmentos gerados pela quebra de estrutura secundária, selecionando as regiões de *loops*.

Para validação da BPE foi proposta sua utilização em uma metaheurística, SADE (QIN; SUGANTHAN, 2005; QIN; HUANG; SUGANTHAN, 2009), aplicada ao problema de predição da estrutura 3D de proteínas. Assim, a metodologia foi implementada em duas versões: (i) SADE utilizando BPE como conhecimento e (ii) SADE sem BPE. As estruturas geradas por cada versão foram comparadas considerando os critérios de RMSD, GDT_TS e Função de Energia.

Nos resultados obtidos foi possível afirmar que a utilização da BPE na metaheurística SADE consegue gerar estruturas mais próximas das experimentais em relação à sua não utilização no método. Por mais que regiões de *loops* não apresentem padrões comuns, estes dados demonstram que a utilização de uma biblioteca específica pode ajudar no problema de predição de estrutura 3D de proteínas.

Sendo assim, pode-se reconhecer que este trabalho tem como contribuição científica: (i) a criação de uma biblioteca de padrões estruturais para regiões de *loops*, as quais são conhecidas por serem irregulares, variáveis e de difícil predição; (ii) a utilização da abordagem de fragmentos como conhecimento na metaheurística *Self-Adapting Differential Evolution* no problema de predição da estrutura 3D de proteínas.

Como trabalho futuro, seria interessante ampliar a BPE com a inclusão de outros padrões para regiões de *loops*. Além disso, seria importante utilizar o conhecimento da BPE em outras metaheurísticas aplicadas ao problema de predição de estrutura 3D de proteínas. Outro possível trabalho futuro seria modificar a estratégia de fragmentação utilizada na busca por padrões estruturais.

REFERÊNCIAS

ALTMAN, R. B.; DUGAN, J. M. Defining bioinformatics and structural bioinformatics. In: BOURNE, P. E.; WEISSIG, H. (Ed.). **Structural Bioinformatics**. 1. ed. [S.l.]: John Wiley Sons, Inc., 2003. chp. 1, p. 2–14.

ALTSCHUL, S. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, p. 403–410, 1990.

ALTSCHUL, S. Gapped blast and psi-blast: a new generation of protein database search programs. **Nucleic Acid Research**, v. 25, n. 17, p. 3389–3402, 1997.

ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, v. 181, p. 223–230, 1973.

ARNOLD, K. et al. The swiss-model workspace: A web-based environment for protein structure homology modelling. **Bioinformatics**, v. 22, p. 195–201, 2006.

BERMAN, H. et al. The protein data bank. **Nucleic Acids Research**, Oxford University Press, Piscataway, NJ, USA, v. 28, n. 1, p. 235–242, 2000.

BONET, J. et al. ArchDB 2014: structural classification of loops in proteins. **Nucleic Acid Research**, v. 42, p. D315–D319, 2014.

BORGUESAN, B.; INOSTROZA-PONTA, M.; DORN, M. Nias-server: Neighbors influence of amino acids and secondary structures in proteins. **Journal of Computational Biology**, 2016.

BORGUESAN, B. et al. APL: an angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. **Computational Biology and Chemistry**, v. 59, p. 142–157, 2015.

BOUSSAÏD, I.; LEPAGNOT, J.; SIARRY, P. A survey on optimization metaheuristics. **Information Sciences**, v. 237, p. 82–117, 2013.

BROOKS, R. et al. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. **Journal of Computational Chemistry**, v. 4, n. 2.

BURLEY, S. et al. Contributions to the NIH-NIGMS protein structure initiative from the psi production centers. **Structure**, v. 16, p. 5–11, 2008.

CHANDONIA, J. et al. The astral compendium in 2004. **Nucleic Acid Research**, v. 32, p. D189–D192, 2004.

CHAUDHURY, S.; LYSKOV, S.; GRAY, J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. **Bioinformatics**, v. 26, n. 5, p. 689–691, 2010.

CHEN, Y. Introduction to bioinformatics. In: CHEN, Y. (Ed.). **Bioinformatics Technologies**. 1. ed. [S.l.]: Springer, 2005. chp. 1, p. 1–13.

CHOTHIA, C.; LESK, A. The relation between the divergence of sequence and structure in proteins. **The EMBO Journal**, v. 5, p. 823–826, 1986.

CHRISTEN, M. et al. The grooms software for biomolecular simulation: Gromos05. **Journal of Computational Chemistry**, v. 26, p. 1719–1751, 2005.

CORNELL, W. et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. **Journal of the American Chemical Society**, v. 117, n. 19, p. 5179–5197, 1995.

CRESCENZI, P. et al. On the complexity of protein folding. **Journal of Computational Biology**, v. 5, n. 3, p. 423–466, 1998.

DORN, M. et al. Three-dimensional protein structure prediction: Methods and computational strategies. **Computational Biology and Chemistry**, v. 53, p. 251–276, 2014.

FERNANDEZ-FUENTES, N.; OLIVA, B.; FISER, A. ArchPRED: a template based loop structure prediction server. **Nucleic Acid Research**, v. 34, p. W173–W176, 2006.

FISER, A. Comparative protein structure modelling. In: RIGDEN, D. J. (Ed.). **From Protein Structure to Function with Bioinformatics**. 1. ed. [S.l.]: Springer, 2009. chp. 3, p. 57–90.

FISER, A.; SALI, A. ModLoop: automated modeling of loops in protein structures. **Bioinformatics**, v. 19, p. 2500–01, 2003.

GNIEWEK, P. et al. Bioshell-threading: versatile monte carlo package for protein 3d threading. **BMC Bioinformatics**, v. 1, p. 15–29, 2014.

HEINIG, M.; FRISHMAN, D. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. **Nucleic Acids Research**, v. 32, p. W500–2, 2004.

HOVMÖLLER, S.; ZHOU, T.; OHLSON, T. Conformations of amino acids in proteins. **Acta Crystallographica Section D**, Munksgaard International Publishers, v. 58, n. 5, p. 768–776, 2002.

ISMER, J. et al. SL2: an interactive webtool for modeling of missing segments in proteins. **Nucleic Acid Research**, v. 44, p. W390–W394, 2016.

JANSON, G. et al. Pymod 2.0: improvements in protein sequence-structure analysis and homology modeling within pymol. **Bioinformatics**, v. 32, n. 21, 2016.

JONES, D. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. **Journal of Molecular Biology**, v. 287, n. 4, p. 797–815, 1999.

JONES, D. Predicting novel protein folds by using fragfold. **Proteins: Structure, Function, and Bioinformatics**, v. 45, p. 127–132, 2001.

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

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

KALLBERG, M. et al. Raptorx server: a resource for template-based protein structure modeling. **Methods in Molecular Biology**, v. 1137, p. 17–27, 2014.

KELLEY, L.; STERNBERG, M. Protein structure prediction on the web: a case study using the phyre server. **Nature Protocols**, v. 4, p. 363–371, 2009.

KO, J. et al. The FALC-Loop web server for protein loop modeling. **Nucleic Acid Research**, v. 39, 2011.

KOSÍNSKI, J. et al. Template based prediction of three-dimensional protein structures: Fold Recognition and Comparative Modeling. In: BUJNICKI, J. M. (Ed.). **Prediction of Protein Structures, Functions, and Interactions**. [S.l.]: John Wiley Sons, Ltd., 2009. chp. 4, p. 87–116.

KRYSHTAFOVYCH, A.; FIDELIS, K.; MOULT, J. Casp10 results compared to those of previous casp experiments. **Proteins**, v. 82, p. 164–174, 2014.

LEE, J. et al. *Ab Initio* protein structure prediction. In: RIGDEN, D. J. (Ed.). **From Protein Structure to Function with Bioinformatics**. 1. ed. [S.l.]: Springer, 2009. chp. 1, p. 3–25.

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

LESK, A. **Introduction to Protein Science: Architecture, Function and Genomics.** 2. ed. New York, NY, USA: Oxford University Press, 2010.

LEVINTHAL, C. Are there pathways for protein folding? **Journal de Chimie Physique et de Physico-Chimie Biologique**, v. 65, n. 1, p. 44, 1978.

LI, Y. Conformational sampling in template-free protein loop structure modeling: An overview. **Computational and Structural Biotechnology Journal**, v. 5, 2013.

LIANG, S.; ZHANG, C.; ZHOU, Y. LEAP: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. **Journal of Computational Chemistry**, v. 35, p. 335–341, 2014.

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? a proposed definition and overview of the field. **Methods Of Information In Medicine**, New Haven, CT, USA, v. 40, n. 4, p. 346–358, 2001.

LÓPEZ-BLACO, J. et al. RCD+: Fast loop modeling server. **Nucleic Acid Research**, 2016.

MARSDEN, R. et al. Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. **Nucleic Acid Research**, v. 34, p. 1066–1080, 2006.

MESSIH M.A., L. R.; TRAMONTANO, A. LoopIng: a template-based tool for predicting the structure of protein loops. **Bioinformatics**, v. 31, n. 23, p. 3767–3772, 2015.

MURZIN, A. G. et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. **Journal of Molecular Biology**, v. 247, p. 536–540, 1995.

N.A., O. et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. **Nucleic Acids Research**, v. 44, p. D733–45, 2016.

NOTREDAME, C.; HIGGINS, D.; HERINGA, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. **Journal of Molecular Biology**, v. 302, n. 1, p. 205–217, 2000.

OFFMANN, B.; TYAGI, M.; BREVERN, A. G. D. Local Protein Structures. **Current Bioinformatics**, v. 3, p. 165–202, 2007.

PAULING, L.; COREY, R. The pleated sheet, a new layer configuration of polypeptide chains. **Proceedings of the National Academy of Sciences of the United States of America**, 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. **Proceedings of the National Academy of Sciences of the United States of America**, v. 37, n. 4, p. 205–211, 1951.

PEARSON, W.; LIPMAN, D. Improved tools for biological sequence comparison. **PNAS**, v. 85, n. 8, p. 2444–2448, 1988.

PIEPER, U. et al. MODBASE: a database of annotated comparative protein structure models and associated resources. **Nucleic Acid Research**, v. 34, p. D291–295, 2006.

QIN, A. K.; HUANG, V.; SUGANTHAN, P. N. Differential evolution algorithm with strategy adaptation for global numerical optimization. **IEEE Transactions on Evolutionary Computation**, v. 13, n. 2, 2009.

QIN, A. K.; SUGANTHAN, P. N. Self-adaptive differential evolution algorithm for numerical optimization. **IEEE Congress on Evolutionary Computation**, Edinburgh, Scotland, p. 1785–1791, 2005.

RAMACHANDRAN, G.; SASISEKHARAN, V. Conformation of polypeptides and proteins. **Advances In Protein Chemistry**, v. 23, p. 238–438, 1968.

RICHARDSON, J. The anatomy and taxonomy of protein structure. **Biopolymers**, v. 34, p. 167–339, 1981.

ROHL, C. et al. Protein structure prediction using rosetta. **Methods Enzymol**, v. 383, n. 2, p. 66–93, 2004.

RUSSELL, R.; COPLEY, R.; BARTON, G. Protein fold recognition by mapping predicted secondary structures. **Journal of Molecular Biology**, v. 259, n. 3, p. 349, 1996.

SALI, A.; BLUNDELL, T. Comparative protein modelling by satisfaction of spatial restraints. **Journal of Molecular Biology**, v. 234, p. 779–815, 1993.

SCHEEFF, E. D.; FINK, J. L. Fundamentals of protein structure. In: BOURNE, P. E.; WEISSIG, H. (Ed.). **Structural Bioinformatics**. 1. ed. [S.l.]: John Wiley Sons, Inc., 2003. chp. 2, p. 15–39.

SHEHU, A.; KAVRAKI, L. Modeling structures and motions of loops in protein molecules (review). **Entropy**, v. 14, p. 252–290, 2012.

SHMYGELSK, A.; LEVITT, M. Generalized ensemble methods for *de novo* structure prediction. **PNAS**, v. 106, n. 5, p. 1415–1420, 2008.

SILLITOE, I. et al. CATH: comprehensive structural and functional annotations for genome sequences. **Nucleic Acid Research**, v. 43, p. D376–81, 2015.

STORN, R.; PRICE, K. Differential Evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces. **California: ICSI**, 1995.

STORN, R.; PRICE, K. Differential Evolution – a simple and efficient heuristic for global optimization over continuous spaces. **Journal of Global Optimization**, v. 11, p. 341–359, 1997.

TABLI, E. **Metaheuristic: from design to implementation.** Hobouken, NJ: Wiley, 2009.

TANG, K.; ZHANG, J.; LIANG, J. Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth monte carlo method. **PLoS Computational Biology**, v. 10, 2014.

TAYLOR, W. The classification of amino acid conservation. **Journal of Theoretical Biology**, v. 119, n. 2, p. 205–218, 1986.

TEILUM, K.; OLSEN, J.; KRAGELUND, B. Functional aspects of protein flexibility. **CMLS**, v. 66, n. 14, p. 2231–2247, 2009.

THOMPSON, J.; HIGGINS, D.; GIBSON, T. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acid Research**, v. 22, p. 4673–4680, 1994.

TRAMONTANO, A. **Protein Structure Prediction.** [S.l.]: Wiley-VCH, 2006.

VANHEE, P. et al. BriX: a database of protein building blocks for structural analysis, modeling and design. **Nucleic Acid Research**, v. 39, p. D435–D442, 2011.

VERLI, H. Níveis de informação biológica. In: VERLI, H. (Ed.). **Bioinformática da Biologia à flexibilidade molecular**. 1. ed. [S.l.: s.n.], 2014. chp. 1, p. 13–37.

ZHANG, Y. I-tasser server for protein 3d structure prediction. **BMC Bioinformatics**, v. 9, p. 40, 2008.

ZHANG, Y.; SKOLNICK, J. Scoring function for automated assessment of protein structure template quality. **Proteins: Structure, Function, and Bioinformatics**, v. 57, n. 4, p. 702–710, 2004.