UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

BRUNO BORGUESAN

GARTS: um Algoritmo Genético baseado no método de Seleção por Torneio Restrito Adaptativo para o problema de Predição de Estruturas 3D de Proteínas

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Márcio Dorn

Porto Alegre
2016

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Borguesan, Bruno

GARTS: um Algoritmo Genético baseado no método de Seleção por Torneio Restrito Adaptativo para o problema de Predição de Estruturas 3D de Proteínas / Bruno Borguesan. – 2016.

142 f.: il.

Orientador: Márcio Dorn.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2016.

1. Bioinformática Estrutural. 2. Predição da Estrutura 3D de Proteínas. 3. Meta-heurísticas Multimodais. 4. Métodos Baseados em Conhecimento. I. Dorn, Márcio. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Pós-Graduação: Prof. Vladimir Pinheiro do Nascimento
Diretor do Instituto de Informática: Prof. Luís da Cunha Lamb
Coordenador do PPGC: Prof. Luigi Carro
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

*“If a cluttered **desktop** is a sign of a cluttered mind,*
*of what, then, is an empty **desktop** a sign?”*
— Albert Einstein

AGRADECIMENTOS

Ao meu orientador, Professor Doutor Márcio Dorn, por todo o apoio, incentivo e tempo disponibilizado durante o período de mestrado. Nossas conversas me ajudaram muito a concluir esta etapa da minha vida. Agradeço pelo exemplo de pessoa e pesquisador que tu és. Também gostaria de agradecer a Universidade Federal do Rio Grande do Sul, seu corpo docente e funcionários, por todo o suporte oferecido durante a pesquisa, bem como ao CNPq pelo apoio financeiro.

Aos meus amigos e colegas de laboratório por todos os momentos de conversas que com certeza facilitaram a minha trajetória acadêmica. Agradeço também aos meus amigos do Paraná, em especial a toda Família Arapuca, que me ajudou a relaxar durante os períodos estressantes do mestrado.

Aos meus pais, Cerineu Borguesan e Cleuza Aparecida Borguesan, por tudo que batalharam pela nossa família. Por terem me proporcionado a melhor educação e criação possível, moldando a pessoa que eu sou hoje. Agradeço imensamente por todo o amor e incentivo incondicional recebido por vocês em todos esses anos.

Em especial, ao meu irmão Adriano Borguesan (*in memoriam*), que no tempo em que esteve comigo me ensinou muito.

RESUMO

Predizer a estrutura tridimensional de proteínas permanece como um dos problemas mais desafiadores da Bioinformática Estrutural e até o momento continua sem solução. O conhecimento sobre a estrutura tridimensional de um polipeptídeo/proteína proporciona aos pesquisadores uma importante informação para inferir a função desta proteína na célula de um organismo. Entretanto, devido ao fato de que as regras que governam o processo de enovelamento dessas proteínas serem apenas parcialmente conhecidas várias soluções podem acabar sendo consideradas ótimas durante a simulação. Nesta dissertação, é apresentada uma nova abordagem para a predição da estrutura tridimensional de proteínas considerando o problema como multimodal. O método proposto apresenta um Algoritmo Genético (AG) baseado na Seleção por Torneio Restrito Adaptativo (ARTS) com o propósito de evitar a convergência do método em mínimos locais através da diversificação da população em grupos e restringindo suas interações por sua similaridade. O método proposto utiliza conhecimento extraído da base de dados experimental através de uma biblioteca de fragmentos para inicializar a população com bons indivíduos. A abordagem também utiliza a preferência conformacional da vizinhança dos aminoácidos a fim de reduzir o espaço de busca do problema de predição da estrutura tridimensional de proteínas. O método proposto foi testado com um conjunto de 24 sequências de proteínas distintas. Através da análise estrutural realizada para cada uma das proteínas preditas, foi possível verificar que a nova proposta apresenta resultados comparáveis com as estruturas determinadas experimentalmente. Sugerindo assim, que a abordagem produziu bons resultados. O método também foi comparado com o servidor I-TASSER limitando o seu conhecimento para apenas 25% das estruturas com similaridade de sequência. Os resultados mostraram que a nossa abordagem foi capaz de produzir soluções com valores de RMSD e GDT-TS melhores em mais de 50% dos casos de teste, corroborando assim com a efetividade do método proposto.

Palavras-chave: Bioinformática Estrutural. Predição da Estrutura 3D de Proteínas. Meta-heurísticas Multimodais. Métodos Baseados em Conhecimento.

GARTS: a Genetic Algorithm combined with an Adaptive Restricted Tournament Selection approach for the 3D Protein Structure Prediction problem

ABSTRACT

The prediction of three-dimensional structures of proteins represents one of the most important and unsolved problems in Structural Bioinformatics. As the size of a protein increases, the potential energy landscape rapidly becomes very complex with multiple local minima. We propose a new strategy that combines a Genetic Algorithm with a variation of the Adaptive Restricted Tournament Selection approach to avoid the early convergence by differentiating the individuals of the population into groups and limiting their interaction by structural similarity. The proposed method combines a template fragment library obtained from experimental-determined protein structures with conformational preferences of amino acids residues to reduce the search space. The proposed method was tested on a test bed of 24 protein sequences. Structural analysis was performed and achieved results indicate that predicted conformations adopt a fold similar to the experimental ones. We also compared our method with the I-TASSER Server using only 25% of structures with sequence identity, the result showed that our approach achieved comparable solutions with improved values of RMSD and GDT-TS scores of more than 50% of test cases. Thus, corroborating to the effectiveness of our proposal.

Keywords: Protein Structure Prediction. Knowledge-Based Search Methods. Multimodal Metaheuristics. Structural Bioinformatics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
CASP	*Critical Assessment of Structure Prediction*
DSSP	*Define Secondary Structure of Proteins*
RMN	Ressonância Magnética Nuclear
PDB	*Protein Data Bank*
RefSeq	*Reference Sequence Database*
STRIDE	*STRuctural IDEntification*
X-RAY	Cristalografia por difração de Raios-X
NCBI	*National Center for Biotechnology Information*
CCDC	*Cambridge Crystallographic Data Centre*
BNL	*Brookhaven National Laboratory*
RCSB PDB	*Research Collaboratory of Structural Bioinformatics*
wwPDB	*Worldwide PDB*
PDBj	*PDB Japan*
DM	Dinâmica Molecular
FM	*Free Modeling*
TBM	*Template-Based Modeling*
APL	*Angle Probability List*
NPAS	*Neighbors Preferences of Amino Acids and Secondary Structures*
FM-B Lib	*FragMent-Based LIBrary*
AG	Algoritmo Genético
EP	Enxame de Partículas
RTS	*Restricted Tournament Select*
ARTS	*Adaptative Restricted Tournament Select*

RMSD	*Root Mean Square Deviation*
SASA	Área da Superfície Acessível ao Solvente
GDT_TS	*Global Distance Total Score Test*
SAS	*Sequence Annotated by Structure*

SUMÁRIO

1 INTRODUÇÃO

O avanço da tecnologia e o seu emprego em áreas como a Biologia desencadeou um aumento significativo no número de dados biológicos. A análise manual destes dados é em alguns casos um procedimento difícil e custoso. Para aprimorar a análise destes dados, faz-se necessário o desenvolvimento e a aplicação de técnicas computacionais. Essa união entre Biologia e Computação criou um novo campo de pesquisa conhecido como Bioinformática (LESK, 2005).

Um dos principais tópicos de pesquisa da Bioinformática é o estudo do dogma central da Biologia Molecular onde sequências de DNA são transcritas em sequências de RNA, que por sua vez podem ser traduzidas em sequência de proteínas. A aplicação de técnicas computacionais em qualquer uma das etapas do dogma central seja para representação, armazenamento, recuperação ou análise de informações é considerado como Bioinformática (ALTMAN; DUGAN, 2005). Segundo Luscombe et al. (2001) os principais objetivos da Bioinformática podem ser divididos em três:

1. Organização dos dados de uma maneira que permita que pesquisadores tenham acesso fácil a estes dados e possam submeter as novas entradas que são produzidas;
2. Desenvolvimento de ferramentas e recursos que auxiliem os pesquisadores na análise destes dados;
3. Uso destas ferramentas computacionais para analisar os dados e interpretar os resultados.

A grande mobilização em torno do Projeto Genoma, visando desvendar o código genético de um organismo, em conjunto com o avanço no poder computacional, acarretou em um gigantesco número de sequências genômicas. A base de dados *REFerence SEQuence* (*RefSeq*) (PRUITT et al., 2002) é uma das principais com relação ao código genético de organismos, contendo as sequências que representam moléculas de DNA, RNA e proteínas. A sua principal característica é a qualidade das sequências depositadas, sendo uma base de dados onde apenas sequências não-redundantes são adicionadas. O *RefSeq* possui atualmente mais de 16 milhões de sequências de genômicas, 13 milhões de sequências de RNA e mais de 54 milhões de sequências de proteínas[1].

Porém, quando se trata de proteínas conhecer apenas sua sequência não é o suficiente para afirmar qual é a sua função biológica (ANFINSEN, 1973). O paradigma sequência-estrutura-função diz que proteínas globulares podem exercer a sua função biológica somente

[1] Valores obtidos em novembro de 2015. http://www.ncbi.nlm.nih.gov/refseq/

ao enovelar-se em uma estrutura nativa determinada pela sua sequência de resíduos de aminoácidos (ANFINSEN, 1973; BRANDEN; TOOZE, 1999). O conhecimento sobre a estrutura tridimensional (3D) de um polipeptídeo/proteína proporciona aos pesquisadores uma importante informação para inferir a função da proteína na célula (LASKOWSKI; WATSON; THORNTON, 2003; SCHEEF; FINK, 2005). A determinação da estrutura de uma proteína é experimentalmente custosa (devido aos custos associados com técnicas de cristalografia, eletroscopia ou ressonância magnética nuclear) e muitas vezes infactível (TRAMONTANO; LESK, 2006). Atualmente no *Protein Data Bank* (PDB) (BERMAN et al., 2000), que é o mais importantes banco de dados público de estruturas 3D experimentalmente determinadas, contém aproximadamente 113 mil[2] estruturas tridimensionais. Essa discrepância entre o número de sequências e o número de estruturas 3D conhecidas em conjunto com a importância desse conhecimento influenciou no desenvolvimento de uma nova subdisciplina da Bioinformática: a Bioinformática Estrutural (ALTMAN; DUGAN, 2005). A Bioinformática Estrutural tem como principal foco a representação, armazenamento, recuperação, análises e exibição de informações estruturais de macromoléculas biológicas (ALTMAN; DUGAN, 2005).

Um dos principais desafios da Bioinformática Estrutural trata do problema de predição da estrutura 3D de proteínas. O problema de predição consiste em a partir da sequência linear de resíduos de aminoácidos predizer a estrutura 3D de proteínas, que é de fundamental importância no desenvolvimento de compostos químicos e fármacos para, por exemplo, ativar ou inibir a função proteica (CREIGHTON, 1990; DORN et al., 2014). Predizer a estrutura de um polipeptídeo/proteína, apenas a partir de sua sequência linear representa um problema desafiador no campo da Otimização Matemática sendo classificado em complexidade computacional como um problema NP-completo, devido à explosão de possíveis conformações que o processo de enovelamento de proteínas pode gerar (LATHROP, 1994; NGO; MARKS; KARPLUS, 1997; CRESCENZI et al., 1998; GUYEUX et al., 2014).

O estudo do enovelamento de proteínas, do ponto de vista biológico, pode-se dizer que possui duas principais linhas que foram exploradas: a da hipótese da termodinâmica de Anfinsen (ANFINSEN, 1973) e a do paradoxo de Levinthal (LEVINTHAL, 1968). A hipótese da termodinâmica é baseada na observação de que o processo de enovelamento de uma proteína ocorre de forma espontânea até o ponto de mais baixa energia termodinâmica (ANFINSEN, 1973). Enquanto o paradoxo de Levinthal, argumenta que uma estrutura não consegue se enovelar em um tempo factível se não tiver um "caminho" de enovelamento específico até

[2]Em novembro de 2015 esse era o número de estruturas encontrado no site do PDB: http://www.rcsb.org

sua estrutura nativa (LEVINTHAL, 1968). Afirmando assim que essa estrutura nativa não representa necessariamente o mínimo global de energia, mas sim a um mínimo de energia mais acessível no caminho cinético (LEVINTHAL, 1968). A partir dessas duas linhas, surgiu uma nova hipótese que combina essas duas visões, onde a estrutura nativa é realmente o estado termodinâmico mais estável, porém, diferente do caminho único de Levinthal, o enovelamento possui vários caminhos diferentes para chegar a essa mesma estrutura mínima (BRYNGELSON et al., 1995). Essa hipótese chamada de funil de enovelamento ou de panorama energético do enovelamento de proteínas é a que está mais próxima de uma representação do enovelamento proteico (BRYNGELSON et al., 1995). Porém, encontrar esses caminhos que levam a uma estrutura nativa ainda é um processo bastante complexo que segue sem solução, sendo necessário métodos computacionais para tentar obter esses caminhos para assim predizer a estrutura tridimensional nativa dessas proteínas.

Quando um problema como a predição da estrutura 3D de proteínas é encontrado, que não pode ser resolvido de forma ótima por nenhum método determinístico (exato) em um tempo razoável, é possível usar meta-heurísticas para achar soluções satisfatórias. Meta-heurísticas são uma das mais comuns e poderosas técnicas utilizadas em situações onde o conhecimento sobre o problema é restrito e soluções exatas não são atualmente computáveis. Uma meta-heurística pode ser formalmente definida como um processo interativo que guia uma heurística subordinada combinando de forma inteligente diferentes conceitos para explorar o espaço de busca. Neste contexto, estratégias computacionais são utilizadas para estruturar a informação disponível sobre o problema tendo como propósito encontrar, de forma eficiente, soluções ótimas e próximas a solução exata do problema investigado com um esforço computacional reduzido. No contexto do problema de predição da estrutura 3D de proteínas são vários os trabalhos que utilizam alguma meta-heurística, como Algoritmos Genéticos por exemplo, para tentar auxiliar a resolução do problema (ELOFSSON; GRAND; EISENBERG, 1995; PARK, 2005; HOQUE; CHETTY; DOOLEY, 2006; CUTELLO; NARZISI; NICOSIA, 2006; DORN; BURIOL; LAMB, 2011; DORN et al., 2013; BORGUESAN et al., 2015a). Esses algoritmos são tradicionalmente usados para encontrar apenas um resultado ótimo da função objetivo, entretanto, alguns problemas podem ter várias soluções ótimas se a função for multimodal (GLIBOVETS; GULAYEVA, 2013). Este é o caso da predição da estrutura tridimensional de proteínas, que devido ao fato de que as regras que governam o processo de enovelamento dessas proteínas serem apenas parcialmente conhecidas várias soluções podem acabar sendo consideradas ótimas (BRYNGELSON et al., 1995; WOLYNES, 2005; BOEHR; NUSSINOV; WRIGHT, 2009). Diversas técnicas foram desenvolvidas ao longo dos anos para

encontrar múltiplas soluções ótimas (global ou local) durante a execução do método. Essas abordagens normalmente trabalham com a ideia de formar e manter diversas subpopulações a fim de localizar várias possíveis soluções para um mesmo problema (JONG; ALAN, 1975; GOLDBERG; RICHARDSON, 1987; HARIK, 1995; ROY; PARMEE, 1996; WONG; LEUNG; WONG, 2010; GLIBOVETS; GULAYEVA, 2013; ISLAM; CHETTY, 2013).

Apesar dos esforços no desenvolvimento de meta-heurísticas, sempre haverá questões relacionadas a efetividade de uma determinada meta-heurística aplicada para resolver uma ampla gama de problemas. O teorema da inexistência do almoço grátis (*No-Free-Lunch*) (WOLPERT; MACREADY, 1997) diz que todos os algoritmos de otimização têm exatamente o mesmo desempenho, quando faz-se a média através de todos os infinitos problemas possíveis. No entanto, isso não quer dizer que uma determinada meta-heurística não possa ser melhor que outras quando aplicada em um problema específico (BOUSSAÏD; LEPAGNOT; SIARRY, 2013).

1.1 Objetivos

Com o aumento dos dados experimentais de proteínas, as técnicas baseadas em conhecimento para realizar a predição da estrutura 3D de proteínas estão apresentando resultados cada vez mais precisos, com um custo computacional relativamente menor que métodos baseados apenas em princípios físico-químico (LESK, 2005). Nas últimas edições do *Critical Assessment of Structure Prediction*[3] (CASP), vários dos trabalhos publicados utilizaram alguma forma de conhecimento para auxiliar no problema de predição da estrutura 3D de proteínas (KRYSHTAFOVYCH et al., 2014a; KRYSHTAFOVYCH; FIDELIS; MOULT, 2014b; NUGENT; COZZETTO; JONES, 2014). Apesar dos avanços na área de predição da estrutura 3D de proteínas, permanece pequeno o número de trabalhos que aplicam abordagens multimodais para tratar o problema de várias soluções serem consideradas ótimas. Assim, novas meta-heurísticas ainda são necessárias para tentar resolver o problema de percorrer um grande espaço de busca conformacional de uma cadeia polipeptídica considerando os vários caminhos possíveis do enovelamento proteico.

A partir destas informações, foram definidos dois objetivos específicos neste trabalho. O primeiro objetivo é a extração de conhecimento através da análise das informações da base de estruturas determinadas experimentalmente. Enquanto o segundo objetivo é utilizar o conhecimento gerado e combiná-lo com técnicas computacionais que auxiliem a predição da

[3] <http://predictioncenter.org>

estrutura 3D de proteínas.

O objetivo geral desta dissertação é o estudo do problema da Predição da Estrutura 3D de Proteínas e o desenvolvimento de uma meta-heurística com estrutura multimodal que utilize de forma inteligente informações estruturais do PDB. Vale ressaltar que o objetivo não é resolver o problema da predição da estrutura 3D de proteínas, mas desenvolver uma nova abordagem que utilize informações estruturais em conjunto com técnicas multimodais de modo a contribuir para uma futura solução do problema.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte forma: Capítulo 2, no qual são fornecidos conceitos básicos sobre proteínas que são necessários para o entendimento dos próximos capítulos descritos nesta dissertação; Capítulo 3, que apresenta a topologia de métodos relacionados à Predição de Estruturas de Proteínas apresentados no último CASP; Capítulo 4, no qual é apresentado diferentes formas de extrações de conhecimento da base de dados estruturais que foram desenvolvidas durante este trabalho; Capítulo 5, no qual são detalhados os conceitos padrões de Algoritmo Genético e abordagens Multimodais; Capítulo 6, que descreve a metodologia da nova abordagem desenvolvida nesta dissertação; Capítulo 7, no qual são apresentados e discutidos os resultados obtidos; Capítulo 8, que apresenta as conclusões obtidas e os trabalhos futuros e por fim o Capítulo 9, no qual é apresentando as publicações alcançadas durante o mestrado.

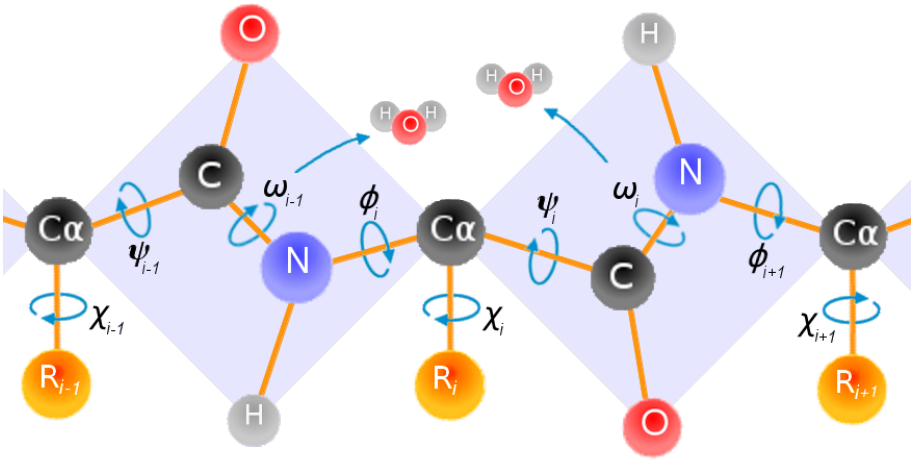
2 PROTEÍNAS

2.1 Introdução

Proteínas ou polipeptídeos são polímeros formados por longas cadeias de resíduos de aminoácidos unidos através de uma ligação peptídica (BRANDEN; TOOZE, 1999; LILJAS et al., 2001; LESK, 2010). Cada proteína é definida por sua sequência única de resíduos de aminoácidos que em condições fisiológicas se enovela em uma forma específica conhecida como estado nativo (ANFINSEN, 1973). Cada proteína é composta por uma cadeia principal igual para todos os aminoácidos e por uma cadeia lateral específica de cada resíduo (LESK, 2005).

Todos os resíduos de aminoácidos apresentam uma região em comum, independente do resíduo. Esta região é denominada esqueleto peptídico ou cadeia principal, a qual é composta por um grupo amino (H_3N^+), por um grupo carboxílico (COO^-) e por um átomo de carbono que liga estes dois grupos, denominado carbono alfa ($C\alpha$) (RICHARDSON, 1981). Cada aminoácido também possui um conjunto de átomos, denominado cadeia lateral (ou radical "R"), ligado ao $C\alpha$. É a cadeia lateral que possibilita a diferenciação entre os aminoácidos. A ligação peptídica é formada quando o grupo carboxílico de um resíduo reage com o grupo amino de outro resíduo, liberando uma molécula de água (BRANDEN; TOOZE, 1999; LESK, 2005). Essa ligação química é apresentada na Figura 2.1, onde é possível verificar a formação de duas ligações peptídicas e a liberação de suas respectivas moléculas de água (H_2O).

Figura 2.1: Ângulos de torção *phi* (ϕ), *psi* (ψ) e *ômega* (ω) presentes na cadeia principal em uma ligação peptídica. O número de ângulos *chi* (χ) dependem do resíduo de aminoácido e é representado na figura pelo grupo R

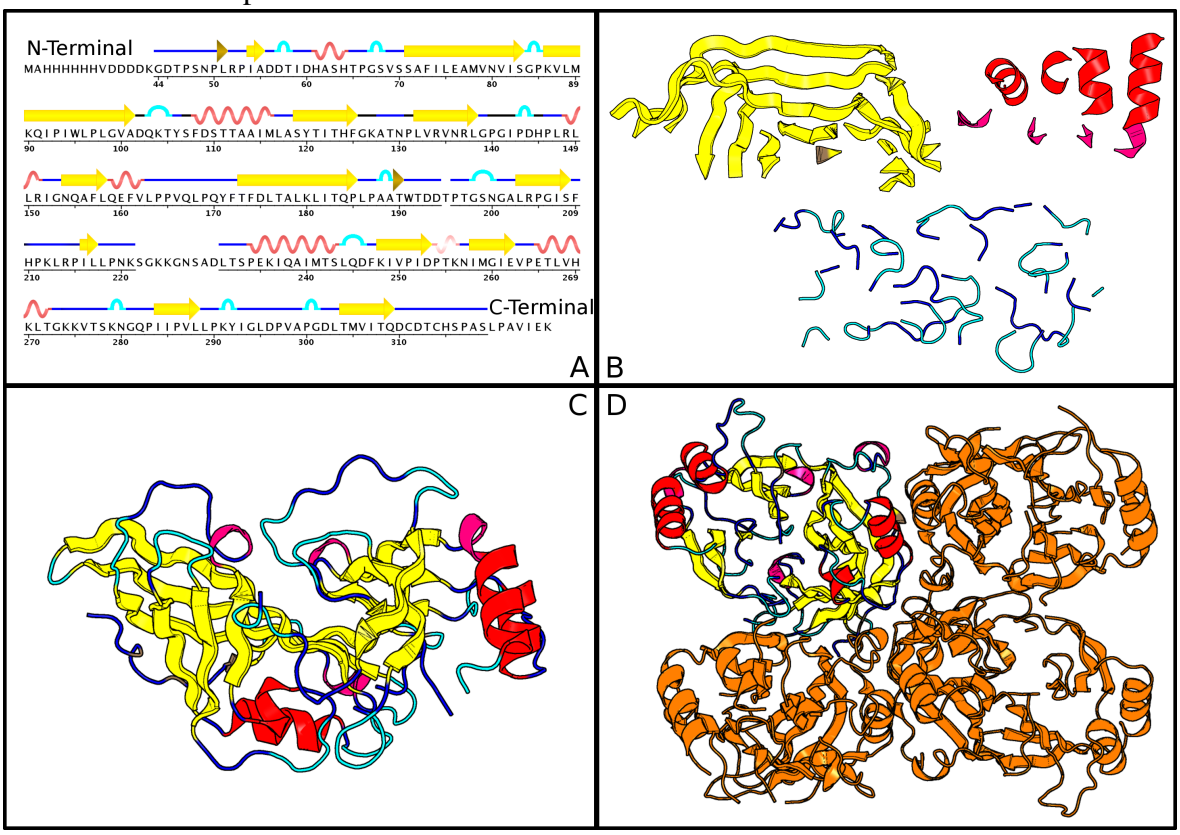


Fonte: Adaptado de Borguesan et al. (2015a).

2.2 Níveis Estruturais

Lesk, (2005) apresenta a divisão das proteínas em quatro níveis estruturais: Estrutura Primária, Estrutura Secundária, Estrutura Terciária e Estrutura Quaternária. Cada um destes níveis estruturais são apresentados nas próximas seções e estão representados na Figura 2.2.:

Figura 2.2: Representação gráfica dos 4 níveis estruturais da proteína depositada no PDB com o código 4LDB. A - Estrutura Primária e Estrutura Secundária na forma linear; B - Estrutura Secundária e seu arranjo espacial; C - Estrutura Terciária; D - Estrutura Quaternária. As estruturas em tons de amarelo representam folhas, em tons de vermelho representam hélices e em tons de azul representam voltas



Fonte: do autor (2016).

2.2.1 Estrutura Primária

A estrutura primária de uma proteína é descrita pela sequência linear de resíduos de aminoácidos. O início da estrutura primária de uma proteína corresponde a sua região N−terminal e o final da estrutura primária é determinada pela região C−terminal (BRANDEN; TOOZE, 1999). A Figura 2.2-A apresenta a sequência de aminoácidos que representam a estrutura primária da proteína com código 4LDB no PDB. A Tabela 2.1 apresenta os 20

aminoácidos principais que constituem as proteínas da maioria dos organismos com suas respectivas abreviações e características das cadeias laterais (SCHEEF; FINK, 2005).

Tabela 2.1: Lista dos 20 aminoácidos, suas abreviações e características das cadeias laterais

Aminoácido	Abreviação		Ângulos Chi	Polaridade
	3 Letras	1 Letra		
Alanina	ALA	A	0	Não-Polar
Arginina	ARG	R	4	Positivo
Asparagina	ASN	N	2	Neutro
Aspartato	ASP	D	2	Negativo
Cisteína	CYS	C	3	Neutro
Fenilalanina	PHE	F	2	Aromático
Glicina	GLY	G	0	Não-Polar
Glutamato	GLU	E	3	Negativo
Glutamina	GLN	Q	3	Neutro
Histidina	HIS	H	2	Positivo
Isoleucina	ILE	I	2	Não-Polar
Leucina	LEU	L	2	Não-Polar
Lisina	LYS	K	4	Positivo
Metionina	MET	M	3	Não-Polar
Prolina	PRO	P	1	Não-Polar
Serina	SER	S	1	Neutro
Tirosina	TYR	Y	2	Aromático
Treonina	THR	T	1	Neutro
Triptofano	TRP	W	2	Aromático
Valina	VAL	V	1	Não-Polar

Fonte: Adaptado de Lehninger et al. (2005).

Conforme comentado no Capítulo 1 desta dissertação, uma das principais base de dados com Estruturas Primárias de proteínas é o *RefSeq* (PRUITT et al., 2002). Esta base contém mais de 54 milhões de sequências de aminoácidos, não redundantes, proveniente de diversos organismos.

2.2.2 Estrutura Secundária

Frequentemente, a disposição espacial dos aminoácidos da estrutura primária apresenta algumas regularidades (TRAMONTANO; LESK, 2006). Este padrão no arranjo estrutural das proteínas é chamado de estrutura secundária. Essas estruturas são definidas pela presença de ligações de hidrogênio entre o grupo carboxílico e o grupo amino de um

polipeptídeo (RICHARDSON, 1981; LESK, 2005). Os arranjos mais conhecidos e estáveis são as conformações em hélice (PAULING; COREY; BRANSON, 1951) e folha (PAULING; COREY, 1951).

Nas conformações em forma de hélices, a cadeia polipeptídica se enovela no formato helicoidal. As ligações de hidrogênio desta conformação organizam-se na parte interna do cilindro e as cadeias laterais dos resíduos ficam voltada para o exterior da hélice (RICHARDSON, 1981). Quando as cadeias polipeptídicas estão arranjadas lado a lado e formam ligações de hidrogênio entre estes segmentos, são formadas as conformação chamadas de folhas (RICHARDSON, 1981). As ligações dessas duas estruturas mais estáveis (hélices e folhas) ocorrem através de um outro tipo de estrutura secundária chamadas de voltas. Este tipo de conformação em geral faz a conexão entre os tipos de hélices e folhas, permitindo o arranjo entre eles e auxiliando na conformação da estrutura terciária das proteínas (BRANDEN; TOOZE, 1999). A Figura 2.2-A apresenta a estrutura primária da proteína com código 4LDB no PDB em conjunto com a sua estrutura secundária no formato linear enquanto a Figura 2.2-B apresenta estas conformações estruturais de hélices (estruturas em tons de vermelho), folhas (estruturas em tons de amarelo) e voltas (estruturas em tons de azul) na sua representação espacial.

Atualmente existem diversos algoritmos para fazer a predição da estrutura secundária apenas a partir de sua sequência primária, como exemplo pode-se citar: PSIPRED (JONES, 1999; BUCHAN et al., 2013), SPINE-X (FARAGGI et al., 2012), SCORPION (YASEEN; LI, 2014a; YASEEN; LI, 2014b) e SPIDER2 (HEFFERNAN et al., 2015). O método mais conhecido e utilizado atualmente é o PSIPRED (ZHANG, 2008) que incorpora duas redes neurais do tipo *feed-forward* fazendo uma análise sobre os dados de saída do alinhamento de sequência feito pelo PSI-BLAST (ALTSCHUL et al., 1997). Nos últimos testes realizados o PSIPRED tem alcançado aproximadamente 80% de acurácia em seus resultados (HEFFERNAN et al., 2015).

A literatura também apresenta várias ferramentas que fazem a atribuição da estrutura secundária de proteínas e os dois mais conhecidos e amplamente utilizados pela comunidade científica são o *Define Secondary Structure of Proteins* (DSSP) (KABSCH; SANDER, 1983a; TOUW et al., 2015) e o *STRuctural IDEntification* (STRIDE) (FRISHMAN; ARGOS, 1995; HEINIG; FRISHMAN, 2004). Porém esses algoritmos necessitam da estrutura tridimensional da proteína para poder analisar suas ligações de hidrogênio e atribuir a estrutura secundária de cada aminoácido da proteína alvo. Nas próximas seções, os métodos de atribuição da estrutura secundária STRIDE e DSSP, utilizados nesta dissertação, são detalhados.

2.2.2.1 STRIDE

O método STRIDE, desenvolvido por Frishman e Argos (1995), usa padrões de ligações de hidrogênio em conjunto com os ângulos diedros da cadeia principal para atribuir a estrutura secundária. O STRIDE faz sua atribuição classificando os aminoácidos em sete diferentes estruturas secundárias que são:

- Dois tipos de folhas: ponte isolada (`B` ou `b`); folha estendida ou folha-β (`E`).
- Três tipos de hélices: hélice-α (`H`); hélice-3_{10} (`G`); hélice-π (`I`);
- Um tipo de volta: volta (`T`);
- Quando não identifica nenhum desses padrões o resíduo de aminoácido é atribuído como região desordenada (`C`).

Para atribuir cada uma dessas estruturas secundárias o STRIDE utiliza algumas regras em conjunto com uma função empírica para calcular a energia da formação da ligação de hidrogênio (ZHANG; SAGUI, 2015). Para atribuir as hélice-α (`H`) o algoritmo analisa a estrutura tridimensional e verifica se contém pelo menos duas ligações de hidrogênios consecutivas entre os resíduos `i` $\rightarrow$ `i + 4`. O padrão é ignorado caso algum dos ângulos diedros internos estiver em uma região desfavorável na análise físico-química (FRISHMAN; ARGOS, 1995; ANDERSEN; ROST, 2005; ZHANG; SAGUI, 2015). Essa definição também é usada para os padrões de hélice-3_{10} (`G`) com ligações de hidrogênios entre os resíduos `i` $\rightarrow$ `i + 3` e para o padrão de hélice-π (`I`) com ligações de hidrogênios entre os resíduos `i` $\rightarrow$ `i + 5`.

Para fazer a atribuição da estrutura secundária folha-β (`E`) o algoritmo do STRIDE procura por, no mínimo, dois pares de resíduos com ligações de hidrogênio entre o grupo amino de um aminoácido com o grupo carboxílico de outro aminoácido. Já a ponte isolada (`B`) é atribuída quando apenas um par de resíduos possui a ligação de hidrogênio, exceto para as ligações de hidrogênio em que um dos resíduos da ligação está alinhado para o lado externo da proteína, o qual é designado como uma variação da ponte isolada (`b`) (FRISHMAN; ARGOS, 1995; ZHANG; SAGUI, 2015).

As regiões de volta (`T`) são atribuídas baseadas nas definições feitas por Richardson (1981) analisando os ângulos diedros internos (*phi* e *psi*), dos resíduos `i + 1` e `i + 2` (FRISHMAN; ARGOS, 1995; ZHANG; SAGUI, 2015). Quando o algoritmo do STRIDE não identifica nenhum dos padrões descritos acima o resíduo de aminoácido é atribuído como uma região desordenada (`C`).

2.2.2.2 DSSP

O DSSP, desenvolvido por Kabsch e Sander (1983a), também utiliza padrões das ligações de hidrogênios, porém é combinado com um modelo eletrostático para fazer sua atribuição (ZHANG; SAGUI, 2015). O DSSP faz sua atribuição classificando os aminoácidos em oito diferentes estruturas secundárias que são:

- Dois tipos de folhas: ponte isolada (`B`); folha estendida ou folha-β (`E`).
- Três tipos de hélices: hélice-α (`H`); hélice-3_{10} (`G`); hélice-π (`I`);
- Dois tipos de voltas: volta (`T`); dobra ou curvatura (`S`);
- Quando não identifica nenhum desses padrões o resíduo de aminoácido é atribuído como região desordenada (`C`).

As definições de hélices do DSSP são bem similares com as do STRIDE com exceção na definição das ligações de hidrogênios. Outra diferença é que o DSSP não aplica as definições de hélices nos resíduos localizados nas duas extremidades das ligações de hidrogênios das hélices (ZHANG; SAGUI, 2015).

Para a atribuição da estrutura secundária folha-β (`E`) é utilizada a mesma definição do STRIDE que precisa no mínimo dois pares de resíduos com ligações de hidrogênio e no mínimo um par para a atribuição da ponte isolada (`B`). Entretanto, o DSSP permite a diferenciação entre folhas no mesmo sentido (paralelas) e em sentidos opostos (antiparalelas) (KABSCH; SANDER, 1983a; ZHANG; SAGUI, 2015).

Para atribuir a estrutura secundária de volta (`T`) é verificado se existe uma ligação de hidrogênio entre os resíduos `i` e `i + n` (`i` $\rightarrow$ `i + n`) (ZHANG; SAGUI, 2015). As regiões de dobra (`S`) são atribuídas quando o ângulo entre átomos $C\alpha_{i-2}-C\alpha_i-C\alpha_{i+2}$ for menor que 110° (KABSCH; SANDER, 1983a; ZHANG; SAGUI, 2015). Quando o algoritmo do DSSP não identifica nenhum dos padrões descritos acima o resíduo de aminoácido também é atribuído como região desordenada (`C`).

2.2.3 Estrutura Terciária

A estrutura terciária de uma proteína é representada pelo arranjo tridimensional (3D) dos átomos dos aminoácidos que compõem a estrutura primária (RICHARDSON, 1981). É a forma 3D de uma proteína composta pela combinação das estruturas secundárias arranjadas no espaço. Essa estrutura formada também é chamada de estrutura nativa ou estrutura funcional (SCHEEF;

FINK, 2005). A Figura 2.2-C apresenta um exemplo da estrutura terciária da proteína com código no PDB de 4LDB, onde é possível identificar as conformações de hélices, folhas e voltas arranjadas tridimensionalmente.

Proteínas são geralmente classificadas em três grupos, baseado em sua forma e solubilidade. São as proteínas Fibrosas, que são normalmente esticadas e não são solúveis em água; As proteínas Globulares, que são estruturas em um formato que lembra uma esfera, com os aminoácidos hidrofóbicos na parte interna da esfera e os hidrofílicos na parte externa da esfera, sendo uma estrutura bastante solúvel; E por último as proteínas de membranas, que são proteínas não solúveis que interagem com membranas biológicas.

As proteínas globulares ainda possuem subdivisões baseado na sua conformação estrutural. Essa classificação é feita em quatro grupos de topologias. O primeiro grupo são as estruturas α que são proteínas que só contém hélices em sua conformação. O segundo grupo são as estruturas β que representam apenas estruturas com conformação de folhas. O terceiro grupo são as estruturas α/β que alterna entre estruturas em hélices e em folhas de forma regular. E o último grupo são as estruturas α+β que alterna entre estruturas em hélices e em folhas de forma irregular. A estrutura de código PDB 4LDB apresentada na Figura 2.2 representa a proteína codificada pelo vírus ebola (BORNHOLDT et al., 2013). Esta estrutura é classificada como α+β devido ao fato de sua conformação estrutural alternar entre hélices e folhas de maneira irregular.

Conhecendo a estrutura tridimensional (estrutura terciária) de uma proteína é possível analisar ou inferir qual é a função dessa proteína na célula de um organismo (RICHARDSON, 1981; BRANDEN; TOOZE, 1999). Esse arranjamento 3D também permite identificar o sítio ativo de enzimas, encontrar as cavidades para atracamento molecular, entre outras (LEHNINGER; NELSON; COX, 2005; SCHEEF; FINK, 2005; LESK, 2010).

2.2.3.1 Métodos Experimentais para Determinar a Estrutura Tridimensional

Nas estruturas depositadas no PDB, os métodos utilizados para poder determinar a estrutura tridimensional de uma proteína em 99% dos casos foram através de cristalografia por difração de Raios-X (*X-Ray*) ou espectroscopia por Ressonância Magnética Nuclear (RMN). O restante foi determinado por Microscopia Eletrônica, métodos híbridos ou outros (BERMAN et al., 2000).

Cristalografia por Difração de Raios-X:

A maioria das estruturas depositadas no PDB, aproximadamente 90%, foram obtidas através do método de cristalografia por difração de Raios-X. Esta abordagem consegue provir informações com bastante detalhes sobre o posicionamento atômico de proteínas, permitindo a visualização de ligantes, inibidores, íons e outras moléculas que foram incorporadas no cristal. Entretanto, esse processo de cristalização possui dificuldades para fazer a determinação de proteínas que não são rígidas, pois o método depende de muitas moléculas alinhadas na mesma orientação (MCPHERSON, 2004; BERMAN et al., 2000). A acurácia das estruturas determinadas pelo processo de cristalografia por difração de Raios-X é estritamente dependente da qualidade com que esses cristais foram gerados. Quando cristais de boa qualidade são gerados, é possível ter uma maior confiança de que os átomos da estrutura refletem corretamente a estrutura da proteína (BERMAN, 2008; MCPHERSON, 2004). Com isso, duas métricas de qualidade são bastante aceitas pela comunidade científica que são a Resolução, que é a quantidade de detalhes que pode ser observado na estrutura determinada, e o Fator-R, que mede a correspondência dos padrões de difração simulado contra os padrões observados experimentalmente (BERMAN et al., 2000; HOVMÖLLER; ZHOU; OHLSON, 2002). Anualmente, a maioria das novas estruturas depositadas no PDB ainda são determinadas através do método de cristalografia por difração de Raios-X (BERMAN et al., 2012).

Espectroscopia por Ressonância Magnética Nuclear (RMN):

A determinação da estrutura tridimensional de proteínas também pode ser feita através do método de RMN que representa pouco mais de 9% das estruturas depositadas no PDB. Esse baixo número de estruturas acontece devido à limitação do método em só conseguir determinar estruturas de proteínas pequenas, em razão de que proteínas maiores apresentam problemas de picos sobrepostos no espectro do RMN (BERMAN et al., 2000). A maior vantagem da determinação de estruturas utilizando o método de RMN é que o processo é feito com a proteína em solução, ao contrário dos outros métodos, o que permite a determinação de proteínas flexíveis. A estrutura determinada pelo RMN normalmente inclui um conjunto de modelos, que possuem uma grande similaridade para regiões mais rígidas e resultados diferentes para regiões mais flexíveis da proteína (WUTHRICH, 1989; BERMAN et al., 2000).

2.2.4 Estrutura Quaternária

A estrutura quaternária de uma proteína é o arranjamento de várias estruturas terciárias. As subunidades que compõem uma estrutura quaternária podem ser oriundas de distintas cadeias polipeptídicas ou de várias subunidades contendo a mesma estrutura primária (RICHARDSON, 1981). Atualmente, o maior banco de dados de estruturas em nível terciário e quaternário, que foram determinadas experimentalmente é o *Protein Data Bank* (PDB) (BERMAN et al., 2000). A Figura 2.2-D apresenta uma estrutura quaternária composta de quatro estruturas terciárias iguais a Figura 2.2-C.

2.3 Representação da Estrutura 3D de Proteína

Há muitas formas de se representar a estrutura tridimensional de uma proteína computacionalmente. Esta representação está relacionada ao número de variáveis que são usadas para representar a proteína. As representações mais detalhadas são as que incluem todos os átomos da proteína e as moléculas de solvente circundantes, por exemplo, moléculas de água (H_2O). Usar a abordagem que usa todos os átomos (*all-atom*) para representar a proteína é o método mais detalhado, porém, dependendo do objetivo pode ser computacionalmente custoso pela explosão de variáveis (SCHEEF; FINK, 2005).

Outra abordagem utilizada e com um custo computacional reduzido é a de átomo unido (*united-atom*) (LILJAS et al., 2001). Esta abordagem consiste em transformar os grupos específicos de átomos em pseudo-átomos, reduzindo assim o número de variáveis a serem descritas (SCHEEF; FINK, 2005). A abordagem *coarse-grained* tem o mesmo princípio, porém ela faz um número maior de transformações de átomos em partículas reduzindo ainda mais o custo computacional que a abordagem de átomo unido (CHIVIAN et al., 2005). Porém essas abordagens, devido a essas transformações, podem perder informações importantes referentes ao posicionamento correto dos átomos.

A maioria dos métodos de enovelamento de proteínas usam algum modelo geométrico simplificado, pois reduz o número de variáveis e com isso permite um menor custo computacional (MOULT, 2005). Por isso, a maioria dos trabalhos utiliza o modelo com ângulos de torção que diminuem o número de variáveis se comparado ao modelo de posicionamento 3D de cada átomo, mas conseguem manter um certo realismo ao modelo (HOVMÖLLER; ZHOU; OHLSON, 2002; SCHEEF; FINK, 2005; LESK, 2005; BORGUESAN et al., 2015a).

Sabe-se que a partir da ligação peptídica entre aminoácidos três ângulos de torção são

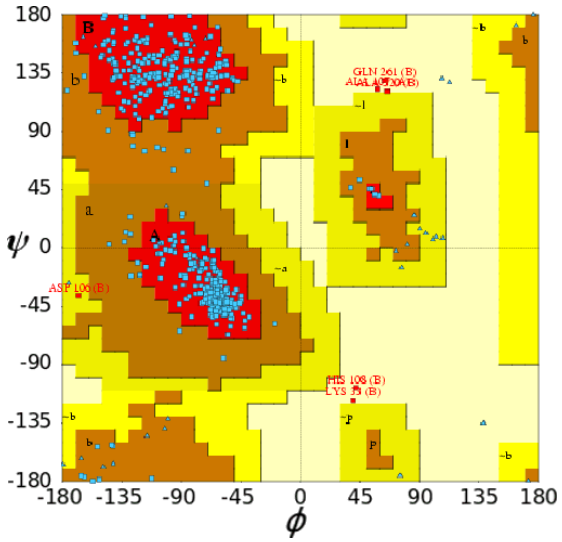
formados na cadeia principal, esses ângulos são chamados de *phi* (ϕ), *psi* (ψ) e *ômega* (ω). A Figura 2.1 mostra os ângulos de torção que ocorrem em uma ligação peptídica.

O ângulo *ômega* tem valores limitados próximos ou iguais a $-180°$ ou a $180°$, devido ao caráter parcial de dupla ligação que a mantém praticamente plana (RICHARDSON, 1981). Enquanto, os ângulos *phi* e *psi* são livres para fazer qualquer ângulo de rotação de $-180°$ até $180°$. Em razão dessa liberdade, os ângulos *phi* e *psi* são considerados os principais responsáveis na determinação da organização espacial (conformação) da estrutura da proteína (RICHARDSON, 1981). Métodos como o STRIDE e o DSSP além de fazerem a atribuição da estrutura secundária, também computam os ângulos diedros *phi* e *psi* de estruturas no formato do PDB.

Entretanto, mesmo com essa liberdade nos valores de ângulos *phi* e *psi*, alguns pares de combinações ϕ e ψ são proibidos devido à interferências estéricas entre os átomos da cadeia lateral (HOVMÖLLER; ZHOU; OHLSON, 2002). Os valores permitidos e proibidos para os ângulos de torção ϕ e ψ são graficamente demonstrados pelo mapa, ou gráfico, de Ramachandran (RAMACHANDRAN; SASISEKHARAN, 1968). A Figura 2.3 mostra dois exemplos de adaptações do mapa de Ramachandran (LASKOWSKI et al., 1996).

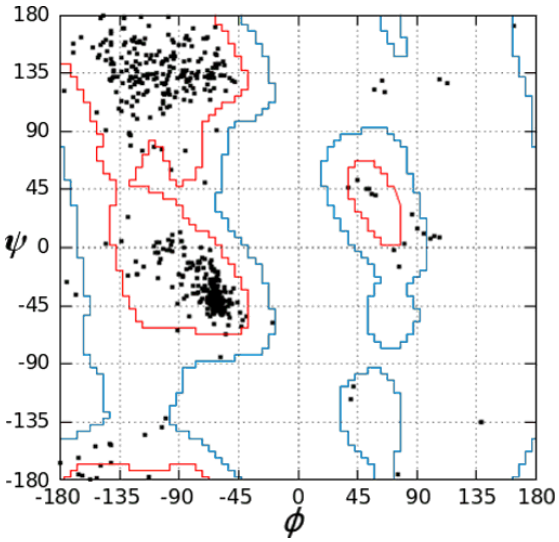
Figura 2.3: Exemplos de mapas de Ramachandan. (a) gerado pelo software PROCHECK (LASKOWSKI et al., 1996). (b) gerado com base no trabalho de Lovell et al. (2003)

(a)



Fonte: Laskowski et al. (1996).

(b)



Fonte: Lovell et al. (2003).

A Figura 2.3a mostra o gráfico de Ramachandran gerado pelo software PROCHECK (LASKOWSKI et al., 1996) com o resultado da proteína cujo código no

PDB é 3JLS. O mapa é separado em quatro regiões: região mais favorável; região permitida; região ainda aceitável; região proibida. A região mais favorável é representada pela área vermelha, a região permitida é apresentada com a cor marrom, a região ainda aceitável é apresentada pela cor amarelo e a região não permitida pela cor amarelo claro. Os pontos em azul representam os aminoácidos em regiões permitidas, já os pontos em vermelho representam aminoácidos em regiões proibidas. Essas regiões representam a probabilidade de ocorrer choques estereoquímicos entre resíduos de aminoácidos, onde a região proibida representa uma grande probabilidade de choques enquanto a região mais favorável uma baixa probabilidade de choques estereoquímicos.

A Figura 2.3b mostra uma adaptação atualizada do mapa de Ramachandran apresentado por Lovell et al. (2003). O mapa é separado em três de regiões: região mais favorável (vermelho), região permitida (azul) e região proibida (branco). Os pontos neste mapa também são referentes aos aminoácidos da proteína cujo código no PDB é 3JLS.

O posicionamento das cadeias laterais também é definido por um conjunto de ângulos de torção, denominados *chi*'s (χ). O número de ângulos χ depende do número de átomos desta cadeia lateral (RICHARDSON, 1981). Os ângulos de torção χ também podem variar entre $-180°$ até $180°$ e são apresentados na Figura 2.1. Esses ângulos influenciam diretamente na conformação global da proteína, devido às forças de atração e repulsão presentes em todos os átomos (DUNBRACK JR; KARPLUS, 2003). Além disto, esta conformação influência na função que a proteína exerce na célula de um organismo.

2.4 Funções das Proteína

Proteínas possuem uma grande variedade de diferentes tipos de estruturas tridimensionais que consequentemente, devido à função estar diretamente relacionada com a estrutura 3D, também possuem uma grande variação de funções que podem exercer dentro de um organismo (ALBERTS et al., 2007). Algumas das principais funções exercidas pelas proteínas segundo Lesk (2010) são:

Proteínas estruturais: são as mais abundantes dentre todos os tipos de proteínas do corpo humano. Essas proteínas fibrosas possuem a função de promover a sustentação estrutural aos tecidos do organismo. Exemplos vão desde a queratina da camada mais externa da nossa pele até a camada de revestimento dos vírus;

Proteínas catalisadoras: possuem a principal função de acelerar a velocidade de uma reação química do metabolismo. As enzimas são exemplos de proteínas do tipo catalisadoras;

Proteínas de defesa: possuem a função de reconhecer e repelir agentes patogênicos invasores. Seu principal exemplo são os anticorpos;

Proteínas reguladoras: como o nome já diz, são proteínas que tem como função regular atividades metabólicas no organismo bem como controlar a transcrição do código genético. O exemplo mais conhecido de proteína reguladora é a insulina;

Proteínas sensoriais: atuam como sensores de sinais gerados dentro de nossos corpos e de sinais externos, permitindo que células mudem seu comportamento caso necessário;

Proteínas transportadoras: possuem a função de transportar moléculas para dentro e para fora das células. O maior exemplo é a hemoglobina que transporta o oxigênio para os tecidos do corpo;

Existem várias outras funções que proteínas podem exercer e até mesmo proteínas que podem exercer mais de uma função, como o caso de proteínas virais. Entretanto, independente da proteína, conhecer a sua estrutura tridimensional *a priori* é de extrema importância para determinar sua função individual (BRANDEN; TOOZE, 1999; SCHEEF; FINK, 2005; LESK, 2010).

2.5 Base de Dados

Como mencionado na Seção 2.2, as proteínas são normalmente apresentadas em quatro níveis estruturais. Algumas das principais base de dados que representam a estrutura primária de proteínas são o *GenBank* (BENSON et al., 2013) e o *RefSeq* (PRUITT et al., 2002). Para a estrutura tridimensional de proteínas, a principal base de dados é o PDB (BERMAN et al., 2000).

O banco de dados *GenBank*, criado em 1992 pelo *National Center for Biotechnology Information* (NCBI), é uma base de dados de acesso público sendo uma das mais completas no que diz respeito à sequência genética (BENSON et al., 2013). Porém a abordagem de submissão destas sequências no *GenBank* permite que ocorra adição de sequências redundantes. Atualmente o número de sequências depositadas no *GenBank* é mais de 188 milhões[1].

A base de dados *RefSeq*, criado em 2002 também pelo NCBI, contém as sequências que representam moléculas de DNA, RNA e proteínas (PRUITT et al., 2002). Sua principal diferença para o *GenBank* é que apenas sequências não-redundantes são adicionadas na base de dados. O *RefSeq* possui atualmente mais de 54 milhões[2] de sequências de proteínas não-

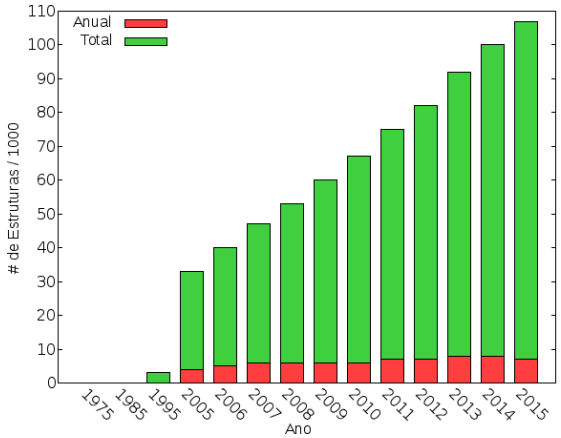
[1] Valor obtido do *GenBank* em novembro de 2015: http://www.ncbi.nlm.nih.gov/genbank/
[2] Valor obtido do RefSeq em novembro de 2015: http://www.ncbi.nlm.nih.gov/refseq/

redundantes.

O PDB (*Protein Data Bank*), foi estabelecido em 1971, em conjunto com o *Cambridge Crystallographic Data Centre* (CCDC) e o *Brookhaven National Laboratory* (BNL), porém somente em 1976 que as primeiras estruturas foram depositadas (BERMAN, 2008). Em 1999, já com quase 10 mil estruturas, a administração do PDB passou para o consorcio *Research Collaboratory of Structural Bioinformatics* (RCSB PDB), composto por *Rutgers, The State University of New Jersey*, o *San Diego Supercomputer Center* e o *National Institute of Standards and Technology* (BERMAN et al., 2000). Em 2003, com pouco mais de 21 mil estruturas, o *Worldwide PDB* (wwPDB) formalizou a colaboração internacional entre o RCSB PDB, o *Macromolecular Structure Database* e *PDB Japan* (PDBj) (BERMAN, 2008). Isto abriu o caminho para o PDB se transformar no maior e mais importante banco de dados público de estruturas tridimensionais, possuindo atualmente mais de 105 mil[3] estruturas 3D de proteínas. A Figura 2.4 apresenta o crescimento no número de estruturas disponíveis no PDB desde a sua criação em 1972, até novembro de 2015.

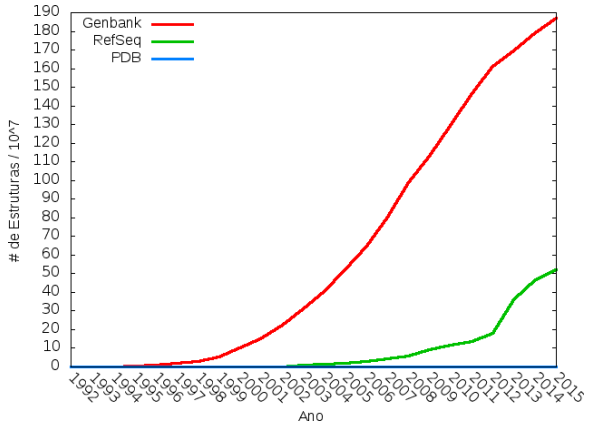
Entretanto, é possível observar uma grande lacuna entre o número de sequências de proteínas não-redundantes depositadas no *RefSeq* e o número de proteínas com a estrutura 3D conhecida. Se considerarmos apenas o número de estruturas 3D também não redundantes essa lacuna piora, pois enquanto o *RefSeq* possui mais de 54 milhões de sequências de proteínas não redundantes o PDB possui apenas 64 mil estruturas 3D de proteínas sem redundância, o que representa apenas 0.11% de conhecimento. A Figura 2.5 apresenta essa diferença entre

Figura 2.4: Gráfico de crescimento do número de proteínas depositadas no PDB desde 1972 até novembro de 2015



Fonte: Adaptado de Berman et al. (2012).

Figura 2.5: Gráfico de crescimento do número de sequências desde 1992 até novembro de 2015



Fonte: do autor (2016).

[3]Em novembro de 2015 esse era o número de estruturas encontrado no site do PDB: http://www.rcsb.org/pdb/

o número de sequências genéticas no *GenBank* (mais de 180 milhões) e o número de sequências de proteínas no *RefSeq* (mais de 54 milhões) quando comparados com o número de estruturas tridimensionais não redundantes no PDB (aproximadamente 64 mil). Devido a essa lacuna, vários métodos são desenvolvidos todos anos para tentar predizer a estrutura tridimensional de proteínas sem depender da custosa e muitas vezes infactível determinação experimental dessas proteínas.

2.6 Resumo do Capítulo

Neste capítulo foram apresentados os conceitos básicos da Bioinformática Estrutural que são necessários para o entendimento dos próximos capítulos descritos nesta dissertação. Entre os conceitos citados foram discutidas as proteínas e as suas unidades primárias (os aminoácidos), os níveis hierárquico estruturais, algumas das principais funções exercidas pelas proteínas, as representações de proteína mais conhecidas bem como as principais bases de dados de proteínas. No próximo capítulo são abordadas as principais metodologias para predição da estrutura 3D de proteínas que se destacaram nas últimas competições mundiais de predição e a topologia desses métodos.

3 MÉTODOS DE PREDIÇÃO DA ESTRUTURA 3D DE PROTEÍNAS

3.1 Introdução

Como descrito no capítulo anterior, proteínas são longas sequências formadas através da combinação de 20 diferentes resíduos de aminoácidos que em condições fisiológicas adotam uma estrutura 3D única (ANFINSEN, 1973; CREIGHTON, 1990). O conhecimento da estrutura de proteínas possibilita a investigação de processos biológicos de forma mais direta, com alta resolução e grande nível de detalhes (LODISH et al., 1990). O paradigma sequência-estrutura-função diz que proteínas globulares podem obter a sua função biológica somente ao enovelar-se em uma estrutura única determinada pela sua sequência de resíduos de aminoácidos (RICHARDSON, 1981). O conhecimento sobre a estrutura 3D de um polipeptídeo/proteína proporciona aos pesquisadores importante informação para inferir a função da proteína na célula permitindo também o desenvolvimento de fármacos para ativar ou inibir essa função (BRANDEN; TOOZE, 1999).

Determinar experimentalmente a estrutura de uma proteína é um processo demorado e custoso (devido aos custos associados com a cristalografia, eletroscopia ou ressonância magnética nuclear) (TRAMONTANO; LESK, 2006). A dificuldade em determinar a estrutura 3D de proteínas gerou uma enorme discrepância entre o volume de dados de sequências de resíduos de aminoácidos e o número de estruturas 3D conhecidas. Esta situação não somente ilustra a necessidade, mas também motiva futuras pesquisas no campo de desenvolvimento de métodos computacionais para a predição da estrutura tridimensional de proteínas (DORN, 2012).

Ao longo dos últimos anos, diversos métodos foram propostos como solução ao problema da predição da estrutura tridimensional de proteínas. Estes métodos podem ser divididos em quatro classes (DORN et al., 2014): 1) métodos de primeiros princípios que não utilizam informação da base (OSGUTHORPE, 2000); 2) métodos de primeiros princípios que utilizam informação da base experimental (SRINIVASAN; ROSE, 2002; ROHL et al., 2004); 3) métodos de reconhecimento de enovelamento (*threading*) (BOWIE; EISENBERG, 1994; BRYANT; ALTSCHUL, 1995); e 4) métodos de modelagem comparativa (SÁNCHEZ; SALI, 1997; MARTÌ-RENOM et al., 2000).

O problema da determinação do enovelamento de uma proteína é classificado em complexidade computacional como um problema NP completo, isto é, ele está entre os mais difíceis problemas em termos de requisitos computacionais (CRESCENZI et al., 1998).

Esta complexidade deve-se ao fato que o processo de enovelamento de uma proteína é extremamente seletivo. Uma longa cadeia de resíduos de aminoácidos acaba assumindo um imenso número de conformações (LEVINTHAL, 1968). Métodos de primeiros princípios (que não utilizam informação da base experimental) podem obter novos e desconhecidos enovelamentos de proteínas. Entretanto, a complexidade e a alta dimensionalidade do espaço de busca conformacional, mesmo para uma pequena molécula torna o problema infactível computacionalmente. A simulação do processo de enovelamento em nível atômico, como o utilizado em Dinâmica Molecular (DM) (GUNSTEREN; BERENDSEN, 1990) não é possível (para proteínas grandes que são de interesse médico e científico) devido ao alto custo computacional, apesar dos esforços visando a construção de plataformas de alto desempenho. Por outro lado, a modelagem comparativa não apresenta tais problemas; entretanto, estes métodos podem somente predizer estruturas de proteínas com sequência que são similares ou bem próximas com a sequência de outras proteínas com estrutura já conhecida. Métodos de reconhecimento de enovelamento via *threading* também estão limitados à base de dados de enovelamentos derivada do PDB. Porém, extrair conhecimento de bases com dados biológicos pode ser uma tarefa complicada (KRYSHTAFOVYCH et al., 2014a; NUGENT; COZZETTO; JONES, 2014).

Independente da classificação do método, uma maneira frequentemente utilizada para validação das abordagens desenvolvidas é testar o poder de predição do método sobre estruturas que foram recentemente adicionadas no PDB. Essas estruturas foram determinadas experimentalmente, mas ainda não tiveram seus resultados publicados. Essa proteínas ficam com a estrutura no estado de *espera* permitindo assim que métodos tentem fazer a predição da sua estrutura 3D sem conhecimento prévio do seu enovelamento. A cada dois anos, desde 1994, uma competição mundial é feita para descobrir qual método tem uma maior acurácia em um conjunto de proteínas que estão no estado de *espera* no PDB, chamado de *Critical Assessment of Structure Prediction*[1] (CASP) (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014b).

3.2 CASP

O CASP possui apenas duas classes de estruturas que são utilizadas na competição, são as *Free Modeling* (FM) e as *Template-Based Modeling* (TBM). O primeiro grupo, são as sequências que não possuem nenhuma outra proteína similar a ela que já possua sua estrutura tridimensional conhecida. O segundo grupo, corresponde as sequências que possuem estruturas

[1] <http://predictioncenter.org>

tridimensional relacionadas conhecidas. Essa divisão é realizada, pois proteínas que já possuem estruturas com uma alta taxa de similaridade de sequência (TBM), são consideradas proteínas de menor dificuldade de predição (ZHANG, 2007).

A Tabela 3.1 apresenta os resultados do último CASP (11 edição) realizado em 2014. Os servidores de predição Zhang-Server (ZHANG et al., 2015, in press), QUARK (XU; ZHANG, 2012), BAKER-ROSETTASERVER (KIM; CHIVIAN; BAKER, 2004) e RaptorX (PENG; XU, 2011) foram os métodos que receberam um maior destaque nas explicações a seguir por serem os servidores que obtiveram os melhores resultados nas últimas edições do CASP (KRYSHTAFOVYCH et al., 2014a; NUGENT; COZZETTO; JONES, 2014).

Tabela 3.1: Resultados dos servidores no último CASP (11 edição) mostrando a posição e o nome do método/servidor. Os métodos com * também estavam presentes entre os melhores resultados nas edições anteriores do CASP

Pos.	FM	TBM
	Método	Método
1	Zhang-Server*	Zhang-Server*
2	QUARK*	nns
3	RBO_Aleph	QUARK*
4	MULTICOM-CONSTRUCT	BAKER-ROSETTASERVER*
5	MULTICOM-REFINE	RaptorX*

Fonte: Adaptado de Kryshtafovych et al. (2014b).

QUARK e Zhang-Server

Ambos os métodos QUARK[2] (XU; ZHANG, 2012) e Zhang-Server[3] (ZHANG et al., 2015, in press), foram desenvolvidos pelo *Zhang Lab* da Universidade de Michigan. QUARK é um método desenvolvido para o problema da predição da estrutura 3D de proteínas classificado como um método de primeiros princípios que utiliza conhecimento (XU; ZHANG, 2012). Esta abordagem utiliza fragmentos de tamanho 1-20 resíduos de aminoácidos de proteínas de boa qualidade depositadas no PDB. Estruturas são então formadas através de um arranjamento dos fragmentos e aplicadas a um método de Monte Carlo (MC) que faz réplicas das estruturas, por meio da substituição de fragmentos (XU; ZHANG, 2012). O MC é guiado por um campo de força composto por padrões físico-químicos e conhecimentos extraídos das bases de dados estruturais derivados da sequência primária da proteína alvo (ângulos de torções, acessibilidade ao solvente, empacotamento de estruturas secundárias, entre outros). As

[2]<http://zhanglab.ccmb.med.umich.edu/QUARK/>
[3]<http://zhanglab.ccmb.med.umich.edu/I-TASSER/>

estruturas geradas durante a simulação são então agrupadas baseada no método de clusterização SPICKER (ZHANG; SKOLNICK, 2004b) e são refinadas utilizando Dinâmica Molecular guiada por fragmentos (FG-MD) (ZHANG; LIANG; ZHANG, 2011). Quando estruturas com alta taxa de similaridade são encontradas (TBM) o QUARK utiliza as estruturas retornadas pelo LOMETS (WU; ZHANG, 2007), através do processo de reconhecimento de enovelamento para inicializar o MC e prosseguir com a execução do método.

O método Zhang-Sever é uma combinação da abordagem I-TASSER (ZHANG, 2008) para estruturas com enovelamento similares disponíveis (TBM) em conjunto com o método QUARK (XU; ZHANG, 2012) para proteínas que não obtiveram estruturas similares encontrada (FM) (ZHANG et al., 2015, in press). O I-TASSER procura estruturas de enovelamento similares com alto nível de confiança pelo método LOMETS (WU; ZHANG, 2007) (*threading*) selecionando as melhores estruturas. Essas estruturas são utilizadas para inicializar o MC no lugar das estruturas arranjadas através dos fragmentos como é feito no QUARK (ZHANG et al., 2015, in press). As estruturas geradas são então agrupadas pelo método SPICKER e refinadas pelo FG-MD (ZHANG, 2008). Para estruturas com baixa taxa de similaridade o Zhang-Sever utiliza os melhores resultados do QUARK, combinando com estruturas de enovelamento similares a esses resultados para inicializar o método (ZHANG et al., 2015, in press). O Zhang-Sever também utiliza a biblioteca de fragmentos SEGMER (WU; ZHANG, 2010) e os mapas de contados gerados pelo SVM-SEQ (WU; ZHANG, 2008) para melhorar a predição de proteínas com baixa taxa de similaridade.

Devido a esta combinação entre os métodos, onde o I-TASSER obtêm melhores resultados para estruturas TBM e o QUARK para estruturas FM o Zhang-Sever vem obtendo os melhores resultados em ambas as classes de estruturas (FM e TBM) ao longo dos últimos anos de CASP (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014b; ZHANG et al., 2015, in press). Entretanto, os autores dos métodos afirmam que para obterem resultados ainda melhores, serão necessários avanços no processo de empacotamento das estruturas e a correta predição das estruturas secundárias para que não influenciem negativamente na predição da estrutura 3D de proteínas (ZHANG et al., 2015, in press).

BAKER-ROSETTASERVER

O BAKER-ROSETTASERVER[4] (KIM; CHIVIAN; BAKER, 2004) é um servidor para predição da estrutura tridimensional de proteínas que consiste em dois passos principais:

[4] <http://robetta.bakerlab.org>

identificação do limite de domínio de atuação e modelagem da estrutura.

Identificação do limite de domínio de atuação: Nesta fase do algoritmo, métodos de reconhecimento de enovelamento como o HHSearch (SODING, 2005), Sparks (YANG et al., 2011) e RaptorX (PENG; XU, 2011) são aplicados para gerar alinhamentos e identificar modelos com similaridade de sequência (TBM). As estruturas são então rearranjadas e agrupadas a fim de identificar e ranquear as melhores estruturas (KIM; CHIVIAN; BAKER, 2004).

Modelagem da estrutura: Para cada uma das melhores estruturas preditas na etapa anterior o protocolo de modelagem comparativa do RosettaCM (SONG et al., 2013) é utilizado, e os modelos gerados que possuírem falhas são corrigidos utilizando uma combinação de fragmentos em conjunto com uma minimização no espaço de busca dos ângulos de torções. Após a modelagem concluída, a amostragem é realizada usando as funções de baixa resolução do Rosetta (LEAVER-FAY et al., 2011) com restrição de espaço de busca gerado para cada grupo da amostragem. Para estruturas FM o método BAKER-ROSETTASERVER também utiliza a abordagem baseada em conjuntos de fragmentos, porém de tamanhos de 3 e 9 resíduos de aminoácidos (KIM; CHIVIAN; BAKER, 2004). Todos os modelos são então refinados utilizando o protocolo de relaxamento estrutural do Rosetta (CONWAY et al., 2014) que minimiza a estrutura pela sua função de energia que utiliza todos os átomos. As estruturas finais são selecionadas a partir da clusterização estrutural escolhendo os 100 melhores indivíduos de cada topologia e gerando uma estrutura média que é refinada novamente (KIM; CHIVIAN; BAKER, 2004).

RaptorX

O método RaptorX[5] (PENG; XU, 2011) consiste nos seguintes principais componentes: reconhecimento de estruturas similares classificadas por sua qualidade para estruturas TBM (PENG; XU, 2010) e a abordagem de fragmentos utilizada para as estruturas FM (ZHAO; PENG; XU, 2010).

Para realizar o reconhecimento de estrutura similares, o método RaptorX utiliza um sistema de perfil estrutural para identificar estruturas de boa taxa de similaridade e um sistema de pontuação que utiliza uma sofisticada correlação entre diversas características da proteínas para identificar mais facilmente até mesmo estruturas de baixa taxa de similaridade. Para classificação de qualidade das estruturas o RaptorX utiliza informações extraídas da modelagem

[5] <http://velociraptor.ttic.edu>

realizada pelo MODELLER (ESWAR et al., 2008). O RaptorX também utiliza uma abordagem de fragmentos que permite restringir as possíveis conformações de proteínas com baixa taxa de similaridade permitindo assim encontrar estruturas com enovelamentos diferentes (PENG; XU, 2011).

Outros

Os outros servidores de predição que também ficaram entre os melhores resultados do último CASP foram: MULTICOM[6] (CHENG et al., 2012; LI et al., 2014), RBO_Aleph* (MABROUK et al., 2015) e *nns** (JOO et al., 2014; JOO et al., 2015, in press).

O método MULTICOM é baseado em uma abordagem de combinação multinível aplicada em uma rede neural recursiva que tenta melhorar os vários passos da predição de estrutura de proteínas (LI et al., 2014). Esse método possui diversas variações, entre as que tiveram melhores resultados no último CASP estão os métodos MULTICOM-REFINE e MULTICOM-CONSTRUCT (CHENG et al., 2012). O método MULTICOM-REFINE utiliza características de cada resíduo como perfis a partir da estrutura primária, estrutura secundária e acessibilidade ao solvente (FM). Essas características são utilizadas no treinamento de um comitê de redes neurais para tentar predizer os contatos gerais dos resíduos e o emparelhamento espacial dos resíduos (CHENG et al., 2012). O método MULTICOM-CONSTRUCT utiliza uma variação da técnica MULTICOM-REFINE para inicializar o método e aplica uma abordagem de mapa de contato para tentar melhorar o arranjamento espacial da estrutura e o contato geral entre os resíduos da proteína (CHENG et al., 2012; LI et al., 2014). O servidor RBO_Aleph também tem enfoque em mapa de contatos dos resíduos pertencente a proteína alvo em conjunto com propriedades evolutivas e físico-química (FM) desses resíduos (MABROUK et al., 2015). O servidor *nns* (JOO et al., 2015, in press) é uma variação do método LEE (JOO et al., 2014) que aplica otimização no alinhamento estrutural e rearranjamento da cadeia lateral da proteína. O *nns* utiliza um processo de arrefecimento simulado ao longo do espaço conformacional que foi reduzido baseado em estruturas com enovelamento similares (TBM).

3.3 Desenvolvimento de Novos Métodos de Predição

No último CASP (edição 11), 44 servidores de predição da estrutura 3D de proteínas participaram da competição. Entretanto, devido à complexidade do problema, os

[6] <http://sysbio.rnet.missouri.edu/multicom_toolbox/>

*Serviço online ainda não está disponível.

resultados desses métodos para muitas proteínas testadas obtiveram estruturas bem distintas das experimentais (KINCH et al., 2016, in press). Essa dificuldade, ainda motiva muitos pesquisadores de diversas áreas a desenvolverem novos métodos que possam ao menos auxiliar na resolução do problema da predição da estrutura 3D de proteínas.

Analisando os métodos desenvolvidos no último CASP, que conseguiram fazer a correta predição de algumas proteínas, é possível identificar semelhanças em suas abordagens. A principal delas é que todos os métodos desenvolvidos, independente da classe de estruturas que estão tratando (FM ou TBM), utilizam variações das classes de métodos 2, 3 e 4 (apresentadas na introdução deste capítulo), para tentar resolver o problema. Por exemplo, todos os métodos desenvolvidos utilizam técnicas para extrair conhecimento da base experimental a fim de diminuir o espaço de busca conformacional do problema. Também é possível analisar que vários métodos utilizam alguma abordagem baseada em fragmentos de estruturas que são recombinados para tentar predizer uma proteína alvo. Outra abordagem bastante utilizada é a de agrupar as estruturas, a fim de identificar grupos estruturais que são formados durante a execução dos métodos de predição.

A partir destas informações, foram montadas duas linhas de pesquisa nesse trabalho. A primeira linha é baseada na extração de conhecimento através da análise das informações da base de estruturas determinadas experimentalmente. Essa linha de pesquisa é apresentada no Capítulo 4 desta dissertação. A segunda linha de pesquisa é como utilizar o conhecimento gerado da primeira pesquisa e combinar com técnicas computacionais que permitam a predição da estrutura 3D de proteínas. Essa segunda linha de pesquisa é apresentada no Capítulos 5 desta dissertação. A combinação dessas duas linhas de pesquisas realizadas nesse trabalho, permitiu o desenvolvimento de uma nova técnica, apresentada no Capítulo 6.

3.4 Resumo do Capítulo

Neste Capítulo foram apresentadas a topologia de métodos de predição da estrutura 3D de proteínas, bem como a competição mundial que tenta encontrar qual o melhor método para realizar a correta predição do estado funcional de uma proteína (CASP). Foi realizada uma análise sobre os servidores que obtiveram melhores resultados no último CASP (edição 11), mostrando que independente da classe de estruturas (FM ou TBM) a serem preditas, os servidores que obtiveram mais sucesso combinavam técnicas de várias classes de métodos que utilizam conhecimento das bases de dados.

Levando em consideração as vantagens e desvantagens apresentadas pelos atuais

métodos de predição, o Capítulo 4 apresentada diferentes abordagens para tentar extrair conhecimento a fim de diminuir o espaço de busca que métodos de predição da estrutura 3D proteínas possuem.

4 EXTRAÇÃO DE CONHECIMENTO DA BASE DE DADOS

4.1 Introdução

O principal banco de dados relacionado a estruturas de proteínas é o PDB (BERMAN et al., 2000), porém determinar estruturas experimentalmente além de complexo é dependente de muitos fatores como qualidade de equipamentos; pessoal qualificado; protocolos corretos; etc. Devido a esses fatores algumas estruturas disponíveis no PDB podem ter uma maior qualidade/certeza que outras (HOVMÖLLER; ZHOU; OHLSON, 2002). O sucesso de abordagens, que usam conhecimento, para o problema de predição da estrutura 3D de proteínas está estritamente relacionado com a qualidade e em como é usado esse conhecimento (SCHEEF; FINK, 2005; DORN et al., 2014; BORGUESAN et al., 2015a). Analisando os dados do último CASP, apresentados no capítulo anterior, é possível verificar que todos os trabalhos que obtiveram bons resultados utilizavam alguma abordagem para extrair o conhecimento da base de dados (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014b; ZHANG et al., 2015, in press). Ao longo dos anos, foram várias as abordagens para extrair informações relevantes a partir das estruturas 3D determinadas experimentalmente e disponíveis no PDB. No Capítulo 3 foram apresentados algumas destas abordagens como biblioteca de fragmentos (PENG; XU, 2011; KIM; CHIVIAN; BAKER, 2004), preferências conformacionais de partes de estruturas (XU; ZHANG, 2012), entre outros. A partir deste conhecimento obtido é possível desenvolver técnicas para tratar esses dados de modo a facilitar sua utilização em meta-heurísticas aplicadas para resolução do problema da predição da estrutura 3D de proteínas. É com essa ideia que as abordagens *Angle Probability List* (APL) (BORGUESAN et al., 2015a), *Neighbors Preferences of Amino Acids and Secondary Structures* (NPAS) (BORGUESAN; INOSTROZA-PONTA; DORN, 2015b) e a *FragMent-Based LIBrary* (FM-B Lib) foram desenvolvidas durante o período do mestrado e são apresentadas nas próximas seções.

4.2 APL

Conforme comentado anteriormente, abordagens que utilizam conhecimento para realizar a predição da estrutura 3D de proteínas constantemente estão entre os melhores resultados das últimas edições do CASP (KRYSHTAFOVYCH et al., 2014a). Isto acontece, devido à complexidade do problema que necessita de conhecimento para reduzir o espaço de busca conformacional que a predição da estrutura 3D de proteínas possui. São vários os

trabalhos que analisam a existência de preferências conformacionais entre os aminoácidos de um conjunto de proteínas (HOVMÖLLER; ZHOU; OHLSON, 2002; DUNBRACK JR; KARPLUS, 2003; DORN et al., 2013; BORGUESAN et al., 2015a). Esta análise é feita através da inspeção da ocorrência de padrões estruturais em determinadas conformações de proteínas depositadas no PDB. Hovmöller et al. (HOVMÖLLER; ZHOU; OHLSON, 2002) procurou essa preferência para todos os 20 resíduos de aminoácidos padrões em mais de mil estruturas de proteínas depositadas no PDB. Dorn et al. (DORN et al., 2013) também fez essa análise e aplicou o conhecimento extraído em uma meta-heurística para realizar a predição da estrutura 3D de proteínas. Entretanto, nenhum destes trabalhos combinou as informações dos aminoácidos com suas estruturas secundárias para tentar extrair conhecimento sobre preferências conformacionais de proteínas. Para contornar isto, a metodologia APL (*Angle Probability List*) foi desenvolvida, a fim de extrair informações relevantes sobre a preferência conformacional de proteínas e utilizar este conhecimento para restringir o espaço de busca conformacional do problema da predição da estrutura 3D de proteínas.

4.2.1 Base de Dados

APL é uma metodologia baseada na frequência relativa de pares de aminoácidos (aa) e sua estrutura secundária (ss) observados em proteínas de boa qualidade depositadas no PDB. Com essa frequência é possível analisar a existência de preferências conformacionais sobre esses pares (aa,ss). Para realizar essa análise foi selecionado um conjunto de proteínas determinadas experimentalmente através da técnica X-Ray, que representam aproximadamente 90% do total de estruturas do PDB. Como o PDB possui muitas estruturas parecidas um filtro de similaridade foi utilizado, resultando em apenas estruturas com no máximo 30% de similaridade de sequência e que foram depositadas no PDB até dezembro de 2014.

Estruturas determinadas por X-Ray possuem alguns critérios de qualidade comumente utilizado pela comunidade científica, como resolução e fator-R (HOVMÖLLER; ZHOU; OHLSON, 2002; MCPHERSON, 2004). O fator-R permite avaliar o sucesso no processo de refinamento da estrutura, que consiste na medida de concordância entre o modelo construído e os dados experimentais. Esse fator-R é constantemente associado a estruturas de boa qualidade quando fica entre valores 15% e 20%. Para esse trabalho foi utilizado o fator-R observado de no máximo 20%. Já o valor de resolução indica o nível de detalhamento com o qual a proteína foi determinada. Para esse trabalho dois valores de resolução foram testados. O primeiro foi utilizando resolução $\leq$ 2.0Å que retornou um total de 6,650 estruturas 3D de proteínas. Esses

resultados foram apresentado no trabalho de Borguesan et al. (2015a), entretanto alguns padrões ficaram com poucos dados. Para contornar esse problema, foi utilizado então um segundo valor de resolução $\leq$ 2.5Å que ainda são consideradas estruturas de boa qualidade, o que aumentou a base de estruturas selecionadas para 11,130.

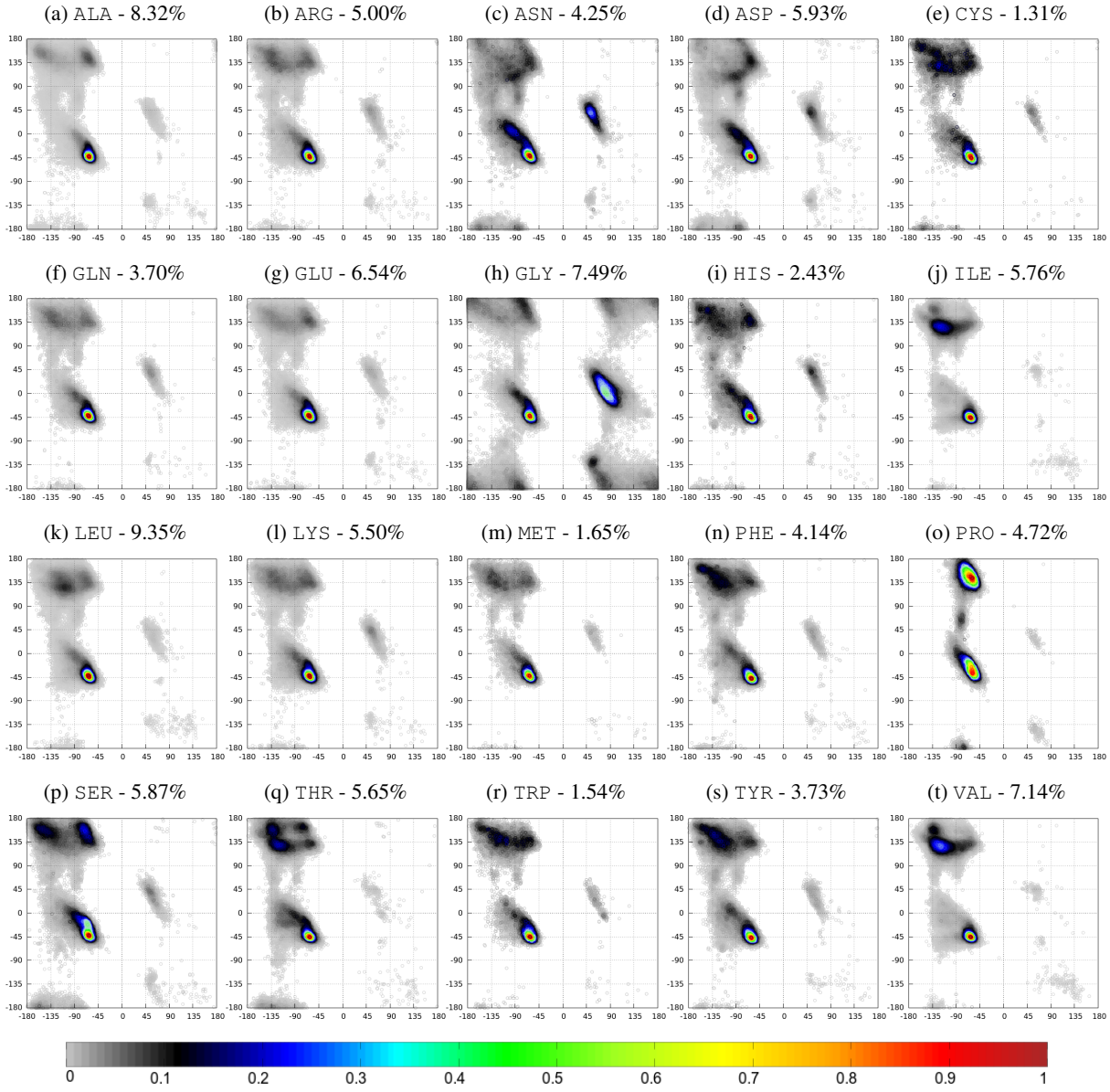
Com as estruturas 3D de proteínas selecionadas a base de dados matinha um total de 5,255,768 resíduos de aminoácido com valor médio de ocupância entre os átomos da cadeia principal igual a 1. A ocupância representa a probabilidade do correto posicionamento espacial do átomo. O valor 1 representa a certeza do posicionamento. Parâmetros similares a esses para filtrar a base de dados foram usados em outros trabalhos (HOVMÖLLER; ZHOU; OHLSON, 2002; BORGUESAN et al., 2015a).

Com esse conjunto de dados definidos foi possível então computar os ângulos diedros formados entre as ligações peptídicas desses aminoácidos. Para computar esses ângulos foi utilizado o algoritmo STRIDE (FRISHMAN; ARGOS, 1995; HEINIG; FRISHMAN, 2004), explicado no Capítulo 2 deste trabalho, que além da atribuição da estrutura secundária também computa os dois principais ângulos de torção de proteínas, os ângulos ϕ e ψ. Com essa combinação entre aminoácido e ângulos diedros foi possível analisar a preferência conformacional dos resíduos através dos gráficos de Ramachandran (RAMACHANDRAN; SASISEKHARAN, 1968). Todos os gráficos de Ramachandran apresentados neste capítulo utilizam a configuração do ângulo ϕ no eixo das abscissas e o ângulo ψ no eixo das ordenadas.

4.2.2 Preferência Conformacional

A Figura 4.1 apresenta esses gráficos mostrando a preferência dos 20 resíduos de aminoácidos, bem como a frequência (%) de ocorrência de cada aminoácido da base de dados desenvolvida (BORGUESAN et al., 2015a). Essa preferência é analisada pela densidade no gráfico de Ramachandran, onde a região em vermelho escuro representa a maior preferência e a região branca a de nenhuma ocorrência. Essa figura permite analisar que os aminoácidos Prolina (PRO - Fig. 4.1-o) e Glicina (GLY - Fig. 4.1-h) possuem um espaço conformacional totalmente diferente dos outros aminoácidos. Esse efeito sobre a prolina ocorre pelo fato de ser um aminoácido especial que possui um anel heterocíclico entre as suas cadeias lateral e principal o que restringe sua conformação (ANFINSEN, 1973). Enquanto a Glicina, que estruturalmente é o aminoácido mais simples sendo o único opticamente inativo, possui uma menor restrição que os outros aminoácidos acarretando em uma maior liberdade na escolha de regiões do gráfico de Ramachandran (ANFINSEN, 1973). Os outros aminoácidos possuem uma menor variação nas suas preferências conformacionais.

Figura 4.1: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido e a sua procentagem de ocorrência na base de dados. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran

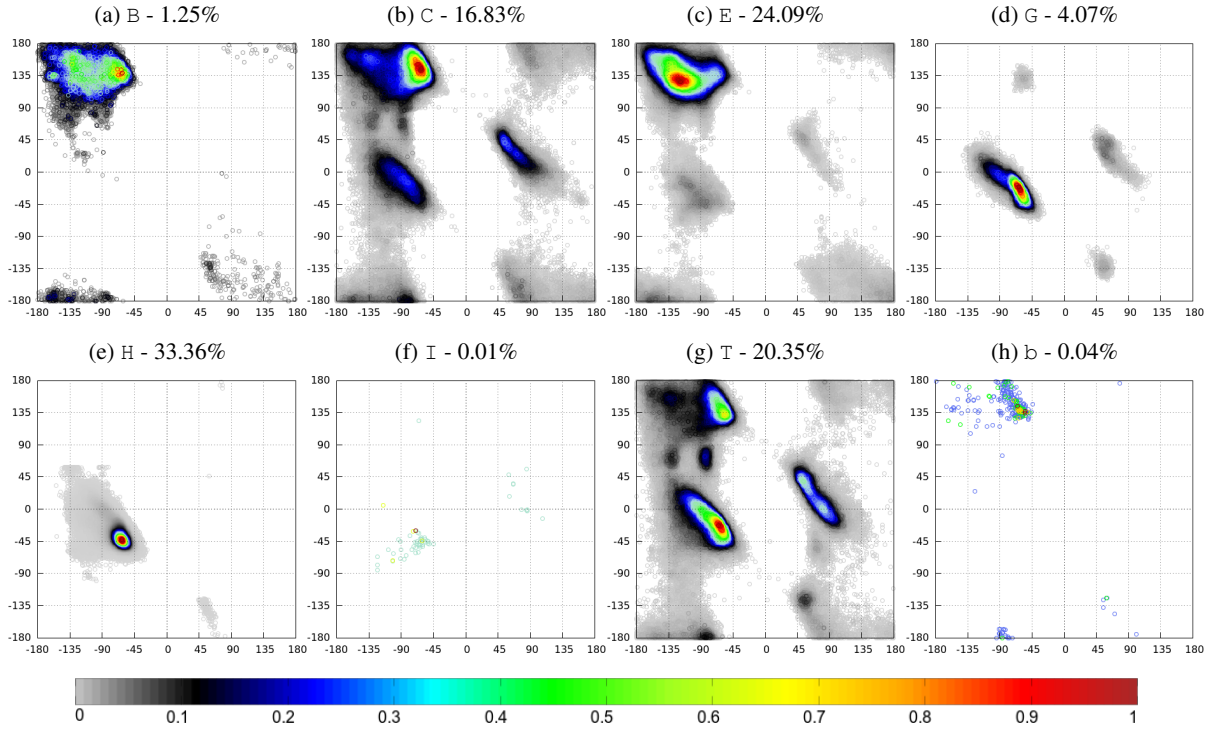


Fonte: Adaptado de Borguesan et al. (2015a).

Outra análise realizada foi gerar os mesmos gráficos de Ramachandran, porém desta vez com o intuito de analisar a preferência conformacional da estrutura secundária atribuída pelo STRIDE (FRISHMAN; ARGOS, 1995; HEINIG; FRISHMAN, 2004) e pelo DSSP (KABSCH; SANDER, 1983a; TOUW et al., 2015). Nesta análise é possível ver claramente a diferença de regiões de preferência que cada uma das estruturas secundárias possui (HOVMÖLLER; ZHOU; OHLSON, 2002; DORN et al., 2013; BORGUESAN et al., 2015a). Também é possível analisar que as regiões de hélices (H, G e I) e folhas (B e E) apresentam similaridade em termos

quantitativos e da distribuição no gráfico de Ramachandran quando comparado os dois métodos de atribuição utilizados, com uma pequena variação nas regiões de menor frequência de cada um (regiões menos densas do gráfico). Já as regiões de voltas (C, T) não apresentam essa mesma similaridade pelo fato do método DSSP atribuir a estrutura secundária de dobra (S) que o método STRIDE não considera. As Figuras 4.2 e 4.3 apresentam a preferência conformacional das estrutura secundárias atribuídas pelos métodos STRIDE e DSSP, respectivamente, bem como a frequência (%) de ocorrência de cada estrutura secundária no conjunto de dados selecionado. Através dessas figuras é possível também verificar que as regiões de hélices possuem uma menor variação nos seus pares *phi* e *psi*, que ocorre devido à sua estrutura helicoidal ser bastante fechada e rígida. As estruturas de voltas já possuem uma variação maior pelo fato de serem estruturas que normalmente ficam mais acessível ao solvente sendo assim estruturas de maior flexibilidade.

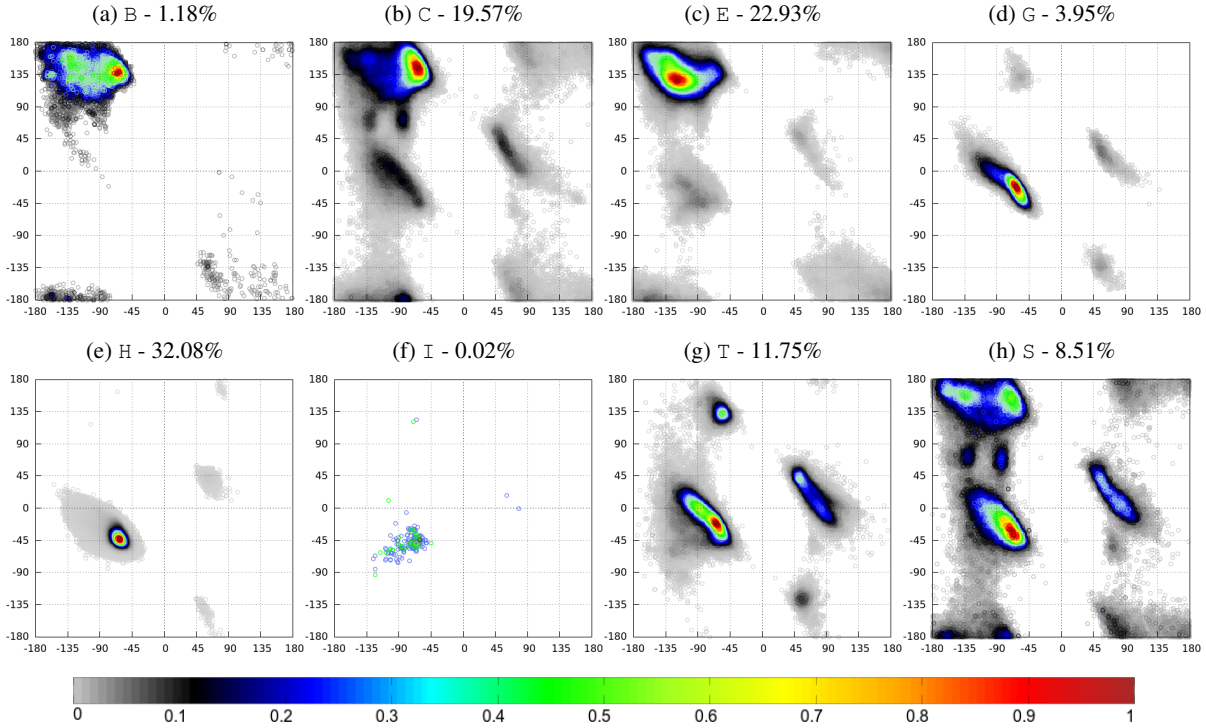
Figura 4.2: Gráficos de Ramachandran para os 8 estruturas secundárias atribuido pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



Fonte: Adaptado de Borguesan et al. (2015a).

As Tabelas 4.1 e 4.2, mostram uma análise quantitativa comparando as estruturas secundárias atribuídas pelo STRIDE e pelo DSSP, respectivamente, em cada um dos 20 resíduos de aminoácidos-padrão. Essa análise permite a comparação do mesmo conjunto de dados com a atribuição da estrutura secundária computada por diferentes métodos.

Figura 4.3: Gráficos de Ramachandran para os 8 estruturas secundárias atribuido pelo DSSP. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



Fonte: Adaptado de Borguesan et al. (2015b).

Tabela 4.1: Distribuição da base de dados entre os 20 aminoácidos principais atribuindo suas estruturas secundárias pelo STRIDE

Aminoácido	B	C	E	G	H	I	T	b	Total
ALA	3,436	53,130	80,338	**19,934**	**211,450**	**17**	69,096	139	437,540
ARG	3,682	41,014	55,811	**10,891**	**105,699**	**32**	45,358	89	262,576
ASN	3,015	**44,909**	33,520	9,627	58,138	18	**73,949**	69	223,245
ASP	3,652	**61,138**	41,871	16,753	84,123	54	**103,675**	141	311,407
CYS	1,248	10,232	22,306	2,342	19,306	0	13,344	24	68,802
GLN	1,883	27,974	38,143	**8,849**	**83,852**	**12**	33,510	111	194,334
GLU	2,354	42,913	58,331	**20,220**	**155,325**	**38**	64,332	157	343,670
GLY	3,637	**110,208**	62,801	11,028	60,344	79	**145,319**	271	393,687
HIS	2,172	22,218	31,062	5,518	38,138	25	28,457	37	127,627
ILE	**4,717**	36,219	**119,789**	6,706	106,485	32	28,462	**79**	302,489
LEU	5,689	59,431	131,631	**19,663**	**218,360**	**99**	56,134	216	491,223
LYS	3,158	46,123	56,787	**12,804**	**112,979**	**44**	57,332	39	289,266
MET	1,082	11,112	22,011	**3,088**	**38,154**	**6**	11,070	20	86,543
PHE	3,652	28,423	74,044	9,040	70,425	37	31,992	49	217,662
PRO	2,337	**81,714**	25,990	14,124	34,623	10	**88,996**	52	247,846
SER	4,952	**65,044**	66,520	15,277	79,057	40	**77,300**	131	308,321
THR	4,830	**61,396**	87,669	8,247	76,278	43	**58,337**	128	296,928
TRP	1,065	9,959	25,289	4,421	27,067	8	13,030	27	80,866
TYR	3,392	25,474	65,833	8,151	62,855	46	30,456	54	196,261
VAL	**5,806**	45,913	**166,292**	7,349	110,485	51	39,391	**188**	375,475
Total	65,759	884,544	1,266,038	214,032	1,753,143	691	1,069,540	2,021	5,255,768

Fonte: Adaptado de Borguesan et al. (2015a).

Por exemplo, analisando a Tabela 4.1 dos dados do STRIDE, é possível verificar que os aminoácidos ALA, ARG, GLN, GLU, LEU, LYS e MET, são os que apresentam uma maior quantidade de dados nas regiões de Hélices (H G e I). Os aminoácidos ILE e VAL apresentam dados mais densos nas regiões de Folhas (E, B e b), enquanto os aminoácidos ASN, ASP, GLY, PRO, SER e THR apresentam uma quantidade maior de ocorrências nas regiões de Voltas (C e T). Os aminoácidos CYS, HIS, PHE, TRP e TYR foram os que tiveram uma quantidade de dados equilibrada entre mais de um tipo de conformação. Enquanto que pra Tabela 4.2 dos dados do DSSP, apenas a preferência da LYS pra hélice trocou para uma representação mais equilibrada dos dados, os outros aminoácidos mantiveram o padrão do STRIDE. Também é possível afirmar que a estrutura secundária de hélice π (I) atribuída pelo STRIDE e pelo DSSP e a estrutura de ponte β isolada de só um resíduo (b) atribuído pelo STRIDE tem uma baixa representatividade dos dados, representando menos de 0.1%.

Tabela 4.2: Distribuição da base de dados entre os 20 aminoácidos principais atribuindo suas estruturas secundárias pelo DSSP

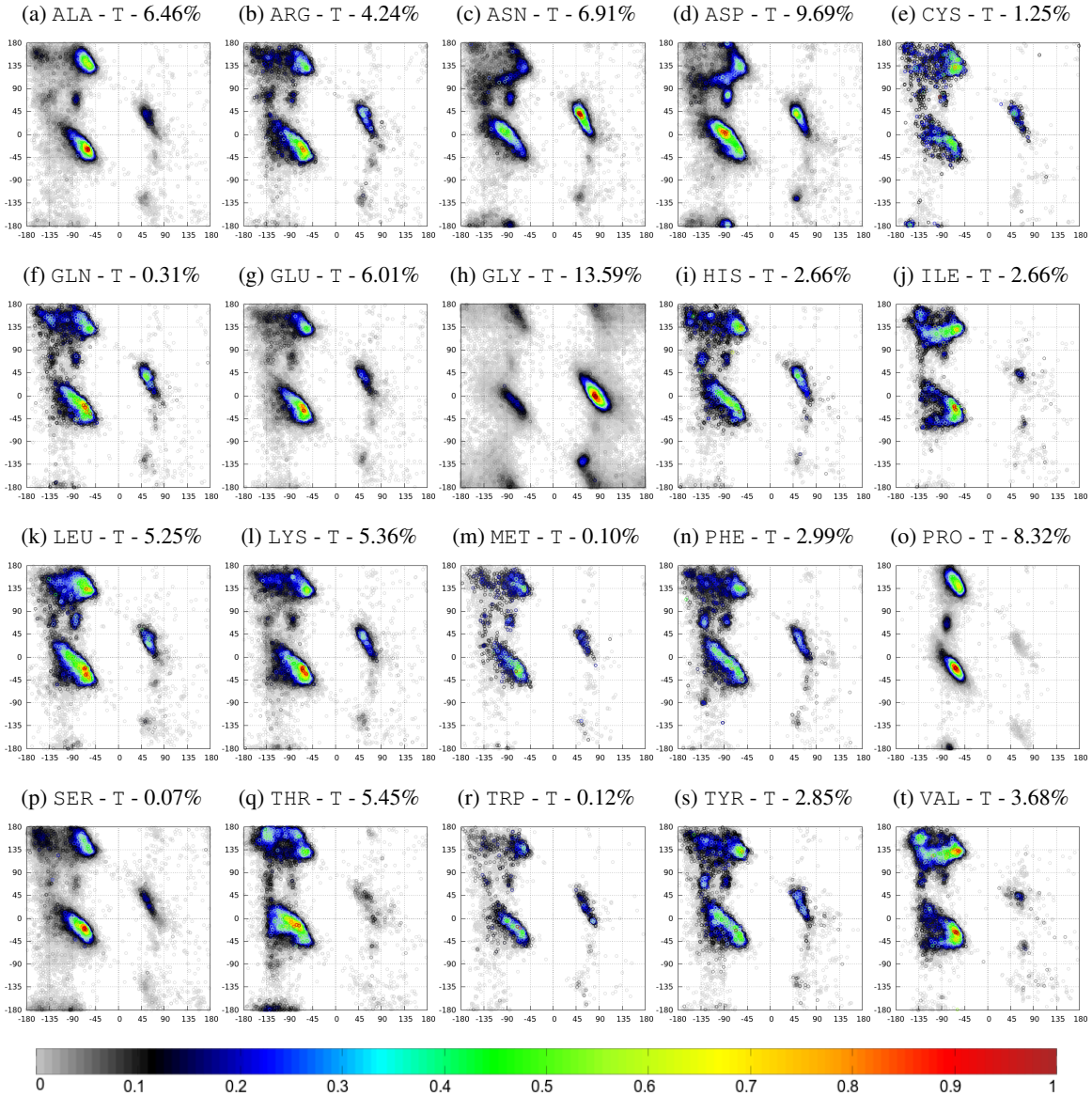
Aminoácido	B	C	E	G	H	I	T	S	Total
ALA	3,142	67,177	76,977	**18,232**	**203,247**	**33**	41,842	26,890	437,540
ARG	3,415	45,012	53,337	**10,449**	**101,398**	**48**	27,542	21,375	262,576
ASN	2,676	**54,947**	31,086	10,017	52,210	30	**46,060**	**26,219**	223,245
ASP	3,253	**81,870**	38,023	16,777	81,866	83	**52,543**	**36,992**	311,407
CYS	1,220	15,324	21,528	2,329	18,281	19	5,426	4,675	68,802
GLN	1,866	32,481	35,550	**8,471**	**79,934**	**33**	21,330	14,669	194,334
GLU	2,137	48,597	54,679	**18,670**	**151,423**	**62**	41,185	26,917	343,670
GLY	3,265	**81,950**	58,989	11,870	56,989	127	**112,165**	**68,332**	393,687
HIS	1,998	27,598	28,893	5,642	35,840	44	15,155	12,457	127,627
ILE	**4,622**	44,874	**115,669**	6,245	104,909	72	12,063	14,035	302,489
LEU	5,581	72,514	127,384	**18,201**	**210,494**	**161**	31,264	25,624	491,223
LYS	3,033	49,777	53,719	12,075	107,801	74	37,022	25,765	289,266
MET	1,036	13,945	21,053	**3,095**	**36,589**	**19**	5,859	4,947	86,543
PHE	3,729	35,035	71,853	8,919	67,830	56	16,731	13,509	217,662
PRO	2,522	**103,454**	23,522	14,221	34,075	14	**46,039**	**23,999**	247,846
SER	3,995	**78,874**	61,826	15,445	73,783	36	**39,534**	**34,828**	308,321
THR	4,410	**72,861**	81,965	8,081	73,418	74	**27,023**	**29,096**	296,928
TRP	1,082	12,337	24,356	4,361	26,651	16	6,825	5,238	80,866
TYR	3,445	31,690	63,553	8,164	60,420	35	15,984	12,970	196,261
VAL	**5,727**	58,297	**161,187**	6,567	108,634	95	16,214	18,754	375,475
Total	62,154	1,028,614	1,205,149	207,831	1,685,792	1,131	617,806	447,291	5,255,768

Fonte: Adaptado de Borguesan et al. (2015b).

Esta análise quantitativa corrobora com estudos que afirmam que aminoácidos podem ter preferência por diferentes estruturas secundárias (WILLIAMS et al., 1987; PETSKO; RINGE, 2004). Assim, combinou-se esses elementos para tentar extrair essas preferências conformacionais. Para isto, foi desenvolvido uma base de dados onde foram combinados os

ângulos de torções de acordo com o seu resíduo de aminoácido (AA) e a sua estrutura secundária (SS). Essa base de dados, chamada de AASS, permitiu computar os mapas de Ramachandran para cada conjunto de aminoácido pertencendo a uma estrutura secundária específica e analisar a sua preferência conformacional. Mostrando assim, que para uma determinada estrutura secundária um mesmo aminoácido pode ter diferentes regiões conformacionais. Isso é claramente observado quando as estruturas secundárias de volta e regiões desordenadas atribuídos pelo STRIDE são comparadas (BORGUESAN et al., 2015a). A Figura 4.4 mostra esta preferência para os 20 resíduos de aminoácidos na estrutura secundária de volta (T).

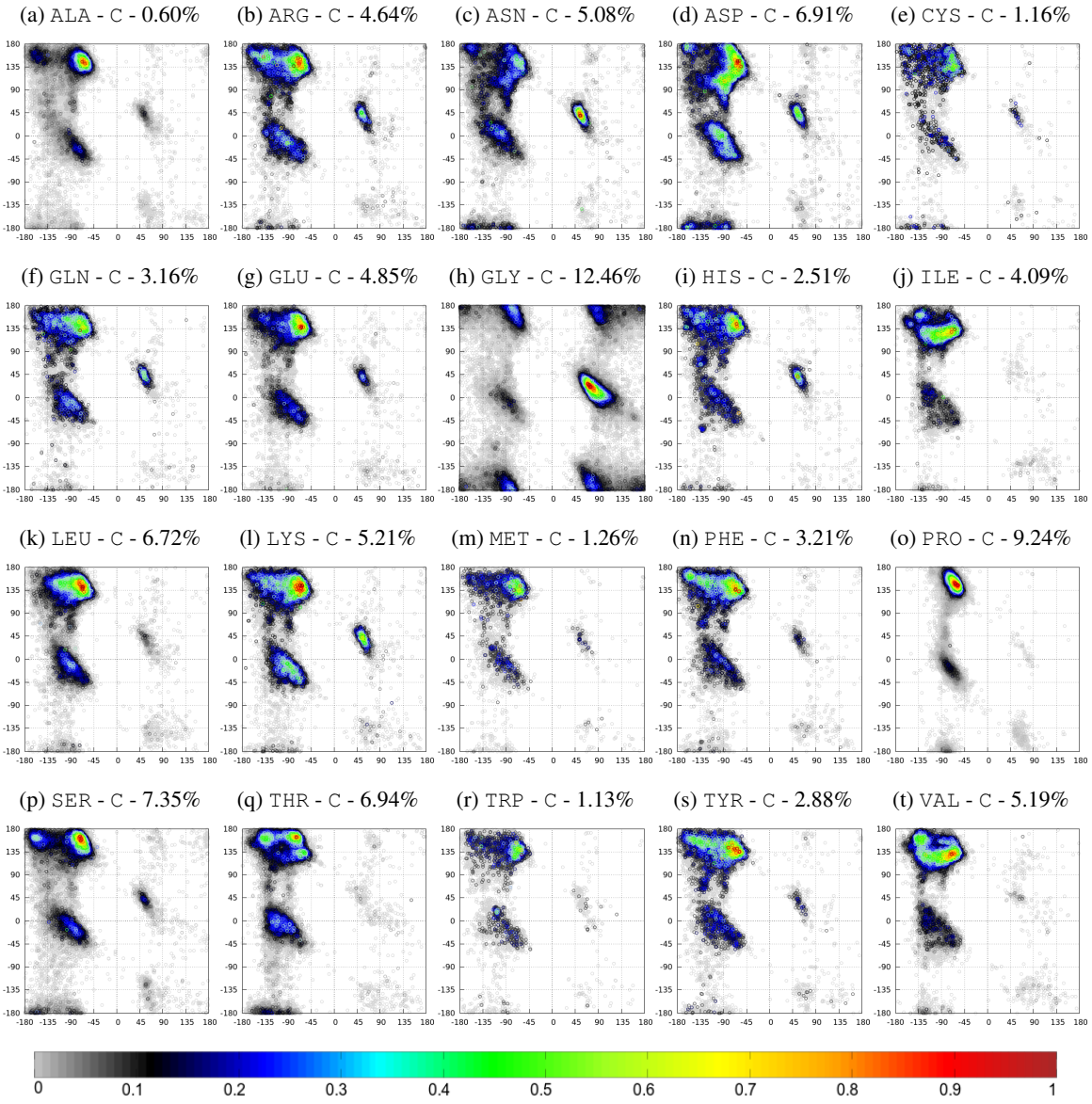
Figura 4.4: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária volta (T) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



Fonte: Adaptado de Borguesan et al. (2015a).

Como pode ser observado, diferentes resíduos de aminoácidos em uma mesma estrutura secundária possuem particularidades na sua preferência conformacional (ϕ and ψ). A Figura 4.5 mostra os mesmos 20 resíduos de aminoácidos-padrão que a figura anterior, porém agora somente os que foram atribuídos como regiões desordenadas (C) pelo STRIDE.

Figura 4.5: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária de regiões desordenadas (C) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



Fonte: Adaptado de Borguesan et al. (2015a).

Analisando essas figuras é possível corroborar com vários estudos que indicam a presença de padrões conformacionais em sequências de aminoácidos e suas estruturas secundárias (XIA; XIE, 2002; MOELBERT; EMBERLY; TANG, 2004; TING et al., 2010a).

Esse tipo de informação estrutural pode ser de grande auxilio para reduzir o espaço de busca conformacional do problema de predição da estrutura 3D de proteínas (DORN et al., 2013; BORGUESAN et al., 2015a). Assim, como os métodos apresentados no Capítulo 3, este tipo de informação é amplamente utilizada como conhecimento aplicado a meta-heurísticas para tentar contornar a grande dimensionalidade do problema da predição da estrutura 3D de proteínas. As preferências conformacionais para as outras estruturas secundárias (Hélice α, Hélice 3_{10}, Folha β e Ponte β Isolada), podem ser encontrados no Apêndice A desta dissertação. Essas figuras do Apêndice A permitem verificar que existe uma grande varições na preferência conformacional dos aminoácido quando foram atribuídos em diferentes estruturas secundárias. Se analisarmos a preferência conformacional dos aminoácidos para uma mesma estrutura secundária ainda assim é possível verificar uma variação, porém menor nesses casos.

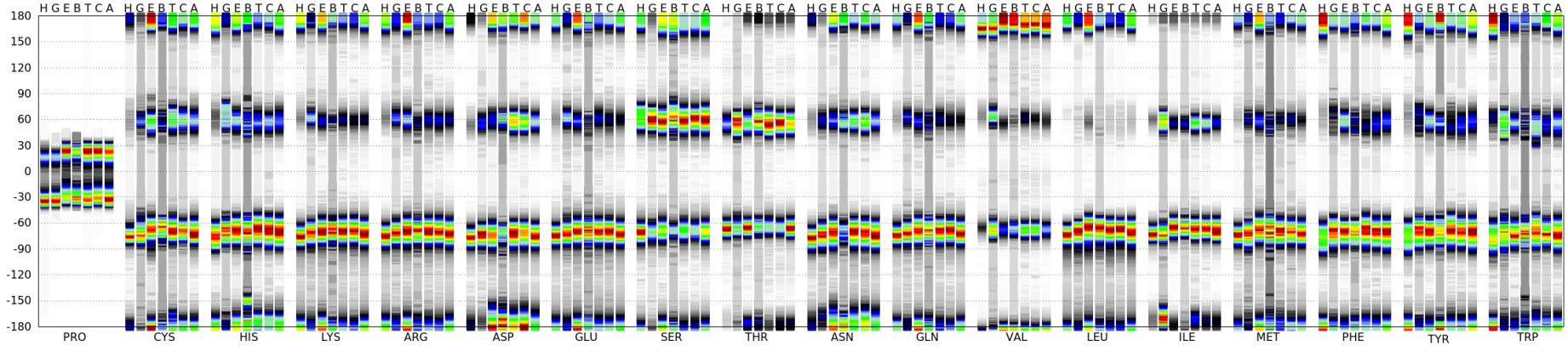
4.2.3 Extração de Conhecimento

Para usar essa informação uma abordagem chamada de *Angle Probability List* (`APL`) foi desenvolvida onde a frequência de um determinado aminoácido (aa) em uma determinada estrutura secundária (ss) é computado $F_{aa,ss}$ para um conjunto de $[-180, 180] \times [-180, 180]$ células do mapa de Ramachandran. Cada célula (i,j) tem o número de vezes que um par (aa,ss) ocorreu com aquele valor de (i,j) que representa os ângulos diedros (phi,psi). Assim, para cada par (aa,ss) foi computado a $APL_{aa,ss}$ (Eq. 4.1) que representa a frequência normalizada de cada par. Permitindo assim, por exemplo, que uma meta-heurística possa atribuir maior chance de selecionar os ângulos diedros (phi,psi) que possuírem uma alta frequência associada.

$$APL_{aa,ss}(i,j) = \frac{F_{aa,ss}(i,j)}{\sum(F_{aa,ss})}, \tag{4.1}$$

Mesmo os ângulos diedros (phi,psi) conseguindo modelar o enovelamento principal de uma proteína, outras informações são necessárias para que métodos desenvolvidos possam modelar corretamente a estrutura 3D de proteínas, como dos ângulos de torção da cadeia lateral (χ's) e o ângulo ômega da cadeia principal (ω). Para isso, foi adicionado a base de dados (AASS) essas informações geradas pelo PROMOTIF (HUTCHINSON; THORNTON, 1996). Assim para cada um dos 5,255,768 resíduos de aminoácidos além dos ângulos principais (phi,psi) já existentes, foram adicionados os ângulos diedros da cadeia lateral (χ's) e o ângulo ômega da cadeia principal (ω). Conforme comentado no Capítulo 2, o ângulo ômega tem uma variação pequena no seu valor (sempre um valor próximo a $-180°$ ou $180°$), entretanto os ângulos diedros da cadeia lateral possuem uma maior variação. A Figura 4.6 apresenta a

Figura 4.6: Preferência conformacional do ângulo Chi_1 para os 18 aminoácidos que possuem cadeia lateral combinados com as 6 estruturas secundárias com maior quantidade de dados atribuído pelo STRIDE. A estrutura secundária "A" representa todas as estruturas secundárias combinadas

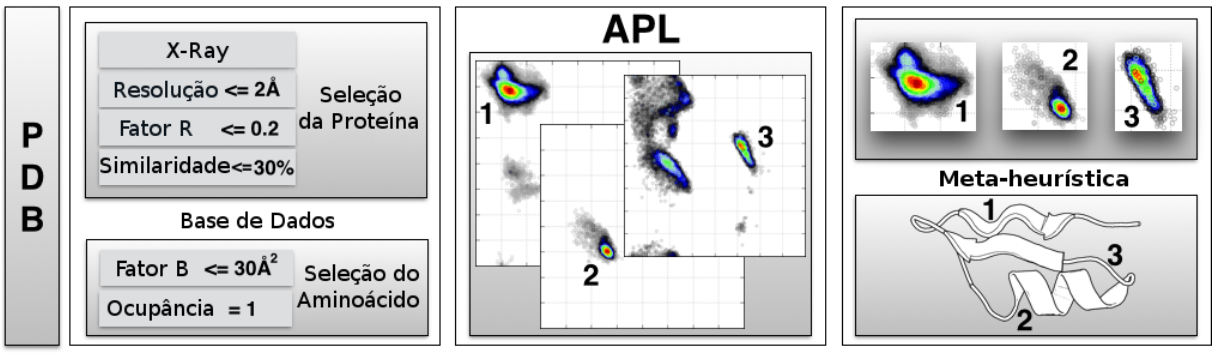


Fonte: do autor (2016).

preferência conformacional que o primeiro ângulo diedro da cadeia lateral (Chi_1) possui para os 18 aminoácidos que possuem cadeia lateral, essa preferência é representada na figura pela coluna "A" de cada aminoácido. Esta figura ainda permite analisar a existência da preferência quando o aminoácido é combinado com as 6 estruturas secundárias com maior quantidade de dados atribuído pelo STRIDE (H, G, E, B, T, C). A análise dessa figura permite verificar a existência também de preferências conformacionais nas cadeias laterais, corroborando assim para a importância em adicionar esses dados a base de dados desenvolvida (AASS). Essa combinação de ângulos permite uma representação precisa de proteínas com um número pequeno de variáveis se comparado com posicionamento cartesiano. Esse número de variáveis é importante pois, quanto maior for este valor, maior será o tempo necessário para que uma meta-heurística consiga percorrer o espaço de busca do problema (GOLDBERG, 1989; LARRAÑAGA et al., 2006; WITTEN; FRANK; HALL, 2011). Da mesma forma quando aplicamos conhecimento sobre uma variável, delimitando assim sua região de busca e evitando que o método percorra regiões onde uma possível solução não está presente.

A combinação entre a APL desenvolvida e uma meta-heurística para percorrer esse espaço de busca pode ser uma importante informação para a predição da estrutura 3D de proteínas (BORGUESAN et al., 2015a). A Figura 4.7, demonstra como foi realizada a construção do modelo APL, extraindo informações do PDB e aplicando isso em uma meta-heurística para diminuir a complexidade do problema.

Figura 4.7: Esquema da abordagem APL aplicado ao problema da predição da estrutura 3D de proteínas



Fonte: Adaptado de Borguesan et al. (2015a).

Na próxima seção é apresentado como meta-heurísticas podem tirar proveito do conhecimento extraído da abordagem APL aplicado ao problema de predição da estrutura 3D de proteínas.

4.2.4 Meta-heurísticas em Conjunto com Abordagem APL

Para validar a abordagem APL e o conhecimento produzido pela mesma, duas meta-heurísticas foram aplicadas: um Enxame de Partículas (EP) (KENNEDY, 2003) e um Algoritmo Genético (AG) (HOLLAND, 1975; GOLDBERG, 1989). No método EP as possíveis soluções, chamadas também de partículas, "voam" pelo espaço de busca do problema seguindo sempre a melhor solução do momento. Enquanto o AG é um algoritmo baseado em conceitos evolutivos, onde a aplicação de operadores de mutação e de recombinação podem gerar melhores indivíduos conforme a evolução da população. Os resultados encontrados nesta etapa foram publicados em Borguesan et al. (2015a). Neste trabalho foi testada uma versão padrão destas duas meta-heurísticas utilizando APL como conhecimento para realizar a predição de 8 diferentes estruturas de proteínas. Quando comparadas essas duas meta-heurísticas, para esse problema, foi considerado que o Algoritmo Genético obteve melhores resultados que o Enxame de Partículas. Assim, foi selecionado o AG para detalhar a utilização do conhecimento da APL no problema da predição da estrutura 3D de proteínas.

O algoritmo desenvolvido recebe como entrada a sequência primária e a sequência secundária da estrutura alvo a ser predita. Com esses dados de entrada é possível então selecionar as APLs geradas como descrito anteriormente. A seguir, o algoritmo seleciona os ângulos de torção para montar o indivíduo, dando maior chance de escolha para o ângulo que possuir maior frequência, porém também é dada uma chance pequena de selecionar o ângulo de menor frequência (BORGUESAN et al., 2015a). O tamanho da população inicial do AG desenvolvido foi de 100 indivíduos. Com a população completa o algoritmo então avalia esses indivíduos utilizando uma função de aptidão do PyRosetta (CHAUDHURY; LYSKOV; GRAY, 2010) que foi implementada baseada no programa Rosetta (ROHL et al., 2004), um dos programas mais famosos para modelagem molecular. Essa população é então ordenada por esse valor de aptidão a fim de gerar 3 grupos, onde os 10% melhores indivíduos da população eram chamados de classe A, os 50% seguintes eram considerados da classe B e os últimos 40% eram da classe C. Para gerar a próxima população o método automaticamente promove os indivíduos da classe A para a próxima população, utilizando a abordagem de elitismo. Os outros 50% seguintes são gerados através do operador de reprodução uniforme, onde o método seleciona um indivíduo da classe A, chamado de pai A, e um indivíduo que vem ou classe B ou da classe C, chamado de pai B+C, para aplicar o operador de reprodução. O indivíduo gerado por essa reprodução tem $50-70\%$ de chance de vir do pai A e o resto de vir do pai B+C. E a classe C é gerada da mesma forma que a população inicial, a fim de manter a diversidade do método.

Esse método foi executado 15 vezes por um período de 24 horas para um conjunto de 20 proteínas. Para validação da APL, foi testado o algoritmo desenvolvido utilizando a APL, contra o mesmo algoritmo, porém utilizando a população inicial com valores aleatórios variando entre $[-180, 180]$. Para todos os 20 casos testados a melhor estrutura foi encontrada através do método que utilizava conhecimento da APL, corroborando assim para o esquema proposto (BORGUESAN et al., 2015a). Os resultados da validação da metodologia APL foram publicados em um periódico de Biologia Computacional (*Computational Biology and Chemistry*) (BORGUESAN et al., 2015a).

Entretanto, o método APL utiliza o conhecimento apenas local (par aa_i, ss_i), porém vários trabalhos afirmam que a informação da vizinhança é de extrema importância para o enovelamento de uma proteína (KABAT; WU, 1973; CRASTO; FENG, 2001; TING et al., 2010b). Para solucionar essa deficiência da APL o servidor *Neighbors Preferences of Amino Acids and Secondary Structures* (NPAS) (BORGUESAN; INOSTROZA-PONTA; DORN, 2015b) foi desenvolvido e é apresentando na próxima seção.

4.3 NPAS

Como falado anteriormente, a vizinhança de uma estrutura primária e secundária tem fundamental importância para tentar entender o enovelamento global de uma proteína (ANFINSEN, 1973). Kabat e Wu (1973) fizeram um dos primeiros trabalhos que estudam a influência que os aminoácidos n-1 e n+1 possuem sobre o aminoácido n, na conformação de hélices-α e de folhas-β. O estudo de Crasto e Feng (2001) também analisa o efeito que os aminoácidos n-1 e n+1 possuem sobre o aminoácido n, porém focado nas regiões de voltas. O Dunbrack Lab desenvolveu um trabalho onde analisa separadamente a influência do aminoácidos n-1 no aminoácidos n e a influência do n+1 no n, apresentando evolução na predição da conformação de voltas de proteínas (TING et al., 2010b). Todos esses trabalhos afirmam que os aminoácidos vizinhos possuem importantes informações na conformação da estrutura secundária ou até mesmo do enovelamento de proteínas.

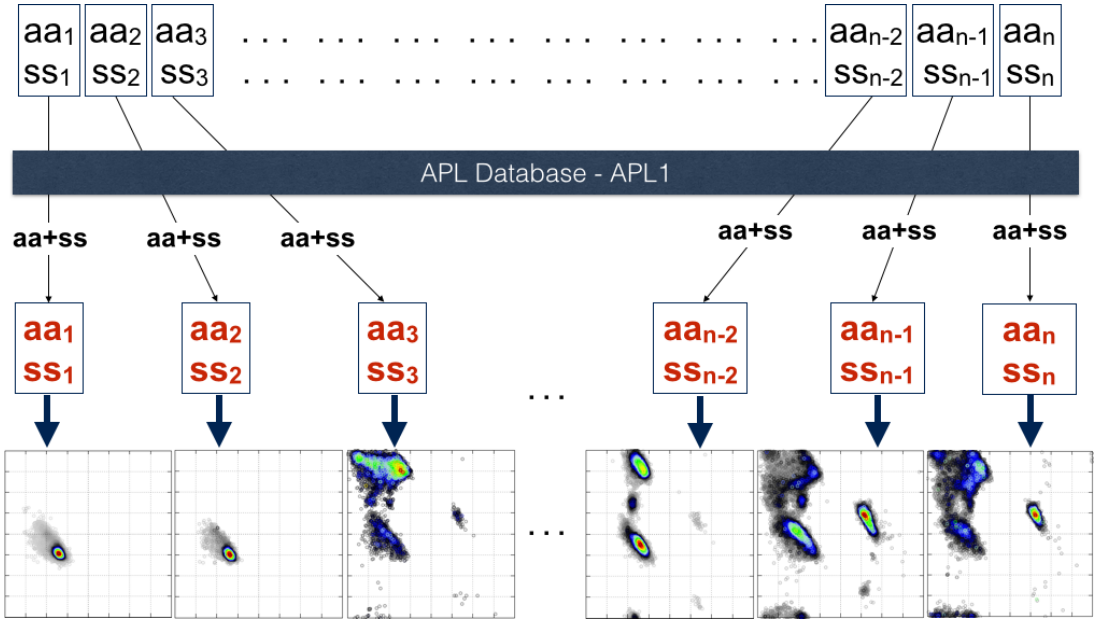
4.3.1 Base de Dados

Baseado nesses trabalhos, um servidor foi desenvolvido que permite gerar a APL apresentada na seção anterior e mais 3 diferentes grupos de APL que consideram a vizinhança da sequência primária e secundária de uma proteína alvo. Para gerar esses grupos, foi utilizado

o mesmo conjunto de dados (AASS) contendo 11,130 estrutura de proteínas com 5,255,768 resíduos de aminoácidos apresentados na seção anterior. Esses grupos foram chamados de: APL2 que considera um único vizinho da direita ou da esquerda; APL3 que considera a influência completa (esquerda e direita); e a APLCentral que mantém apenas o aminoácido central fixo e a sequência completa da estrutura secundária de tamanhos 5, 7 e 9. A seguir a implementação de cada grupo gerado pelo NPAS são detalhados.

APL*$_1$*: A Figura 4.8 apresenta como o servidor NPAS recebe uma sequência primária de uma proteína e a sua sequência secundária e gera a APL1, explicada na seção anterior, para cada um dos pares (aa, ss) da sequência informada. Nesta estrutura da APL1 a influência dos aminoácidos vizinhos não é levada em consideração.

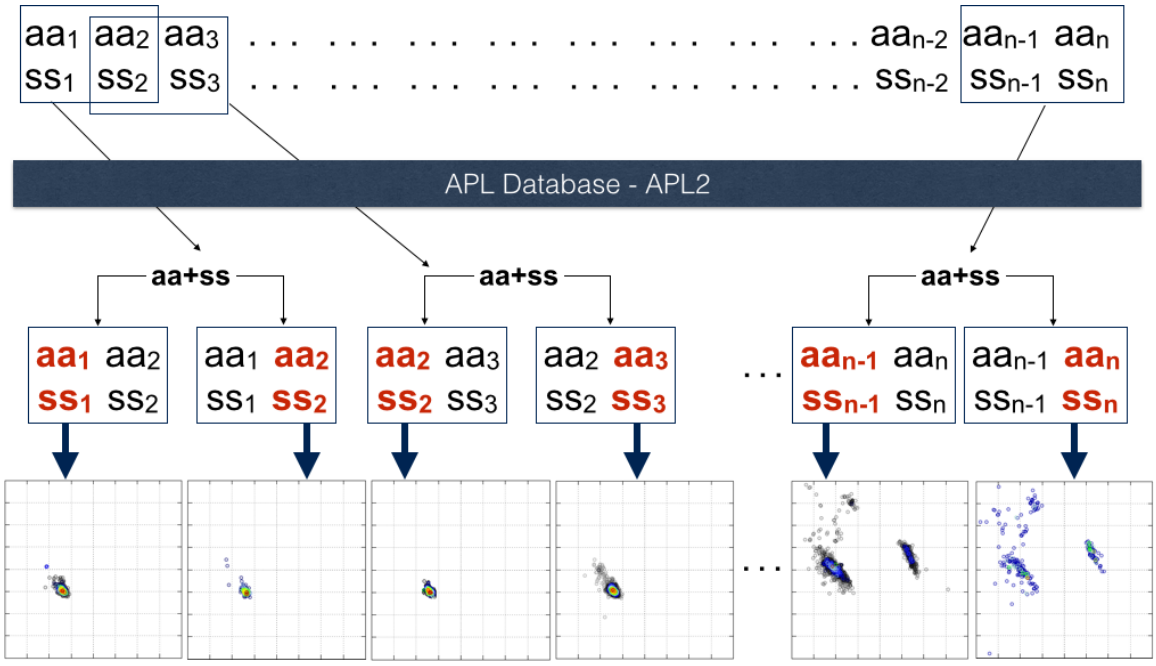
Figura 4.8: Representação gráfica do modelo APL1 desenvolvido e implementado no servidor NPAS



Fonte: Borguesan et al. (2015b).

APL*$_2$*: esse grupo considera um único vizinho para analisar a influência. Diferente do APL1 que considera apenas o par (aa_i,ss_i) o APL2 gera dois grupos de APL. O primeiro é analisando qual a influência do par da direita (aa_{i+1},ss_{i+1}) sobre o par anterior (aa_i,ss_i). O segundo é a influência do par da esquerda (aa_{i-1},ss_{i-1}) sobre o par posterior (aa_i,ss_i). Devido a essa divisão, sempre dois arquivos são formados: APL2-Esquerda e APL2-Direita. Está divisão é baseada no trabalho de Ting et al. (2010b), onde é verificado a importância de cada vizinho (esquerda e direita) separadamente. A Figura 4.9 mostra a estrutura utilizada pelo servidor NPAS para gerar a APL2.

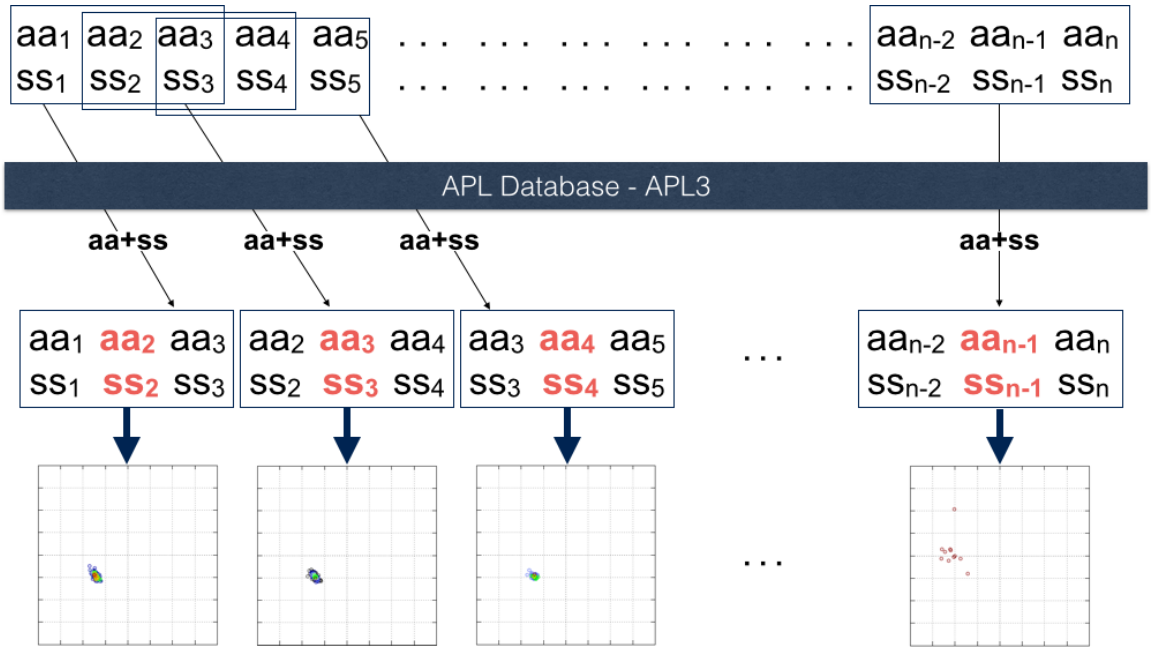
Figura 4.9: Representação gráfica do modelo APL2 desenvolvido e implementado no servidor NPAS



Fonte: Borguesan et al. (2015b).

***APL*$_3$:** esse grupo é chamado de influência completa, pois considera o efeito que os pares da esquerda (aa_{i-1},ss_{i-1}) e da direita (aa_{i+1},ss_{i+1}) tem sobre o par central (aa_i,ss_i). A Figura 4.10 mostra a estrutura utilizada pelo servidor NPAS para gerar a APL3.

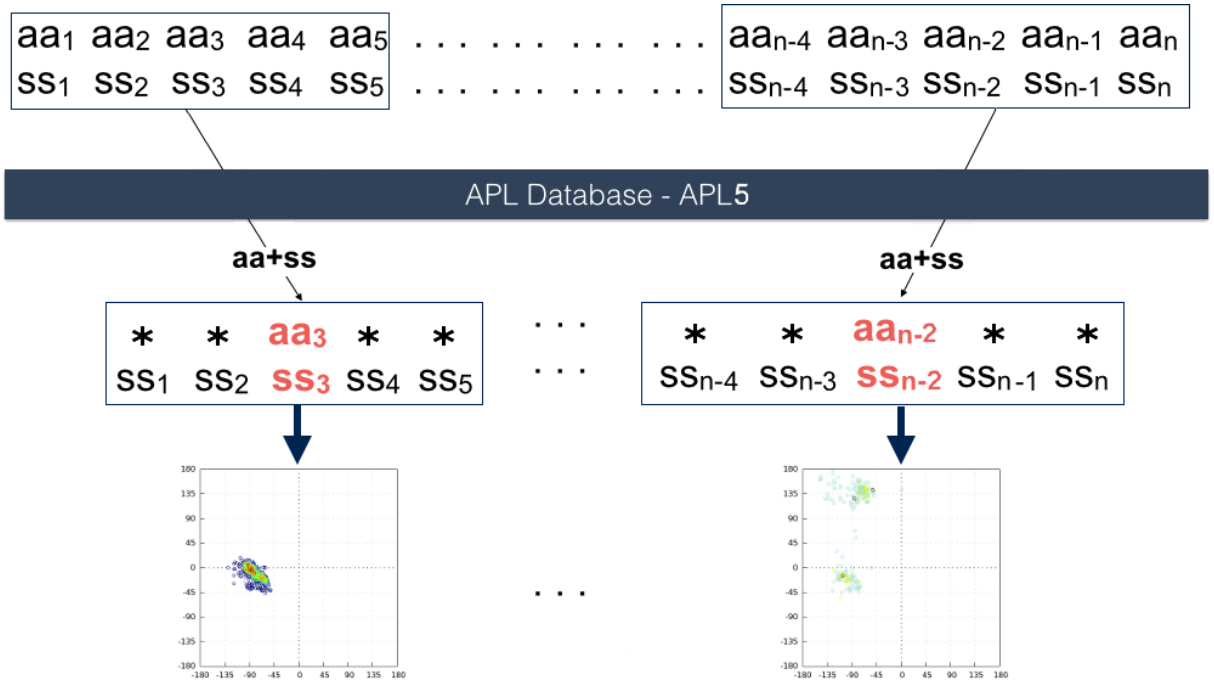
Figura 4.10: Representação gráfica do modelo APL3 desenvolvido e implementado no servidor NPAS



Fonte: Borguesan et al. (2015b).

Como o grupo APL3 procura exatamente a trinca ($(aa_{i-1}$,$ss_{i-1})$, $(aa_i$,$ss_i)$, $(aa_{i+1}$,$ss_{i+1})$), ele tem uma informação bem mais específica sobre a proteína alvo. Porém, quanto mais específica for esta informação, menor será a quantidade de dados resultantes. Em alguns casos essa informação acaba sendo muito específica, ocasionando a perda da generalidade do método. Devido a essa especificidade, em alguns casos, nenhum dado contendo exatamente a mesma trinca inicial é encontrado. Para tentar contornar esse problema foi desenvolvido o método ***APLCentral*** que mantém apenas o aminoácido central fixo e a sequência completa da estrutura secundária de tamanhos 5, 7 e 9. Assim, para gerar o APLCentral5 foi considerado a sequência [(*,ss_{i-2}), (*,ss_{i-1}), $(aa_i$,$ss_i)$, (*,ss_{i+1}), (*,ss_{i+2})] onde o * pode ser qualquer aminoácido. A Figura 4.11 mostra como é a estrutura utilizada pelo servidor NPAS para gerar a APLCentral5.

Figura 4.11: Representação gráfica do modelo APLCentral5 desenvolvido e implementado no servidor NPAS



Fonte: Borguesan et al. (2015b).

Todos os modelos de APLs gerados contém a informação estrutural dos ângulos diedros (phi, psi) com a sua respectiva frequência de ocorrência. Esses arquivos também contém os ângulos de torção chis e ômega de cada ocorrência, atribuídos pelo PROMOTIF, conforme comentado na Seção 4.2.3.

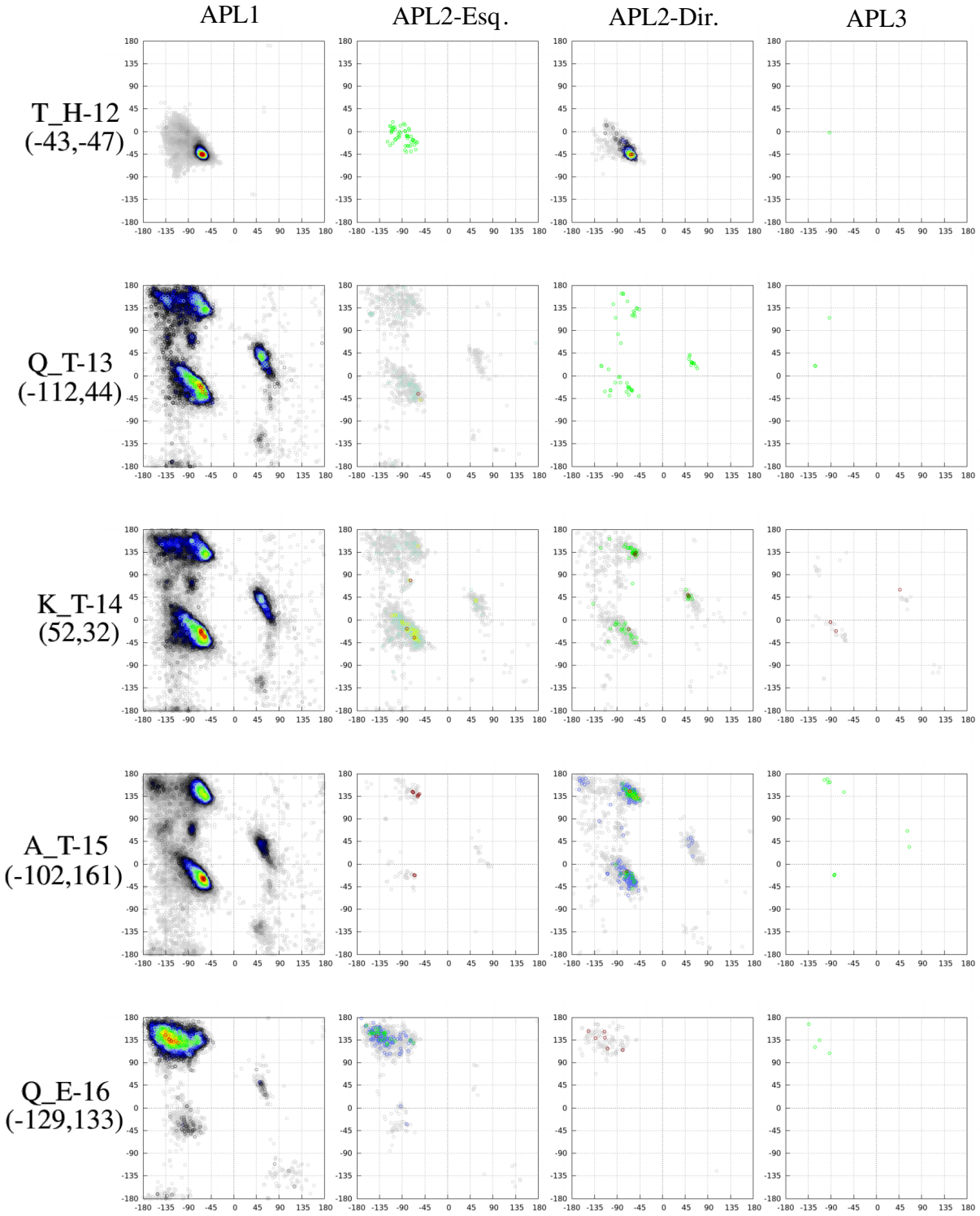
4.3.2 Extração de Conhecimento

Entretanto, uma questão crucial para poder utilizar este conhecimento gerado pelo servidor NPAS é entender que quanto maior o modelo de APL (APL3 e APLCentral9), mais difícil será encontrar dados similares na base de dados. Da mesma forma que quanto menor o modelo de APL (APL1 e APLCentral5), maior será o número de dados similares encontrados na base de dados AASS. As Figuras 4.12 e 4.13 apresentam um exemplo desta questão onde a proteína com o código PDB 1ACW foi selecionada e teve os arquivos das APL1, APL2, APL3 e APLCentral5, APLCentral7 e APLCentral9 gerados, utilizando as sequências primárias e secundárias entre os resíduos 12 e 16 desta proteína. A sequência primária dessa proteína é "VSCEDCPEHCSTQKAQAKCDNDKCVCEPI" e a sua sequência secundária é "CCCHHHHHHHHHTTTEEEEETTEEEEECC". A Figura 4.12 apresenta a preferência conformacional dos aminoácidos entre as posições 12 e 16 para as APL1 na primeira coluna, APL2-Esquerda na segunda coluna, APL2-Direita na terceira coluna e a APL3 na última coluna. A Figura 4.13 apresenta a preferência conformacional dos aminoácidos entre as posições 12 e 16 para a APLCentral5 na primeira coluna, APLCentral7 na segunda coluna e APLCentral9 na última coluna.

A coluna inicial das duas Figuras 4.12 e 4.13, representa o aminoácido central com a sua estrutura secundária atribuído pelo STRIDE e o índice desse par na sequência da estrutura 1ACW. O valor abaixo desse índice são os ângulos de torção (phi, psi) presentes na estrutura experimental. Com isto, é possível analisar que mesmo aumentando a especificidade da APL o método ainda consegue chegar em uma região próxima do valor da experimental.

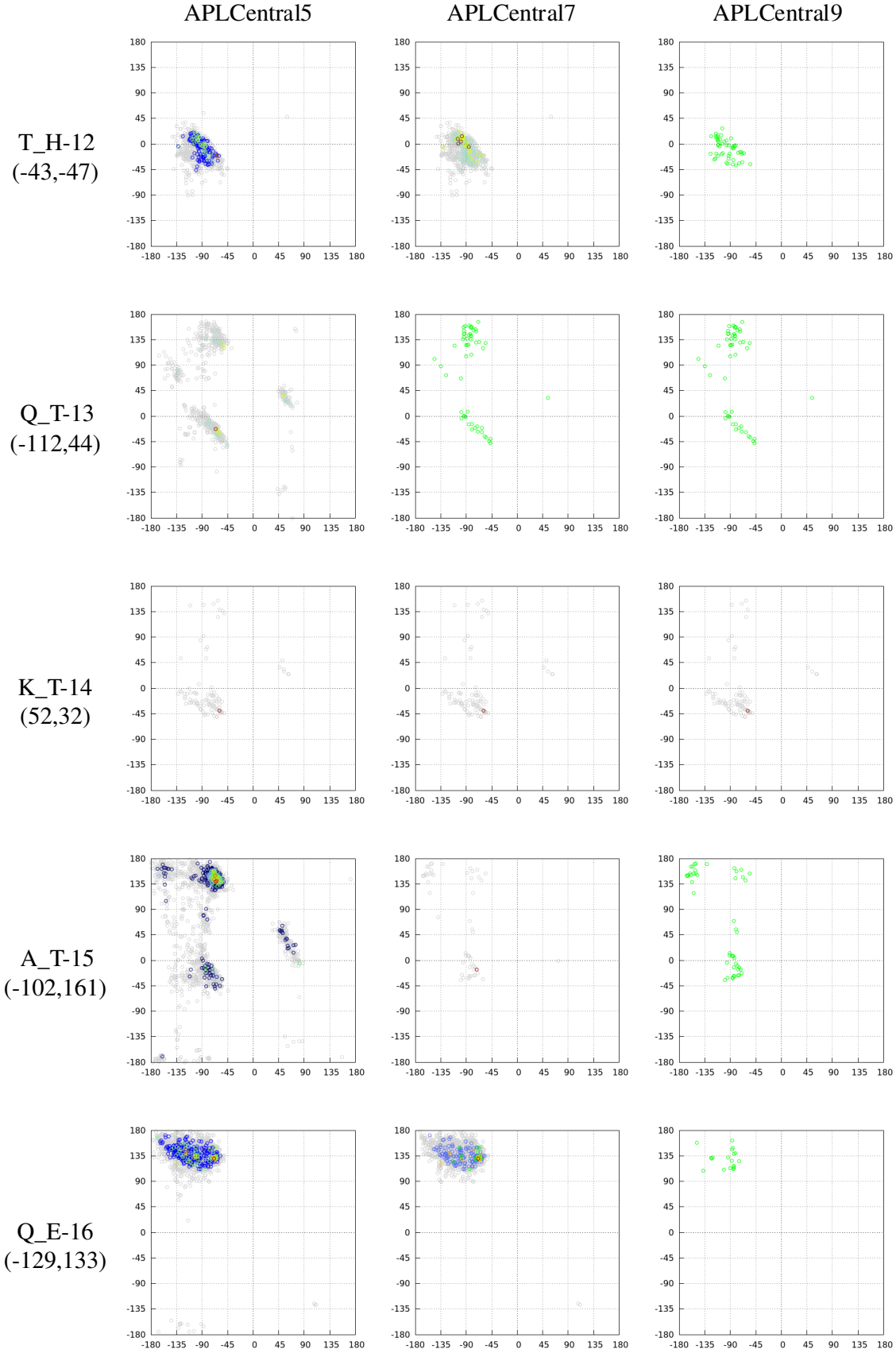
As Figuras 4.12 e 4.13, também mostram que as APL3 e APLCentral9 são as que apresentam o menor conjunto de pontos, enquanto as APL1 e APLCentral5 são as que contém a maior quantidade de pontos do modelo. Permitindo assim, afirmar que a melhor maneira de utilizar esse conhecimento em uma meta-heurística para predição da estrutura 3D de proteínas seja combinar essas APLs, a fim de manter a generalidade das APLs mais densas e a especificidade das APLs mais restritas. As preferências conformacionais para os outros índices dessa estrutura 1ACW, podem ser encontrados no Apêndice B desta dissertação. As figuras do Apêndice B mostram que regiões de hélices (que são mais regulares) são bem definidas independente da APL utilizada. Entretanto, as regiões de folhas e principalmente de voltas apresentam uma maior variação, ficando mais restrito conforme a APL que utilizar.

Figura 4.12: Comparativo entre os arquivos de APL1, APL2-Esq., APL2-Dir. e APL3 para a estrutura com o código no PDB de 1ACW entre os aminoácidos 12 e 16. Os valores na primeira coluna entre os parênteses representam o valor de PHI e PSI esperados



Fonte: Borguesan et al. (2015b).

Figura 4.13: Comparativo entre os arquivos de APLCentral5, APLCentral7 e APLCentral9 para a estrutura com o código no PDB de 1ACW entre os aminoácidos 12 e 16. Os valores na primeira coluna entre os parênteses representam o valor de PHI e PSI esperados



Fonte: Borguesan et al. (2015b).

4.4 FM-B Lib

Outra abordagem bastante utilizada atualmente pelos grupos de pesquisa que tentam resolver o problema da predição da estrutura 3D de proteínas é a baseada em fragmentos (PARK, 2005; GRONT et al., 2011; OLIVEIRA; SHI; DEANE, 2015). Considerando apenas o último CASP, como apresentado no Capítulo 3, é possível analisar que a maioria dos métodos que ficaram entre os melhores resultados utilizaram alguma abordagem baseada em fragmentos, seja para refinar estruturas (ZHANG; LIANG; ZHANG, 2011), ou para gerar novos indivíduos (KIM; CHIVIAN; BAKER, 2004; ZHAO; PENG; XU, 2010; XU; ZHANG, 2012).

4.4.1 Base de Dados

Baseado nesses trabalhos, uma nova biblioteca de fragmentos foi desenvolvida chamada de *FragMent-Based LIBrary* (FM-B Lib). Para desenvolver essa abordagem, foi utilizado a mesma base de dados das seções anteriores (AASS), onde 11,130 estruturas de boa qualidade depositadas no PDB foram selecionadas e tiveram suas estrutura secundárias atribuídas pelo STRIDE.

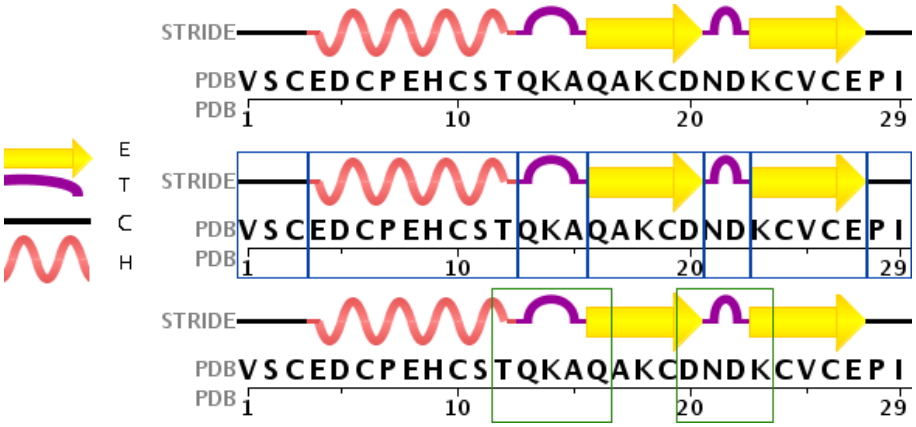
4.4.2 Extração de Conhecimento

Para gerar os fragmentos a nossa abordagem é dividida em dois grupos. O primeiro grupo consiste em fragmentos de uma mesma estrutura secundária. Por exemplo, uma estrutura como a 1ACW, apresentada na seção anterior, que contém a estrutura secundária igual a "CCCHHHHHHHHHTTTEEEEETTEEEEECC", nesse grupo ela seria fragmentada em sete partes: CCC, HHHHHHHHH, TTT, EEEEE, TT, EEEEE e CC. Cada uma dessas partes são consideradas como índices que vão conter a informação dos aminoácidos bem como os ângulos diedros da sua cadeia principal e da lateral.

O segundo grupo tem como prioridade fragmentos de voltas e regiões desordenadas, onde ocorre a união entre duas estruturas secundárias do tipo de hélice ou folha. Por exemplo, usando a mesma estrutura da 1ACW, que contém a estrutura secundária igual a "CCCHHHHHHHHHTTTEEEEETTEEEEECC", nesse grupo ela seria fragmentada em duas partes: HTTTE e ETTE. Cada uma dessas partes são consideradas como índices e também vão conter a informação dos aminoácidos bem como os ângulos diedros da sua cadeia principal e da

lateral. A Figura 4.14 a seguir mostra como é feita essas duas fragmentações da base de dados.

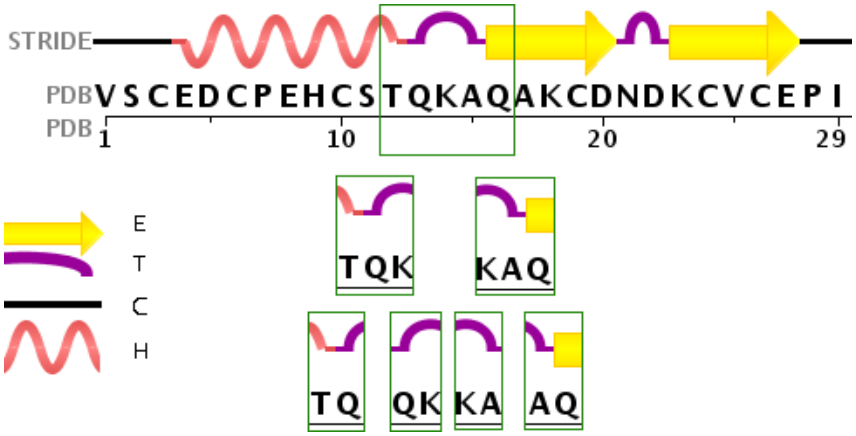
Figura 4.14: Modelo de como ocorre a fragmentação do método FM-B Lib aplicado na estrutura 1ACW. A seleção em azul representa a fragmentação sequêncial feita pelo primeiro grupo e a seleção em verde representa a fragmentação por conectores de voltas e regiões desordenadas feito pelo segundo grupo



Fonte: Adaptado de Berman et al. (2000).

Entretanto, essa técnica de fragmentação acaba limitando o método de predição para estruturas que já contém algum fragmento similar na base de dados (TBM). Para contornar esse problema e permitir que esse conhecimento possa ser utilizado em métodos para predição de estruturas FM também foram fragmentados os índices gerados pelos dois grupos anteriores. Deste modo, um índice do tipo HTTTE, é fragmentado em outros dois índices: HTT e TTE. Esses dois índices são então fragmentados em outros 4 índices: HT, TT, TT e TE. Deste modo, é possível gerar pequenos fragmentos que possibilitam a construção de qualquer estrutura de proteína. A Figura 4.15 apresenta como é feita essa fragmentação intermediária gerada a partir dos fragmentos maiores.

Figura 4.15: Modelo de como ocorre a fragmentação intermediária gerada a partir dos fragmentos maiores, aplicado na estrutura 1ACW



Fonte: do autor (2016).

Os fragmentos são limitados em pelo menos 2 resíduos para serem adicionados a base. Com isso a base ficou com mais de 5,000 tipos diferentes de fragmentos com tamanhos de 2 até 30 resíduos de aminoácidos contendo mais de 10 milhões de fragmentos de proteínas.

4.5 Resumo do Capítulo

Neste capítulo, foram apresentadas diferentes formas de transformar os dados do PDB a fim de extrair conhecimento que possa ser útil para predição da estrutura 3D de proteínas. Entre esses estudos foram citados a APL (que extrai conhecimento local de proteínas), o NPAS (que extrai o conhecimento da vizinhança de proteínas) e o FM-B Lib (que extrai conhecimento de fragmentos de proteínas). Porém, transformar a base de modo a facilitar o uso dos dados é apenas o primeiro passo no desenvolvimento de uma solução para o problema de predição da estrutura 3D de proteínas. O segundo passo é utilizar esses dados transformados de modo inteligente aplicando a um meta-heurísticas a fim de reduzir o espaço de busca conformacional que o problema possui.

No próximo capítulo são apresentados os conceitos padrões de Algoritmos Genéticos e técnicas para resolver problemas multimodais, que irão servir de base para o desenvolvimento de uma nova meta-heurística capaz de utilizar o conhecimento apresentado nesse capítulo para tentar auxiliar a resolução do problema da predição da estrutura 3D de proteínas.

5 ALGORITMO GENÉTICO E PROBLEMAS MULTIMODAIS

5.1 Introdução

Conforme comentado no Capítulo 4, transformar a base de modo a facilitar o uso dos dados é apenas o primeiro passo no desenvolvimento de uma solução para o problema da predição da estrutura tridimensional de proteínas. O segundo passo é utilizar esses dados transformados de modo inteligente aplicando à meta-heurísticas a fim de reduzir o espaço de busca conformacional que o problema em questão possui. Meta-heurísticas designam uma classe de métodos computacionais aproximativos que otimizam um problema através de um processo iterativo guiado por uma heurística combinando de forma inteligente diferentes conceitos para explorar o espaço de busca (LUKE, 2009; MUCHERINO; SEREF, 2009; BOUSSAÏD; LEPAGNOT; SIARRY, 2013). Meta-heurísticas fazem algumas ou nenhuma suposição sobre o problema a ser otimizado e podem buscar grandes espaços de soluções candidatas. Entretanto, não há garantias de que estas encontrarão uma solução ótima ao problema (BLUM; ROLI, 2003). As técnicas mais comuns aplicadas ao problema da predição da estrutura 3D de proteínas são as baseadas em Algoritmos Evolutivos (UNGER, 2004; DORN; BURIOL; LAMB, 2014). Esses algoritmos são tradicionalmente usados para encontrar apenas um resultado ótimo da função objetivo, entretanto, alguns problemas podem ter várias soluções ótimas se a função for multimodal (GLIBOVETS; GULAYEVA, 2013). É o caso da predição da estrutura 3D de proteínas, onde, devido à rugosidade do panorama energético, um mesmo valor de aptidão pode representar diferentes conformações estruturais para a mesma proteína alvo (BRYNGELSON et al., 1995; WOLYNES, 2005; BOEHR; NUSSINOV; WRIGHT, 2009).

Neste capítulo são apresentados os conceitos básicos de Algoritmos Genéticos, sua organização padrão e trabalhos onde o Algoritmo Genético foi aplicado para resolver o problema da predição da estrutura tridimensional de proteínas. Também são apresentadas as principais técnicas para tratar problemas multimodais e sua aplicabilidade para resolver o problema da predição de estrutura 3D de proteínas.

5.2 Algoritmo Genético

Algoritmos Genéticos (AG) são heurísticas de busca baseados nos conceitos evolucionários de Darwin e da seleção natural (HOLLAND, 1975). AG são modelados através

do uso de populações de indivíduos, onde cada indivíduo representa uma solução possível para um determinado problema. Para cada indivíduo é calculado um valor de aptidão (função de *fitness*) que indica quão bem o indivíduo resolve o problema. Para cada iteração do algoritmo, chamada de geração, diferentes indivíduos são selecionados (Pais) e a recombinação desses indivíduos gera novas soluções (Filhos) que são inseridas na população. O valor de aptidão é quem indica quais indivíduos ficam ou saem da população. Esses indivíduos ainda podem sofrer influências de operadores de mutações (GOLDBERG, 1989). O algoritmo continua até que algum critério de parada seja satisfeito. É também bastante comum a utilização de técnicas para manter a diversidade da população durante as gerações e também abordagens que evitem que o algoritmo fique preso em mínimos locais (SRINIVAS; PATNAIK, 1994; LUKE, 2009). O Algoritmo 1 a seguir apresenta o pseudocódigo de um Algoritmo Genético padrão contendo os seus principais componentes (Inicialização, Avaliação, Seleção, Recombinação e Mutação). Na próxima seção foram detalhados cada um desses operadores genéticos, bem como alguns conceitos mais utilizados.

Algoritmo 1: Pseudocódigo para um Algoritmo Genético Padrão.

```
1 Inicializa a população de indivíduos;
2 Avalia cada indivíduo da população;
3 repita
4 |   Seleciona indivíduos (Pais);
5 |   Recombina esses Pais gerando novos indivíduos (Filhos);
6 |   Aplica Mutação nos Filhos gerados;
7 |   Avalia os novos indivíduos;
8 |   Seleciona quais indivíduos serão mantidos para a próxima geração;
9 até Alguma condição for satisfeita
```

5.2.1 Organização

Pode-se dizer que a maioria dos métodos chamados de Algoritmos Genéticos possuem pelo menos os seguintes elementos em comum: População de indivíduos; Avaliação da População; Seleção de indivíduos; Recombinação de indivíduos; e Mutação aleatória sobre novos indivíduos. Nas próximas seções são apresentados detalhes sobre cada um desses operadores genéticos.

5.2.2 Inicialização

Indivíduos das implementações padrões de AG são constantemente representados como um conjunto de dados binários (0 ou 1). Cada indivíduo é considerado como uma possível solução no espaço de busca. Entretanto, com o passar dos anos, diversas outras formas de indivíduos foram testadas como alfabetos, valores reais, pontos cartesianos, entre outros (PEDERSEN; MOULT, 1996; MITCHELL, 1999; HOQUE; CHETTY; SATTAR, 2009). Essa definição do formato do indivíduo depende do problema que se está tentando resolver. Trabalhos de predição da estrutura 3D de proteínas, por exemplo, constantemente representam indivíduos como um vetor de ângulos de rotação (DORN et al., 2013; BORGUESAN et al., 2015a).

5.2.3 Avaliação

Como o AG é baseado na ideia de sobrevivência dos indivíduos mais adaptados, faz-se necessário uma forma de avaliar a aptidão desses indivíduos (MITCHELL, 1999). A função de Avaliação do AG então atribui um valor de aptidão (*fitness*) para cada indivíduo da população atual. Este valor depende de quão bem o indivíduo resolve o problema. A AG pode ser implementado para procurar valores mínimos ou máximos da função de Avaliação que representem o melhor indivíduo da população atual (MANNING; SLEATOR; WALSH, 2013). Trabalhos de predição da estrutura 3D de proteínas, por exemplo, utilizam como função de Avaliação uma combinação de energias calculadas através das interações interatômicas da proteína alvo (DORN; BURIOL; LAMB, 2011; BORGUESAN et al., 2015a).

5.2.4 Seleção

Uma vez que a população foi avaliada através de uma função de aptidão, várias técnicas podem ser aplicadas para selecionar os indivíduos que serão usados para produzir soluções que serão adicionadas na próxima população. Os métodos de seleção tendem a sempre favorecer a escolha de soluções com melhores valores de aptidão, garantindo assim que boas soluções permaneçam ou que gerem também boas soluções para a próxima população (GOLDBERG, 1989; XIE; ZHANG, 2013). Dentre as várias técnicas existentes é possível destacar a Seleção por Roleta (HOLLAND, 1992; MITCHELL, 1999), Seleção por Ranqueamento (BAKER,

1985; MITCHELL, 1999), Seleção por Torneio (GOLDBERG; DEB, 1991; MITCHELL, 1999) e Elitismo (JONG; ALAN, 1975; GONÇALVES; RESENDE, 2011; LENO et al., 2013).

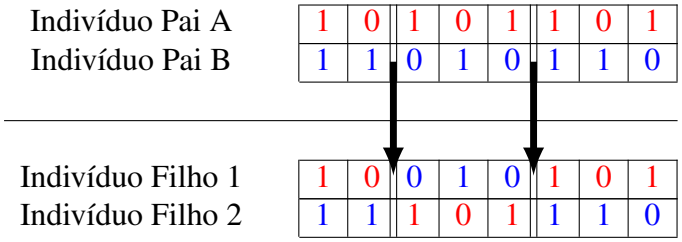
- **Seleção por Roleta**: corresponde a ideia de que cada indivíduo possui um setor em uma roleta baseado no quão apto é esse indivíduo. Assim, quanto melhor o valor de aptidão, maior será a região da roleta atribuída a esse indivíduo. Garantindo também que indivíduos de baixa aptidão tenham ao menos uma pequena probabilidade de ser selecionado, mantendo a diversidade do método;
- **Seleção por Ranqueamento**: tem como característica principal o ordenamento dos indivíduos, baseado em seus valores de aptidão e sua probabilidade de seleção é referente ao seu posicionamento nesta população ordenada;
- **Seleção por Torneio**: consiste em selecionar um conjunto de indivíduos da população aleatoriamente de tamanho menor que o tamanho da população. Então o indivíduo de melhor aptidão desse conjunto é selecionado;
- **Elitismo**: é uma abordagem que tem como ideia principal forçar o AG a manter conjunto com os melhores indivíduos para as próximas populações. Evitando assim que boas soluções da população atual sejam alteradas durante o processo de mutação ou recombinação.

Entretanto, segundo Goldberg e Deb (1991), os métodos de seleção se tornam equiparáveis quando corretamente parametrizados. Assim, métodos de seleção podem ser escolhidos de acordo com o problema e a implementação (GOLDBERG; DEB, 1991; MITCHELL, 1999).

5.2.5 Recombinação

Para gerar novos indivíduos, operadores genéticos como recombinação (*crossover*) são utilizados. Esse operador tem como principal objetivo a combinação dos valores de dois indivíduos, chamados de pais, gerando novos indivíduos, chamados de filhos (MITCHELL, 1999). A Figura 5.1 apresenta um modelo simples de recombinação entre dois indivíduos (Pai A e Pai B) que são recombinados a partir de dois pontos de corte aleatórios gerando dois novos indivíduos (Filho 1 e Filho 2) (GOLDBERG, 1989). Outra variedade desse operador é a recombinação uniforme, que consiste em sortear de qual Pai (A ou B) cada posição do indivíduo Filho irá receber (MITCHELL, 1999). Existe uma variedade de operadores de recombinação que, geralmente, são escolhidos baseados no problema a ser otimizado (MITCHELL, 1999).

Figura 5.1: Representação simples da recombinação entre dois indivíduos com conjunto de dados binários utilizando dois pontos de corte

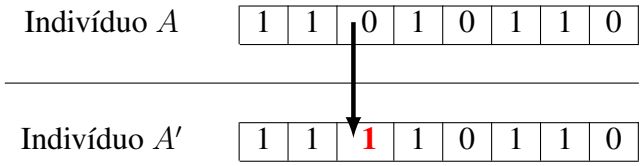


Fonte: do autor (2016).

5.2.6 Mutação

Quando a população é criada, através de seleção e da recombinação, ela é modificada utilizando um operador de mutação. A mutação é relacionada como uma modificação em uma pequena parte de um determinado indivíduo (HOLLAND, 1975). Geralmente o indivíduo escolhido para mutação é o filho gerado a partir da recombinação. Um exemplo de uma mutação simples é apresentado na Figura 5.2, onde apenas uma posição do indivíduo A é alterado, gerando um novo indivíduo A'.

Figura 5.2: Representação simples de uma mutação sobre um indivíduos com conjunto de dados binários



Fonte: do autor (2016).

Assim como os operadores de recombinação, também existe uma grande variedade de abordagens para mutação dependendo da forma de representação do indivíduo e do problema a ser resolvido (GOLDBERG, 1989; MITCHELL, 1999; MANNING; SLEATOR; WALSH, 2013).

5.2.7 Outros Operadores

Além dos operadores descritos acima, AG constantemente utilizam outras abordagens, principalmente para controlar a diversidade do AG. Esses operadores trabalham com a

ideia de prevenir que o AG fique preso em uma região local, geralmente ocasionado por uma convergência precoce da população em um mínimo que não é o global da função (MITCHELL, 1999). Outro operador bastante utilizado é o de reinício da população, onde somente uma parte dos melhores indivíduos são mantidos, enquanto os outros indivíduos são gerados novamente do mesmo modo que na inicialização, influenciando na diversidade do método (HOUCK; JOINES; KAY, 1996).

5.2.8 Critérios de Parada e Próxima População

Definir quando um AG deve terminar pode ser uma tarefa bastante complicada que dependente do problema investigado. Se o algoritmo roda por pouco tempo, pode representar uma solução que não é o mínimo global, ou se rodar por muito tempo pode ocasionar em um desperdício de poder computacional (MITCHELL, 1999; SAFE et al., 2004). Assim, o critério de parada comumente utilizado em AG é determinado pelo número de avaliações da função de aptidão ou por um tempo de execução pré-definido. Esse critério é normalmente combinado com a convergência da população. Assim, quando todos os indivíduos da população forem iguais, ou muito similares, o AG para. Porém, novamente, o critério de parada deve ser definido baseado no problema que se está tentando resolver (GOLDBERG, 1989; MITCHELL, 1999).

Enquanto o critério de parada não é satisfeito, uma próxima população é criada utilizando todos os operadores apresentados acima. Comumente o tamanho da população se mantém inalterado e é feita apenas a substituição de indivíduos da população, buscando sempre evoluir os indivíduos mais aptos sem perder a diversidade da população (GOLDBERG, 1989; MITCHELL, 1999).

5.3 Algoritmo Genético Aplicado no Problema da Predição da Estrutura de Proteínas

Métodos baseados em conceitos evolucionários, como o AG, são aplicados com uma alta taxa de sucesso em um grande número de problemas da Bioinformática Estrutural. No contexto do problema de predição da estrutura 3D de proteínas são vários os trabalhos que utilizam alguma variação de AG para tentar solucionar o problema (UNGER, 2004). Elofsson et al. (1995) apresentam um Algoritmo Genético com busca local no espaço conformacional de ângulos diedros de uma proteína, buscando predizer a sua estrutura nativa. Unger (2004) apresenta uma revisão de como Algoritmos Genéticos podem ser utilizados para

auxiliar o problema de predição da estrutura de proteínas. Park (2005) por exemplo, usa o Algoritmo Genético para recombinar fragmentos de proteínas procurando a conformação de menor energia. Cutello e Park (2006) utilizam Algoritmos Genéticos para resolver uma representação multiobjetivo do problema de determinação da estrutura 3D de uma proteína. Hoque et al. (2006) apresenta um Algoritmo Genético guiado ao problema de determinação do enovelamento de proteínas utilizando o modelo de hidrofobicidade. Dorn et al. (2013) e Borguesan et al. (2015a) utilizam um AG combinado com informações estruturais para reduzir o espaço de busca conformacional do problema da predição da estrutura 3D de proteínas.

Dentre esses métodos, os que obtiveram maior sucesso, estão os que usaram uma representação de indivíduos na forma de ângulos de torção, combinados com função de aptidão para tentar representar ao máximo as interações físico-química que ocorrem no processo de enovelamento de proteínas (DORN; BURIOL; LAMB, 2011; DORN et al., 2013; BORGUESAN et al., 2015a).

Porém, o Algoritmo Genético tradicionalmente é usado para encontrar apenas um resultado ótimo da função objetivo (problema unimodais), entretanto, alguns problemas podem ter várias soluções ótimas (problemas multimodais) (GLIBOVETS; GULAYEVA, 2013). Pode ser considerado o caso da predição da estrutura 3D de proteínas, onde um mesmo valor de aptidão pode representar diferentes conformações estruturais para a mesma proteína alvo. Ou até mesmo, ter vários caminhos de enovelamentos com um mesmo valor de aptidão. Isso ocorre pelo fato, já comentado anteriormente, de que o processo de enovelamento proteico é apenas parcialmente conhecido, impossibilitando uma função de aptidão que represente com 100% de confiança o processo de enovelamento, ocasionando na alta rugosidade da superfície da função de aptidão (BRYNGELSON et al., 1995; WOLYNES, 2005; BOEHR; NUSSINOV; WRIGHT, 2009). Na próxima seção são apresentadas abordagens multimodais para contornar esse problema, onde subpopulações evoluem paralelamente não permitindo a convergência da população para um mínimo local.

5.4 Abordagens Multimodais

Embora os AG utilizem a ideia de população de indivíduos, é possível que quando duas soluções diferentes mas com o mesmo valor de aptidão apareçam em uma mesma população, assim o AG padrão dá prioridade para um deles nos processos de seleção e recombinação. Esse efeito, chamado de deriva genética, faz com que a população convirja para um único mínimo. Diversas técnicas foram criadas durante os anos para contornar esse problema a

fim de transformar um algoritmo que somente resolve problemas unimodais em algoritmos que também resolvam problemas multimodais. As técnicas mais conhecidas para realizarem essas tarefas são: a baseada em Nichos (*crowding*) (JONG; ALAN, 1975); a baseada em Compartilhamento de Recursos (*Fitness Sharing*) (GOLDBERG; RICHARDSON, 1987); e a baseada na Seleção por Torneio Restrito (RTS) (HARIK, 1995).

5.4.1 Nichos (*Crowding*)

No modelo de Nichos (*Crowding*), proposto por Jong (JONG; ALAN, 1975), grupos (nichos) são formados baseado na similaridade dos indivíduos permitindo uma investigação paralela de vários mínimos/máximos de funções. Assim, quando um novo indivíduo é gerado para adicionar a população, o método primeiro procura a qual nicho de indivíduos ele pertence e substitui o indivíduo desse grupo que for menos apto. Se um indivíduo menos apto não for encontrado, a solução pode ser então descartada (JONG; ALAN, 1975; MENGSHOEL; GOLDBERG, 2008).

5.4.2 Compartilhamento de Recursos (*Fitness Sharing*)

Com mesma ideia de similaridade entre os indivíduos sobre qual o método baseado em Nichos foi desenvolvido, a abordagem de Compartilhamento de Recursos (*Fitness Sharing*) (GOLDBERG; RICHARDSON, 1987) onde recursos dos indivíduos são compartilhados com seus vizinhos definidos pela sua similaridade. O objetivo é reduzir o valor de aptidão de grupos de indivíduos com um alto grau de semelhança, diminuindo assim a chance de seleção nos operadores genéticos. Com essa abordagem o crescimento descontrolado de indivíduos similares dentro de uma população é evitado.

5.4.3 Seleção por Torneio Restrito (RTS)

Outra variação para problemas multimodais são métodos baseados em evolução que utilizam técnicas de Seleção por Torneio Restrito (*Restricted Tournament Select*) onde a competição só ocorre entre indivíduos similares (HARIK, 1995). Essa abordagem utiliza o mesmo conceito da técnica de Nichos, onde novos indivíduos devem substituir indivíduos similares da população. Essa abordagem favorece o surgimento de novos nichos, permitindo

que eles evoluam e não sejam comparados com um indivíduos diferentes com maior valor de aptidão no momento (ótimo local) (HARIK, 1995).

Uma variação da abordagem RTS foi proposta por Roy e Parmee (1996), que desenvolveram uma versão adaptativa do modelo RTS, chamada de ARTS (*Adaptative Restricted Tournament Select*). Essa versão tem o objetivo de agrupar os indivíduos similares em grupos (nichos) durante as gerações de um Algoritmo Genético. Para adicionar o Filho gerado da recombinação o método procura em qual nicho esse indivíduo se encaixa, então procura o indivíduo menos apto deste nicho para substituir pelo Filho gerado. O Algoritmo 2 apresenta o pseudocódigo dessa abordagem.

Algoritmo 2: Pseudocódigo para um Algoritmo Genético utilizando a abordagem ARTS proposta por Roy e Parmee (1996).

```
1  Inicializa a população de indivíduos;
2  Avalia cada indivíduo da população;
3  repita
4  |  Agrupa os indivíduos da população em grupos (nichos);
5  |  Seleciona indivíduos (Pais);
6  |  Recombinar esses Pais gerando novos indivíduos (Filhos);
7  |  Aplica Mutação nos Filhos gerados;
8  |  Avalia os novos indivíduos;
9  |  Procura o grupo mais próximo para adicionar Filho;
10 |  se encontrar um indivíduo nesse grupo com valor de aptidão menor que o Filho
   |  gerado então
11 |  |  substitui o indivíduo pelo Filho gerado;
12 |  senão
13 |  |  descarta o Filho gerado;
14 |  fim
15 até Alguma condição for satisfeita
```

5.5 Estratégias Multimodais Aplicadas no Problema da Predição da Estrutura Tridimensional de Proteínas

Essa abordagem também vem sendo aplicada para tentar resolver o problema de predição da estrutura tridimensional de proteínas. O trabalho de Chen e Hu (2009), apresenta um Algoritmo Evolutivo que combina técnicas de Nichos e uma variação do modelo ARTS sobre o problema da predição da estrutura de proteínas utilizando o modelo de hidrofobicidade para representação da proteína. Seus resultados mostraram que utilizar técnicas de multimodais

melhoram as soluções quando comparados com o Algoritmo Evolutivo padrão. Wong et al. (2010) por exemplo, apresentou um Algoritmo Evolutivo que considera a similaridade estrutural das proteínas como critério no processo de recombinação e mutação dos indivíduos da população. Islam e Chetty (2013) apresentam um Algoritmo Memético onde é feito a clusterização dos indivíduos baseado em suas estruturas e aplica no problema de predição da estrutura de proteínas. Custódio, Barbosa e Dardenne (2014), produziram um algoritmo que combina um AG com um mecanismo de Nichos para tentar solucionar o problema de predição da estrutura de proteínas utilizando modelo hidrofóbico-hidrofílico para representação das proteínas. O método permitiu uma investigação melhor referente ao panorama da minimização da energia e melhora na probabilidade de encontrar a estrutura nativa, mesmo ela não sendo a solução mínima global. Assim, devido à alta rugosidade da superfície da função de aptidão para o problema da predição da estrutura 3D de proteínas, técnicas para preservação da diversidade da população através da investigação de múltiplas regiões, podem ser de grande utilidade por reduzirem a chance de convergência para mínimos locais.

5.6 Resumo do Capítulo

Neste capítulo, foram apresentados os conceitos padrões de Algoritmos Genéticos (Inicialização, Avaliação, Seleção, Recombinação e Mutação) bem como as principais técnicas para tratar problemas onde a função objetivo é multimodal (Nichos, Compartilhamento de Recursos e a Seleção por Torneio Restrito). Esses conceitos são de fundamental importância para o entendimento do próximo capítulo onde o algoritmo desenvolvido é apresentado. Este algoritmo combina um Algoritmo Genético (AG) baseado no método de Seleção por Torneio Restrito Adaptativo (ARTS) utilizando todos os conhecimentos apresentados no Capítulo 4 para tentar solucionar o problema da predição da estrutura 3D de proteínas de forma multimodal.

6 PREDIÇÃO DA ESTRUTURA 3D DE PROTEÍNAS BASEADA EM ALGORITMO GENÉTICO MULTIMODAL - GARTS

6.1 Introdução

Conforme comentado nos capítulos anteriores, o conhecimento estrutural de uma proteína é de fundamental importância para inferir a sua função bem como permitir a produção de novos fármacos para inibir ou ativar suas funções. Porém, devido a vários fatores, determinar a estrutura de proteínas experimentalmente ainda é um processo infactível para grandes polipeptídios. Com isso, técnicas para predição da estrutura tridimensional de proteínas tem grande importância. Para tentar diminuir o espaço de busca do problema de predição da estrutura 3D de proteínas diversos trabalhos com bom desempenho no problema (Capítulo 3) utilizam conhecimento extraído das bases de dados de proteínas determinadas experimentalmente (Capítulo 4). Porém, diferente da maioria dos problemas onde existe apenas uma solução ótima, a predição da estrutura tridimensional de proteínas pode ter diversas soluções consideradas ótimas (Capítulo 5). Isto ocorre, devido ao fato de que as regras que governam o processo de enovelamento dessas proteínas serem apenas parcialmente conhecidas, assim várias soluções podem acabar sendo consideradas ótimas. Entretanto, conforme comentado nos Capítulos 3 e 5, ainda são poucos os trabalhos relacionados que consideram o problema de predição da estrutura 3D de proteínas como multimodal. Com base nestas informações, uma nova meta-heurística foi desenvolvida para auxiliar no problema de predição da estrutura 3D de proteínas. Este método utiliza conhecimento da base de dados experimental aplicado a um Algoritmo Genético (AG) baseado no método de Seleção por Torneiro Restrito Adaptativo (ARTS) dando origem assim ao método proposto, GARTS. Desenvolvendo assim um abordagem que considera a multimodalidade do problema de predição da estrutura 3D de proteínas em conjunto com as vantagens que técnicas baseadas em conhecimento possuem. Na próxima seção é detalhado como se comporta cada parte do método desenvolvido.

6.2 Método Proposto

Todas as rotinas implementadas no desenvolvimento do método foram escritas na linguagem de programação Python versão 2.7. Para poder focar no desenvolvimento da nova meta-heurística as rotinas de representação, transformação e avaliação da aptidão das estruturas geradas pelo método foram implementadas pela plataforma PyRosetta (CHAUDHURY;

LYSKOV; GRAY, 2010) (desenvolvida pela Universidade Johns Hopkins) baseado no método Rosetta (ROHL et al., 2004) (desenvolvido pelo *Baker Lab* apresentado no Capítulo 3). Essa plataforma foi escolhida por ser baseada no programa Rosetta que frequentemente obtêm bons resultados no CASP (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014b).

Como o objetivo do método é auxiliar no problema da predição da estrutura 3D de proteínas, os dados de entrada para execução do algoritmo são a estrutura primária em conjunto com a estrutura secundária de uma proteína alvo, evitando assim que o método necessite fazer a predição da estrutura secundária. Assim, com esses dados de entradas é possível buscar conhecimento utilizando as técnicas apresentadas no Capítulo 4 (APL, NPAS e FM-B Lib) para poder auxiliar uma meta-heurística através da redução do espaço conformacional que o problema da predição da estrutura 3D de proteínas possui. Devido ao complexo panorama energético que este problema possui, com vários mínimos e várias soluções com o mesmo valor de aptidão, foi desenvolvido uma meta-heurística que combina um AG com a técnica ARTS, apresentados no Capítulo 5 desta dissertação. Esta combinação, permitiu aplicar variações nos operadores do AG utilizando informações referentes a características do método ARTS. O Algoritmo 3 apresenta o pseudocódigo do método GARTS desenvolvido. Nas subseções seguintes são apresentados os operadores e as características deste algoritmo.

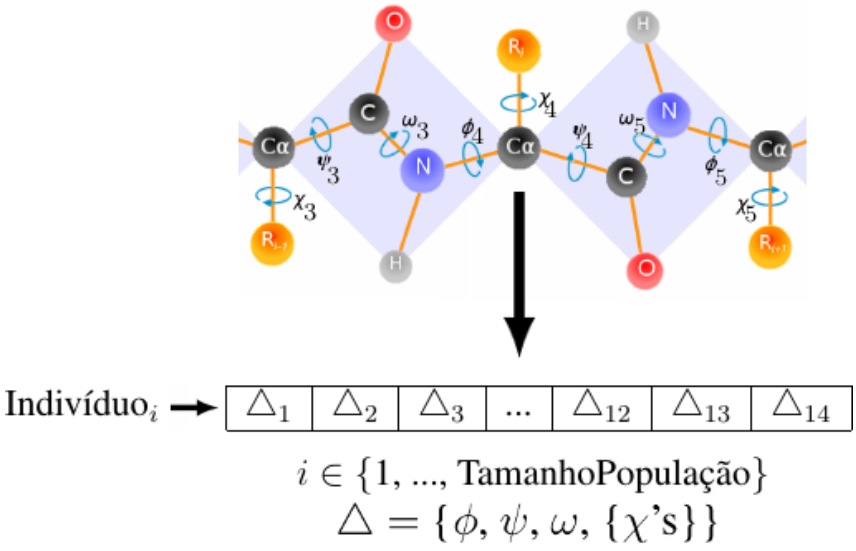
Algoritmo 3: Pseudocódigo do método GARTS proposto.

```
1  Inicializa a população de indivíduos utilizando conhecimento;
2  Agrupa os indivíduos da população em nichos e define um indivíduo central;
3  Avalia cada indivíduo da população;
4  repita
5      repita
6          Seleciona indivíduos (Pais);
7          Recombina esses Pais gerando novos indivíduos (Filhos);
8          Aplica Mutação nos Filhos gerados utilizando conhecimento;
9      até Controle de Diversidade ser satisfeito
10     Avalia os novos indivíduos;
11     Procura o grupo com o indivíduo central mais similar para adicionar o Filho;
12     se encontrar um indivíduo nesse grupo que seja menos apto que o Filho então
13         substitui o indivíduo pelo Filho gerado;
14     senão
15         descarta o Filho gerado;
16         se descartar o Filho gerado um número X de vezes então
17             Reinicia a População;
18         fim
19     fim
20     Agrupa os indivíduos da população em nichos e define um indivíduo central;
21 até Condição de Parada ser satisfeita
```

6.2.1 Inicializa a População de Indivíduos

Conforme explicado no Capítulo 5, Algoritmos Genéticos são formados por populações de indivíduos, e como comentado nos Capítulos 2 e 4, ângulos de torção conseguem representar com bastante detalhes a conformação tridimensional de uma proteína. Com isso, é possível então pensar em um vetor de ângulos de torção que representem o posicionamento e a conformação de uma determinada proteína. Por exemplo, para uma estrutura alvo de 14 resíduos de aminoácidos é gerado um vetor com 14 posições de ângulos onde cada posição tem seus respectivos ângulos *phi*, *psi*, *omega* e *chi's* (lembrando que os ângulos chi's variam entre 4 ou nenhum ângulo, dependendo do aminoácido). A Figura 6.1 apresenta a organização deste indivíduo com 14 conjuntos de ângulos de torção. Para converter esse conjunto de ângulos em uma estrutura tridimensional foi utilizado rotinas de rotação e translação disponibilizados pela plataforma PyRosetta que cria um objeto chamado de *Pose* contendo as coordenadas cartesianas dos átomos que representam a estrutura 3D da proteína.

Figura 6.1: Modelo de um indivíduo contendo 14 conjuntos de ângulos de torção que representam a conformação de uma proteína de 14 resíduos

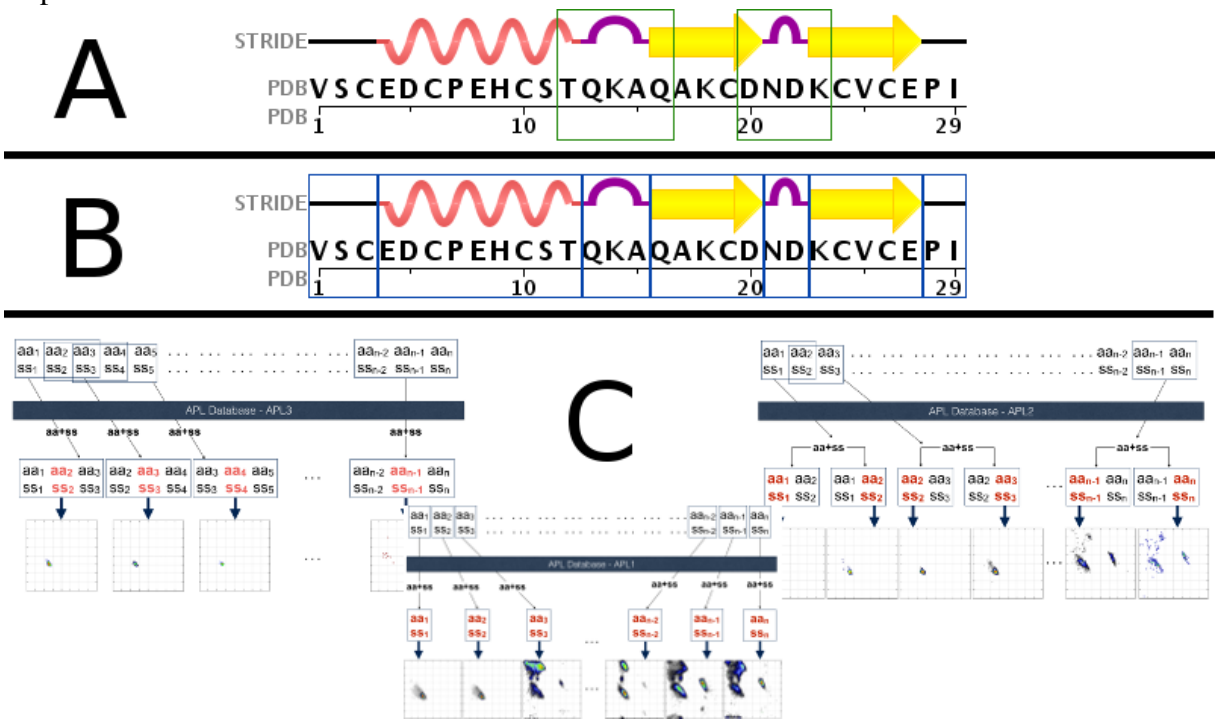


Fonte: Adaptado de Borguesan et al. (2015a).

Para atribuir esses conjuntos de ângulos de torção é utilizada a abordagem de fragmentos *FM-B Lib* detalhada na Seção 4.4. Para isso, o método aplica a abordagem de fragmentar os dados de entrada da proteína alvo (estrutura primária e secundária) para poder pesquisar na base de dados do *FM-B Lib* fragmentos compatíveis que permitem montar o indivíduo. A

Figura 6.2 mostra como é feito esse processo. A prioridade do método é primeiro adicionar fragmentos que vem da região de volta que une duas estruturas regulares como hélice ou folha e encontrar na base de dados fragmentos com a mesma sequência de aminoácidos que dos dados de entrada (Fig. 6.2-A). A segunda etapa do método é adicionar fragmentos de uma mesma estrutura secundária que ainda não foram adicionados (Fig. 6.2-B). E a última parte, caso não encontre algum fragmento na base de dados do *FM-B Lib*, é complementar os espaços em branco da estrutura utilizando ângulos de torção do método APL_3, APL_2 ou APL_1 apresentados nas Seções 4.2 e 4.3 (Fig. 6.2-C). Vale ressaltar que o método, em cada uma das partes (A, B e C), sempre tenta atribuir primeiro o fragmento mais próximo ao início da sequência, utilizando a hipótese de enovelamento sequencial de proteínas (HUARD; DEANE; WOOD, 2006). Caso dois fragmentos comecem na mesma posição o fragmento de maior tamanho (de maior conhecimento) tem a prioridade. Caso o fragmento comece na mesma posição e possua o mesmo tamanho que outro fragmento, o método escolhe aleatoriamente algum deles para montar o indivíduo.

Figura 6.2: Abordagem de inicialização dos indivíduos do método GARTS, combinando a biblioteca de fragmentos FM-B Lib em conjunto com as APL, ambos apresentados no Capítulo 4



Fonte: do autor (2016).

6.2.2 Agrupar Indivíduos da População

Para poder agrupar os indivíduos, antes de escolher o método de agrupamento, é necessário definir qual métrica utilizar para identificar a similaridade entre eles. A métrica mais conhecida e comumente utilizada para identificar a semelhança entre estruturas tridimensionais de proteínas é o RMSD (do inglês *Root Mean Square Deviation*, ou Desvio Médio Quadrático) (ZHANG; SKOLNICK, 2004a). O RMSD pode ser calculado de duas maneiras: utilizando todos os átomos de uma proteína (ENGH; HUBER, 1991) ou utilizando somente um subconjunto de átomos (somente o C_α, por exemplo) (KABSCH; SANDER, 1983b). Para este trabalho, foi utilizado a variação que computa apenas a distância entre os C_α de duas estruturas (Eq. 6.1), considerado uma métrica suficiente para identificar a similaridade do enovelamento global entre duas estruturas (KABSCH; SANDER, 1983b). Para cálculo do RMSD é utilizado a rotina do PyRosetta, onde primeiro alinha-se o centro de massa das duas estruturas, então uma dessas estruturas é rotacionada até chegar em uma superposição ótima, permitindo assim o correto cálculo da distância entre os átomos das duas estruturas.

$$\mathbf{RMSD}(a,b) = \sqrt{\left(\sum_{i=1}^{n} \|da_{C_{\alpha i}} - db_{C_{\alpha i}}\|^2\right)/n}, \tag{6.1}$$

onde $da_{C_{\alpha i}}$ e $db_{C_{\alpha i}}$ são vetores que representam o posicionamento (i) do átomo (C_α) em duas estruturas distintas (a e b) de tamanho n.

Com a métrica para identificar a similaridade entre estruturas definida, é possível escolher o método para agrupar esses indivíduos. Um dos algoritmos comumente utilizado na literatura é o agrupamento hierárquico (JOHNSON, 1967). Essa técnica é utilizada pra agrupar tanto fragmentos de proteínas (TENDULKAR et al., 2004) como modelos de proteínas completos (GRONT; KOLINSKI, 2005; JAMROZ; KOLINSKI, 2013; DANG et al., 2014). O agrupamento hierárquico organiza os itens em hierarquia com uma estrutura no formato de árvore baseado na distância ou semelhança entre os itens (JOHNSON, 1967). Esse agrupamento pode ser aglomerativo ou divisivo. A versão aglomerativa inicia com cada indivíduo em seu grupo e vai agrupando até ficar em apenas um ou k grupos. Já a versão divisiva, inicia com um único grupo e vai dividindo até chegar em cada indivíduo em seu grupo ou em k grupos (JOHNSON, 1967). Existem diversas técnicas para definir a similaridade entre os grupos, porém para este trabalho foi escolhido a forma completa, onde a similaridade entre dois grupos é baseada nos pontos menos similares (mais distantes) entre os dois grupos. Essa

técnica foi escolhida por ser menos suscetível a ruídos (JOHNSON, 1967). No problema em questão, esses ruídos são estruturas que se diferenciam de toda população e devido à forma de similaridade entre dois grupos ser baseada nas estruturas mais distintas entre eles, esses ruídos acabam sendo isolados.

Como o objetivo do trabalho não é o desenvolvimento de um método de agrupamento, foi utilizado um pacote de rotinas SciPy[1] que já contém uma implementação otimizada para o agrupamento hierárquico aglomerativo (MÜLLNER, 2013). Essa rotina é de grande utilidade quando o objetivo é agrupar estruturas similares, usando como entrada apenas um limiar de similaridade (MILLMAN; AIVAZIS, 2011; JAEGER et al., 2014; COCHEZ; MOU, 2015). A matriz de distância, nesse caso, é definida como uma matriz de RMSD entre todas as estruturas da população. Para definir um mínimo de similaridade (limiar) entre duas estruturas com o mesmo número de resíduos, foi utilizado a Equação 6.2, apresentando por Maiorov e Crippen (1994). Essa equação consegue, em média, representar um bom limiar de proximidade entre estruturas globulares quando o RMSD é utilizado como critério de similaridade (REVA; FINKELSTEIN; SKOLNICK, 1998).

$$\mathbf{limiar} = N^{1/3}, \tag{6.2}$$

onde *N* é o número de resíduos de aminoácidos presentes na proteína.

O método também permite definir um número mínimo ou máximo de grupos, entretanto, para isso é necessário uma variação no limiar. Deste modo, se com o limiar $N^{1/3}$ esse critério não for satisfeito, uma constante é adicionada nesse limiar de 0.1 até atingir um número mínimo ou máximo de grupos pré-definido. Todos os grupos que contiverem somente 1 indivíduo são adicionados a um novo grupo chamado “restos”, podendo assim, ser considerado como um grupo de diversidade do método.

Para definir qual é a estrutura central de cada grupo, de modo a facilitar a inserção de novos indivíduos na população, um somatório do RMSD entre todas as estruturas internas do grupo é realizado. Assim, a estrutura que comparada contra todas as outras do seu grupo tiver o menor valor global (menor diferença entre todas) é considerada como indivíduo central do grupo. O Algoritmo 4 apresenta o pseudocódigo utilizado para desenvolver o agrupamento apresentado nesta seção.

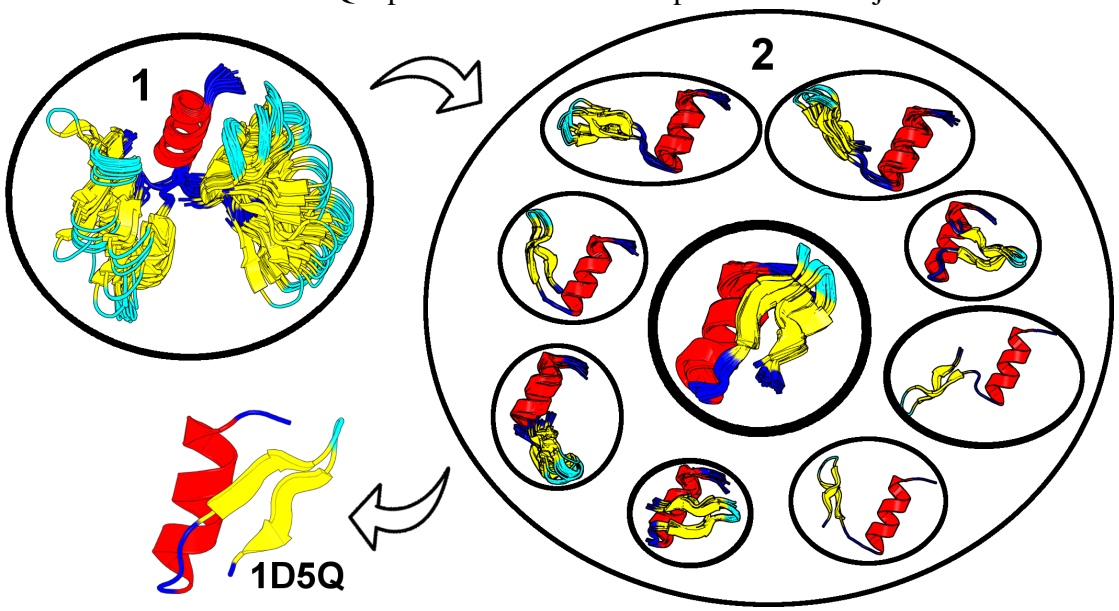
[1]SciPy (JONES et al., 2001) é um conjunto de ferramentas para computação científica desenvolvida na linguagem de programação Python (*SCIentific PYthon*)

Algoritmo 4: Pseudocódigo do agrupamento.

```
Computa a matriz de distância da população de indivíduos utilizando o RMSD;
limiar ← N^(1/3);
repita
    Realiza o agrupamento hierárquico completo do SciPy;
    limiar ← (N^(1/3)) + 0.1;
até condição do número de grupos mínimo ou máximo ser satisfeita
Agrupa indivíduos solitários em um grupo de "restos";
para cada grupo faça
    Encontrar a estrutura de menor diferença entre todas para definir como central;
fim
```

A Figura 6.3 apresenta o objetivo desta abordagem de agrupamento, onde a primeira parte (1) da figura mostra um conjunto de estruturas geradas durante a evolução de um algoritmo competindo pelo mesmo espaço. Enquanto a segunda parte (2) mostra este mesmo conjunto de estruturas porém, desta vez agrupadas por sua similaridade estrutural (valor de RMSD). É possível assim, analisar que são várias conformações que ocorrem durante o processo de predição da estrutura 3D de proteínas. Esta abordagem de agrupamento permite que vários grupos de possíveis soluções evoluam separadamente. Neste caso, a estrutura central da segunda parte (2) da Figura 6.3, representa o nicho mais populoso e também com resultados mais próximos da estrutura experimental desejada (1D5Q). Entretanto, cada subgrupo pode evoluir para um possível resultado melhor e mais próximo da estrutura desejada que esse agrupamento central.

Figura 6.3: Representação gráfica do agrupamento de estruturas por sua similaridade. Onde 1 representa as estruturas competindo por um mesmo espaço, 2 representa as estruturas agrupadas pelo seu enovelamento e 1D5Q representa a estrutura experimental desejada



Fonte: do autor (2016).

6.2.3 Função de Avaliação

A qualidade de uma solução é avaliada pela sua função de aptidão. Essa função avalia o grau de aptidão de um indivíduo para resolver um problema em questão. Neste trabalho utilizamos a função de energia implementada pelo PyRosetta ($E_{PyRosetta}$) (CHAUDHURY; LYSKOV; GRAY, 2010), tendo seus termos e pesos constantemente otimizados (LEAVER-FAY et al., 2013; O'MEARA et al., 2015). Porém, para auxiliar a função na avaliação da formação de proteínas globulares outros dois termos foram adicionados: *Área da Superfície Acessível ao Solvente* (E_{SASA}) e *Reforço da Formação da Estrutura Secundária* (E_{SS}). Assim, o objetivo global do algoritmo é tentar minimizar o somatório destas três partes: $E_{PyRosetta}+E_{SASA}+E_{SS}$. Cada uma destas partes são detalhadas nas próximas subseções.

6.2.3.1 Função de Energia do PyRosetta

A função de energia do PyRosetta utilizada como padrão pela plataforma é a mesma implementada pelo Rosetta utilizando a adaptação dos termos *talaris2013* (ROHL et al., 2004; CHAUDHURY; LYSKOV; GRAY, 2010; LEAVER-FAY et al., 2013; O'MEARA et al., 2015). Essa função de energia, que considera todos os átomos (*all-atom*), é composta por 15 termos e mais um conjunto de pesos específicos para cada um dos 20 aminoácidos (*ref* e *METHOD_WEIGHTS*). Segundo Combs et al. (2013), a maioria dos termos da função de energia do Rosetta foram gerados a partir de potenciais baseados em conhecimento. Para todas as interações de *van der Waals* os potencias de Lennard-Jones 6-12 de atração (*fa_atr*) e repulsão (*fa_rep* e *fa_intra_rep*) são computados. O potencial de solvatação de Lazaridis-Karplus (*fa_sol*) representa um método de solvatação implícita para redução do custo computacional. Interações interatômicas são avaliadas através do potencial eletrostático de Coulomb (*fa_elec*). Também são computados os potencias de energia das ligações de hidrogênio para interação entre cadeias principais próximas (*hbond_sr_bb*), cadeias principais distantes (*hbond_lr_bb*), cadeias laterais (*hbond_sc*) e entre as cadeias laterais e cadeias principais (*hbond_bb_sc*). A função também considera a energia dos ângulos de torções *phi* e *psi* (*p_aa_pp* e *rama*), para o ângulo omega (*omega*) e para os ângulos da cadeia lateral (*fa_dun*). O método possui também valor de energia específico sobre os aminoácidos próximos ao anel da prolina (*pro_close*), assim como o potencial sobre a ligação dissulfeto (*dslf_fa13*) que estabiliza a proteína. A Tabela 6.1 apresenta os termos da função do PyRosetta (*talaris2013*).

Tabela 6.1: Termos de energia utilizados pela função do PyRosetta (talaris2013). Cada um desses termos tem seu objetivo específico e um peso atribuído pelo Rosetta

Termo	**Descrição**
fa_atr *fa_rep* *fa_intra_rep*	Forças de Lennard-Jones (NERIA; FISCHER; KARPLUS, 1996) (KUHLMAN; BAKER, 2000)
fa_sol	Energia de solvatação de Lazaridis-Karplus (LAZARIDIS; KARPLUS, 2000)
fa_elec	Potencial eletrostático de Coulomb (LEAVER-FAY et al., 2013; O'MEARA et al., 2015)
hbond_sr_bb *hbond_lr_bb* *hbond_bb_sc* *hbond_sc*	Energia da ligação de hidrogênio (KORTEMME; MOROZOV; BAKER, 2003) (GORDON; MARSHALL; MAYOT, 1999)
dslf_fa13	Energia potencial da ponte de dissulfeto (LEAVER-FAY et al., 2013; O'MEARA et al., 2015)
omega *rama* *fa_dun* *p_aa_pp* *pro_close*	Energia referente aos ângulos de torções (RAMACHANDRAN; SASISEKHARAN, 1968) (DUNBRACK; COHEN, 1997) (DUNBRACK JR; KARPLUS, 2003) (WEDEMEYER; BAKER, 2003)
ref *METHOD_WEIGHTS*	Energia referente a cada aminoácido

Fonte: Adaptado de Combs et al. (2013).

6.2.3.2 Área da Superfície Acessível ao Solvente

A outra parte da função aptidão deste trabalho consiste em adicionar o modelo empírico de cálculo da Área da Superfície Acessível ao Solvente (do inglês, Solvent Accessible Surface Area (SASA)). Nesse modelo, a energia livre da solvatação é resultada da contribuição atômica de todos os resíduos que contém áreas expostas ao solvente. Essa área reflete principalmente na hidrofobicidade ou hidrofilicidade do átomo em questão (CONNOLLY, 1983; ROUX; SIMONSON, 1999). Essa abordagem permite verificar se uma proteína globular está ou não enovelada auxiliando assim no empacotamento desta proteína (ROUX; SIMONSON, 1999). Para realizar este cálculo foi utilizado uma rotina já presente na plataforma PyRosetta, que utiliza um raio de 1.5 para testar superfície da molécula.

6.2.3.3 Reforço da Formação da Estrutura Secundária

A terceira parte adicionada para formar a função de aptidão é tentar auxiliar o método a produzir a mesma estrutura secundária que a informada como dado de entrada

do método. Este reforço parte do pressuposto de enovelamento chamado de *Framework model*, onde primeiro estruturas secundárias são formadas, para só então a estrutura enovelar-se em seu estado nativo (DAGGETT; FERSHT, 2003). Essa abordagem utiliza uma rotina do PyRosetta que implementa uma versão simplificada da abordagem de atribuição do DSSP (KABSCH; SANDER, 1983b). Essa abordagem identifica apenas três conformações de estruturas secundárias: Hélice, Folha e Volta. Assim, quando o indivíduo gerado pelo método realizar corretamente a predição da estrutura secundária de cada resíduo um reforço positivo é computado ao seu valor de aptidão. Entretanto, quando a estrutura secundária for diferente da esperada um reforço negativo é adicionada ao valor de aptidão.

Assim, o método tenta encontrar uma estrutura final com menor energia pelo PyRosetta ($E_{PyRosetta}$), menor valor de área da superfície acessível ao solvente que ajuda a encontrar estruturas globulares (E_{SASA}) e tentar garantir a mesma conformação da estrutura secundária informada como entrada do método (E_{SS}). A Equação 6.3 representa a Função de Avaliação utilizada para guiar o método desenvolvido, chamada apenas de Energia.

$$\textbf{Energia} = E_{PyRosetta} + E_{SASA} + E_{SS} \tag{6.3}$$

6.2.4 Seleção e Recombinação dos Indivíduos

Após todos os indivíduos da população terem seus valores de aptidão atribuídos o método pode aplicar os operadores de seleção e recombinação para, seguindo a lógica dos conceitos apresentados no Capítulo 5, produzirem melhores indivíduos. Para efetuar a seleção, o método utiliza como base a abordagem de Seleção por Torneio, onde apenas uma parte da população é selecionada, nesse caso um grupo/nicho é selecionado aleatoriamente, e então o indivíduo com o melhor valor de aptidão é escolhido. Neste problema o indivíduo com o menor valor de Energia é considerado o melhor. Com isso, é possível garantir que ao menos um indivíduo com bom valor de aptidão, mesmo localmente, seja selecionado. Para selecionar o segundo indivíduo e permitir assim fazer a recombinação, o método utiliza uma porcentagem informada como entrada do algoritmo para definir a probabilidade de seleção ocorrer do mesmo grupo que o primeiro selecionado ou se o segundo indivíduo é selecionado de um grupo/nicho diferente. Se o método entrar no primeiro caso, um indivíduo é escolhido aleatoriamente do mesmo grupo. Se o método cair no segundo caso, ele pode selecionar com a mesma probabilidade o indivíduo mais apto de um grupo diferente, ou um indivíduo aleatório do grupo de "resto" para incentivar a diversidade do método.

Com esses dois indivíduos selecionados é possível aplicar o operador de recombinação. Para este trabalho foi utilizado o método de Recombinação Uniforme, dando uma probabilidade maior para o método utilizar os dados do primeiro indivíduo selecionado e o restante vem do segundo indivíduo selecionado. Esta probabilidade foi definida entre 50% a 70% por serem valores constantemente utilizados (GONÇALVES; RESENDE, 2011; BORGUESAN et al., 2015a). Assim o método garante que o indivíduo gerado contenha pelo menos metade de seus dados provenientes de um bom indivíduo (mais apto do seu grupo), porém com ao menos 30% dos dados de um indivíduo diferente. Através desta abordagem, é possível garantir que indivíduos mais aptos possam passar seus dados para a próxima população, entretanto sempre mantendo um mínimo de diversidade nestas recombinações.

6.2.5 Mutação dos Indivíduos

Conforme apresentado no Capítulo 5, o objetivo principal da mutação é a introdução e manutenção da diversidade genética da população, para garantir que todo o espaço conformacional seja analisado. Para isso o método desenvolvido aplica duas formas de mutação.

- **Mutação 1**: é aplicada sobre uma pequena porcentagem dos indivíduos gerados pela recombinação. Essas mutações ocorrem nas regiões de mesma estrutura secundária utilizando conhecimento das APLCentral9, APLCentral7 e APLCentral5 apresentadas no Capítulo 4. O método sempre da maior chance de escolha para a APLCentral de tamanho maior. Desta forma, se esta mutação for aplicada, o método substitui uma parte da região de mesma estrutura secundária de um indivíduo, gerado após a recombinação, por valores obtidos através da APLCentral de tamanho maior (ordem de prioridade 9, 7 e 5). Essa mutação permite que o método utilize o conhecimento da metodologia de APLCentral para incrementar a diversidade estrutural da população;
- **Mutação 2**: também é aplicada em uma pequena parcela dos indivíduos resultantes da recombinação, porém nesta etapa somente nas regiões de troca de estrutura secundária. Assim, quando a mutação é aplicada, resíduos das regiões de voltas são atualizados utilizando conhecimento das APL3, APL2 e APL1, também apresentadas no Capítulo 4. Mantendo o mesmo critério de maior chance de escolha para a APL de maior tamanho (ordem de prioridade 3, 2 e 1). Esse critério garante que mesmo a APL1 tenha chance de alterar algum indivíduo de modo a percorrer todo o espaço conformacional, mas sempre mantendo a abordagem baseada em conhecimento.

6.2.6 Evolução da População

Para adicionar o indivíduo gerado na população atual é utilizado o mesmo conceito do Torneio por Seleção Restrita apresentado no Capítulo 5. Assim, primeiro é verificado com qual estrutura central dos grupos, definido pelo Algoritmo 4, esse novo indivíduo possui maior similaridade (menor RMSD). Quando essa estrutura é selecionada, o método então verifica se dentro do mesmo grupo desta estrutura central existe um resultado que possua uma aptidão pior para resolver o problema que o novo indivíduo gerado. No método desenvolvido, tenta-se encontrar um indivíduo que possua energia maior que a do novo indivíduo gerado. Se o método não encontrar nenhuma solução para substituir, então o indivíduo gerado é descartado e o processo de criação de uma nova solução é reiniciado. Com isso, garantimos que somente indivíduos mais aptos serão adicionados nos grupos, colaborando assim para evolução da população. Esta abordagem é apresentada no Algoritmo 3 do método desenvolvido nas linhas 11 até a 15.

6.2.7 Controle de Diversidade e Reinício da População

A perda de diversidade da população ou no caso de problemas multimodais, de uma subpopulação (nicho), ocorre quando a população converge precocemente para um mínimo que não é o global da função (MITCHELL, 1999). Se a população perder a sua diversidade, pode ocorrer do valor de aptidão não melhorar ou necessitar de um número de evolução da população muito grande para apresentar melhora. No caso da predição, a população pode convergir antecipadamente para uma estrutura que não é o mínimo no panorama energético do problema, perdendo assim a diversidade e impossibilitando o método de escapar deste mínimo local. Com isso, o Controle de Diversidade é realizado para contornar esse problema, onde é verificado se já existe um indivíduo na população com os mesmos valores que o novo indivíduo gerado após a mutação. Se não existir, o indivíduo pode ser adicionado, caso contrário um novo indivíduo tem que ser gerado, evitando assim que indivíduos idênticos sejam adicionadas na população. Esta abordagem de Controle de Diversidade é apresentada no Algoritmo 3 do método desenvolvido nas linhas 5 até a 9.

O reinício da população utiliza a mesma ideia de diversidade da população. Esta técnica consiste em evitar que um grupo de indivíduos utilize todo os recursos, não permitindo a diversidade e a inclusão de novos indivíduos. Assim, no método proposto, é utilizada uma abordagem de reinício após um número de tentativas em adicionar um novo indivíduo na

população tenha falhado. Quando este número de tentativas é atingido, então o método gera novas estruturas da mesma forma que a população inicial (Seção 6.2.1) para adicionar a próxima população, substituindo sempre os piores indivíduos de cada grupo. Desta forma é garantido que os melhores indivíduos de cada grupo não sejam removidos, mas sempre permitindo a inserção de novos indivíduos, contribuindo para diversidade estrutural da população. Esta abordagem de Reinício da População é apresentada no Algoritmo 3 do método desenvolvido nas linhas 16 até a 18.

6.2.8 Condições de Parada

Conforme comentado no capítulo anterior, definir quando um algoritmo deve terminar pode ser uma tarefa bastante complicada que dependente do problema investigado. Neste caso, o método desenvolvido trabalha com dois critérios de parada: o método termina se a população estabilizar; e termina se atingir um determinado tempo de execução. Para a população estabilizar é verificado se o agrupamento da população satisfez os critérios de parada apresentados no Algoritmo 4 utilizando o limiar igual a $N^{1/3}$. Esse critério garante que o método atingiu um número mínimo ou máximo de grupos esperados (Seção 6.2.2) utilizando um limiar de similaridade baixo ($N^{1/3}$), significando na convergência da população. Para evitar que o algoritmo nunca termine (devido à complexidade da estrutura a ser predita a população pode não estabilizar) um critério de tempo máximo de execução é definido no início do método.

6.3 Resumo do Capítulo

Neste capítulo foi apresentado um novo método para o problema da predição da estrutura 3D de polipeptídeos. O método proposto utiliza conhecimento extraído da base experimental (Capítulo 4) para conseguir reduzir o espaço de busca conformacional do problema. O método, chamado de GARTS, combina uma técnica de Algoritmo Genético em conjunto com uma variação da abordagem de Torneio de Seleção Restrita para conseguir tratar o problema da predição da estrutura 3D de proteínas de forma multimodal. Assim os indivíduos da população do Algoritmo Genético são agrupados utilizando uma Clusterização Hierárquica Completa, a fim de tentar evoluir cada grupo sem que o método estabilize em um mínimo local. Essa abordagem de clusterização da população do Algoritmo Genético conforme a geração evolui é a grande contribuição desta dissertação, pois trata o problema de predição de forma multimodal

permitindo assim formar e manter múltiplas subpopulações com o objetivo de localizar várias possíveis soluções ao final da execução.

No próximo capítulo são apresentados vários experimentos realizados para validação do algoritmo proposto. Esses experimentos serão divididos em 3 Grupos. O primeiro visa descobrir o melhor conjunto de parâmetros do método. O segundo para verificar se a técnica GARTS realmente produz melhores resultados que uma técnica padrão normalmente utilizada. E por último é realizado uma bateria de testes contra o servidor de predição de estrutura 3D de proteína I-TASSER, para comparar a qualidade das estrutura geradas pelo modelo desenvolvido nesta dissertação.

7 RESULTADOS E DISCUSSÃO

7.1 Introdução

Neste capítulo, a metodologia desenvolvida e apresentada no capítulo anterior, é validada para o problema da predição da estrutura 3D de proteínas. Nas próximas seções os critérios utilizados para avaliação dos resultados produzidos pelos algoritmos serão detalhados. Também são detalhados os diferentes conjuntos de parâmetros que foram aplicados ao método GARTS desenvolvido a fim de obter os melhores resultados. Para validar o método GARTS, foi realizada a comparação com um Algoritmo Genético utilizando os mesmos critérios de inicialização, recombinação e mutação do método proposto, apresentado no capítulo anterior. A última validação realizada, foi comparar os resultados obtidos pelo GARTS com os resultados gerados pelo servidor de predição I-TASSER apresentado no Capítulo 3 como um dos melhores para predição da estrutura 3D de proteínas.

7.2 Critérios de Validação dos Resultados

Como o método desenvolvido trabalha com população de indivíduos, seu resultado final é um conjunto de possíveis soluções. Assim, para validar esse trabalho, foi analisada toda a população final para encontrar o melhor indivíduo. Para selecionar a melhor solução, quatro critérios para validação dos resultados foram considerados: Valor de Aptidão, RMSD, GDT_TS e Q-index. Esses critérios permitem a validação da estrutura gerada em termos estereoquímicos (Valor de Aptidão) e em termos estruturais (RMSD, GDT_TS e Q-index). Cada um desses critérios foram detalhados nas próximas seções.

7.2.1 Valor de Aptidão

O Valor de Aptidão foi apresentado na Seção 6.2.3, onde o melhor indivíduo é a solução com o menor valor de *Energia* seguindo a Equação 7.1 onde é feita a soma entre os pesos: Função de Energia do PyRosetta ($E_{PyRosetta}$); Área da Superfície Acessível ao Solvente (E_{SASA}); Reforço da Formação da Estrutura Secundária (E_{SS}).

$$\mathbf{Energia} = E_{PyRosetta} + E_{SASA} + E_{SS} \tag{7.1}$$

7.2.2 RMSD

O RMSD também foi apresentado no capítulo anterior, entretanto, para validação da estrutura predita ela é comparada contra a estrutura experimental depositada no PDB. Outra alteração no cálculo do RMSD é retirar os dois resíduos terminais da estrutura, devido sua flexibilidade. Assim, a Equação 6.1 apresentada no capítulo anterior foi modificada para a Equação 7.2. Resultados de RMSD igual a 0 correspondem a indivíduos idênticos em termos estruturais. O cálculo de RMSD é executado por uma rotina implementada pelo PyRosetta (CHAUDHURY; LYSKOV; GRAY, 2010).

$$\mathbf{RMSD}(p, e) = \sqrt{\left(\sum_{i=2}^{n-1} \|dp_{C_{\alpha i}} - de_{C_{\alpha i}}\|^2\right)/n}, \tag{7.2}$$

onde $dp_{C_{\alpha i}}$ e $de_{C_{\alpha i}}$ são vetores que representam o posicionamento (i) do átomo (C_α) em duas estruturas p (predita) e e (experimental) de tamanho n.

7.2.3 GDT_TS

O terceiro critério de validação é o GDT_TS (do inglês, *Global Distance Total Score Test*) (ZHANG; SKOLNICK, 2004a), métrica constantemente utilizada para representar a qualidade estrutural das proteínas preditas na competição do CASP, apresentada no Capítulo 3 desta dissertação (KRYSHTAFOVYCH; FIDELIS; MOULT, 2014b). O valor GDT_TS de uma estrutura é uma porcentagem que representa o quanto a predição convergiu para a estrutura nativa (ZHANG; SKOLNICK, 2004a). O cálculo do GDT_TS é dado pela fórmula descrita na Equação 7.3. Assim, um valor de GDT_TS igual a 100% significa que cada resíduo da estrutura predita é alinhado com a estrutura experimental com distância menor que 1Å. O cálculo de GDT_TS é executado por uma rotina disponibilizada[1] pelo Zhang Lab (ZHANG; SKOLNICK, 2004a).

$$GDT\_TS = (GDT_{P1} + GDT_{P2} + GDT_{P4} + GDT_{P8})/4 \tag{7.3}$$

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

[1] <http://zhanglab.ccmb.med.umich.edu/TM-score/>

7.2.4 Q-index

Outra métrica que é utilizada nesse trabalho para validação da correta formação da estrutura secundária é o Q-index (Eq. 7.4). Essa métrica consiste na porcentagem da correta classificação de cada resíduo da estrutura predita, comparando com a estrutura secundária esperada pelo método (ZHANG; ZHANG, 2001). Nesse esquema, foi utilizado o STRIDE para atribuição da estrutura secundária, porém essas estruturas foram agrupadas em 4 grupos seguindo o esquema a seguir: `H` e `G` ficou `H`; `E` e `B` ficou `E`; `T` continuou como `T`; e todos os outros estados foram representados como `C`. Assim, a Equação 7.5 representa a métrica Q_4 utilizada para validação da formação correta da estrutura secundária do método.

$$Q_i(\%) = \frac{\#\ de\ AA\ Corretamente\ Preditos_i}{\#\ de\ AA\ no\ grupo_i} \times 100 \tag{7.4}$$

$$Q4(\%) = \frac{\sum \#\ de\ AA\ Corretamente\ Preditos_i}{\sum \#\ de\ AA\ no\ grupo_i} \times 100 \tag{7.5}$$

onde $i\ \epsilon\ \{H, E, T, C\}$, enquanto $\#deAA$ representam o número de aminoácidos.

7.3 Base de Teste

Para validação dos resultados deste trabalho, duas base de testes foram utilizadas. A primeira base foi criada com um número menor de estruturas permitindo fazer a validação dos parâmetros do método de forma mais eficiente. A segunda base possui o triplo do tamanho, permitindo que a comparação com outros métodos seja estatisticamente mais acurada. As bases de testes foram escolhidas baseadas em diferentes padrões conformacionais para tentar abranger todas as topologias básicas de proteínas globulares. Devido à complexidade do problema, aumentar o tamanho das estruturas selecionadas implicaria em também aumentar o tempo mínimo necessário para que o método encontrasse boas soluções. Assim, apenas estruturas de proteínas pequenas, de 14 até 76 resíduos, foram selecionadas.

Vale também ressaltar que todas as estruturas de proteínas testadas passaram por um processo de purificação da base de dados. Esse processo, garante que o método desenvolvido não utilize o conhecimento prévio de nenhuma estrutura com similaridade de sequência entre

a proteína alvo. Com isso, foi possível testar o método GARTS como se estivesse tentando fazer a predição de uma estrutura com um novo enovelamento que ainda não está presente na base de dados do PDB. Esse processo foi realizado através do servidor SAS[2] (do inglês, *Sequence Annotated by Structure*) (MILBURN; LASKOWSKI; THORNTON, 1998). Esse servidor permite identificar estruturas até mesmo com baixa similaridade de sequência usando informações estruturais de uma proteína alvo. Assim, todas as estruturas que o método SAS considerou similar foram retiradas das bases de dados utilizadas pelo método GARTS (APL, NPAS e FM-B Lib, todas descritas no Capítulo 4). Todos os testes realizados nesta dissertação utilizaram um ambiente computacional Linux X86_64 dedicado com processador Intel Xeon E5-310 QuadCore de 1.6GHz com 8MB de memória cache e 16GB de memória RAM.

7.3.1 Base de Teste I

A primeira Base de Teste (I) foi utilizada para validação dos parâmetros iniciais utilizados no método GARTS desenvolvido. A Tabela 7.1 apresenta essas estruturas, onde a primeira coluna representa a topologia do padrão conformacional (apresentados na Seção 2.2.3), a segunda coluna é o código da estrutura no PDB, a terceira coluna é a referência da proteína, a quarta coluna é o número de resíduos da proteína e a última coluna é o conteúdo da estrutura secundária da proteína que o método pretende obter ao final do processo de predição.

Tabela 7.1: Proteínas utilizadas para testar a abordagem desenvolvida com o número de resíduos variando entre `14-63` aminoácidos

Topologia	Código PDB	Referência	Tamanho	Conteúdo da SS
α	`1L2Y`	(NEIDIGH; FESINMEYER; ANDERSEN, 2002)	20	2 hélices
α	`1ZDD`	(STAROVASNIK; BRAISTED; WELLS, 1997)	34	2 hélices
β	`1K43`	(PASTOR et al., 2002)	14	1 folha
β	`1DFN`	(HILL et al., 1991)	30	2 folhas
α+β	`1D5Q`	(VITA et al., 1999)	27	1 hélice 1 folha
α+β	`1ACW`	(BLANC et al., 1996)	29	1 hélice 1 folha
α+β	`2P5K`	(GARNETT et al., 2007)	63	3 hélices 1 folha
α/β	`1AB1`	(YAMANO; HEO; TEETER, 1997)	46	2 hélices 1 folha

Fonte: Adaptado de Berman et al. (2000).

7.3.2 Base de Teste II

A segunda Base de Teste (II) foi utilizada para fins de validação do método desenvolvido, comparando-o com um AG padrão e o servidor I-TASSER. A Tabela 7.2 apresenta as estruturas

[2]<http://www.ebi.ac.uk/thornton-srv/databases/sas/>

selecionadas, onde a primeira coluna representa a topologia do padrão conformacional (apresentados na Seção 2.2.3), a segunda coluna é o código da estrutura no PDB, a terceira coluna é a referência da proteína, a quarta coluna é o número de resíduos da proteína e a última coluna é o conteúdo da estrutura secundária da proteína.

Tabela 7.2: Proteínas utilizadas para testar a abordagem desenvolvida com tamanho variando entre 14−76 resíduos. Estruturas 3P7K e 2PMR não tiveram seus artigos publicados

Topologia	Código PDB	Referência	Tamanho	Conteúdo da SS
α	2MTW	(CIFUENTES et al., 2005)	20	1 helix
α	3P7K	n/a	45	1 hélice
α	1L2Y	(NEIDIGH; FESINMEYER; ANDERSEN, 2002)	20	2 hélice
α	1WQC	(CHAGOT et al., 2005)	26	2 hélices
α	1ZDD	(STAROVASNIK; BRAISTED; WELLS, 1997)	34	2 hélices
α	2P81	(RELIGA et al., 2007)	44	2 hélices
α	3V1A	(DER et al., 2012)	48	2 hélices
α	1ROP	(BANNER; KOKKINIDIS; TSERNOGLOU, 1987)	56	2 hélices
α	2MR9	(NOWICKA et al., 2015)	44	3 hélices
α	2P6J	(SHAH et al., 2007)	52	3 hélices
α	1ENH	(CLARKE et al., 1994)	54	3 hélices
α	2KDL	(ALEXANDER et al., 2009)	56	3 hélices
α	1AIL	(LIU et al., 1997)	70	3 hélices
α	2PMR	n/a	76	3 hélices
α	2JUC	(BONET; RAMIREZ-ESPAIN; MACIAS, 2008)	59	4 hélices
α	1UTG	(MORIZE et al., 1987)	70	5 hélices
β	1K43	(PASTOR et al., 2002)	14	1 folha
β	1DFN	(HILL et al., 1991)	30	2 folhas
α+β	1D5Q	(VITA et al., 1999)	27	1 hélice 1 folha
α+β	1ACW	(BLANC et al., 1996)	29	1 hélice 1 folha
α+β	1Q2K	(CAI et al., 2004)	31	1 hélice 1 folha
α+β	2P5K	(GARNETT et al., 2007)	63	3 hélices 1 folha
α/β	1AB1	(YAMANO; HEO; TEETER, 1997)	46	2 hélices 1 folha
α/β	1CRN	(TEETER, 1984)	46	2 hélices 1 folha

Fonte: Adaptado de Berman et al. (2000).

7.4 Escolha dos Parâmetros

Como apresentado no capítulo anterior, o método GARTS necessita de alguns parâmetros iniciais para poder executar a predição da estrutura 3D de proteínas. Os conjunto de parâmetros do método podem ser divididos em duas partes. A primeira parte são os parâmetros do AG propriamente dito que são: tamanho da população (1) e porcentagem de mutação sobre os novos indivíduos (2). O outro conjunto de parâmetros pertence ao método desenvolvido que são: número mínimo de grupos (3), probabilidade da seleção do segundo indivíduo ser do mesmo grupo (4), número de tentativas para executar o reinício da população (5) e quantos indivíduos serão adicionados na população após o reinício (6). Os valores testados para cada

um dos parâmetros são apresentados na Tabela 7.3 a seguir. Esses valores foram definidos arbitrariamente baseados no tamanho da população. Para evitar uma explosão de combinações, os valores dos parâmetros foram agrupados em três conjuntos: P1, P2 e P3.

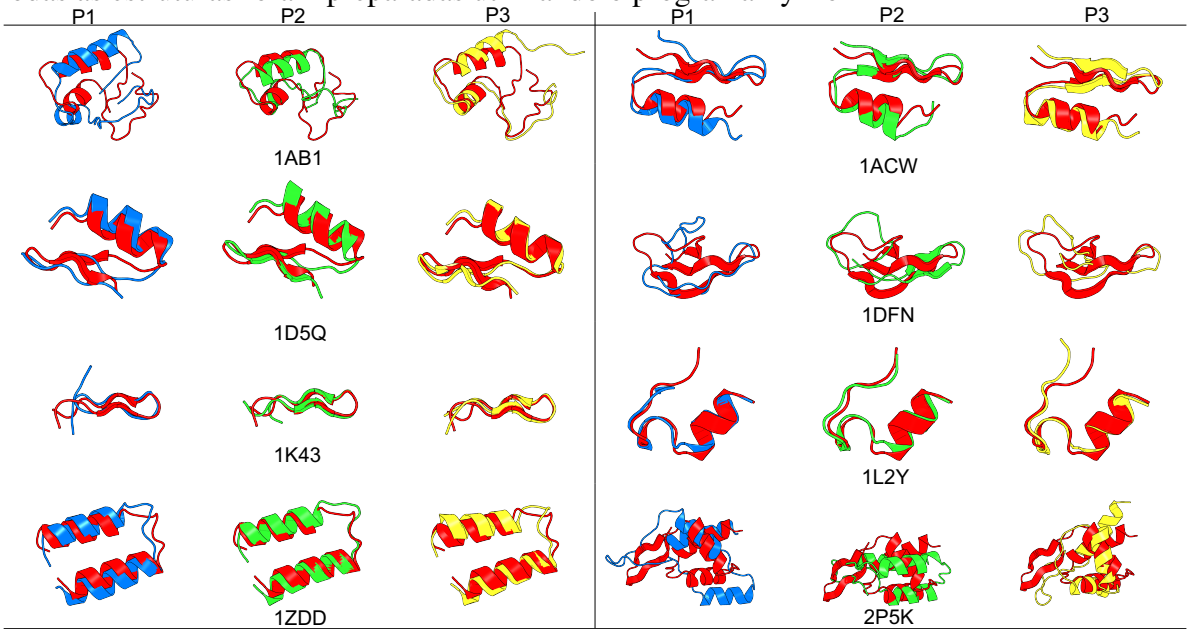
Tabela 7.3: Valores dos parâmetros testados no método para cada um dos conjuntos P1, P2 e P3.

Parâmetro	P1	P2	P3
1	100	200	400
2	5%	10%	20%
3	5	8	10
4	50%	60%	70%
5	500	1000	1500
6	10	15	20

Fonte: do autor (2016).

Para validação desses parâmetros o método GARTS foi executado por 6 horas para cada um dos três conjuntos de valores (P1, P2 e P3). Para uma validação estatística, cada um desses testes foram executados por 8 vezes utilizando as 8 estruturas da Base de Teste I e a máquina descrita na Seção 7.3. A Figura 7.1 apresenta os resultados desta execução. As estruturas em vermelho, representam a estrutura experimental, enquanto as estruturas em azul, verde e amarelo representam o resultado de menor RMSD para os conjuntos de parâmetros P1, P2 e P3, respectivamente. Através desta figura é possível analisar que o método gerou estruturas próximas das experimentais, independente do conjunto de parâmetros utilizados.

Figura 7.1: Comparativo de RMSD entre os 3 Conjuntos de parâmetros. Estrutura vermelha representa a estrutura experimental. As estruturas em azul, verde e amarelo representam o resultado de menor RMSD para os conjuntos de parâmetros P1, P2 e P3, respectivamente. Todas as estruturas foram preparadas utilizando o programa PyMol



Fonte: do autor (2016).

Entretanto, essa similaridade entre os parâmetros não pode ser completamente observada quando os critérios de avaliação são analisados (aptidão, RMSD e GDT_TS). A Tabela 7.4 apresenta as estruturas com o melhor valor de aptidão entre os indivíduos da última população do método GARTS para os três conjuntos de valores dos parâmetros (P1, P2 e P3). Analisando essa tabela é possível observar que tanto a menor energia encontrada como a energia média entre as 8 execuções tiveram os menores valores observados entre os conjuntos de parâmetros P1 e P2. Porém, também é possível verificar que nos casos 1ACW, 1D5Q, 1DFN e 1L2Y o menor valor de energia não representa o menor valor de RMSD. O mesmo ocorre se analisarmos o menor valor de energia com o maior valor de GDT_TS. Isto ocorre principalmente, devido ao fato, já comentado durante esta dissertação, de que as regras que controlam o processo de enovelamento de proteínas não serem totalmente conhecidas, o que não permite uma função de aptidão que represente 100% esse processo. Assim, o menor valor de energia da última população, pode

Tabela 7.4: Comparativo do menor valor de aptidão entre os três conjuntos de parâmetros (P1, P2 e P3). Os valores de RMSD, GDT_TS e Avaliações de Energia são referentes a estrutura de Menor valor de Energia. A Média de Energia é referente aos valores de menor energia das 8 execuções. Os valores entre parênteses representam o desvio padrão

Código PDB	Menor Energia	RMSD Å	GDT_TS %	Avaliações de Energia	Energia Média
1AB1-P1	-5944,80	5,24	52,72	332 859	-5145,52 ±(512,37)
1AB1-P2	**-8474,40**	4,08	55,43	225 499	**-5184,41** ±(1337,26)
1AB1-P3	-3934,04	11,56	41,85	92 662	-3679,68 ±(122,48)
1ACW-P1	**-12404,83**	8,03	48,28	259 879	-11725,82 ±(480,16)
1ACW-P2	-12363,50	7,27	54,31	115 067	**-11775,26** ±(505,60)
1ACW-P3	-11734,92	6,77	50,86	66 799	-10698,28 ±(772,98)
1D5Q-P1	**-16763,50**	3,17	69,44	179 668	-16627,63 ±(121,87)
1D5Q-P2	-16732,03	3,43	70,37	114 911	**-16654,35** ±(67,32)
1D5Q-P3	-16533,19	2,28	79,63	41 357	-16421,79 ±(75,89)
1DFN-P1	**-7247,94**	6,78	37,50	341 832	-3808,15 ±(1929,09)
1DFN-P2	-6569,93	7,42	39,17	215 486	**-4518,12** ±(1285,50)
1DFN-P3	-3188,02	6,58	38,33	95 051	-2621,55 ±(592,41)
1K43-P1	-4477,74	1,13	78,57	308 462	-4410,03 ±(46,60)
1K43-P2	**-4486,46**	0,54	92,86	436 956	**-4436,29** ±(28,58)
1K43-P3	-4416,47	0,59	85,71	131 216	-4386,43 ±(33,94)
1L2Y-P1	**-2938,87**	2,73	72,50	58 435	**-2822,47** ±(68,45)
1L2Y-P2	-2897,57	2,79	73,75	702 533	-2775,91 ±(84,80)
1L2Y-P3	-2851,89	2,32	78,75	381 568	-2748,86 ±(51,45)
1ZDD-P1	**-22415,00**	3,62	44,12	70 342	**-21765,79** ±(360,95)
1ZDD-P2	-22027,68	4,91	41,91	44 662	-21535,22 ±(216,28)
1ZDD-P3	-21735,11	7,37	45,59	19 928	-21308,09 ±(263,74)
2P5K-P1	**-38189,15**	11,26	31,35	89 382	**-27836,99** ±(8445,45)
2P5K-P2	-33847,04	17,99	30,56	324 661	-27041,32 ±(5942,75)
2P5K-P3	-28153,10	15,92	29,76	134 373	-21143,15 ±(3297,65)

Fonte: do autor (2016).

não ser realmente a estrutura mais próxima da experimental. Também é possível notar essa diferença analisando os valores de energia das estruturas de menor RMSD e maior GDT_TS apresentado nas Tabelas 7.5 e 7.6, respectivamente. Se analisarmos o valor médio, apenas a estrutura 1L2Y obteve um padrão de menor média de energia, menor média de RMSD e maior média de GDT_TS para um mesmo conjunto de parâmetros.

A Tabela 7.5, além de mostrar o menor valor de RMSD entre as 8 execuções para cada um dos conjuntos de parâmetros (P1, P2 e P3) em conjunto com seus valores de energia e GDT_TS, também apresenta o valor médio do menor valor de RMSD das 8 execuções. Essa tabela, mostra que o menor valor de RMSD não apresenta um padrão em relação a qual conjunto de parâmetros é melhor, tendo destaque em todos os conjuntos (P1, P2 e P3). Diferente disso e da Tabela 7.4, onde os conjuntos de parâmetros P1 e P2 foram melhores, a tabela de RMSD apresenta um padrão onde 87.50% das estruturas testadas tiveram um valor de RMSD médio

Tabela 7.5: Comparativo do menor valor de RMSD entre os três conjuntos de parâmetros (P1, P2 e P3). Os valores de Energia, GDT_TS e Avaliações de Energia são referentes a estrutura de Menor valor de RMSD. A Média de RMSD é referente aos valores de menor RMSD das 8 execuções. Os valores entre parênteses representam o desvio padrão

Código PDB	Menor RMSD Å	Energia	GDT_TS %	Avaliações de Energia	RMSD Médio Å
1AB1-P1	4,77	-5439,32	51,63	192 657	6,37 ±(1,28)
1AB1-P2	**4,00**	-5563,21	57,07	241 520	5,57 ±(1,49)
1AB1-P3	4,16	-1719,38	60,87	86 739	**5,41** ±(0,77)
1ACW-P1	**1,52**	-11908,07	75,86	271 341	3,23 ±(1,27)
1ACW-P2	1,80	-11154,37	75,00	89 888	2,58 ±(0,65)
1ACW-P3	1,61	-8047,61	77,59	72 754	**2,36** ±(0,66)
1D5Q-P1	1,60	-16517,02	82,41	182 792	2,16 ±(0,54)
1D5Q-P2	1,60	-16489,03	83,33	114 911	2,27 ±(0,69)
1D5Q-P3	**1,25**	-15395,45	84,26	45 056	**1,91** ±(0,37)
1DFN-P1	4,42	-2731,39	45,00	332 227	5,30 ±(0,87)
1DFN-P2	3,64	-3190,14	49,17	205 626	5,32 ±(0,98)
1DFN-P3	**3,60**	6220,93	46,67	92 749	**4,62** ±(0,58)
1K43-P1	0,32	-4327,79	83,93	473 016	0,63 ±(0,19)
1K43-P2	**0,31**	-4302,31	89,29	422 844	0,58 ±(0,17)
1K43-P3	0,47	-4230,57	92,86	131 216	**0,56** ±(0,10)
1L2Y-P1	**0,99**	-2654,87	86,25	50 647	**1,39** ±(0,40)
1L2Y-P2	1,15	-2713,40	87,50	341 198	1,64 ±(0,36)
1L2Y-P3	1,10	-2705,41	83,75	372 481	1,74 ±(0,47)
1ZDD-P1	1,81	-21526,13	43,38	73 330	3,53 ±(1,20)
1ZDD-P2	**1,31**	-21474,35	42,65	50 515	2,76 ±(1,20)
1ZDD-P3	2,10	-21093,74	42,65	20 141	**2,46** ±(0,47)
2P5K-P1	6,12	-23534,18	43,65	86 790	10,19 ±(2,47)
2P5K-P2	6,44	-25122,01	39,68	347 871	9,29 ±(2,09)
2P5K-P3	**5,03**	-6562,21	41,67	146 820	**6,15** ±(0,63)

Fonte: do autor (2016).

mais baixo quando o conjunto de parâmetros P3 foi utilizado. A única estrutura fora desse padrão foi a 1L2Y, entretanto com uma diferença no RMSD médio entre o melhor e o pior conjunto de parâmetros (P1 e P3) foi de apenas 0.35Å.

A Tabela 7.6, apresenta o maior valor de GDT_TS entre as 8 execuções para cada uma das proteínas da Base de Teste I e em cada um dos conjuntos de parâmetros (P1, P2 e P3). As colunas seguintes representam os valores de energia, RMSD e o número de avaliações de energias realizados pela estrutura de maior GDT_TS. Essa tabela também apresenta a média do maior valor de GDT_TS das 8 execuções. Se analisarmos essa tabela é possível observar que em 75% das estruturas testadas, os melhores resultados foram obtidos utilizando o conjunto de parâmetros P3, tanto em maior valor de GDT_TS como em maior valor médio de GDT_TS entre as 8 execuções. Assim, o conjunto de parâmetros escolhido para executar

Tabela 7.6: Comparativo do maior valor de GDT_TS entre os três conjuntos de parâmetros (P1, P2 e P3). Os valores de Energia, RMSD e Avaliações de Energia são referentes a estrutura de Maior valor de GDT_TS. A Média de GDT_TS é referente aos valores de maior GDT_TS das 8 execuções. Os valores entre parênteses representam o desvio padrão

Código PDB	Maior GDT_TS %	Energia	RMSD Å	Avaliações de Energia	GDT_TS Médio %
1AB1-P1	55,43	-4527,25	7,03	86 379	50,95 ±(3,04)
1AB1-P2	57,07	-5563,21	4,00	241 520	53,40 ±(2,72)
1AB1-P3	**63,59**	-1839,70	4,32	86 739	**56,18** ±(4,85)
1ACW-P1	**77,59**	-11133,23	2,15	271 341	65,73 ±(8,48)
1ACW-P2	**77,59**	-11178,22	2,05	182 078	70,15 ±(3,88)
1ACW-P3	**77,59**	-8047,61	1,61	68 058	**71,66** ±(4,06)
1D5Q-P1	83,33	-16657,99	1,99	192 509	78,47 ±(4,99)
1D5Q-P2	83,33	-16370,98	1,72	114 911	76,85 ±(5,19)
1D5Q-P3	**85,19**	-8727,29	1,69	29 700	**80,09** ±(4,29)
1DFN-P1	47,50	-3303,77	4,52	344 527	44,48 ±(2,31)
1DFN-P2	50,00	-3271,40	4,58	215 438	44,06 ±(4,57)
1DFN-P3	**50,83**	5870,22	6,40	96 094	**47,40** ±(2,76)
1K43-P1	94,64	-4268,71	0,45	302 345	88,84 ±(3,79)
1K43-P2	94,64	-4346,62	0,52	422 844	**91,74** ±(2,69)
1K43-P3	**96,43**	-4207,22	0,50	129 558	90,85 ±(3,08)
1L2Y-P1	**90,00**	-2683,04	1,07	50 647	**85,62** ±(3,95)
1L2Y-P2	87,50	-2713,40	1,15	626 151	83,28 ±(2,67)
1L2Y-P3	86,25	-2526,56	1,59	423 834	82,50 ±(3,13)
1ZDD-P1	47,06	-16627,72	10,59	73 330	44,85 ±(1,24)
1ZDD-P2	47,79	-21132,57	6,22	44 154	45,77 ±(1,56)
1ZDD-P3	**49,26**	-20608,25	9,16	20 114	**48,16** ±(1,04)
2P5K-P1	**44,05**	-23536,16	6,13	317 974	36,06 ±(3,77)
2P5K-P2	40,87	-30175,05	12,95	347 871	36,66 ±(2,92)
2P5K-P3	42,86	-5742,32	6,39	144 098	**41,67** ±(1,00)

Fonte: do autor (2016).

os próximos testes de validação do método GARTS foi o P3, pois, em média, conseguiu uma estrutura mais próxima da experimental em termos de RMSD e GDT_TS na maioria das proteínas da Base de Teste I.

7.5 Validação da Abordagem de Agrupamento do Método GARTS

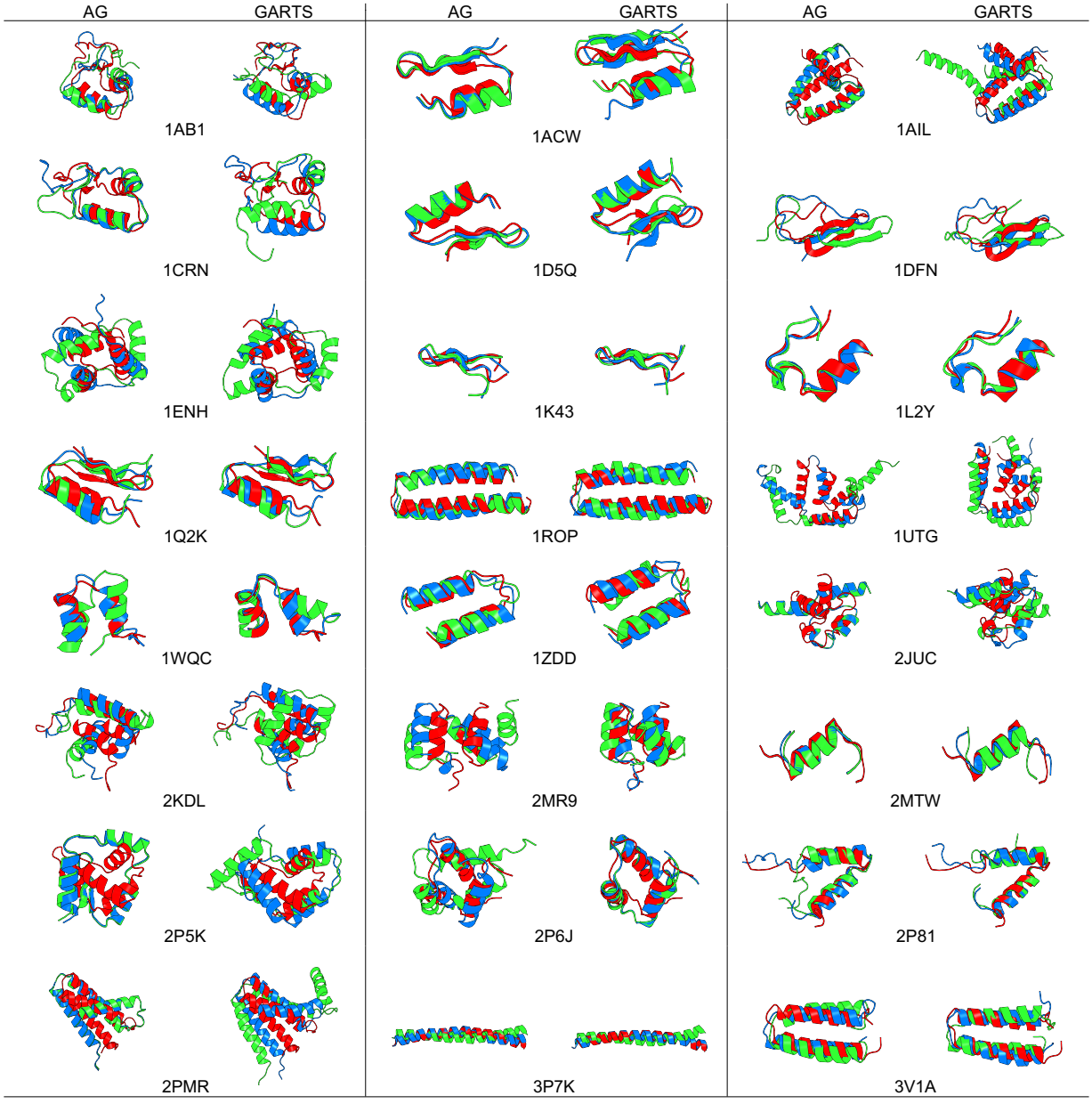
Com o propósito de verificar as melhorias obtidas pelo método GARTS, foi necessário comparar se o diferencial da técnica, que é o agrupamento das estruturas durante a evolução da população, realmente produz melhores resultados em comparação com outro método que não utilize essa abordagem. Para isso, foi desenvolvido um Algoritmo Genético (AG) com os mesmos modelos de inicialização, avaliação, recombinação, mutação e parâmetros que os utilizados pelo método GARTS apresentados no Capítulo 6. A única diferença é que o AG desenvolvido não utiliza a abordagem de agrupamento de indivíduos que o método GARTS. Com isso o método de seleção e os critérios de paradas, que utilizavam conhecimento sobre o agrupamento gerado para definir suas características, também foram alterados. Para seleção dos indivíduos foi utilizada uma abordagem onde o primeiro indivíduo é selecionado aleatoriamente dentre as 10 melhores soluções da população atual, enquanto o segundo indivíduo é selecionado aleatoriamente do resto da população. Assim, garante-se que o primeiro indivíduo sempre vai estar entre as 10 melhores soluções da população atual, enquanto o segundo indivíduo selecionado contribuirá para a diversidade do método. O critério de parada também foi alterado a fim do AG executar até atingir um número máximo de avaliações de energia. Esse valor é definido utilizando o maior número de avaliações de energia que o método GARTS atingiu para uma proteína alvo específica. Essa alteração no critério de parada garante que o método AG desenvolvido não execute um número maior de avaliações de energia que o método GARTS, permitindo uma comparação balanceado por esse número de avaliações realizadas.

Para comparar os métodos, o tempo máximo de execução do método GARTS foi aumentado para 24 horas, mas mantendo o número de 8 execuções para a mesma proteína para avaliação estatística. Os parâmetros utilizados foram os mesmos apresentados na seção anterior do conjunto de valores P3. Com isso, foi possível obter o valor máximo de avaliações de energias para poder executar o AG também por 8 vezes para cada proteína testada até atingir esse valor máximo. O conjunto de proteínas testadas foram as 24 estruturas da Base de Teste II, apresentado na Seção 7.3.

A Figura 7.2, mostra os resultados obtidos na última população por cada um dos métodos, AG e GARTS. As estruturas nessa figura, seguem a seguinte configuração: estrutura

vermelha representa a estrutura experimental, a verde representa a estrutura de menor energia e a azul representa a estrutura de menor RMSD encontrado. As estruturas desta figura estão sobrepostas alinhadas pelos C_α, assim, se a estrutura de menor energia representar também a estrutura de menor RMSD, apenas uma delas seria apresentadas. Entretanto, em ambos os métodos, isso não ocorreu, comprovando assim que para esses casos de teste a menor energia não representa a estrutura de menor RMSD. Analisando essa figura, também é possível observar

Figura 7.2: Representação gráfica da comparação entre o AG comparado com o GARTS. A estrutura em vermelho representa estrutura experimental depositada no PDB enquanto a estrutura em azul representa a solução de menor RMSD e a verde a de menor energia encontradas ao final da execução de ambas as meta-heurísticas. Todas as estruturas foram preparadas utilizando o programa PyMol



Fonte: do autor (2016).

que os dois métodos chegaram em estruturas próximas da estrutura experimental depositada no PDB. Porém, analisando visualmente as estruturas de menor RMSD (estrutura em azul) de ambos os métodos é difícil de afirmar qual deles gerou resultados mais próximos da estrutura disponível no PDB (estrutura em vermelho).

Para tentar contornar essa indefinição e permitir encontrar qual método obteve melhores resultados foram geradas todas as tabelas dos critérios de validação apresentados na Seção 7.2. As Tabelas 7.7, 7.8 e 7.9 representam os resultados comparativo entre ambos os métodos utilizando os critérios de validação de energia, RMSD e GDT_TS, respectivamente. As Tabelas 7.10 e 7.11 representam os resultados do método Q-Index para o método AG e GARTS, respectivamente.

Na tabela da energia do método (Tab. 7.7) é possível observar que apenas nas estruturas 1DFN e 1L2Y que a média da menor energia foi atingida utilizando o método GARTS. Entretanto, como mostrado na Seção 7.4 onde são comparados diferentes conjuntos de parâmetros e na Figura 7.2 o menor valor de energia não está representando a estrutura mais próxima da experimental. E essa afirmação é comprovada analisando os valores médios das tabelas de menor RMSD (Tab. 7.8) e de maior GDT_TS (Tab. 7.9), onde apenas as estruturas 1ROP e 3V1A não obtiveram a melhor média utilizando o método GARTS desenvolvido. Assim, em termos de RMSD médio o método GARTS foi melhor em 95% das estruturas testadas e em termos de GDT_TS médio o método GARTS foi melhor em 92% das estruturas testadas, quando comparados com o Algoritmo Genético desenvolvido sem a abordagem de agrupamento. Também foi realizada uma análise sobre a formação da estrutura secundária esperada, utilizando o critério de avaliação Q-Index. O resultado tanto para o método AG (Tab. 7.10) quanto para o método GARTS (Tab. 7.11) tiveram bons resultados para as regiões de hélices (>99%) e regiões desordenadas (>85%). Porém, regiões de voltas e folhas, que representam apenas 15% do total dos resíduos, ficaram próxima de 60%.

Analisando os resultados do Q-Index é possível observar que o método AG (Tab. 7.10) tem uma taxa de acerto geral entre todas as estruturas secundárias de 90.8%. Enquanto o método GARTS (Tab. 7.11) obteve uma pequena vantagem, atingindo uma taxa de acerto geral de 91.6% entre todas as estruturas testadas. Essa similaridade entre os métodos na formação da estrutura secundária ocorre principalmente devido aos dois métodos utilizarem a mesma função de avaliação que contém o reforço da formação da estrutura secundária. Entretanto, como é possível analisar pelas tabelas do RMSD e GDT_TS (Tab. 7.8, 7.9), essa similaridade da estrutura secundária não representa uma similaridade no enovelamento da proteína, onde o método GARTS atingiu resultados mais próximos da estrutura experimental.

Tabela 7.7: Comparativo do menor valor de energia entre os métodos AG e GARTS. Os valores de RMSD, GDT_TS e Avaliações de Energia são referentes a estrutura de Menor valor de Energia. E a Média de Energia é referente aos valores de menor energia das 8 execuções. Os melhores valores de média para cada proteína foram destacados em negrito

Código PDB	Menor Energia	RMSD Å	GDT_TS %	Avaliações de Energia	Energia Média
1AB1-AG	-8.609,45	6,44	48,91	319 619	**-5.703,53** ±(1.678,69)
1AB1-GARTS	-4.820,49	9,79	44,02	319 570	-4.690,43 ±(84,66)
1ACW-AG	-12.403,00	2,04	73,28	757 619	**-11.949,92** ±(525,76)
1ACW-GARTS	-12.561,22	4,04	62,07	658 711	-11.145,78 ±(2.313,05)
1AIL-AG	-53.023,79	6,22	50,71	104 477	**-52.738,47** ±(255,63)
1AIL-GARTS	-52.232,18	15,65	41,07	104 438	-51.892,45 ±(184,33)
1CRN-AG	-6.788,32	5,06	52,72	299 868	**-4.320,03** ±(1.268,53)
1CRN-GARTS	-3.891,94	9,00	39,13	230 907	-3.432,62 ±(279,14)
1D5Q-AG	-16.805,51	1,51	83,33	1 087 305	**-16.718,00** ±(75,27)
1D5Q-GARTS	-16.818,14	3,45	71,30	1 086 392	-16.704,65 ±(97,94)
1DFN-AG	-5.498,11	8,57	47,50	648 673	-3.833,59 ±(1.336,60)
1DFN-GARTS	-7.025,58	6,97	50,83	638 773	**-4.872,02** ±(1.244,84)
1ENH-AG	-31.824,09	10,28	43,06	157 581	**-31.612,07** ±(183,05)
1ENH-GARTS	-31.343,26	11,69	37,04	139 516	-31.162,71 ±(93,60)
1K43-AG	-4.551,36	1,02	82,14	2 475 863	**-4.503,61** ±(32,02)
1K43-GARTS	-4.553,34	0,76	85,71	2 174 983	-4.486,95 ±(41,56)
1L2Y-AG	-2.959,58	1,29	85,00	1 500 057	-2.806,49 ±(74,76)
1L2Y-GARTS	-2.923,57	1,58	86,25	1 184 259	**-2.876,44** ±(32,63)
1Q2K-AG	-16.360,87	3,81	64,52	882 874	**-14.471,87** ±(1.874,52)
1Q2K-GARTS	-16.170,89	3,99	67,74	882 266	-14.108,76 ±(2.101,34)
1ROP-AG	-46.600,14	2,23	70,98	166 622	**-46.490,86** ±(79,21)
1ROP-GARTS	-46.150,77	3,49	62,95	166 309	-45.963,48 ±(116,47)
1UTG-AG	-43.121,89	15,75	37,86	92 705	**-42.633,96** ±(377,61)
1UTG-GARTS	-42.401,63	11,12	37,86	92 659	-41.988,29 ±(331,62)
1WQC-AG	-13.164,33	4,93	56,73	612 446	**-12.161,05** ±(444,07)
1WQC-GARTS	-12.095,76	4,14	61,54	612 016	-12.000,63 ±(70,56)
1ZDD-AG	-22.687,87	3,16	69,85	450 588	**-22.322,23** ±(261,19)
1ZDD-GARTS	-22.537,35	3,78	66,91	423 784	-21.954,14 ±(346,62)
2JUC-AG	-28.579,29	7,65	36,36	129 133	**-28.176,75** ±(277,30)
2JUC-GARTS	-27.828,36	7,19	40,45	128 906	-27.520,08 ±(192,66)
2KDL-AG	-31.409,54	7,82	42,86	163 906	**-31.212,08** ±(163,72)
2KDL-GARTS	-30.842,87	11,43	36,61	153 069	-30.737,14 ±(85,68)
2MR9-AG	-24.749,34	8,50	40,34	315 907	**-24.434,43** ±(366,20)
2MR9-GARTS	-24.674,90	4,80	51,70	315 719	-24.402,91 ±(132,82)
2MTW-AG	-10.063,82	1,41	81,25	341 825	**-9.992,22** ±(42,85)
2MTW-GARTS	-10.027,68	2,10	80,00	340 917	-9.971,18 ±(34,61)
2P5K-AG	-38.191,25	8,59	34,92	117 895	**-35.120,88** ±(1.263,33)
2P5K-GARTS	-34.383,93	9,87	36,11	117 349	-33.568,00 ±(515,64)
2P6J-AG	-28.311,69	7,36	45,19	167 839	**-28.033,14** ±(133,12)
2P6J-GARTS	-27.651,12	4,58	59,13	167 431	-27.440,85 ±(146,50)
2P81-AG	-22.909,45	5,29	53,41	285 800	**-22.700,44** ±(115,92)
2P81-GARTS	-22.723,34	5,63	68,18	285 731	-22.511,95 ±(130,38)
2PMR-AG	-54.548,36	6,50	48,03	76 581	**-54.012,27** ±(226,87)
2PMR-GARTS	-53.207,33	12,73	38,16	69 155	-53.030,39 ±(122,18)
3P7K-AG	-37.810,29	2,25	80,00	163 155	**-37.750,53** ±(58,35)
3P7K-GARTS	-37.714,40	2,70	76,67	163 154	-37.628,47 ±(55,51)
3V1A-AG	-32.756,47	3,73	64,06	249 719	**-32.596,11** ±(86,92)
3V1A-GARTS	-32.571,63	4,51	59,90	248 737	-32.388,81 ±(125,88)

Fonte: do autor (2016).

Tabela 7.8: Comparativo do menor valor de RMSD entre os métodos AG e GARTS. Os valores de Energia, GDT_TS e Avaliações de Energia são referentes a estrutura de Menor valor de RMSD. E a Média de RMSD é referente aos valores de menor RMSD das 8 execuções. Os melhores valores de média para cada proteína foram destacados em negrito

Código PDB	Menor RMSD Å	Energia	GDT_TS %	Avaliações de Energia	RMSD Médio Å
1AB1-AG	6,08	-8.063,91	48,37	320 186	7,47 ±(1,19)
1AB1-GARTS	5,11	-4.168,32	53,80	243 454	**6,39** ±(0,87)
1ACW-AG	1,78	-12.296,00	77,59	757 619	3,85 ±(2,10)
1ACW-GARTS	2,52	-12.323,68	68,97	757 401	**3,08** ±(0,49)
1AIL-AG	6,01	-52.922,46	51,07	104 477	9,79 ±(2,70)
1AIL-GARTS	5,96	-50.881,00	53,21	95 885	**7,53** ±(1,08)
1CRN-AG	4,01	-6.360,30	54,89	299 868	7,26 ±(2,08)
1CRN-GARTS	5,07	-2.755,33	51,63	299 274	**6,82** ±(1,27)
1D5Q-AG	1,38	-16.753,64	87,04	1 087 305	2,62 ±(0,71)
1D5Q-GARTS	1,46	-16.679,71	85,19	568 473	**2,04** ±(0,31)
1DFN-AG	5,59	-4.484,77	54,17	648 344	7,22 ±(1,03)
1DFN-GARTS	3,92	-5.273,29	59,17	648 243	**5,23** ±(0,88)
1ENH-AG	5,72	-29.044,91	51,39	157 350	7,81 ±(1,46)
1ENH-GARTS	3,99	-30.705,02	57,87	157 319	**5,79** ±(1,57)
1K43-AG	0,51	-4.510,18	91,07	2 475 452	0,90 ±(0,19)
1K43-GARTS	0,40	-4.411,86	89,29	2 475 381	**0,58** ±(0,11)
1L2Y-AG	1,20	-2.715,46	83,75	1 499 780	2,21 ±(1,15)
1L2Y-GARTS	0,90	-2.865,86	86,25	1 499 709	**1,44** ±(0,45)
1Q2K-AG	3,27	-13.769,90	63,71	882 972	4,71 ±(2,10)
1Q2K-GARTS	2,68	-9.178,98	70,16	762 101	**3,53** ±(0,50)
1ROP-AG	2,14	-46.393,17	79,91	166 427	**2,48** ±(0,37)
1ROP-GARTS	1,80	-45.879,86	78,57	165 387	2,53 ±(0,50)
1UTG-AG	7,60	-42.683,37	44,29	93 391	10,17 ±(2,52)
1UTG-GARTS	5,79	-41.447,10	54,29	90 064	**7,58** ±(1,30)
1WQC-AG	1,77	-12.321,64	80,77	612 094	3,50 ±(0,82)
1WQC-GARTS	1,45	-12.011,58	82,69	388 920	**2,33** ±(0,59)
1ZDD-AG	2,62	-22.220,68	75,00	450 688	5,33 ±(2,78)
1ZDD-GARTS	1,85	-21.566,47	77,94	450 216	**2,89** ±(0,82)
2JUC-AG	7,57	-28.514,02	36,36	129 133	10,54 ±(3,15)
2JUC-GARTS	6,66	-27.703,45	39,55	124 691	**8,28** ±(0,97)
2KDL-AG	6,01	-31.143,62	48,21	163 679	9,09 ±(1,69)
2KDL-GARTS	4,68	-30.671,88	54,46	163 372	**7,81** ±(1,57)
2MR9-AG	6,57	-24.512,72	44,32	316 314	8,35 ±(1,07)
2MR9-GARTS	3,10	-24.329,66	66,48	274 744	**4,20** ±(0,54)
2MTW-AG	1,32	-10.033,61	81,25	341 825	1,89 ±(0,32)
2MTW-GARTS	1,04	-9.729,03	83,75	160 220	**1,24** ±(0,16)
2P5K-AG	8,35	-38.103,91	34,52	117 895	11,00 ±(2,05)
2P5K-GARTS	6,08	-17.359,77	44,05	117 349	**7,91** ±(1,53)
2P6J-AG	4,18	-28.011,60	57,69	167 486	7,01 ±(1,57)
2P6J-GARTS	4,51	-27.451,79	60,58	167 431	**6,08** ±(1,23)
2P81-AG	3,35	-22.571,16	63,07	286 052	5,43 ±(1,06)
2P81-GARTS	2,34	-22.325,78	75,00	285 731	**3,81** ±(0,82)
2PMR-AG	6,42	-54.485,20	49,01	76 581	12,44 ±(3,69)
2PMR-GARTS	6,75	-52.646,87	44,74	76 295	**8,08** ±(1,28)
3P7K-AG	0,95	-37.624,31	92,78	163 416	1,83 ±(0,63)
3P7K-GARTS	0,67	-37.451,50	97,22	127 669	**0,84** ±(0,15)
3V1A-AG	2,96	-32.532,81	67,71	249 116	3,47 ±(0,44)
3V1A-GARTS	1,81	-32.031,38	75,52	244 116	**3,32** ±(0,68)

Fonte: do autor (2016).

Tabela 7.9: Comparativo do maior valor de GDT_TS entre os métodos AG e GARTS. Os valores de Energia, RMSD e Avaliações de Energia são referentes a estrutura de Maior valor de GDT_TS. A Média de GDT_TS é referente aos valores de maior GDT_TS das 8 execuções. Os melhores valores de média para cada proteína foram destacados em negrito

Código PDB	Maior GDT_TS %	Energia	RMSD Å	Avaliações de Energia	GDT_TS Médio %
1AB1-AG	51,63	-7.249,78	6,21	319 619	46,74 ±(3,13)
1AB1-GARTS	55,43	-4.127,77	5,20	243 454	**50,20** ±(2,50)
1ACW-AG	77,59	-12.304,04	1,81	757 619	65,95 ±(8,72)
1ACW-GARTS	71,55	-11.484,59	2,83	750 559	**68,00** ±(4,16)
1AIL-AG	51,43	-52.943,21	6,21	104 477	46,30 ±(4,00)
1AIL-GARTS	58,21	-50.957,37	15,77	88 391	**55,40** ±(1,66)
1CRN-AG	55,43	-6.361,37	4,04	299 868	48,29 ±(4,35)
1CRN-GARTS	54,35	-3.171,86	6,13	285 371	**49,32** ±(3,23)
1D5Q-AG	87,04	-16.770,90	1,41	1 087 305	75,81 ±(6,01)
1D5Q-GARTS	85,19	-16.679,71	1,46	568 473	**79,98** ±(4,37)
1DFN-AG	54,17	-4.485,34	5,59	648 344	48,54 ±(4,45)
1DFN-GARTS	60,00	-5.319,11	3,95	648 243	**55,83** ±(2,99)
1ENH-AG	62,50	-31.705,55	8,23	157 384	50,40 ±(6,32)
1ENH-GARTS	59,72	-31.052,05	4,43	136 702	**54,69** ±(5,15)
1K43-AG	91,07	-4.479,42	1,02	2 475 627	84,37 ±(5,63)
1K43-GARTS	96,43	-4.447,50	0,43	2 410 580	**89,96** ±(3,93)
1L2Y-AG	87,50	-2.922,08	1,31	1 500 057	80,00 ±(6,61)
1L2Y-GARTS	88,75	-2.856,76	1,06	1 476 045	**84,22** ±(4,17)
1Q2K-AG	70,16	-16.299,48	3,65	882 874	62,67 ±(6,38)
1Q2K-GARTS	70,97	-11.708,77	2,71	762 101	**67,24** ±(2,47)
1ROP-AG	79,91	-46.392,89	2,24	166 427	**73,89** ±(4,01)
1ROP-GARTS	78,57	-45.939,63	1,82	165 387	73,60 ±(2,84)
1UTG-AG	45,00	-42.687,84	8,59	93 391	40,45 ±(3,54)
1UTG-GARTS	54,29	-41.447,10	5,79	90 064	**46,52** ±(3,80)
1WQC-AG	80,77	-12.321,64	1,77	612 094	66,10 ±(7,08)
1WQC-GARTS	83,65	-11.993,50	1,54	388 920	**72,60** ±(6,29)
1ZDD-AG	75,00	-22.220,68	2,62	450 688	62,41 ±(9,60)
1ZDD-GARTS	77,94	-21.461,63	2,28	550 174	**72,33** ±(4,73)
2JUC-AG	45,00	-28.099,48	8,39	129 005	37,01 ±(4,44)
2JUC-GARTS	41,82	-27.568,34	7,10	128 906	**39,66** ±(2,28)
2KDL-AG	49,55	-31.136,58	6,09	163 679	43,08 ±(3,19)
2KDL-GARTS	54,46	-30.671,88	4,68	163 372	**46,26** ±(4,10)
2MR9-AG	51,70	-17.170,10	9,21	315 773	44,67 ±(3,73)
2MR9-GARTS	67,61	-24.466,02	3,13	274 744	**59,66** ±(5,01)
2MTW-AG	83,75	-9.912,73	1,95	340 968	80,78 ±(1,76)
2MTW-GARTS	86,25	-9.793,75	1,49	201 193	**84,84** ±(1,70)
2P5K-AG	42,86	-34.615,19	10,44	117 403	36,91 ±(2,77)
2P5K-GARTS	51,98	-23.647,54	8,87	117 349	**45,19** ±(4,33)
2P6J-AG	58,17	-27.975,39	4,23	167 486	50,06 ±(4,64)
2P6J-GARTS	60,58	-27.087,96	5,36	173 840	**53,01** ±(5,89)
2P81-AG	67,61	-22.513,70	6,05	285 962	56,68 ±(8,07)
2P81-GARTS	75,57	-22.360,05	2,36	285 731	**64,20** ±(5,98)
2PMR-AG	49,67	-54.487,57	6,44	76 581	41,61 ±(4,00)
2PMR-GARTS	52,63	-51.045,62	16,60	69 155	**48,81** ±(3,30)
3P7K-AG	93,89	-37.624,55	1,01	163 416	83,96 ±(6,73)
3P7K-GARTS	97,22	-37.447,16	0,72	113 129	**94,93** ±(2,21)
3V1A-AG	71,35	-32.610,13	3,20	248 767	**66,74** ±(2,94)
3V1A-GARTS	75,52	-32.031,38	1,81	244 116	66,08 ±(4,40)

Fonte: do autor (2016).

Tabela 7.10: Análise da formação da estrutura secundária entre a estrutura Predita (P) pelo método AG e a Experimental (E), utilizando o critério de avaliação Q-Index

PDB ID (Tam.)	% Q_H (P/E)	% Q_E (P/E)	% Q_T (P/E)	% Q_C (P/E)	% Q4
1AB1 (46)	95.0% (19/20)	0.0% (0/4)	0.0% (0/5)	64.7% (11/17)	65.2%
1ACW (29)	100.0% (9/9)	100.0% (10/10)	40.0% (2/5)	100.0% (5/5)	89.7%
1AIL (70)	96.67% (58/60)	–	–	100.0% (10/10)	97.1%
1CRN (46)	100.0% (20/20)	0.0% (0/4)	40.0% (2/5)	94.1% (16/17)	82.6%
1D5Q (27)	100.0% (11/11)	100.0% (8/8)	100.0% (2/2)	100.0% (6/6)	100.0%
1DFN (30)	–	75.0% (12/16)	77.8% (7/9)	80.0% (4/5)	76.7%
1ENH (54)	100.0% (38/38)	–	–	100.0% (16/16)	100.0%
1K43 (14)	–	100.0% (6/6)	40.0% (2/5)	100.0% (3/3)	78.6%
1L2Y (20)	100.0% (12/12)	–	–	50.0% (4/8)	80.0%
1Q2K (31)	100.0% (11/11)	0.0% (0/8)	100.0% (4/4)	62.5% (5/8)	64.5%
1ROP (56)	100.0% (51/51)	–	–	100.0% (5/5)	100.0%
1UTG (70)	100.0% (56/56)	–	–	100.0% (14/14)	100.0%
1WQC (26)	100.0% (18/18)	–	–	87.5% (7/8)	96.2%
1ZDD (34)	100.0% (26/26)	–	80.0% (4/5)	100.0% (3/3)	97.1%
2JUC (55)	100.0% (35/35)	–	61.5% (8/13)	85.7% (6/7)	89.1%
2KDL (56)	100.0% (36/36)	–	80.0% (4/5)	86.7% (13/15)	94.6%
2MR9 (44)	100.0% (30/30)	–	33.3% (3/9)	100.0% (5/5)	86.4%
2MTW (20)	100.0% (12/12)	–	–	50.0% (4/8)	80.0%
2P5K (63)	97.22% (35/36)	80.0% (8/10)	100.0% (3/3)	78.6% (11/14)	90.5%
2P6J (52)	100.0% (33/33)	–	42.9% (3/7)	100.0% (12/12)	92.3%
2P81 (44)	100.0% (27/27)	–	40.0% (2/5)	83.3% (10/12)	88.6%
2PMR (76)	98.41% (62/63)	–	75.0% (3/4)	77.8% (7/9)	94.7%
3P7K (45)	100.0% (44/44)	–	–	100.0% (1/1)	100.0%
3V1A (48)	100.0% (38/38)	–	66.7% (4/6)	75.0% (3/4)	93.8%
Geral	99.3% (681/686)	66.7% (44/66)	57.6% (53/92)	85.4% (181/212)	90.8%

Fonte: do autor (2016).

Tabela 7.11: Análise da formação da estrutura secundária entre a estrutura Predita (P) pelo método GARTS e a Experimental (E), utilizando o critério de avaliação Q-Index.

PDB ID (Tam.)	% Q_H (P/E)	% Q_E (P/E)	% Q_T (P/E)	% Q_C (P/E)	% $Q4$
1AB1 (46)	100.0% (20/20)	0.0% (0/4)	100.0% (5/5)	88.2% (15/17)	87.0%
1ACW (29)	100.0% (9/9)	100.0% (10/10)	40.0% (2/5)	100.0% (5/5)	89.7%
1AIL (70)	100.0% (60/60)	–	–	60.0% (6/10)	94.3%
1CRN (46)	100.0% (20/20)	0.0% (0/4)	100.0% (5/5)	100.0% (17/17)	91.3%
1D5Q (27)	100.0% (11/11)	100.0% (8/8)	100.0% (2/2)	100.0% (6/6)	100.0%
1DFN (30)	–	75.0% (12/16)	88.9% (8/9)	60.0% (3/5)	76.7%
1ENH (54)	100.0% (38/38)	–	–	75.0% (12/16)	92.6%
1K43 (14)	–	100.0% (6/6)	40.0% (2/5)	100.0% (3/3)	78.6%
1L2Y (20)	100.0% (12/12)	–	–	87.5% (7/8)	95.0%
1Q2K (31)	100.0% (11/11)	50.0% (4/8)	100.0% (4/4)	62.5% (5/8)	77.4%
1ROP (56)	100.0% (51/51)	–	–	100.0% (5/5)	100.0%
1UTG (70)	100.0% (56/56)	–	–	85.7% (12/14)	97.1%
1WQC (26)	100.0% (18/18)	–	–	87.5% (7/8)	96.2%
1ZDD (34)	100.0% (26/26)	–	80.0% (4/5)	100.0% (3/3)	97.1%
2JUC (55)	100.0% (35/35)	–	53.9% (7/13)	85.7% (6/7)	87.3%
2KDL (56)	100.0% (36/36)	–	80.0% (4/5)	66.7% (10/15)	89.3%
2MR9 (44)	96.7% (29/30)	–	88.9% (8/9)	80.0% (4/5)	93.2%
2MTW (20)	100.0% (12/12)	–	–	100.0% (8/8)	100.0%
2P5K (63)	100.0% (36/36)	0.0% (0/10)	0.0% (0/3)	92.9% (13/14)	77.8%
2P6J (52)	100.0% (33/33)	–	42.9% (3/7)	91.7% (11/12)	90.4%
2P81 (44)	100.0% (27/27)	–	100.0% (5/5)	100.0% (12/12)	100.0%
2PMR (76)	100.0% (63/63)	–	0.0% (0/4)	100.0% (9/9)	94.7%
3P7K (45)	100.0% (44/44)	–	–	100.0% (1/1)	100.0%
3V1A (48)	100.0% (38/38)	–	0.0% (0/6)	75.0% (3/4)	85.4%
Geral	99.9% (685/686)	60.6% (40/66)	64.1% (59/92)	86.3% (183/212)	91.6%

Fonte: do autor (2016).

Com essas análises realizadas, é possível afirmar que o método GARTS utilizando a abordagem de agrupamento de estruturas, consegue gerar, em média, estruturas mais próximas das experimentais do que o método AG desenvolvido que não utiliza esse tipo de abordagem. Desta forma, é possível comparar o método desenvolvido com abordagens reconhecidas descritas na literatura, a fim de analisar o real potencial do método GARTS.

7.6 Comparativo Entre o Método GARTS e o Servidor I-TASSER

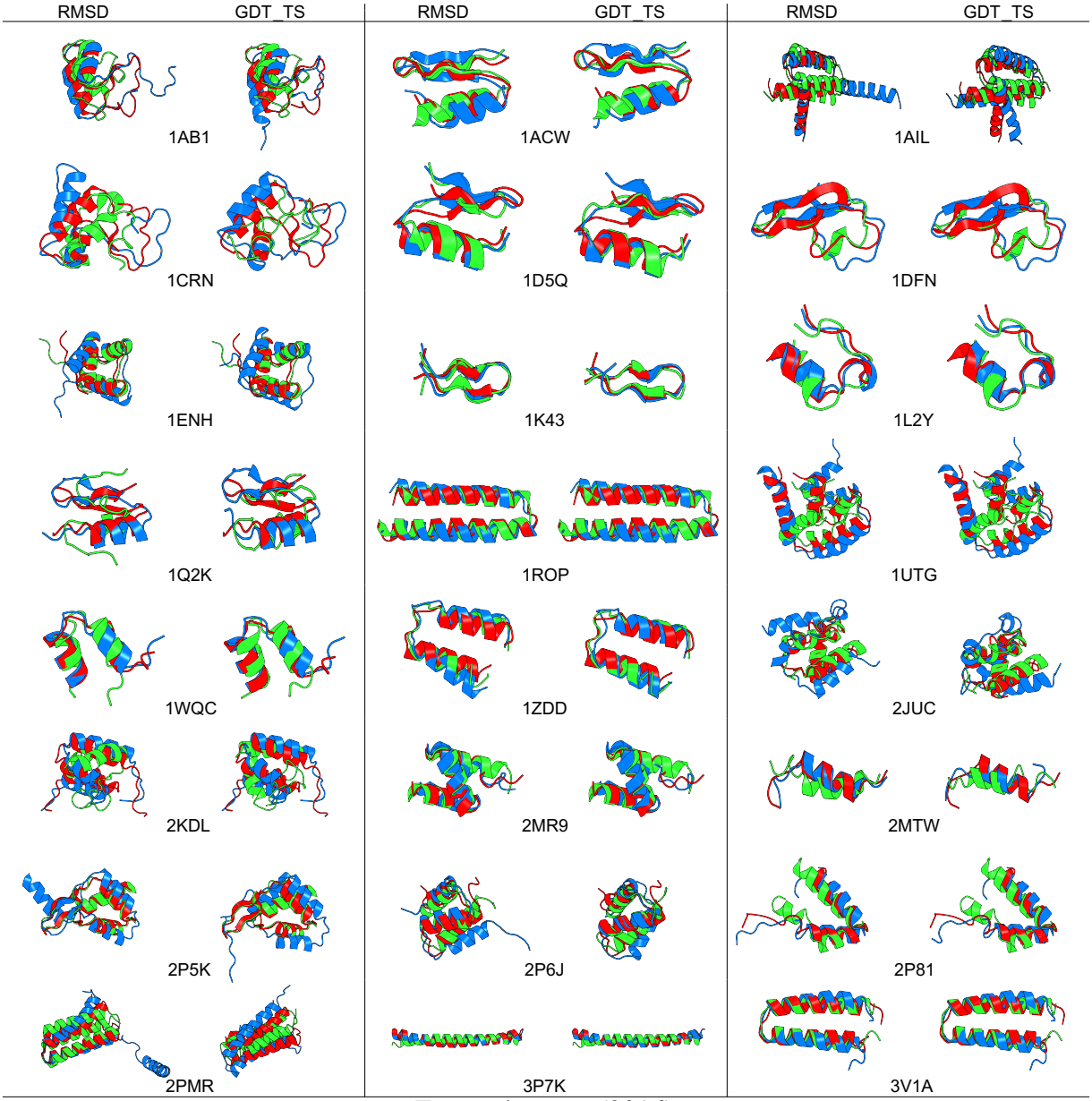
Para validação final do método GARTS realizou-se uma comparação com um método relevante na área de predição da estrutura 3D de proteínas. Para realizar essa comparação foi utilizado o servidor que apresentou os melhores resultados no último CASP (edição 11) apresentado no Capítulo 3. Dentre os melhores servidores estão o I-TASSER, QUARK e ROBETTA. Entretanto, os servidores QUARK e ROBETTA não permitem uma configuração mais específica que permitisse uma comparação justa com o método GARTS desenvolvido (definir a estrutura secundária esperada e limitar o conhecimento sobre a proteína alvo), por esse motivo o servidor I-TASSER foi escolhido.

O servidor I-TASSER é um dos mais utilizados e que constantemente obtém bons resultados na competição CASP. Para poder fazer essa comparação de uma forma justa com o método GARTS desenvolvido, foi necessário remover as estruturas de proteínas que o servidor SAS (Seção 7.3) considerou com similaridade de sequência das bases de dados do I-TASSER, do mesmo modo que foi feito para o método GARTS. Porém, o servidor I-TASSER ainda mantém um mínimo de conhecimento de 25% dessas estruturas. Para garantir que a predição da estrutura secundária que ocorre no servidor I-TASSER não influenciasse no resultado, utilizou-se a mesma entrada de estrutura secundária informado no método GARTS. Assim, os dois métodos tem a mesma entrada de dados e não utilizam estruturas com alta taxa de similaridade de sequência para fazer a predição, com a diferença apenas na maneira que são utilizadas as informações do PDB e o método desenvolvido. O servidor I-TASSER retorna até cinco modelos de estruturas como possíveis soluções para a predição, onde o modelo 1 é a solução que o método considera a mais provável. Entretanto, quando o método estabiliza, apenas um modelo é considerado como resultado. Assim, para comparação com o método GARTS desenvolvido, todos os modelos retornados foram considerados para aplicar os critérios de validação de RMSD, GDT_TS e Q-Index.

A Figura 7.3 mostra a comparação gráfica entre as melhores estruturas encontradas na última população do método GARTS (estruturas em azul) e a melhor estrutura entre os modelos

retornados pelo servidor I-TASSER (estruturas em verde) em comparação com estruturas experimentais (estruturas em vermelho). Esse comparativo mostra o melhor resultado dos métodos em termos de RMSD e GDT_TS para as 24 estruturas de proteínas da Base de Testes II. Analisando essa figura é possível notar que ambos os métodos atingiram estruturas próximas da experimental. Porém, como aconteceu com as outras comparações gráficas entre estruturas apresentadas anteriormente, uma análise visual sobre qual método obteve melhores resultados é uma tarefa bastante complicada.

Figura 7.3: Representação gráfica da comparação entre o GARTS e o servidor I-TASSER. A estrutura em vermelho representa estrutura experimental depositada no PDB enquanto a estrutura em azul e a estrutura em verde representam a melhor solução encontrada ao final da execução do método GARTS e do servidor I-TASSER, respectivamente. Todas as estruturas foram preparadas utilizando o programa PyMol



Fonte: do autor (2016).

Para verificar qual método obteve os melhores resultados é necessário adotar outros parâmetros para análise estrutural dos resultados. A Tabela 7.12 apresenta os resultados da análise da estrutura secundária entre a estrutura predita pelo servidor I-TASSER e a estrutura secundária da proteína experimental. Nessa tabela é possível verificar que o método I-TASSER obteve uma taxa de acerto geral de aproximadamente 70%, ficando bem abaixo dos 91.6% (Tab. 7.11) obtido pelo método GARTS. Também foi comparado qual método obteve uma estrutura com enovelamento mais próximo à proteína no seu estado nativo, através dos critérios de validação de RMSD e GDT_TS. A Tabela 7.13 apresenta esses resultados, contendo a comparação dos valores de RMSD e GDT_TS entre o método GARTS e o servidor I-TASSER. Analisando o menor valor de RMSD encontrado ao final de cada método, é possível verificar que em aproximadamente 63% dos casos testados o método GARTS obteve um valor menor de RMSD que o servidor I-TASSER. Quando essa mesma análise é feita, procurando o maior valor de GDT_TS, é possível observar que em 54% dos casos o melhor resultado foi obtido utilizando o método GARTS desenvolvido.

Tabela 7.12: Análise da formação da estrutura secundária entre a estrutura Predita (P) pelo servidor I-Tasser e a Experimental (E), utilizando o critério de avaliação Q-Index

PDB ID (Tam.)	% Q_H (P/E)	% Q_E (P/E)	% Q_T (P/E)	% Q_C (P/E)	% Q4
1AB1 (46)	90.0% (18/20)	0.0% (0/4)	80.0% (4/5)	29.4% (5/17)	58.7%
1ACW (29)	100.0% (9/9)	100.0% (10/10)	40.0% (2/5)	40.0% (2/5)	79.3%
1AIL (70)	80.0% (48/60)	–	–	30.0% (3/10)	72.9%
1CRN (46)	0.0% (0/20)	0.0% (0/4)	80.0% (4/5)	23.5% (4/17)	17.4%
1D5Q (27)	100.0% (11/11)	50.0% (4/8)	100.0% (2/2)	83.3% (5/6)	81.5%
1DFN (30)	–	25.0% (4/16)	55.6% (5/9)	20.0% (1/5)	33.3%
1ENH (54)	97.4% (37/38)	–	–	56.3% (9/16)	85.2%
1K43 (14)	–	100.0% (6/6)	40.0% (2/5)	100.0% (3/3)	78.6%
1L2Y (20)	58.3% (7/12)	–	–	75.0% (6/8)	65.0%
1Q2K (31)	0.0% (0/11)	0.0% (0/8)	100.0% (4/4)	50.0% (4/8)	25.8%
1ROP (56)	98.0% (50/51)	–	–	40.0% (2/5)	92.9%
1UTG (70)	73.2% (41/56)	–	–	57.1% (8/14)	70.0%
1WQC (26)	88.9% (16/18)	–	–	37.5% (3/8)	73.1%
1ZDD (34)	84.6% (22/26)	–	80.0% (4/5)	100.0% (3/3)	85.3%
2JUC (55)	100.0% (35/35)	–	30.8% (4/13)	57.1% (4/7)	78.2%
2KDL (56)	63.9% (23/36)	–	60.0% (3/5)	26.7% (4/15)	53.6%
2MR9 (44)	100.0% (30/30)	–	33.3% (3/9)	80.0% (4/5)	84.1%
2MTW (20)	58.3% (7/12)	–	–	37.5% (3/8)	50.0%
2P5K (63)	91.7% (33/36)	40.0% (4/10)	100.0% (3/3)	50.0% (7/14)	74.6%
2P6J (52)	100.0% (33/33)	–	28.6% (2/7)	41.7% (5/12)	76.9%
2P81 (44)	92.6% (25/27)	–	0.0% (0/5)	50.0% (6/12)	70.5%
2PMR (76)	68.3% (43/63)	–	0.0% (0/4)	44.4% (4/9)	61.8%
3P7K (45)	93.2% (41/44)	–	–	0.0% (0/1)	91.1%
3V1A (48)	94.7% (36/38)	–	66.7% (4/6)	75.0% (3/4)	89.6%
Geral	82.4% (565/686)	42.4% (28/66)	50.0% (46/92)	46.2% (98/212)	69.8%

Fonte: do autor (2016).

Tabela 7.13: Tabela comparativa entre os melhores valores de RMSD e GDT_TS obtidos pelo método GARTS e pelo servidor I-TASSER. Os melhores valores para cada proteína foram destacados em negrito

PDB ID	Menor RMSD Å	GDT_TS %	Maior GDT_TS %	RMSD Å
1AB1-GARTS	**5,11**	53,80	**55,43**	5,20
1AB1-I-TASSER	6,30	52,72	52,72	6,30
1ACW-GARTS	2,52	68,97	71,55	2,83
1ACW-I-TASSER	**1,09**	86,21	**86,21**	1,09
1AIL-GARTS	**5,96**	53,21	**58,21**	15,77
1AIL-I-TASSER	7,53	47,14	47,14	7,53
1CRN-GARTS	**5,07**	51,63	**54,35**	6,13
1CRN-I-TASSER	7,43	36,96	44,57	8,46
1D5Q-GARTS	**1,46**	85,19	**85,19**	1,46
1D5Q-I-TASSER	2,11	75,00	78,70	2,33
1DFN-GARTS	3,92	59,17	60,00	3,95
1DFN-I-TASSER	**2,50**	72,50	**72,50**	2,50
1ENH-GARTS	3,99	57,87	59,72	4,43
1ENH-I-TASSER	**1,88**	88,43	**88,43**	1,88
1K43-GARTS	**0,40**	89,29	**96,43**	0,43
1K43-I-TASSER	0,72	87,50	91,07	1,10
1L2Y-GARTS	**0,90**	86,25	**88,75**	1,06
1L2Y-I-TASSER	3,31	70,00	70,00	3,31
1Q2K-GARTS	**2,68**	70,16	**70,97**	2,71
1Q2K-I-TASSER	5,08	46,77	48,39	5,33
1ROP-GARTS	1,80	78,57	78,57	1,82
1ROP-I-TASSER	**1,57**	81,25	**81,25**	1,57
1UTG-GARTS	**5,79**	54,29	**54,29**	5,79
1UTG-I-TASSER	6,73	52,50	52,50	6,73
1WQC-GARTS	**1,45**	82,69	**83,65**	1,54
1WQC-I-TASSER	2,13	71,15	72,12	2,61
1ZDD-GARTS	1,85	77,94	77,94	2,28
1ZDD-I-TASSER	**1,79**	79,41	**79,41**	1,79
2JUC-GARTS	6,66	39,55	41,82	7,10
2JUC-I-TASSER	**1,72**	80,45	**80,45**	1,72
2KDL-GARTS	**4,68**	54,46	**54,46**	4,68
2KDL-I-TASSER	9,93	34,82	36,61	10,66
2MR9-GARTS	3,10	66,48	67,61	3,13
2MR9-I-TASSER	**2,28**	81,25	**81,25**	2,28
2MTW-GARTS	**1,04**	83,75	**86,25**	1,49
2MTW-I-TASSER	3,23	66,25	73,75	3,83
2P5K-GARTS	6,08	44,05	51,98	8,87
2P5K-I-TASSER	**1,06**	94,84	**94,84**	1,06
2P6J-GARTS	4,51	60,58	60,58	5,36
2P6J-I-TASSER	**2,60**	65,38	**66,35**	2,87
2P81-GARTS	**2,34**	75,00	**75,57**	2,36
2P81-I-TASSER	3,93	63,64	63,64	3,93
2PMR-GARTS	**6,75**	44,74	52,63	16,60
2PMR-I-TASSER	8,28	43,42	**53,95**	8,47
3P7K-GARTS	**0,67**	97,22	**97,22**	0,72
3P7K-I-TASSER	1,85	80,56	80,56	1,85
3V1A-GARTS	**1,81**	75,52	75,52	1,81
3V1A-I-TASSER	1,91	85,42	**85,42**	1,91

Fonte: do autor (2016).

Com isso, é possível verificar que o método GARTS desenvolvido, aplicado a Base de Teste II conseguiu gerar estruturas equiparáveis ou até melhores que os resultados obtidos através do servidor I-TASSER, quando a base de dados é limitada para não utilizar conhecimento de estruturas com alta taxa de similaridade de sequência com a proteína alvo. Assim, pode-se dizer que o método GARTS, quando testado para proteínas que não possuem similaridade de sequência com nenhuma outra estrutura da base (*Free Modeling*), consegue gerar resultados para pequenas proteínas globulares, em média, próximos ou até melhores que o servidor I-TASSER.

7.7 Resumo do Capítulo

Neste capítulo foram apresentados os critérios de validação dos resultados e os testes realizados com o método de predição desenvolvido. Primeiro foi testado qual o conjunto de parâmetros aplicado ao método GARTS produzia melhores resultados para um conjunto de teste de 8 proteínas. O segundo teste foi para comparar o método GARTS com um Algoritmo Genético sem a abordagem de agrupamento para mostrar a vantagem de utilizar a técnica implementada. O método também foi comparado com o servidor I-TASSER, que é um dos métodos referência na área de predição da estrutura 3D de proteínas ficando em primeiro lugar nos últimos CASP realizados. E os resultados desta comparação mostraram que em 13 das 24 proteínas testadas o método GARTS produziu resultados melhores que o servidor I-TASSER. No próximo capítulo são feitas as considerações finais do trabalho realizado e apresentados alguns trabalhos que podem ser desenvolvidos no futuro para conseguir melhorar os resultados obtidos.

8 CONCLUSÃO

A Bioinformática Estrutural é uma das principais subáreas de pesquisa da Bioinformática e está concentrada principalmente no estudo da estrutura tridimensional (3D) de macromoléculas biológicas tais como as proteínas. O conhecimento sobre a estrutura 3D de uma proteína proporciona aos pesquisadores importante informação para inferir a função da proteína na célula. A determinação da estrutura de uma proteína é demorada e experimentalmente cara (devido aos custos associados com a cristalografia, eletroscopia ou ressonância magnética nuclear). A dificuldade em determinar a estrutura 3D de proteínas gerou uma enorme discrepância entre o volume de dados (sequência de resíduos de aminoácidos) e o número de estruturas 3D conhecidas (apenas cerca de 0.11% das sequências de proteínas possuem estrutura 3D conhecida). Esta situação não somente ilustra a necessidade, mas também motiva futuras pesquisas no campo de desenvolvimento de métodos computacionais para a predição da estrutura tridimensional de proteínas. O problema de determinação do enovelamento de uma proteína é classificado em complexidade computacional como um problema NP completo, isto é, ele está entre os mais difíceis problemas em termos de requisitos computacionais. Esta complexidade deve-se ao fato que o processo de enovelamento de uma proteína ser extremamente seletivo. Uma longa cadeia de resíduos de aminoácidos acaba assumindo um imenso número de conformações.

Analisando os resultados dos últimos CASP é possível observar que métodos que utilizam conhecimento das bases de estruturas experimentais conseguem obter melhores resultados, entretanto ainda estão longe de conseguir solucionar o problema da predição da estrutura 3D de proteínas. Nesta dissertação foi apresentado um novo método na tentativa de solucionar este problema, que combina um Algoritmo Genético (AG) com uma variação do método Seleção por Torneio Restrito Adaptativo (ARTS) utilizado para resolução de problemas multimodais. Esse método, chamado de GARTS, utiliza conhecimento extraídos da base do PDB em forma de biblioteca de fragmentos (FM-B Lib) e preferências conformacionais dos aminoácidos (APL e NPAS).

Nos experimentos realizados os resultados indicaram que o método GARTS desenvolvido conseguiu obter estruturas mais próximas das proteínas determinadas experimentalmente que o método AG que não utiliza a abordagem multimodal para resolver o problema. Estes resultados demonstraram que a aplicação de técnicas multimodais podem auxiliar a contornar o complexo panorama energético do problema de predição da estrutura 3D de proteínas. Quando comparado com o servidor I-TASSER, um dos melhores servidores

para predição da estrutura 3D de proteínas, o método GARTS desenvolvido conseguiu atingir resultados equiparáveis e até melhores na predição de um conjunto de 24 estruturas de proteínas, sem utilizar conhecimento de estruturas similares com as proteínas testadas.

Assim, pode-se afirmar que as maiores contribuições desse trabalho foram: o desenvolvimento de uma nova abordagem que combina características de diferentes técnicas computacionais (AG e ARTS) para tentar auxiliar a resolução do problema da predição da estrutura 3D de proteínas, que ainda desafia pesquisadores de diversas áreas; extrair conhecimento das bases de estruturas determinar experimentalmente (PDB) para reduzir o espaço de buscar do problema através das abordagens de APL, NPAS e FM-B Lib;

8.1 Trabalhos futuros

A partir do desenvolvimento deste trabalho diversas oportunidades de pesquisas foram identificadas, principalmente nas áreas de Bioinformática Estrutural e Ciência da Computação. Por exemplo, em Ciência da Computação, um trabalho que poderia ser desenvolvido seria pesquisar diferentes técnicas de agrupamento ou até mesmo desenvolver uma técnica de agrupamento que permita trabalhar com estruturas tridimensionais, ou que resolvessem o agrupamento de estruturas de maneira mais eficiente. Ou até mesmo testar diferentes métricas para definir a similaridade dessas estruturas. Outra pesquisa para a área de Ciência da Computação é o desenvolvimento de uma abordagem auto-adaptativa que permita a atualização dos parâmetros do método, conforme a evolução dos seus resultados. Também poderia ser interessante, aplicar técnicas de Aprendizado de Máquina para aprender a partir da base de dados qual fragmento deve ser utilizado para poder gerar uma nova estrutura. Como pesquisa na Bioinformática Estrutural, seria interessante desenvolver um protocolo para refinamento/otimização das estruturas resultantes do método, a fim de retirar choques interatômicos. A utilização de informações de mapa de contatos das proteínas também poderia ser incorporada como conhecimento ao método. Outro trabalho futuro é o desenvolvimento de um servidor que permita a comunidade acadêmica utilizar o método proposto para predição da estrutura 3D de proteínas, permitindo assim a participação na competição CASP.

9 PUBLICAÇÕES E PRODUÇÃO TÉCNICA DESENVOLVIDO DURANTE O MESTRADO

9.1 Artigos Completos Publicados em Periódicos

- **BORGUESAN, B.**; BARBACHAN e SILVA, M.; GRISCI, B. I.; INSTROZA-PONTA, M.; DORN, M.. APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. Computational Biology and Chemistry (Print), 2015. **CAPES: A2**

9.2 Trabalhos Completos Publicados em Anais de Congressos

- GRISCI, B. I. ; **BORGUESAN, B.**; DORN, M.; INOSTROZA, M.. Using conformational preferences of amino acid residues and meta-heuristics to predict 3-D protein structures. In: Third International Society for Computational Biology Latin America, 2014, Belo Horizonte. Proceedings of third International Society for Computational Biology Latin America. La Jolla: International Society for Computational Biology, 2014. **CAPES: B4**

9.3 Resumos Expandidos Publicados em Anais de Congressos

- **BORGUESAN, B.**; TORBES, A. R.; VERLI, H.; DORN, M.. CarbM: a web tool to build 3-D structures of carbohydrates. In: Third International Society for Computational Biology Latin America, Belo Horizonte. Proceedings of third International Society for Computational Biology Latin America. La Jolla: International Society for Computational Biology, 2014. **CAPES: B4**
- **BORGUESAN, B.**; BARBACHAN e SILVA, M.; DORN, M.. Side-Chain conformational analysis of the multi-dependent rotamer preferences of proteins. In: Escola Gaúcha de Bioinformática, Porto Alegre. Anais da I Escola Gaúcha de Bioinformática, 2015.
- BOHRER, J. S. ; **BORGUESAN, B.**; DORN, M. . A distributed knowledge-based genetic algorithm for protein structure prediction. In: Escola Gaúcha de Bioinformática, Porto Alegre. Anais da I Escola Gaúcha de Bioinformática, 2015.

- BARBACHAN e SILVA, M. ; **BORGUESAN, B.**; DORN, M. . A knowledge-based particle swarm optimization for the protein structure prediction problem. In: Escola Gaúcha de Bioinformática, Porto Alegre. Anais da I Escola Gaúcha de Bioinformática, 2015.
- **BORGUESAN, B.**; DORN, M.. A MULTI-DEPENDENT SIDE-CHAIN ROTAMER LIBRARY FOR PROTEIN STRUCTURE PREDICTION. In: X-Meeting 2015 - 11th International Conference of the AB3C+Brazilian Symposium of Bioinformatics", São Paulo. Proceedings of 11th International Conference of the AB3C+Brazilian Symposium of Bioinformatics - X-Meeting, 2015. **CAPES: B4**

9.4 Programas de Computador Sem Registro

- **BORGUESAN, B.**; TORBES, A. R.; DORN, M.; VERLI, H.. CarbM: a web tool to build three-dimensional structures of carbohydrates. (http://sbcb.inf.ufrgs.br/carbm.html), 2014.
- **BORGUESAN, B.**; INSTROZA-PONTA, M.; DORN, M.. NPAS-Server: Neighbors Preferences of Amino Acids and Secondary Structures - Server (http://sbcb.inf.ufrgs.br/npas), 2015.

9.5 Trabalhos Completos Em Revisão

- **BORGUESAN, B.**; BOHRER, J. S. ; BARBACHAN e SILVA, M.; CORREA, L. L.; DORN, M.. Improving protein tertiary structure prediction with conformational propensities of amino acid residues. IEEE Congress on Evolutionary Computation CEC - 2016. **CAPES: A2**
- CORREA, L. L.; **BORGUESAN, B.**; FARFAN, C.; INSTROZA-PONTA, M.; DORN, M.. Knowledge-Based Memetic Algorithm for Protein Structure Prediction Problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics. - 2016. **CAPES: B1**

REFERÊNCIAS

ALBERTS, B. et al. **Molecular Biology of the Cell**. 5. ed. [S.l.]: Garland Science, 2007. 1392 p. Hardcover.

ALEXANDER, P. A. et al. A minimal sequence code for switching protein structure and function. **Proceedings of the National Academy of Sciences**, v. 106, n. 50, p. 21149–21154, Dec 2009.

ALTMAN, R. B.; DUGAN, J. M. Defining bioinformatics and structural bioinformatics. In: BOURNE, P. E.; WEISSIG, H. (Ed.). **Structural Bioinformatics**. 1. ed. New York: John Wiley and Sons, Inc., 2005. v. 44, cap. 1, p. 1–14.

ALTSCHUL, S. F. et al. Gapped blast and psi-blast: a new generation of protein database search programs. **Nucleic Acids Research**, Oxford Univ Press, v. 25, n. 17, p. 3389–3402, 1997.

ANDERSEN, C. A. F.; ROST, B. Secondary structure assignment: Structural bioinformatics. In: BOURNE, P. E.; WEISSIG, H. (Ed.). **Structural Bioinformatics**. New York: John Wiley and Sons, Inc., 2005. cap. 17, p. 341.

ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, v. 181, n. 4096, p. 223–230, 1973.

BAKER, J. E. Adaptive selection methods for genetic algorithms. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS AND THEIR APPLICATIONS. **Proceedings...** Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1985. p. 101–111.

BANNER, D.; KOKKINIDIS, M.; TSERNOGLOU, D. Structure of the ColE1 rop protein at 1.7 A resolution. **Journal of Molecular Biology**, v. 196, n. 3, p. 657–675, Aug 1987.

BENSON, D. A. et al. Genbank. **Nucleic Acids Research**, v. 41, n. D1, p. D36–D42, 2013.

BERMAN, H. et al. The protein data bank at 40: Reflecting on the past to prepare for the future. **Structure**, v. 20, n. 3, p. 391 – 396, 2012.

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.

BERMAN, H. M. The Protein Data Bank: a historical perspective. **Acta Crystallographica Section A**, v. 64, n. 1, p. 88–95, Jan 2008.

BLANC, E. et al. Solution structure of P01, a natural scorpion peptide structurally analogous to scorpion toxins specific for apamin-sensitive potassium channel. **Proteins: Structure, Function, and Bioinformatics**, v. 24, n. 3, p. 359–369, Mar 1996.

BLUM, C.; ROLI, A. Metaheuristics in combinatorial optimization: overview and conceptual comparison. **ACM Computing Surveys**, v. 35, p. 268, 2003.

BOEHR, D. D.; NUSSINOV, R.; WRIGHT, P. E. The role of dynamic conformational ensembles in biomolecular recognition. **Nature Chemical Biology**, Nature Publishing Group, v. 5, n. 11, p. 789–796, 2009.

BONET, R.; RAMIREZ-ESPAIN, X.; MACIAS, M. Solution structure of the yeast URN1 splicing factor FF domain: comparative analysis of charge distributions in FF domain structures-FFs and SURPs, two domains with a similar fold. **Proteins: Structure, Function, and Bioinformatics**, v. 73, n. 4, p. 1001–1009, Dec 2008.

BORGUESAN, B.; INOSTROZA-PONTA, M.; DORN, M. **NPAS-Server: Neighbors Preferences of Amino Acids and Secondary Structures**. 2015b. Disponível em: <http://www.sbcb.inf.ufrgs.br/npas>. Acesso em: 17 fev. 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, Part A, p. 142–157, 2015a.

BORNHOLDT, Z. et al. Structural rearrangement of ebola virus vp40 begets multiple functions in the virus life cycle. **Cell**, v. 154, n. 4, p. 763–774, 2013.

BOUSSAÏD, I.; LEPAGNOT, J.; SIARRY, P. A survey on optimization metaheuristics. **Information Sciences**, Elsevier Science Inc., New York, NY, USA, v. 237, p. 82–117, jul. 2013.

BOWIE, J. U.; EISENBERG, D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and empirical guiding fitness function. **Proceedings of the National Academy of Sciences of the United States of America**, v. 91, n. 10, p. 4436–4440, 1994.

BRANDEN, C.; TOOZE, J. **Introduction to Protein Structure**. 2. ed. [S.l.]: Garland Science, 1999.

BRYANT, S. H.; ALTSCHUL, S. Statistics of sequence-structure threading. **Current Opinion In Structural Biology**, v. 5, n. 2, p. 236–244, 1995.

BRYNGELSON, J. D. et al. Funnels, pathways, and the energy landscape of protein folding: A synthesis. **Proteins: Structure, Function, and Bioinformatics**, Wiley Subscription Services, Inc., A Wiley Company, v. 21, n. 3, p. 167–195, 1995. ISSN 1097-0134.

BUCHAN, D. W. et al. Scalable web services for the PSIPRED Protein Analysis Workbench. **Nucleic Acids Research**, v. 41, n. Web Server issue, p. W349–357, Jul 2013.

CAI, Z. et al. Solution structure of BmBKTx1, a new BKCa1 channel blocker from the Chinese scorpion Buthus martensi Karsch. **Biochemistry**, v. 43, n. 13, p. 3764–3771, Apr 2004.

CHAGOT, B. et al. An unusual fold for potassium channel blockers: Nmr structure of three toxins from the scorpion opisthacanthus madagascariensis. **Biochemical Journal**, v. 388, p. 263–271, 2005.

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

CHEN, B.; HU, J. A novel clustering based niching EDA for protein folding. In: WORLD CONGRESS ON NATURE AND BIOLOGICALLY INSPIRED COMPUTING (NaBIC), 2009. **Proceedings...** Washington, D.C., EUA: IEEE Computer Society, 2009. p. 748–753.

CHENG, J. et al. The MULTICOM toolbox for protein structure prediction. **BMC Bioinformatics**, v. 13, p. 65, 2012.

CHIVIAN, D. et al. Ab initio methods. In: BOURNE, P. E.; WEISSIG, H. (Ed.). **Structural Bioinformatics**. 1. ed. New York: John Wiley and Sons, Inc., 2005. v. 44, cap. 27, p. 547–557.

CIFUENTES, G. et al. Evidence supporting the hypothesis that specifically modifying a malaria peptide to fit into HLA-DRbeta1*03 molecules induces antibody production and protection. **Vaccine**, v. 23, n. 13, p. 1579–1587, Feb 2005.

CLARKE, N. et al. Structural studies of the engrailed homeodomain. **Protein Science**, v. 3, n. 10, p. 1779–1787, Oct 1994.

COCHEZ, M.; MOU, H. Twister tries: Approximate hierarchical agglomerative clustering for average distance in linear time. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA (SIGMOD), 2015. **Proceedings...** New York, NY, USA: ACM, 2015. p. 505–517.

COMBS, S. et al. Small-molecule ligand docking into comparative models with rosetta. **Nature Protocols**, Nature Publishing Group, v. 8, n. 7, p. 1277–1298, 2013.

CONNOLLY, M. L. Solvent-accessible surfaces of proteins and nucleic acids. **Science**, American Association for the Advancement of Science, v. 221, n. 4612, p. 709–713, 1983.

CONWAY, P. et al. Relaxation of backbone bond geometry improves protein energy landscape modeling. **Protein Science**, v. 23, n. 1, p. 47–55, Jan 2014.

CRASTO, C. J.; FENG, J.-a. Sequence codes for extended conformation: A neighbor-dependent sequence analysis of loops in proteins. **Proteins: Structure, Function, and Bioinformatics**, John Wiley & Sons, Inc., v. 42, n. 3, p. 399–413, 2001.

CREIGHTON, T. E. Protein folding. **Biochemical Journal**, v. 270, p. 1–16, 1990.

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

CUSTODIO, F. L.; BARBOSA, H. J.; DARDENNE, L. E. A multiple minima genetic algorithm for protein structure prediction. **Applied Soft Computing**, v. 15, p. 88–99, 2014.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. **Journal Of The Royal Society Interface**, v. 3, n. 6, p. 139–151, 2006.

DAGGETT, V.; FERSHT, A. R. Is there a unifying mechanism for protein folding? **Trends in Biochemical Sciences**, Elsevier, v. 28, n. 1, p. 18–25, 2003.

DANG, H.-V. et al. Parallelized clustering of protein structures on cuda-enabled gpus. In: EUROMICRO INTERNATIONAL CONFERENCE, 22, 2014. **Proceedings...** Washington, D.C., EUA: IEEE Computer Society, 2014. p. 1–8.

DER, B. et al. Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. **Journal of the American Chemical Society**, v. 134, n. 1, p. 375–385, 2012.

DORN, M. **MOIRAE: A Computational Strategy to Predict 3-D Structures of Polypeptides.** Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2012.

DORN, M.; BURIOL, L.; LAMB, L. A hybrid genetic algorithm for the 3-d protein structure prediction problem using a path-relinking strategy. In: CONGRESS ON EVOLUTIONARY COMPUTATION (CEC), 2011. **Proceedings...** Washington, D.C., EUA: IEEE Computer Society, 2011. p. 2709–2716.

DORN, M.; BURIOL, L. S.; LAMB, L. C. MOIRAE: A computational strategy to extract and represent structural information from experimental protein templates. **Soft Computing**, v. 18, n. 4, p. 773–795, 2014.

DORN, M. et al. A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: CONGRESS ON EVOLUTIONARY COMPUTATION (CEC), 2013. **Proceedings...** Washington, D.C., EUA: IEEE Computer Society, 2013. p. 1233–1240.

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

DUNBRACK JR, R. L.; KARPLUS, M. Backbone-dependent rotamer library for proteins: application to side-chain prediction. **Journal Of Molecular Biology**, v. 230, n. 2, p. 543–574, 2003.

DUNBRACK, R. L.; COHEN, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. **Protein Science**, Wiley Online Library, v. 6, n. 8, p. 1661–1681, 1997.

ELOFSSON, A.; GRAND, E.; EISENBERG, D. Local moves: An efficient algorithm for simulation of protein folding. **Proteins: Structure, Function, and Bioinformatics**, v. 23, p. 73–82, 1995.

ENGH, R. A.; HUBER, R. Accurate bond and angle parameters for x-ray protein structure refinement. **Acta Crystallographica Section A: Foundations of Crystallography**, International Union of Crystallography, v. 47, n. 4, p. 392–400, 1991.

ESWAR, N. et al. Protein structure modeling with MODELLER. **Methods in Molecular Biology**, v. 426, p. 145–159, 2008.

FARAGGI, E. et al. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. **Journal of Computational Chemistry**, v. 33, n. 3, p. 259–267, Jan 2012.

FRISHMAN, D.; ARGOS, P. Knowledge-based protein secondary structure assignment. **Proteins: Structure Function And Bioinformatics**, v. 23, n. 4, p. 566–579, 1995.

GARNETT, J. et al. A high-resolution structure of the DNA-binding domain of AhrC, the arginine repressor/activator protein from Bacillus subtilis. **Acta Crystallographica Section F: Structural Biology and Crystallization Communications**, v. 63, n. Pt 11, p. 914–917, Nov 2007.

GLIBOVETS, N.; GULAYEVA, N. A review of niching genetic algorithms for multimodal function optimization. **Cybernetics and Systems Analysis**, Springer US, v. 49, n. 6, p. 815–820, 2013.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and Machine Learning**. 1. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

GOLDBERG, D. E.; DEB, K. A comparative analysis of selection schemes used in genetic algorithms. **Foundations of genetic algorithms**, v. 1, p. 69–93, 1991.

GOLDBERG, D. E.; RICHARDSON, J. Genetic algorithms with sharing for multimodal function optimization. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS ON GENETIC ALGORITHMS AND THEIR APPLICATION, 2, 1987. **Proceedings...** Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1987. p. 41–49.

GONÇALVES, J. F.; RESENDE, M. G. Biased random-key genetic algorithms for combinatorial optimization. **Journal of Heuristics**, Springer US, v. 17, n. 5, p. 487–525, 2011.

GORDON, D. B.; MARSHALL, S. A.; MAYOT, S. L. Energy functions for protein design. **Current Opinion in Structural Biology**, Elsevier, v. 9, n. 4, p. 509–513, 1999.

GRONT, D.; KOLINSKI, A. Hcpm—program for hierarchical clustering of protein models. **Bioinformatics**, v. 21, n. 14, p. 3179–3180, 2005.

GRONT, D. et al. Generalized fragment picking in rosetta: Design, protocols and applications. **PLoS ONE**, Public Library of Science, v. 6, n. 8, p. e23294, 08 2011.

GUNSTEREN, W. van; BERENDSEN, H. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. **Angewandte Chemie International**, v. 29, n. 9, p. 992–1023, 1990.

GUYEUX, C. et al. Is protein folding problem really a np-complete one? first investigations. **Journal of Bioinformatics and Computational Biology**, World Scientific, v. 12, n. 01, p. 1350017, 2014.

HARIK, G. Finding multimodal solutions using restricted tournament selection. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 6, 1995. **Proceedings...** Pittsburgh, PA, USA: Morgan Kaufmann, 1995. p. 24–31.

HEFFERNAN, R. et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. **Scientific Reports**, Macmillan Publishers Limited, v. 5, p. 11476, Jun 2015.

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

HILL, C. et al. Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization. **Science**, v. 251, n. 5000, p. 1481–1485, Mar 1991.

HOLLAND, J. **Adaptation in Natural and Artificial Systems**. 1. ed. Ann Arbor, MI, USA: The University of Michigan Press, 1975.

HOLLAND, J. H. **Adaptation in Natural and Artificial Systems**. 2. ed. Cambridge, MA, USA: MIT Press, 1992.

HOQUE, M.; CHETTY, M.; DOOLEY, L. A guided genetic algorithm for protein folding prediction using 3d hydrophobic-hydrophilic model. In: CONGRESS ON EVOLUTIONARY COMPUTATION, (CEC), 2006". **Proceedings...** Washington, D.C., EUA: IEEE Computer Society, 2006. p. 2339–2346.

HOQUE, M. T.; CHETTY, M.; SATTAR, A. Genetic algorithm in *ab initio* protein structure prediction using low resolution model: A review. In: **Biomedical Data and Applications**. [S.l.: s.n.], 2009. v. 224, p. 317–342.

HOUCK, C. R.; JOINES, J. A.; KAY, M. G. Comparison of genetic algorithms, random restart and two-opt switching for solving large location-allocation problems. **Computers & Operations Research**, Elsevier, v. 23, n. 6, p. 587–596, 1996.

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.

HUARD, F. P.; DEANE, C. M.; WOOD, G. R. Modelling sequential protein folding under kinetic control. **Bioinformatics**, v. 22, n. 14, p. e203–e210, 2006.

HUTCHINSON, E.; THORNTON, J. Promotif: A program to identify and analyze structural motifs in proteins. **Protein Science**, v. 5, n. 2, p. 212–220, 1996.

ISLAM, M.; CHETTY, M. Clustered memetic algorithm with local heuristics for ab initio protein structure prediction. **IEEE Transactions on Evolutionary Computation**, v. 17, n. 4, p. 558–576, 2013.

JAEGER, D. et al. pyGCluster, a novel hierarchical clustering approach. **Bioinformatics**, v. 30, n. 6, p. 896–898, 2014.

JAMROZ, M.; KOLINSKI, A. Clusco: clustering and comparison of protein models. **BMC Bioinformatics**, BioMed Central Ltd, v. 14, n. 1, p. 62, 2013.

JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, Springer, v. 32, n. 3, p. 241–254, 1967.

JONES, D. T. Protein secondary structure prediction based on position-specific scoring matrices. **Journal of Molecular Biology**, v. 292, n. 2, p. 195–202, Sep 1999.

JONES, E. et al. **SciPy: Open source scientific tools for Python**. 2001. Disponível em: <http://www.scipy.org/>. Acesso em: 17 fev. 2016.

JONG, D.; ALAN, K. **An Analysis of the Behavior of a Class of Genetic Adaptive Systems.** Tese (Doutorado) — University of Michigan, Ann Arbor, MI, USA, 1975.

JOO, K. et al. Template based protein structure modeling by global optimization in casp11. **Proteins: Structure, Function, and Bioinformatics**, p. n/a–n/a, 2015, in press.

JOO, K. et al. Protein structure modeling for CASP10 by multiple layers of global optimization. **Proteins: Structure, Function, and Bioinformatics**, v. 82 Suppl 2, p. 188–195, Feb 2014.

KABAT, E. A.; WU, T. T. The influence of nearest-neighboring amino acid residues on aspects of secondary structure of proteins. Attempts to locate α-helices and β-sheets. **Biopolymers**, v. 12, n. 4, p. 751–774, Apr 1973.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, Wiley Subscription Services, Inc., A Wiley Company, v. 22, n. 12, 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. 2577–637, 1983.

KENNEDY, J. **The particle swarm: social adaptation of knowledge**. New York: IEEE Press, 2003. 303 p.

KIM, D. E.; CHIVIAN, D.; BAKER, D. Protein structure prediction and analysis using the Robetta server. **Nucleic Acids Research**, v. 32, n. Web Server issue, p. W526–531, Jul 2004.

KINCH, L. N. et al. Casp 11 target classification. **Proteins: Structure, Function, and Bioinformatics**, p. n/a–n/a, 2016, in press.

KORTEMME, T.; MOROZOV, A.; BAKER, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. **Journal of Molecular Biology**, Elsevier, v. 326, n. 4, p. 1239–1259, 2003.

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

KRYSHTAFOVYCH, A. et al. Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th critical assessment of techniques for protein structure prediction experiment CASP10. **Proteins: Structure Function And Bioinformatics**, v. 82, p. 26–42, 2014a.

KUHLMAN, B.; BAKER, D. Native protein sequences are close to optimal for their structures. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 97, n. 19, p. 10383–10388, 2000.

LARRAÑAGA, P. et al. Machine learning in bioinformatics. **Briefings In Bioinformatics**, v. 17, n. 1, p. 86–112, 2006.

LASKOWSKI, R. A. et al. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. **Journal Of Biomolecular NMR**, Kluwer Academic Publishers, v. 8, n. 4, p. 477–486, 1996.

LASKOWSKI, R. A.; WATSON, J. D.; THORNTON, J. M. From protein structure to biochemical function? **Journal of Structural and Functional Genomics**, Springer, v. 4, n. 2-3, p. 167–177, 2003.

LATHROP, R. The protein threading problem with sequence amino acid interaction preferences in np-complete. **Protein Engineering**, v. 7, n. 9, p. 1059–1068, 1994.

LAZARIDIS, T.; KARPLUS, M. Effective energy functions for protein structure prediction. **Current Opinion in Structural Biology**, Elsevier, v. 10, n. 2, p. 139–145, 2000.

LEAVER-FAY, A. et al. Scientific benchmarks for guiding macromolecular energy function improvement. **Methods Enzymol**, NIH Public Access, v. 523, p. 109, 2013.

LEAVER-FAY, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. **Methods Enzymol**, v. 487, p. 545–574, 2011.

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

LENO, I. J. et al. An elitist strategy genetic algorithm for integrated layout design. **The International Journal of Advanced Manufacturing Technology**, Springer-Verlag, v. 66, n. 9-12, p. 1573–1589, 2013.

LESK, A. M. **Introduction to Bioinformatics**. 2. ed. Oxford, NY, USA: Oxford University Press, 2005.

LESK, A. M. **Introduction to Protein Science**. 2. ed. New York: Oxford University Press, 2010. 455 p.

LEVINTHAL, C. Are there pathways for protein folding? **Journal de Chimie Physique et de Physico-Chimie Biologique**, v. 65, n. 1, p. 44–45, 1968.

LI, J. et al. The MULTICOM protein tertiary structure prediction system. **Methods in Molecular Biology**, v. 1137, p. 29–41, 2014.

LILJAS, A. et al. **Textbook of structural biology**. Singapore: World Scientific Printers, 2001. 572 p.

LIU, J. et al. Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. **Nature Structural Biology**, v. 4, n. 11, p. 896–899, Nov 1997.

LODISH, H. et al. **Molecular Cell Biology**. 5. ed. New York, USA: Scientific American Books, W.H. Freeman, 1990. 970 p.

LOVELL, S. C. et al. Structure validation by cα geometry: ϕ, ψ and Cβ deviation. **Proteins: Structure Function And Bioinformatics**, Wiley Subscription Services, Inc., v. 50, n. 3, p. 437–450, 2003.

LUKE, S. **Essentials of metaheuristics**. 1. ed. [S.l.]: Lulu, 2009. 227 p.

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.

MABROUK, M. et al. RBO Aleph: leveraging novel information sources for protein structure prediction. **Nucleic Acids Research**, v. 43, n. W1, p. W343–348, Jul 2015.

MAIOROV, V. N.; CRIPPEN, G. M. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. **Journal of molecular biology**, Elsevier, v. 235, n. 2, p. 625–634, 1994.

MANNING, T.; SLEATOR, R. D.; WALSH, P. Naturally selecting solutions: the use of genetic algorithms in bioinformatics. **Bioengineered**, Taylor & Francis, v. 4, n. 5, p. 266–278, 2013.

MARTÌ-RENOM, M. et al. Comparative protein structure modelling of genes and genomes. **Annual Review of Biophysics and Biomolecular Structure**, v. 29, n. 16, p. 291–235, 2000.

MCPHERSON, A. Introduction to protein crystallization. **Methods**, v. 34, n. 3, p. 254–265, 2004.

MENGSHOEL, O. J.; GOLDBERG, D. E. The crowding approach to niching in genetic algorithms. **Evolutionary computation**, MIT Press, v. 16, n. 3, p. 315–354, 2008.

MILBURN, D.; LASKOWSKI, R. A.; THORNTON, J. M. Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. **Protein engineering**, Oxford Univ Press, v. 11, n. 10, p. 855–859, 1998.

MILLMAN, K. J.; AIVAZIS, M. Python for scientists and engineers. **Computing in Science Engineering**, v. 13, n. 2, p. 9–12, March 2011.

MITCHELL, M. **An Introduction to Genetic Algorithms**. 5. ed. Cambridge: MIT Press, 1999. 158 p.

MOELBERT, S.; EMBERLY, E.; TANG, C. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. **Protein Science**, Cold Spring Harbor Laboratory Press, v. 13, n. 3, p. 752–762, 2004.

MORIZE, I. et al. Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 A resolution. **Journal of Molecular Biology**, v. 194, n. 4, p. 725–739, Apr 1987.

MOULT, J. A Decade of CASP: progress, bottlenecks an prognosis in protein structure prediction. **Current Opinion In Structural Biology**, v. 15, n. 3, p. 285–289, 2005.

MUCHERINO, A.; SEREF, O. Modeling and solving real-life global optimization problems with meta-heuristic methods. In: **Advances in Modeling Agricultural Systems**. [S.l.]: Springer US, 2009, (Springer Optimization and Its Applications, v. 25). p. 403–419.

MÜLLNER, D. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. **Journal of Statistical Software**, v. 53, n. 9, p. 1–18, 2013.

NEIDIGH, J.; FESINMEYER, R.; ANDERSEN, N. Designing a 20-residue protein. **Nature Structural & Molecular Biology**, v. 9, p. 425–430, 2002.

NERIA, E.; FISCHER, S.; KARPLUS, M. Simulation of activation free energies in molecular systems. **The Journal of Chemical Physics**, AIP Publishing, v. 105, n. 5, p. 1902–1921, 1996.

NGO, J.; MARKS, J.; KARPLUS, M. The protein folding problem and tertiary structure prediction. In: JR, K. M.; GRAND, S. (Ed.). **Computational complexity, protein structure prediction and the Levinthal Paradox**. Boston, USA: Birkhauser, 1997. p. 435–508.

NOWICKA, U. et al. DNA-damage-inducible 1 protein (Ddi1) contains an uncharacteristic ubiquitin-like domain that binds ubiquitin. **Structure**, v. 23, n. 3, p. 542–557, Mar 2015.

NUGENT, T.; COZZETTO, D.; JONES, D. T. Evaluation of predictions in the casp10 model refinement category. **Proteins: Structure Function And Bioinformatics**, v. 82, p. 98–111, 2014.

OLIVEIRA, S. H. P. de; SHI, J.; DEANE, C. M. Building a better fragment library for *De Novo* protein structure prediction. **PLoS ONE**, Public Library of Science, v. 10, n. 4, p. e0123998, 04 2015.

O'MEARA, M. et al. A combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. **Journal of Chemical Theory and Computation**, ACS Publications, 2015.

OSGUTHORPE, D. J. Ab initio protein folding. **Current Opinion In Structural Biology**, v. 10, n. 2, p. 146–152, 2000.

PARK, S. A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. **Genome Informatics**, v. 16, n. 2, p. 104–113, 2005.

PASTOR, M. et al. Combinatorial approaches: A new tool to search for highly structured beta-hairpin peptides. **Proceedings of the National Academy of Sciences**, v. 99, n. 2, p. 614–619, 2002.

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.

PEDERSEN, J. T.; MOULT, J. Genetic algorithms for protein structure prediction. **Current Opinion In Structural Biology**, v. 6, n. 2, p. 227–231, 1996.

PENG, J.; XU, J. Low-homology protein threading. **Bioinformatics**, v. 26, n. 12, p. 294–300, Jun 2010.

PENG, J.; XU, J. RaptorX: exploiting structure information for protein alignment by statistical inference. **Proteins**, v. 79 Suppl 10, p. 161–171, 2011.

PETSKO, G. A.; RINGE, D. From sequence to structure. In: **Protein structure and function**. 1. ed. [S.l.]: New Science Press, 2004. cap. 1, p. 1–49.

PRUITT, K. et al. The reference sequence (refseq) database. In: **The NCBI Handbook, Bethesda (MD)**. [S.l.]: The NCBI Handbook, 2002. cap. 18.

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

RELIGA, T. et al. The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain. **Proceedings of the National Academy of Sciences**, v. 104, n. 22, p. 9272–9277, May 2007.

REVA, B. A.; FINKELSTEIN, A. V.; SKOLNICK, J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 å? **Folding and Design**, Elsevier, v. 3, n. 2, p. 141–147, 1998.

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.

ROUX, B.; SIMONSON, T. Implicit solvent models. **Biophysical chemistry**, Elsevier, v. 78, n. 1, p. 1–20, 1999.

ROY, R.; PARMEE, I. Adaptive restricted tournament selection for the identification of multiple sub-optima in a multi-modal function. In: FOGARTY, T. (Ed.). **Evolutionary Computing**. [S.l.]: Springer Berlin Heidelberg, 1996, (Lecture Notes in Computer Science, v. 1143). p. 236–256.

SAFE, M. et al. On stopping criteria for genetic algorithms. In: **Advances in Artificial Intelligence–SBIA 2004**. [S.l.]: Springer, 2004. p. 405–413.

SÁNCHEZ, R.; SALI, A. Advances in comparative protein-structure modeling. **Current Opinion In Structural Biology**, v. 7, n. 2, p. 206–214, 1997.

SCHEEF, E. D.; FINK, J. L. Fundamentals of protein structure. In: BOURNE, P. E.; WEISSIG, H. (Ed.). **Structural Bioinformatics**. 1. ed. New York: John Wiley and Sons, Inc., 2005. v. 44, cap. 2, p. 15–39.

SHAH, P. et al. Full-sequence computational design and solution structure of a thermostable protein variant. **Journal of Molecular Biology**, v. 372, n. 1, p. 1–6, Sep 2007.

SODING, J. Protein homology detection by HMM-HMM comparison. **Bioinformatics**, v. 21, n. 7, p. 951–960, Apr 2005.

SONG, Y. et al. High-resolution comparative modeling with RosettaCM. **Structure**, v. 21, n. 10, p. 1735–1742, Oct 2013.

SRINIVAS, M.; PATNAIK, L. M. Genetic algorithms: A survey. **Computer**, IEEE, v. 27, n. 6, p. 17–26, 1994.

SRINIVASAN, R.; ROSE, G. Ab initio prediction of protein structure using linus. **Proteins-Structure Function And Bioinformatics**, v. 47, n. 4, p. 489–495, 2002.

STAROVASNIK, M. A.; BRAISTED, A. C.; WELLS, J. A. Structural mimicry of a native protein by a minimized binding domain. **Proceedings of the National Academy of Sciences**, v. 94, n. 19, p. 10080–10085, Sep 1997.

TEETER, M. M. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. **Proceedings of the National Academy of Sciences**, v. 81, n. 19, p. 6014–6018, Oct 1984.

TENDULKAR, A. V. et al. Clustering of protein structural fragments reveals modular building block approach of nature. **Journal of molecular biology**, Elsevier, v. 338, n. 3, p. 611–629, 2004.

TING, D. et al. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. **PLOS Computational Biology**, Public Library of Science, v. 6, n. 4, p. e1000763, 2010.

TING, D. et al. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. **PLOS Computational Biology**, v. 6, n. 4, p. e1000763, Apr 2010.

TOUW, W. G. et al. A series of PDB-related databanks for everyday needs. **Nucleic Acids Research**, v. 43, n. Database issue, p. D364–368, Jan 2015.

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

UNGER, R. The genetic algorithm approach to protein structure prediction. In: **Applications of Evolutionary Computation in Chemistry**. [S.l.]: Springer, 2004. p. 153–175.

VITA, C. et al. Rational engineering of a miniprotein that reproduces the core of the cd4 site interacting with hiv-1 envelope glycoprotein. **Proceedings of the National Academy of Sciences**, v. 96, p. 13091–13096, 1999.

WEDEMEYER, W. J.; BAKER, D. Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 53, n. 2, p. 262–272, 2003.

WILLIAMS, R. W. et al. Secondary structure predictions and medium range interactions. **Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology**, Elsevier, v. 916, n. 2, p. 200–204, 1987.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. **Transactions on Evolutionary Computation**, IEEE Press, Piscataway, NJ, USA, v. 1, n. 1, p. 67–82, abr. 1997.

WOLYNES, P. G. Energy landscapes and solved protein–folding problems. **Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences**, The Royal Society, v. 363, n. 1827, p. 453–467, 2005.

WONG, K.-C.; LEUNG, K.-S.; WONG, M.-H. Protein structure prediction on a lattice model via multimodal optimization techniques. In: ANNUAL CONFERENCE ON GENETIC AND EVOLUTIONARY COMPUTATION, 12, 2010. **Proceedings...** New York, NY, USA: ACM, 2010. (GECCO '10), p. 155–162.

WU, S.; ZHANG, Y. LOMETS: a local meta-threading-server for protein structure prediction. **Nucleic Acids Research**, v. 35, n. 10, p. 3375–3382, 2007.

WU, S.; ZHANG, Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. **Bioinformatics**, v. 24, n. 7, p. 924–931, Apr 2008.

WU, S.; ZHANG, Y. Recognizing protein substructure similarity using segmental threading. **Structure**, v. 18, n. 7, p. 858–867, Jul 2010.

WUTHRICH, K. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. **Science**, American Association for the Advancement of Science, v. 243, n. 4887, p. 45–50, 1989.

XIA, X.; XIE, Z. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. **Molecular Biology and Evolution**, v. 19, n. 1, p. 58–67, 2002.

XIE, H.; ZHANG, M. Parent selection pressure auto-tuning for tournament selection in genetic programming. **IEEE Transactions on Evolutionary Computation**, v. 17, n. 1, p. 1–19, Feb 2013.

XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. **Proteins**, v. 80, n. 7, p. 1715–1735, Jul 2012.

YAMANO, A.; HEO, N.; TEETER, M. Crystal structure of Ser-22/Ile-25 form crambin confirms solvent, side chain substate correlations. **Journal of Biological Chemistry**, v. 272, n. 15, p. 9597–9600, Apr 1997.

YANG, Y. et al. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. **Bioinformatics**, v. 27, n. 15, p. 2076–2082, Aug 2011.

YASEEN, A.; LI, Y. Context-based features enhance protein secondary structure prediction accuracy. **Journal of Chemical Information and Modeling**, v. 54, n. 3, p. 992–1002, 2014.

YASEEN, A.; LI, Y. Template-based c8-scorpion: a protein 8-state secondary structure prediction method using structural information and context-based features. **BMC Bioinformatics**, BioMed Central, v. 15, n. 8, 2014.

ZHANG, C.-T.; ZHANG, R. A refined accuracy index to evaluate algorithms of protein secondary structure prediction. **Proteins: Structure, Function, and Bioinformatics**, John Wiley & Sons, Inc., v. 43, n. 4, p. 520–522, 2001.

ZHANG, J.; LIANG, Y.; ZHANG, Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. **Structure**, v. 19, n. 12, p. 1784–1795, Dec 2011.

ZHANG, W. et al. Integration of quark and i-tasser for ab initio protein structure prediction in casp11. **Proteins: Structure, Function, and Bioinformatics**, p. n/a–n/a, 2015, in press.

ZHANG, Y. Template-based modeling and free modeling by i-tasser in casp7. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 69, n. S8, p. 108–117, 2007.

ZHANG, Y. I-tasser server for protein 3d structure prediction. **BMC Bioinformatics**, BioMed Central Ltd, v. 9, n. 1, p. 40, 2008.

ZHANG, Y.; SAGUI, C. Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI. **Journal of Molecular Graphics and Modelling**, v. 55, p. 72 – 84, 2015.

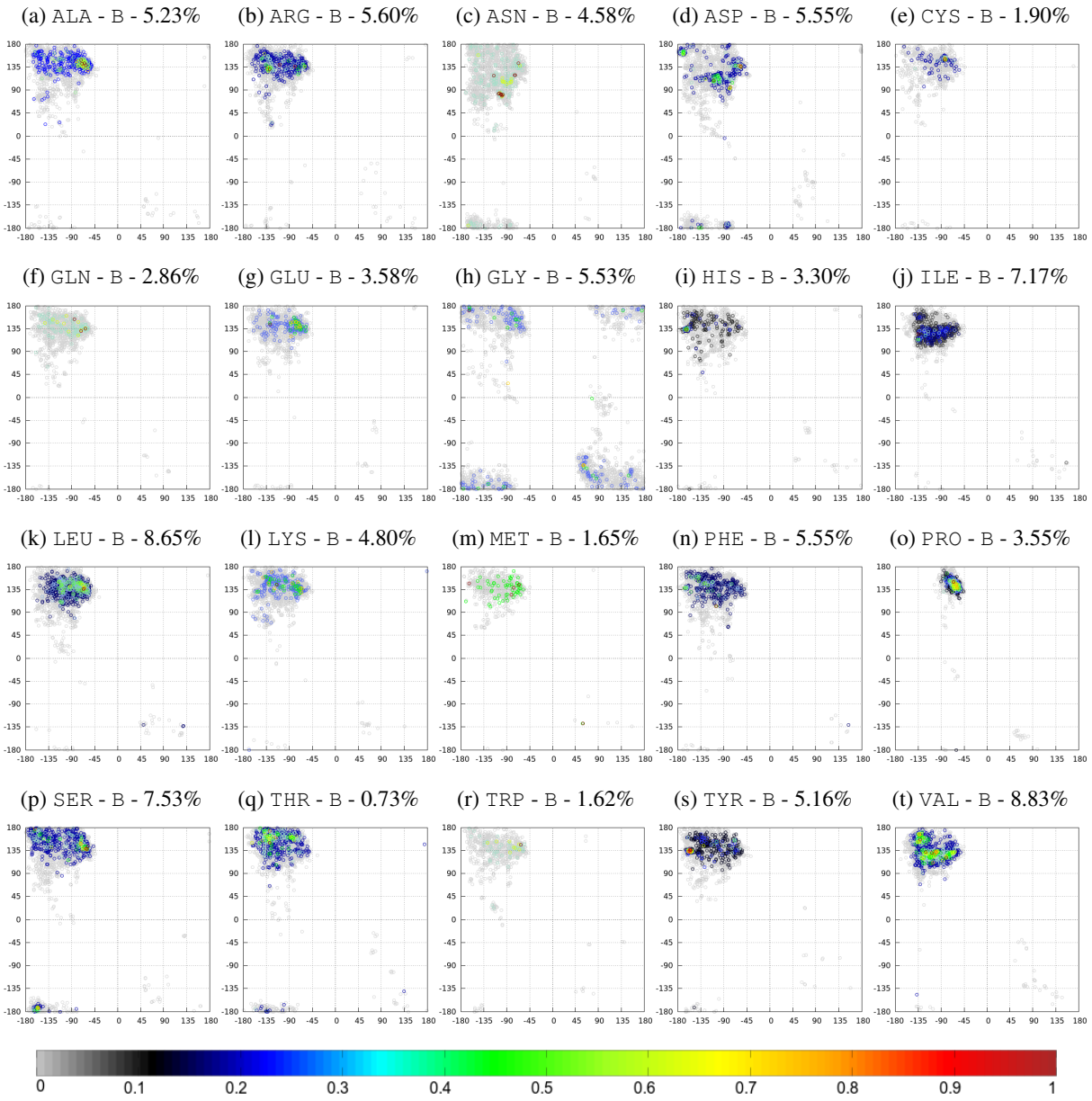
ZHANG, Y.; SKOLNICK, J. Scoring function for automated assessment of protein structure template quality. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 57, n. 4, p. 702–710, 2004.

ZHANG, Y.; SKOLNICK, J. SPICKER: a clustering approach to identify near-native protein folds. **Journal of Computational Chemistry**, v. 25, n. 6, p. 865–871, Apr 2004.

ZHAO, F.; PENG, J.; XU, J. Fragment-free approach to protein folding using conditional neural fields. **Bioinformatics**, v. 26, n. 12, p. i310–317, Jun 2010.

APÊNDICE A - PREFERÊNCIA CONFORMACIONAL DE TODA BASE APL1

Figura 9.1: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária β-Bridge (B) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



Fonte: Adaptado de Borguesan et al. (2015a).

Figura 9.2: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária de regiões desordenadas (C) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran

(a) ALA - C - 0.60% (b) ARG - C - 4.64% (c) ASN - C - 5.08% (d) ASP - C - 6.91% (e) CYS - C - 1.16%

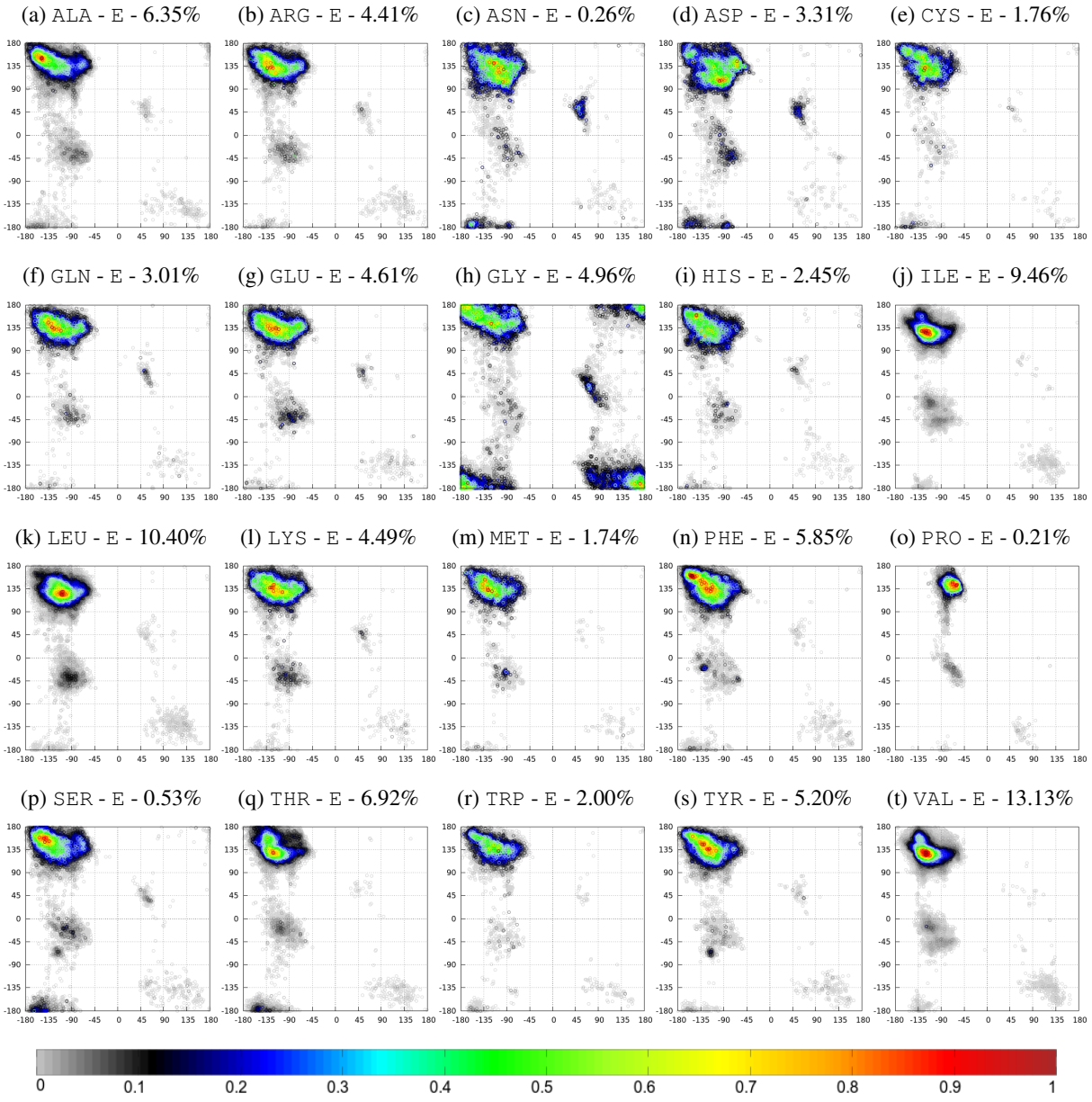
(f) GLN - C - 3.16% (g) GLU - C - 4.85% (h) GLY - C - 12.46% (i) HIS - C - 2.51% (j) ILE - C - 4.09%

(k) LEU - C - 6.72% (l) LYS - C - 5.21% (m) MET - C - 1.26% (n) PHE - C - 3.21% (o) PRO - C - 9.24%

(p) SER - C - 7.35% (q) THR - C - 6.94% (r) TRP - C - 1.13% (s) TYR - C - 2.88% (t) VAL - C - 5.19%

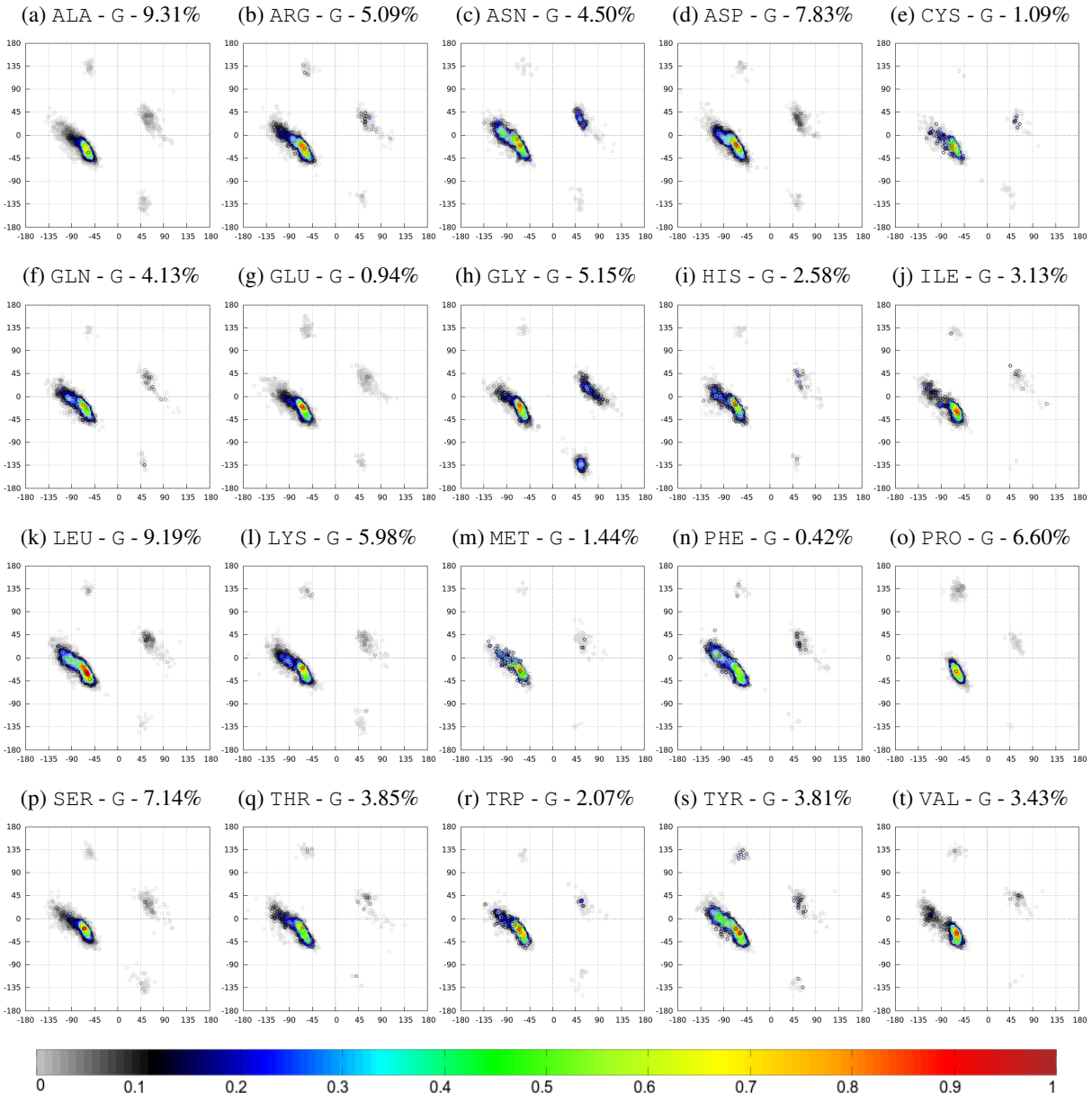
Fonte: Adaptado de Borguesan et al. (2015a).

Figura 9.3: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária folha β (E) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



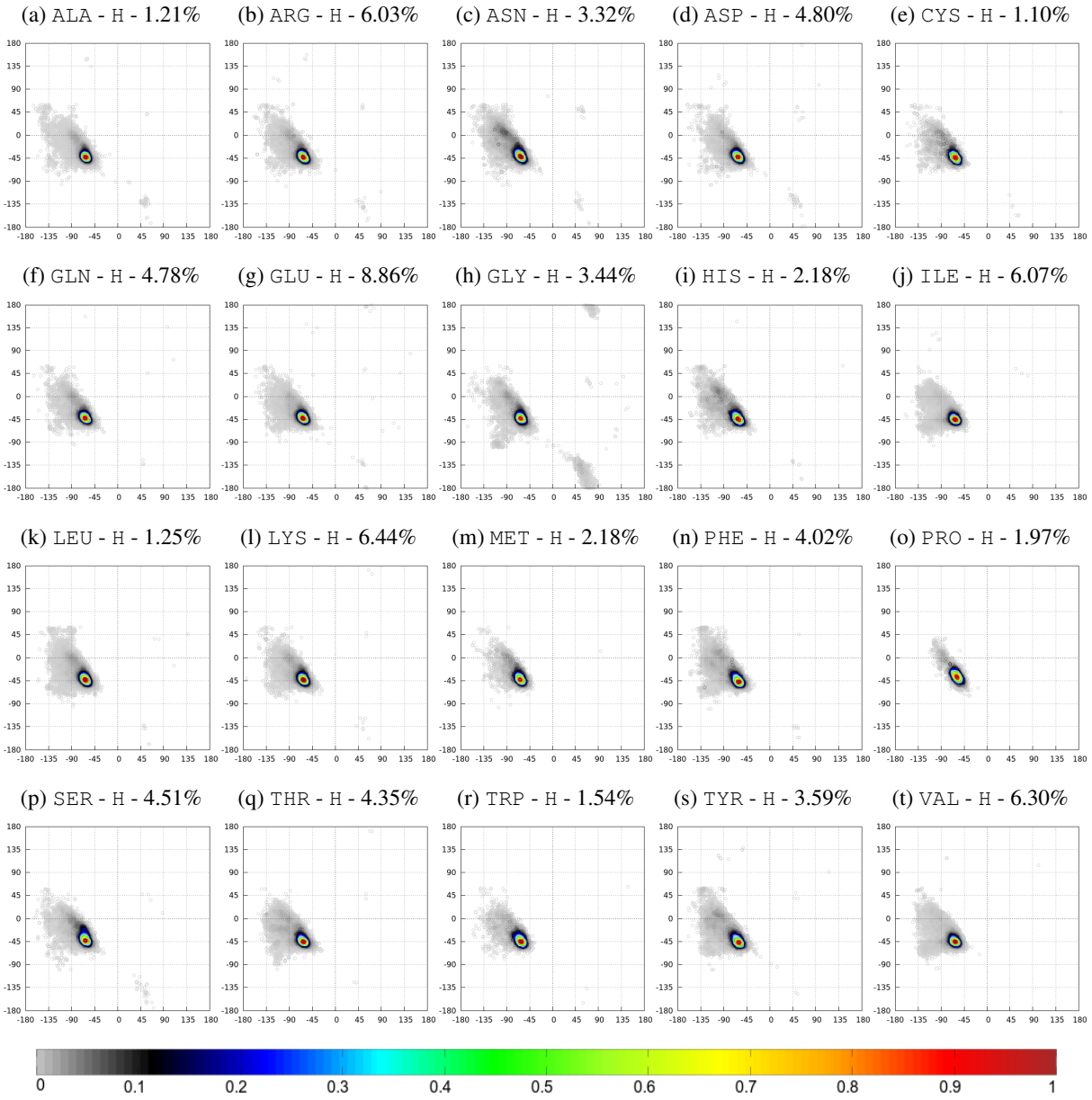
Fonte: Adaptado de Borguesan et al. (2015a).

Figura 9.4: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária hélice 3_10 (G) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



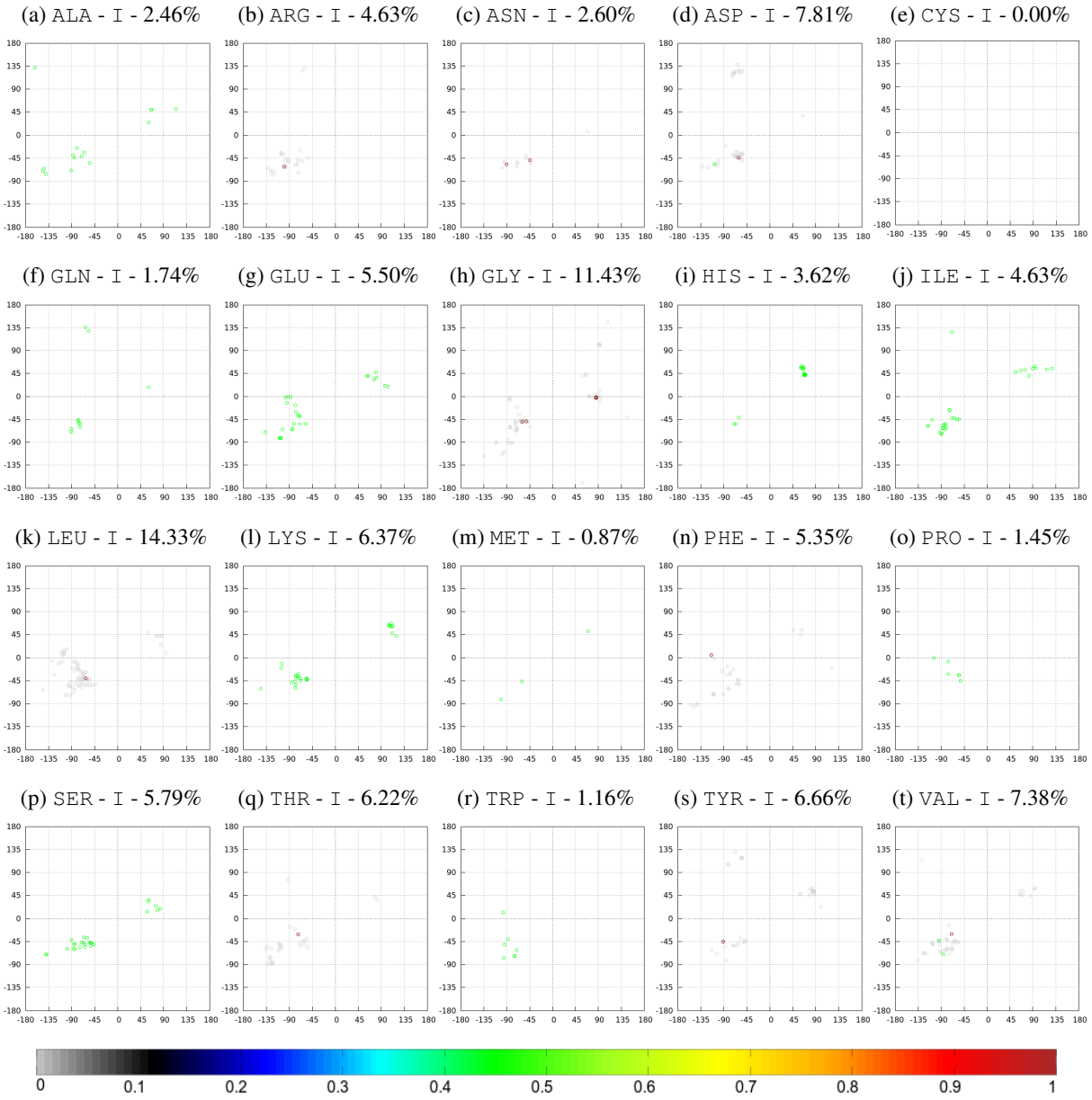
Fonte: Adaptado de Borguesan et al. (2015a).

Figura 9.5: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária hélice α (H) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



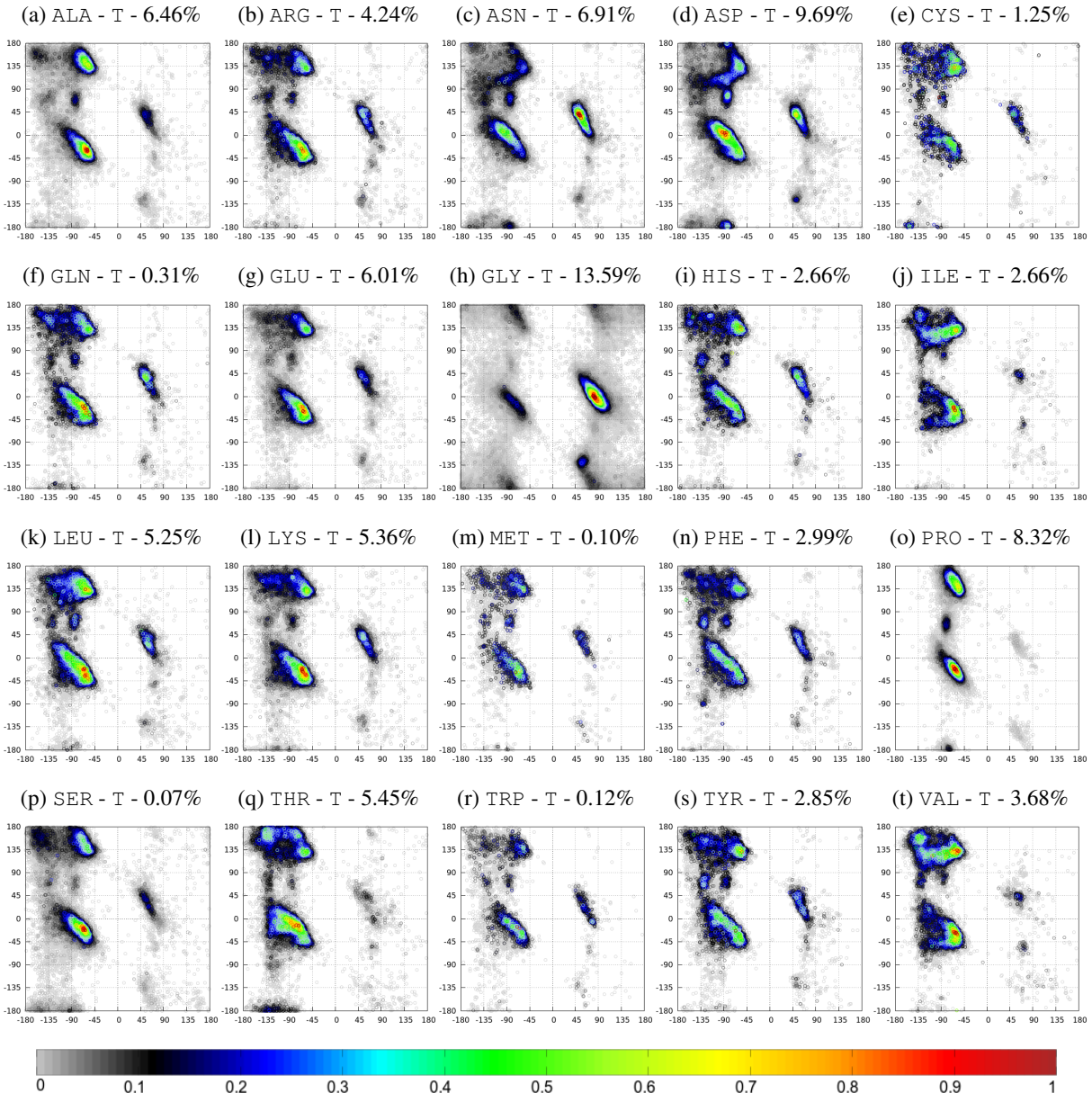
Fonte: Adaptado de Borguesan et al. (2015a).

Figura 9.6: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária hélice π (I) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



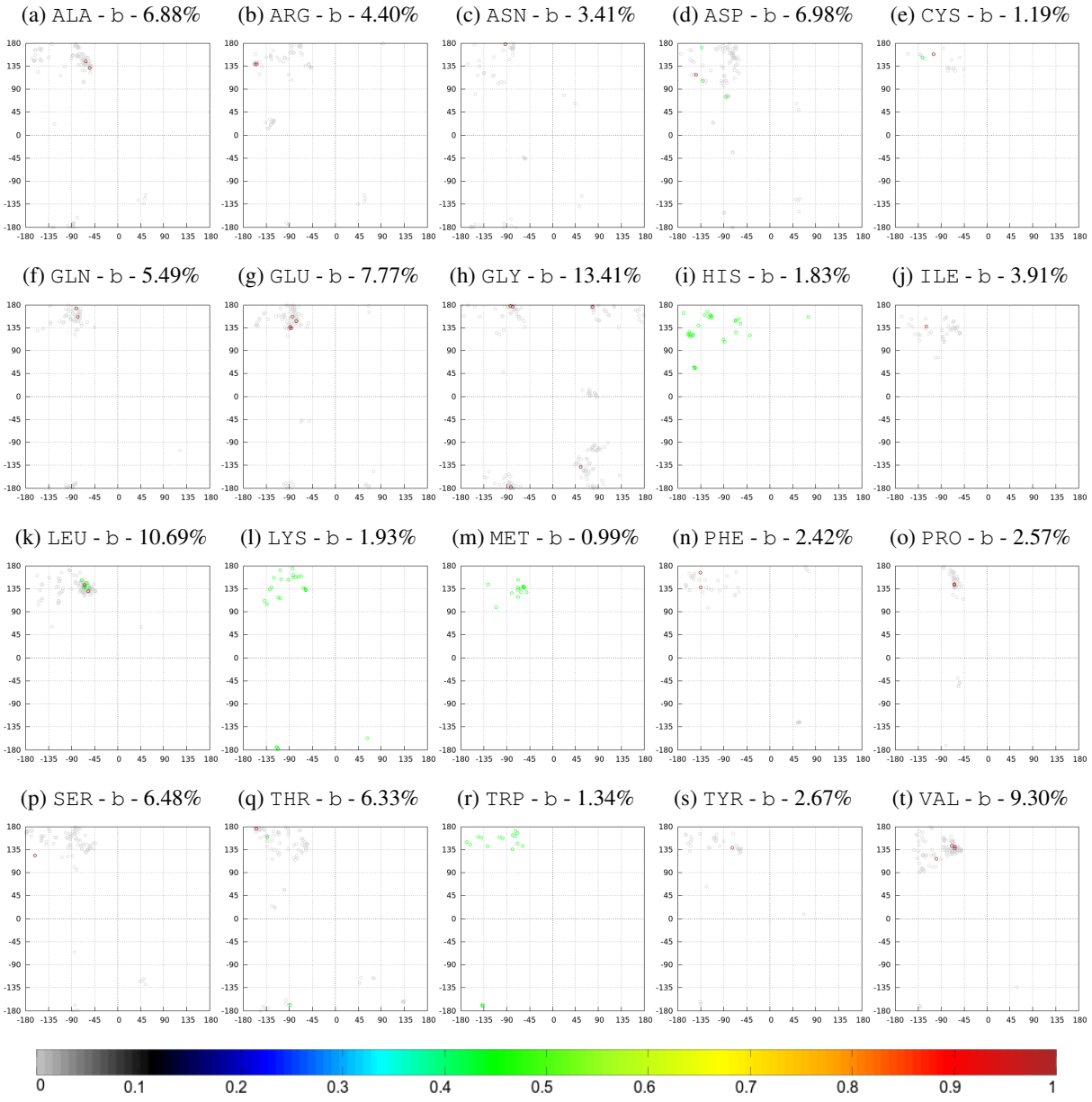
Fonte: Adaptado de Borguesan et al. (2015a).

Figura 9.7: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária volta (T) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran



Fonte: Adaptado de Borguesan et al. (2015a).

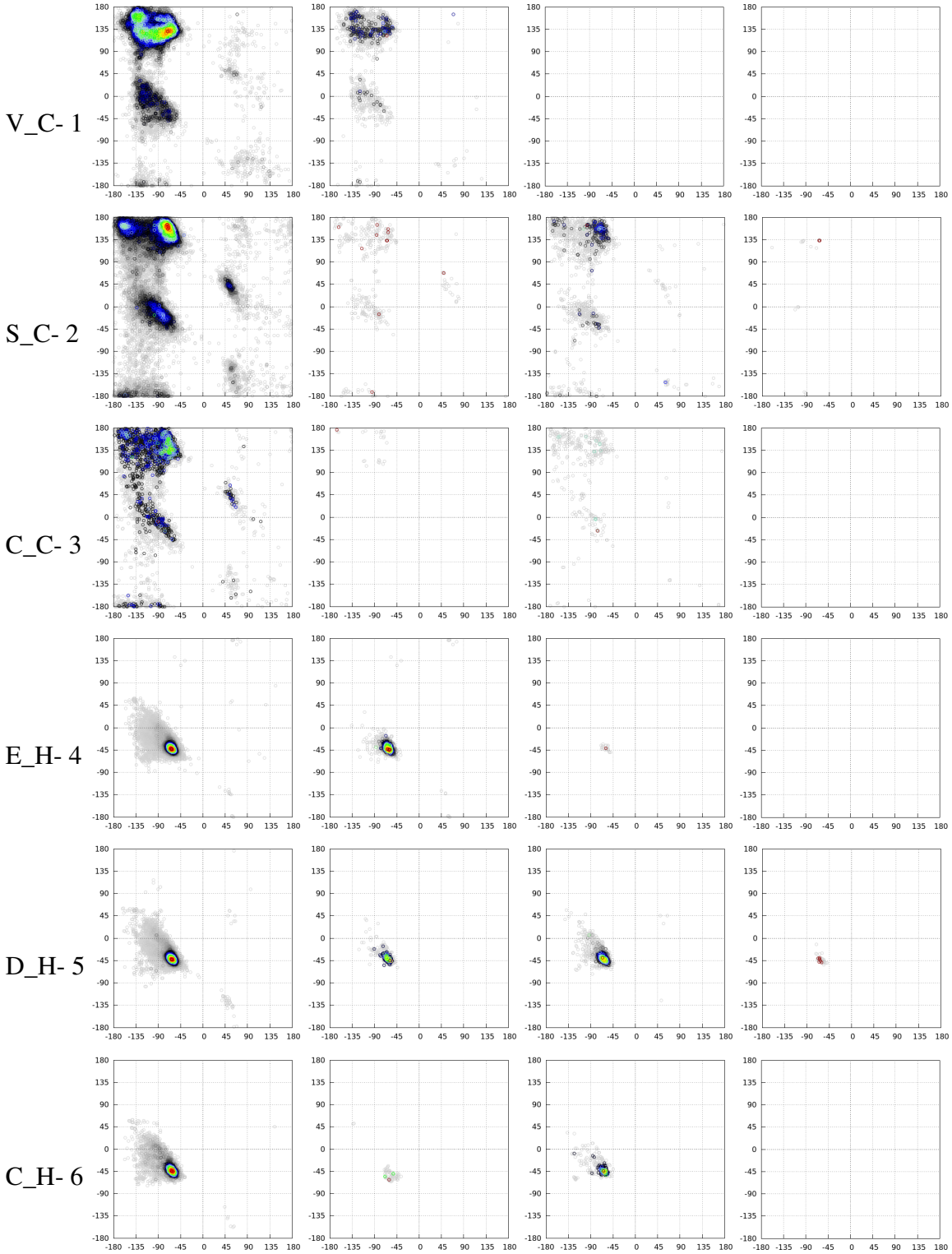
Figura 9.8: Gráficos de Ramachandran para os 20 aminoácidos-padrão. Os gráficos de Ramachandran representam a preferência conformacional do aminoácido para estrutura secundária ponte isolada (b) atribuído pelo STRIDE. A cor vermelho escuro representa a região de maior densidade do gráfico de Ramachandran.



Fonte: Adaptado de Borguesan et al. (2015a).

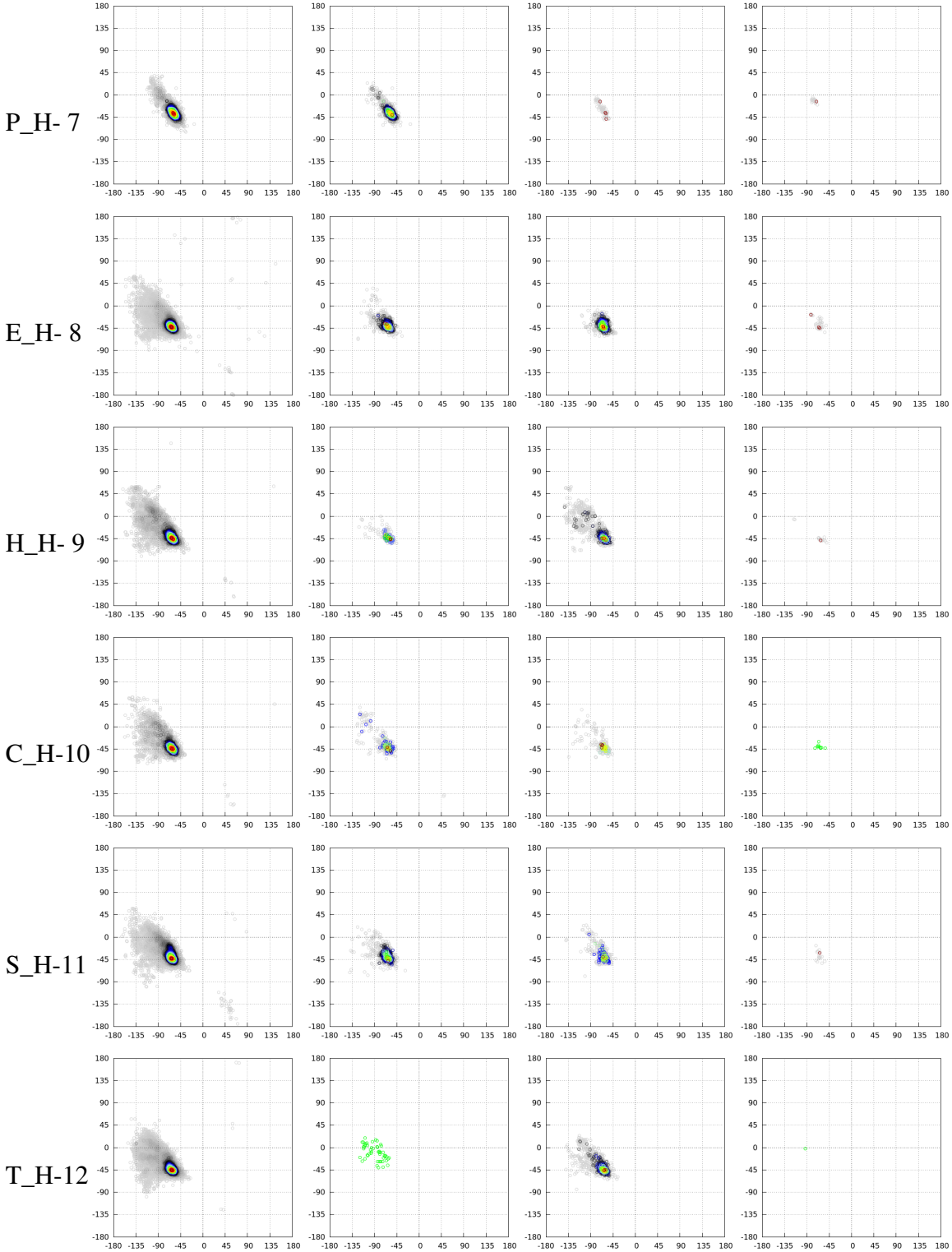
APÊNDICE B - PREFERÊNCIA CONFORMACIONAL DE TODAS AS APL PARA ESTRUTURA DE CÓDIGO PDB 1ACW

Figura 9.9: Comparativo entre os arquivos de APL1, APL2-Esq., APL2-Dir. e APL3 para a estrutura com o código no PDB de 1ACW



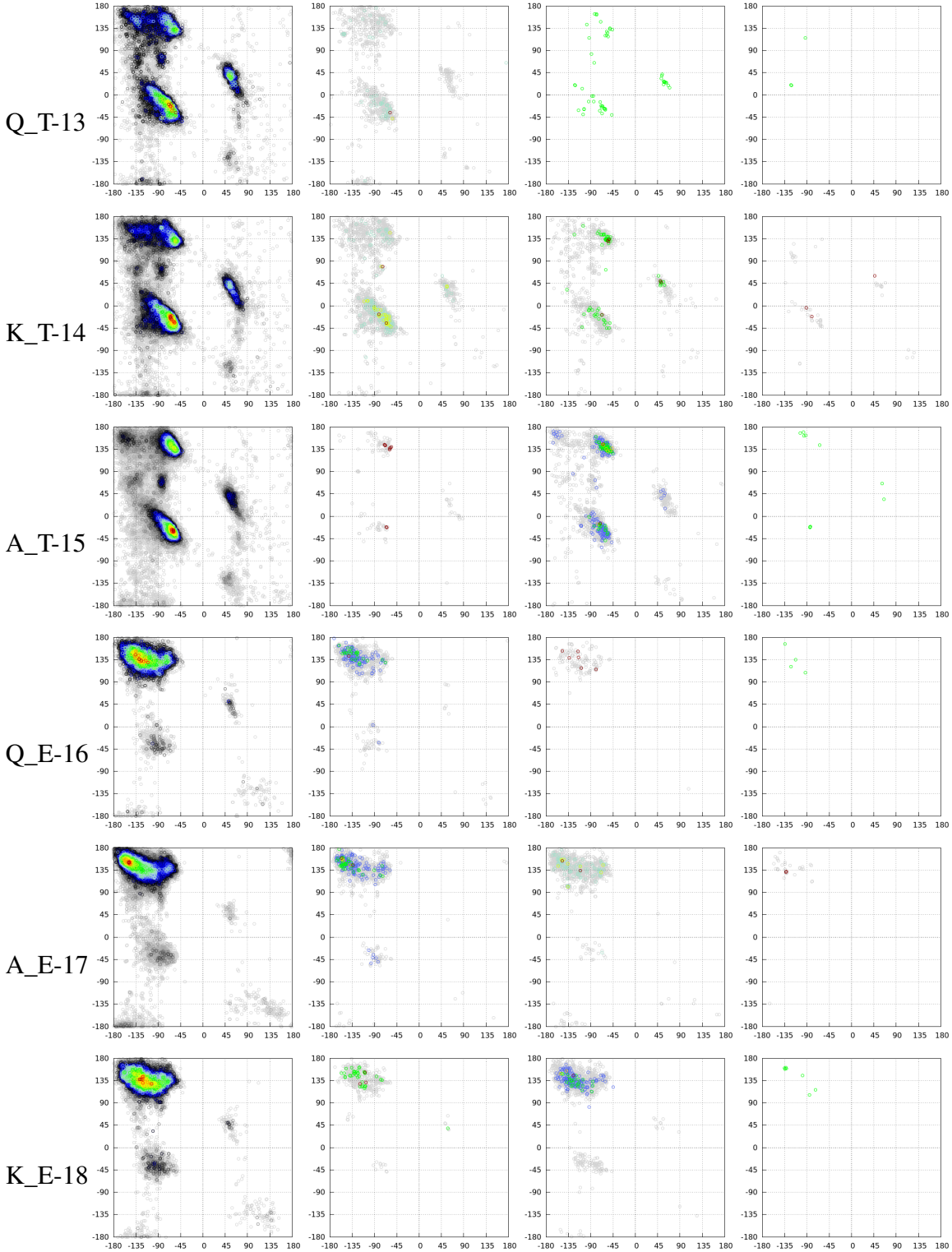
Fonte: Borguesan et al. (2015b).

Figura 9.9: Comparativo entre os arquivos de APL1, APL2-Esq., APL2-Dir. e APL3 para a estrutura com o código no PDB de 1ACW



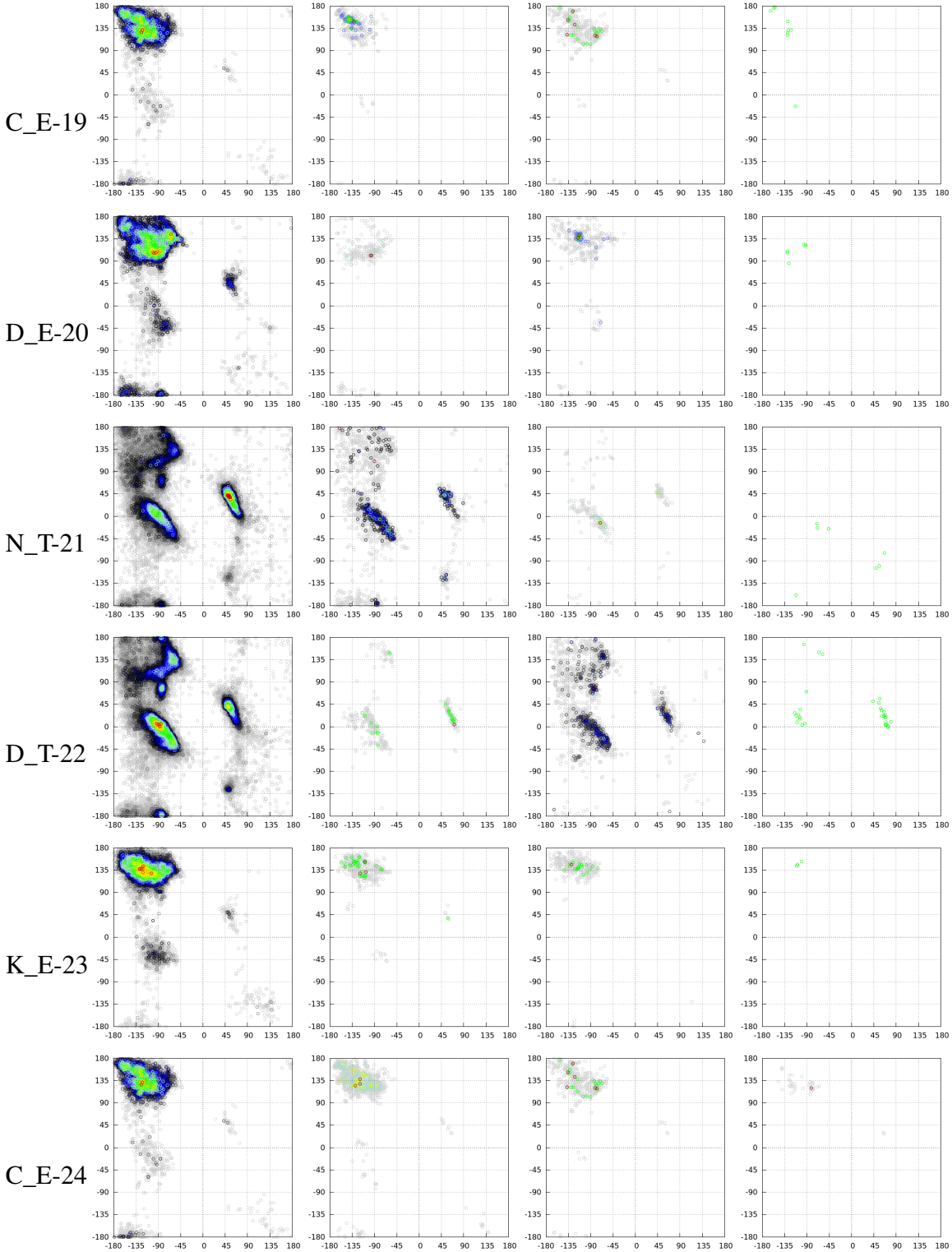
Fonte: Borguesan et al. (2015b).

Figura 9.9: Comparativo entre os arquivos de APL1, APL2-Esq., APL2-Dir. e APL3 para a estrutura com o código no PDB de 1ACW



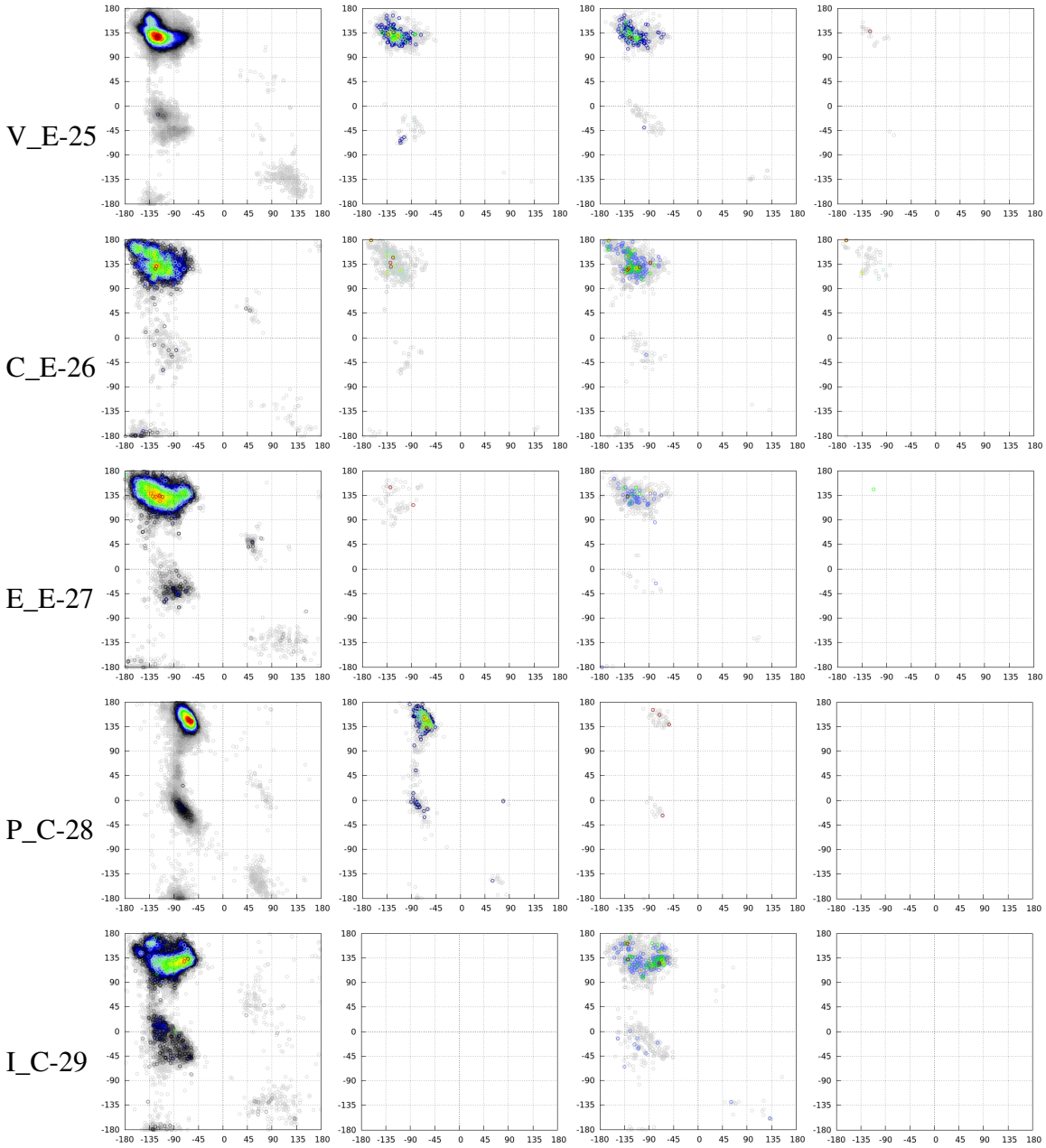
Fonte: Borguesan et al. (2015b).

Figura 9.9: Comparativo entre os arquivos de APL1, APL2-Esq., APL2-Dir. e APL3 para a estrutura com o código no PDB de 1ACW



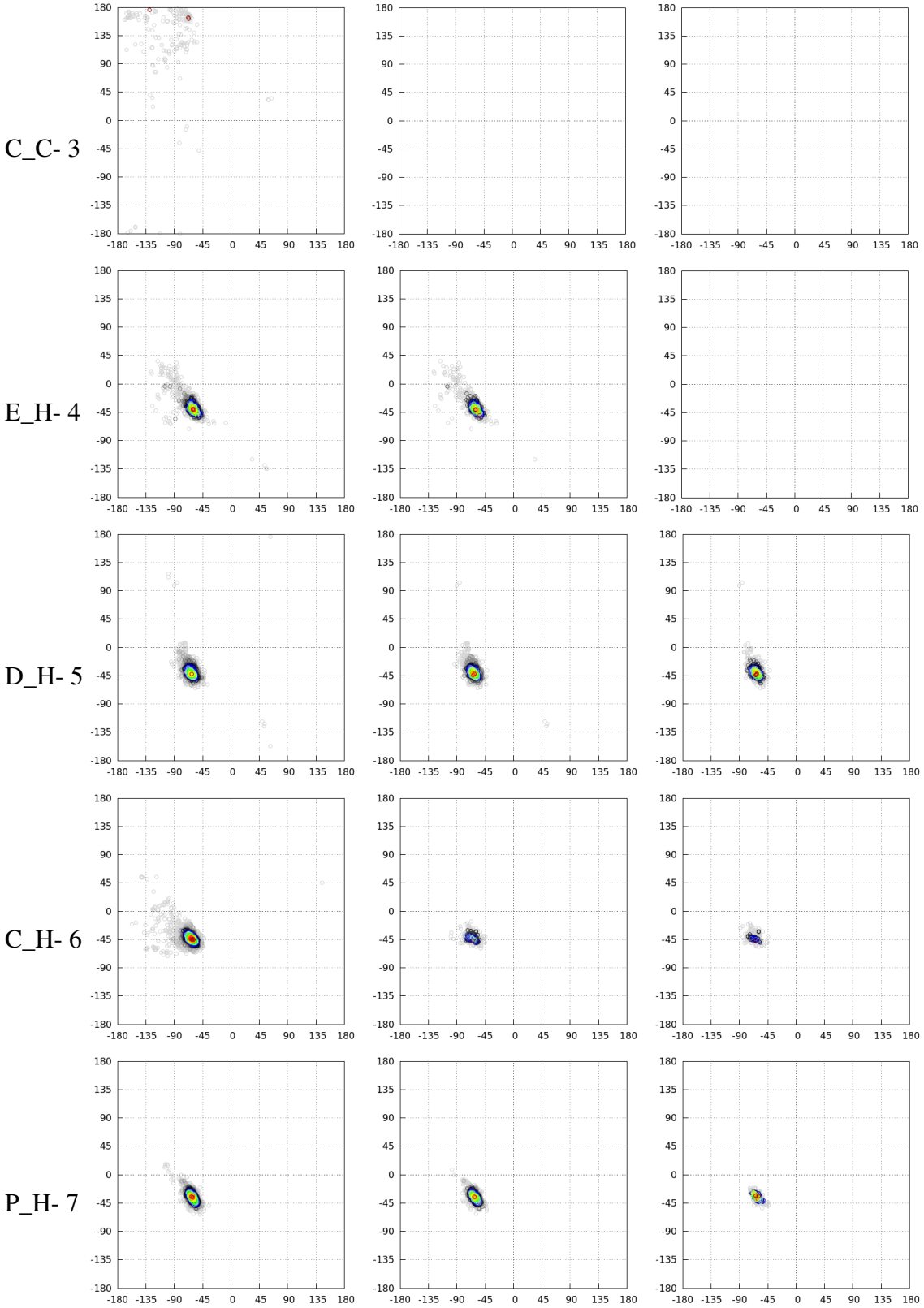
Fonte: Borguesan et al. (2015b).

Figura 9.9: Comparativo entre os arquivos de APL1, APL2-Esq., APL2-Dir. e APL3 para a estrutura com o código no PDB de 1ACW



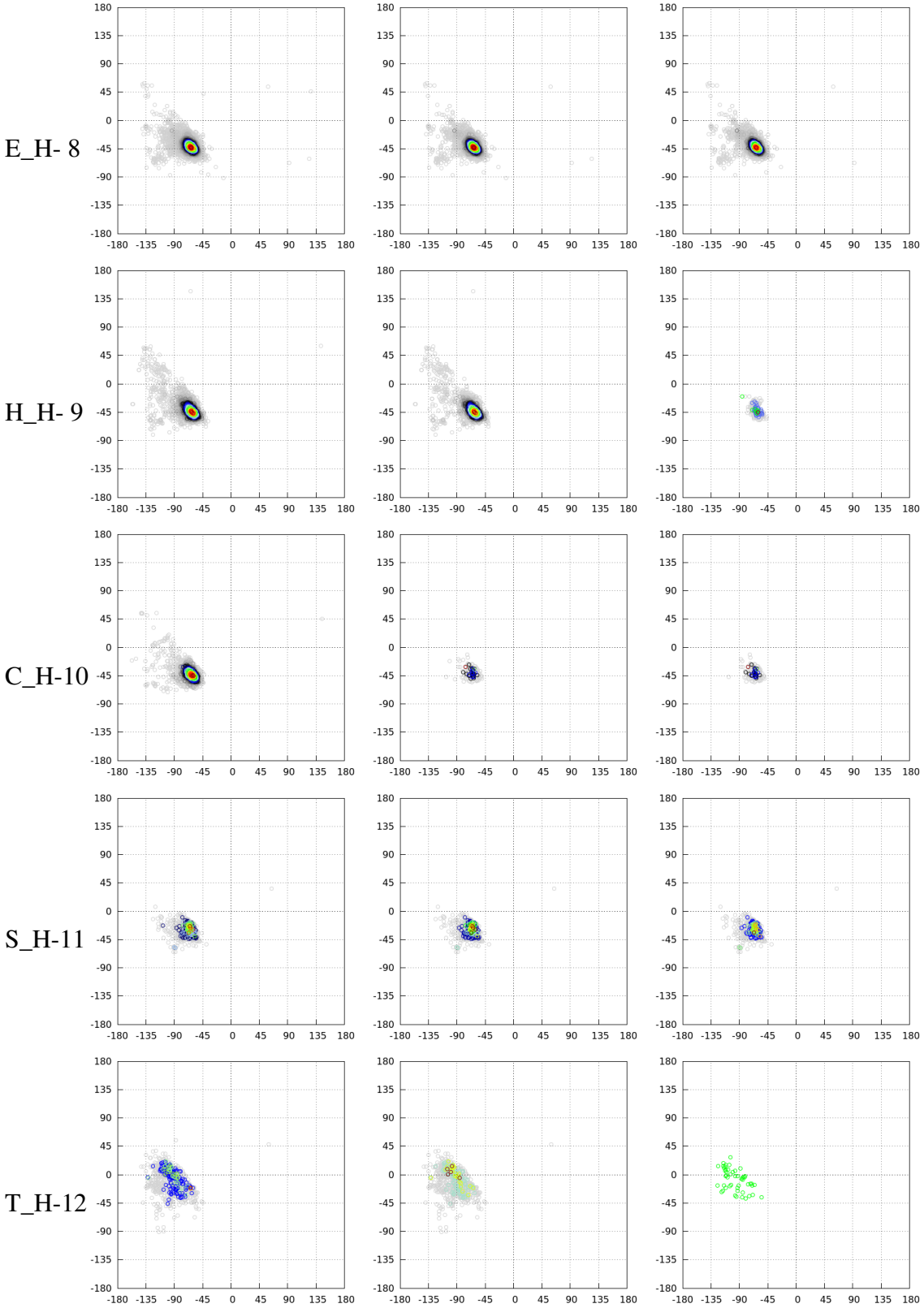
Fonte: Borguesan et al. (2015b).

Figura 9.10: Comparativo entre os arquivos de APLCentral5, APL2-Centroid7 e APLCentral9 para a estrutura com o código no PDB de 1ACW



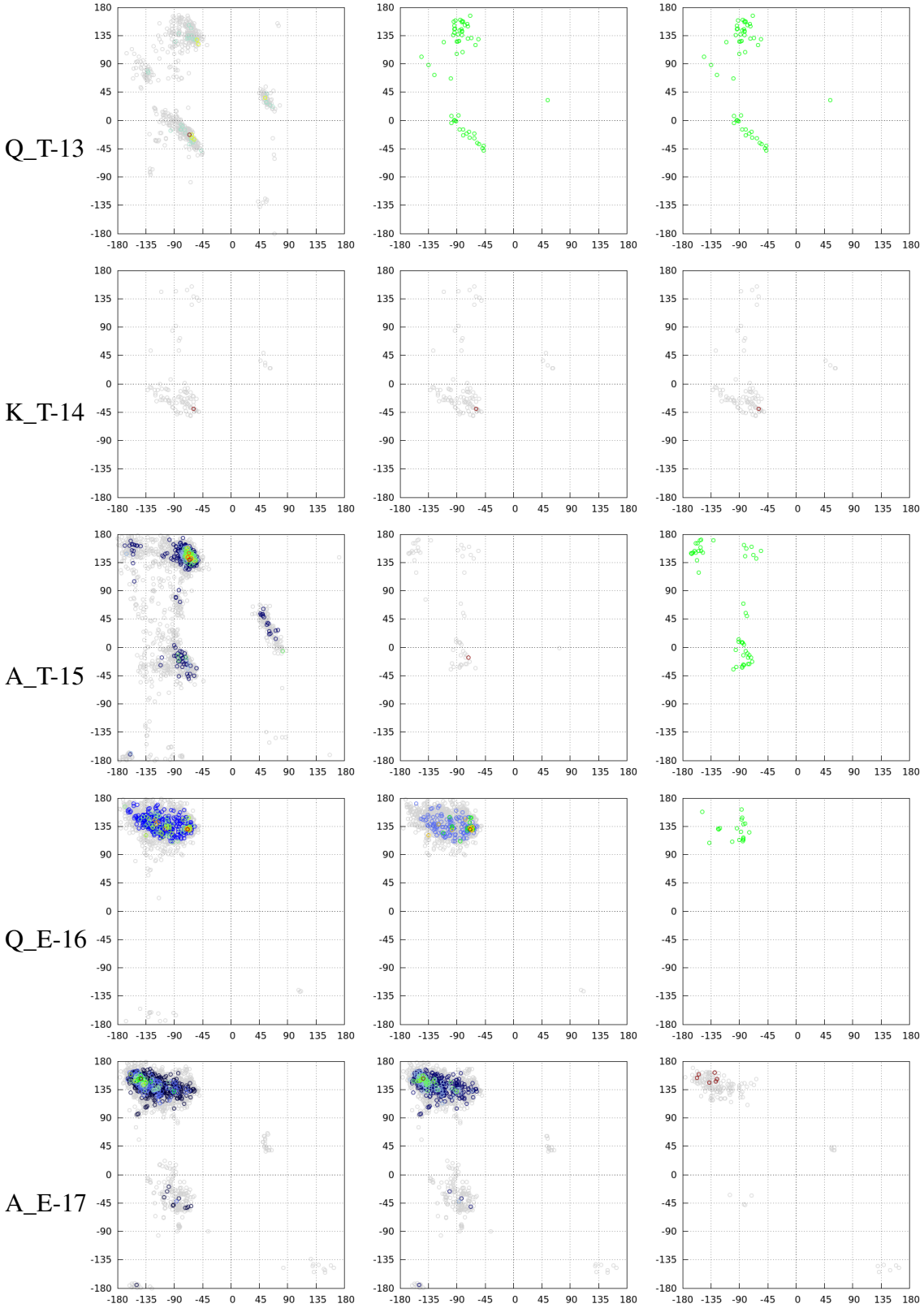
Fonte: Borguesan et al. (2015b).

Figura 9.10: Comparativo entre os arquivos de APLCentral5, APL2-Centroid7 e APLCentral9 para a estrutura com o código no PDB de 1ACW



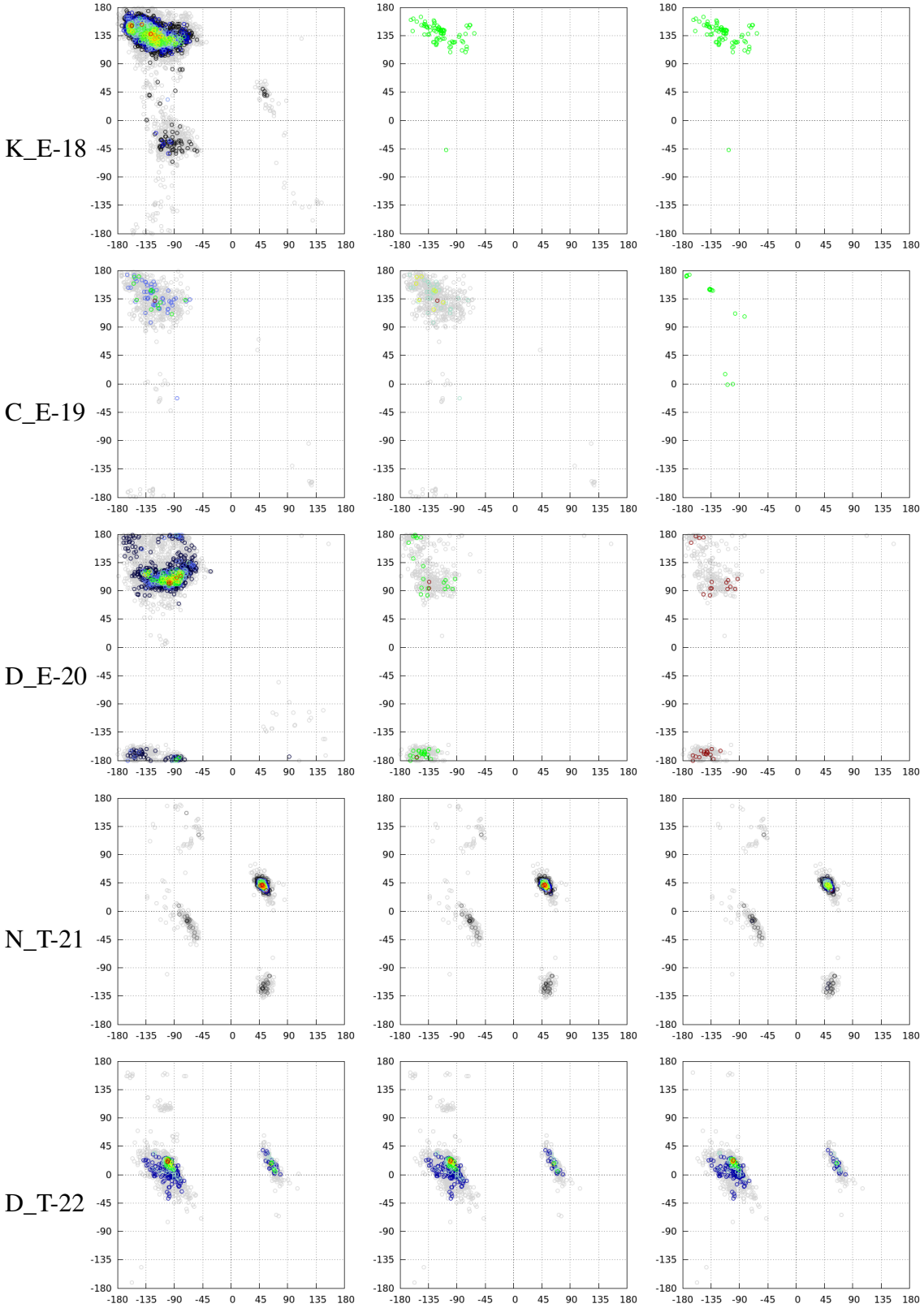
Fonte: Borguesan et al. (2015b).

Figura 9.10: Comparativo entre os arquivos de APLCentral5, APL2-Centroid7 e APLCentral9 para a estrutura com o código no PDB de 1ACW



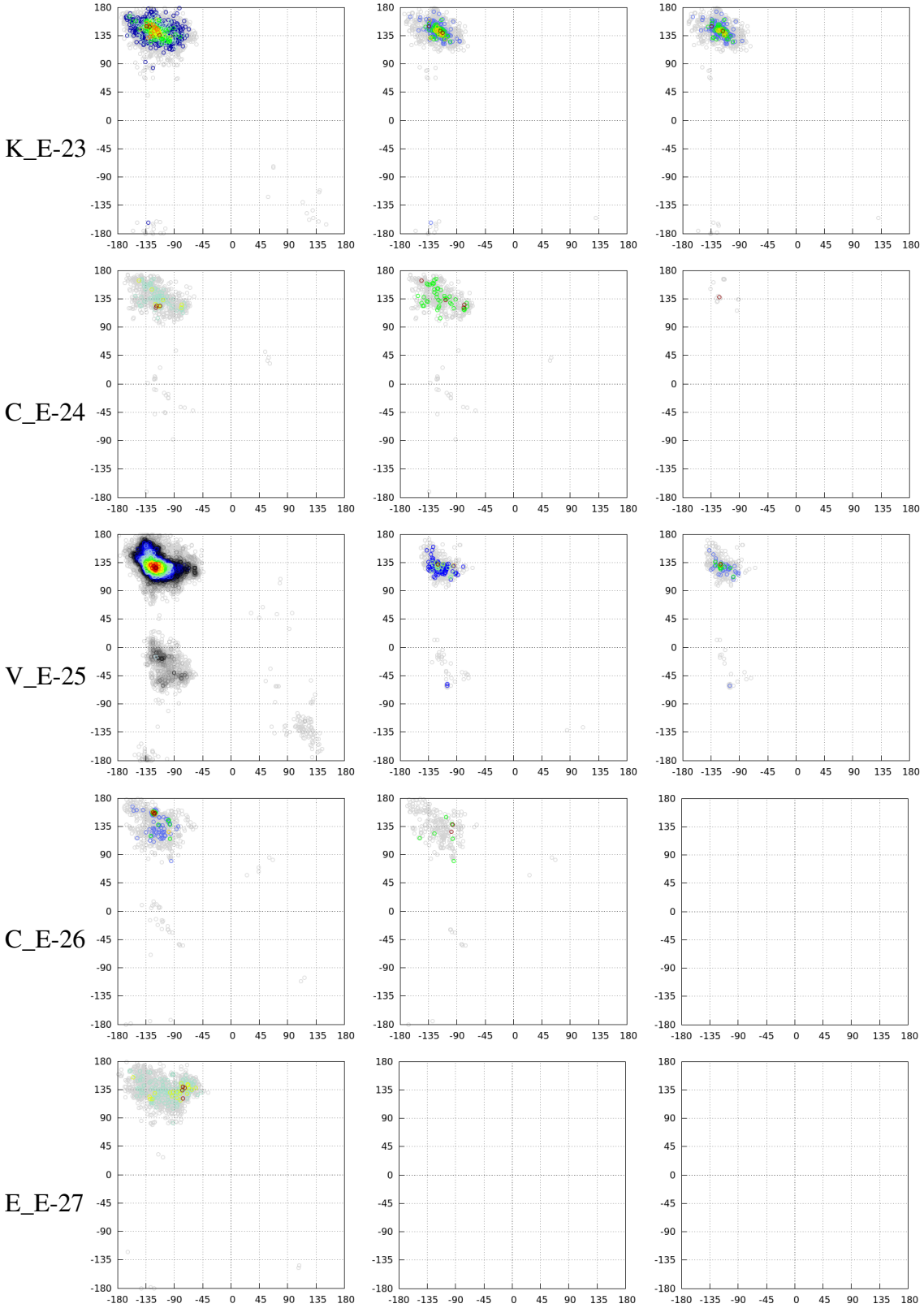
Fonte: Borguesan et al. (2015b).

Figura 9.10: Comparativo entre os arquivos de APLCentral5, APL2-Centroid7 e APLCentral9 para a estrutura com o código no PDB de 1ACW



Fonte: Borguesan et al. (2015b).

Figura 9.10: Comparativo entre os arquivos de APLCentral5, APL2-Centroid7 e APLCentral9 para a estrutura com o código no PDB de 1ACW



Fonte: Borguesan et al. (2015b).