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

BRUNO IOCHINS GRISCI

Predição da flexibilidade de aminoácidos utilizando NeuroEvolução de Topologias Crescentes

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Márcio Dorn
Co-orientador: Prof. Dr. Shan He

Porto Alegre
2016

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Vice-Reitor: Prof. Rui Vicente Oppermann
Pró-Reitor de Graduação: Prof. Sérgio Roberto Kieling Franco
Diretor do Instituto de Informática: Prof. Luis da Cunha Lamb
Coordenador do Curso de Ciência de Computação: Prof. Carlos Arthur Lang Lisbôa
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

“A frase mais excitante de se ouvir na ciência,
a que anuncia novas descobertas,
não é ’Eureka!’,
e sim ’Que engraçado...”’
— Isaac Asimov

AGRADECIMENTOS

Este trabalho foi parcialmente financiado por recursos da FAPERGS (002021-25.51/13) e MCT/CNPq (473692/2013-9), Brasil.

Agradeço ao Prof. Dr. Márcio Dorn pela orientação, motivação e educação ao longo dos projetos desenvolvidos em conjunto, os quais foram de fundamental importância na minha formação, e a todos os membros do *Structural Bioinformatics and Computational Biology Lab* do Instituto de Informática da UFRGS pela ajuda e conselhos quanto à bioinformática. Ao Prof. Dr. Shan He por ter me acolhido em sua pesquisa na *University of Birmingham*, Reino Unido. E à minha família e amigos, pelo apoio e motivação que me permitiram chegar até aqui.

RESUMO

Este trabalho aborda o desafio da predição da estrutura tridimensional de uma dada sequência de aminoácidos, o que foi relatado pertencer à classe dos problemas NP-Completos. É apresentado um novo método baseado na evolução de redes neurais artificiais através de NeuroEvolução de Topologias Crescentes e em agrupamento hierárquico para extração de características estruturais de proteínas determinadas experimentalmente e definir a flexibilidade conformacional de uma sequência de aminoácidos alvo. A técnica proposta manipula informação estrutural do *Protein Data Bank* para gerar intervalos de ângulos de torção com probabilidades associadas para cada aminoácido em uma sequência alvo, representando a sua flexibilidade conformacional. Essa informação pode ser usada para predizer a estrutura tridimensional de sequências proteicas desconhecidas e ajudar na redução do espaço de busca conformacional de moléculas de proteína em métodos de predição da estrutura de proteínas baseados em conhecimento. O método proposto foi testado com uma variedade de proteínas e os resultados indicam que ele de fato é uma opção funcional de representar a flexibilidade de aminoácidos.

Palavras-chave: Bioinformática estrutural. Predição da estrutura tridimensional de proteínas. Redes neurais artificiais. Redes neuroevolutivas.

Prediction of amino acids flexibility using NeuroEvolution of Augmenting Topologies

ABSTRACT

This work addresses the challenge of predicting the three-dimensional structure of a given amino acid sequence, which has been reported to belong to the NP-Complete class of problems. It is presented a new method based on evolving artificial neural networks through NeuroEvolution of Augmenting Topologies and hierarchical clustering to extract structural features from experimentally-determined proteins and determine the conformational flexibility of a target amino acid sequence. The proposed technique manipulates structural information from the Protein Data Bank to generate torsion angles intervals with associated probabilities for each amino acid in a target sequence representing its conformational flexibility. This information may be used to predict the three-dimensional structure of unknown protein sequences and help to reduce the conformational search space of protein molecules in knowledge-based protein structure prediction methods. The method was tested with a variety of proteins and the results indicate that it is indeed a functional way to represent the flexibility of amino acids.

Keywords: Structural bioinformatics. Protein three-dimensional structure prediction. Artificial neural networks. Neuroevolution.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
AA	Aminoácido
AN	Ácido Nucleico
CM	*Comparative Modelling*
ES	Estrutura Secundária
FR	*Fold Recognition*
APL	*Angle Probability List*
DNA	Ácido Desoxirribonucleico
MLP	*Multilayer Perceptron*
NMR	*Nuclear Magnetic Resonance*
PDB	*Protein Data Bank*
PSP	*Protein Structure Prediction*
RNA	Ácido Ribonucleico
XML	*eXtensible Markup Language*
CASP	*Critical Assessment of Structure Prediction*
NEAT	*NeuroEvolution of Augmenting Topologies*
RCSB	*Research Collaboratory for Structural Bioinformatics*

LISTA DE SÍMBOLOS

χ	Ângulo de torção chi
ϕ	Ângulo de torção phi
ψ	Ângulo de torção psi
ω	Ângulo de torção ômega
$C\alpha$	Átomo de carbono central alfa
C	Átomo de carbono
H	Átomo de hidrogênio
N	Átomo de nitrogênio
O	Átomo de oxigênio
$\Phi_k(\cdot)$	Função de ativação não linear do neurônio k
log	Logaritmo
$\|x\|$	Norma euclidiana (comprimento) do vetor x
e	Número de Euler, cujo valor é aproximadamente 2,718
$\sqrt{x}$	Raiz quadrada de x
$\sum_{i=1}^{n}$	Somatório de 1 a n

SUMÁRIO

1 INTRODUÇÃO

Um dos maiores problemas da biologia estrutural e da bioinformática, desafiando há anos biólogos, matemáticos e cientistas da computação, é a predição da estrutura 3D (tridimensional) de proteínas. As proteínas são macromoléculas formadas por uma sequência de moléculas orgânicas chamadas de aminoácidos, que se organizam no espaço assumindo uma estrutura tridimensional. Essa estrutura é essencial para a compreensão da função da proteína nas células dos seres-vivos.

Os métodos experimentais atuais capazes de determinar a estrutura 3D das proteínas são muito caros e demorados, gerando uma defasagem no nosso conhecimento de estruturas de proteínas em relação a descobertas de novas sequências de aminoácidos. Por conta disso, a criação de um método computacional capaz de predizer a estrutura 3D de proteínas tem sido um dos principais objetivos da bioinformática estrutural. Ao longo das últimas décadas diversos algoritmos foram propostos, mas, apesar dos esforços, ainda não há um programa satisfatoriamente capaz de realizar a predição da estrutura 3D de proteínas. O que se tem percebido através de experimentos e comparações é que os métodos que utilizam informação estrutural de bases de dados de proteínas são os que têm obtido melhores resultados. Tais métodos enfrentam dois desafios em especial, a identificação e recuperação de dados de bancos de dados e a criação de estratégias de busca pela estrutura nativa das proteínas.

Neste trabalho, é proposto um método para a recuperação e refinamento dos dados disponíveis no banco de dados *Protein Data Bank*, um dos principais repositórios de informação estrutural de proteínas disponíveis ao público. Através de algoritmos de agrupamento de dados e de redes neurais artificiais treinadas com neuroevolução, torna-se possível a representação dos dados originais em intervalos de busca versáteis de forma automatizada, que consideram a flexibilidade e mobilidade das proteínas e podem ser aplicados como otimização para métodos de busca criados para o problema da predição da estrutura 3D de proteínas.

No Capítulo 2, é feita uma revisão da teoria biológica por trás do problema investigado. São vistas em detalhe as proteínas e seus diferentes níveis estruturais, seus modelos de representação computacional e o problema da predição de suas estruturas 3D. Também são discutidas base de dados biológicas, em especial o *Protein Data Bank*.

O Capítulo 3 traz diferentes técnicas computacionais que são utilizadas pelo método proposto neste trabalho. São vistos algoritmos de agrupamento, em especial o agru-

pamento hierárquico, uma breve discussão sobre algoritmos genéticos e uma revisão de redes neurais artificiais e diferentes modos de treinamento. O algoritmo de neuroevolução de topologias crescentes (NEAT) é visto em detalhes.

O método proposto em si é apresentado no Capítulo 4, unindo os conceitos dos dois capítulos anteriores e mostrando o passo a passo do seu funcionamento. O Capítulo 5 descreve os experimentos realizados e a análise dos resultados. Por fim, o Capítulo 6 faz o encerramento do trabalho.

1.1 Trabalhos relacionados

Antes de se aprofundar no trabalho proposto, convém revisar outras pesquisas relacionadas já publicadas. O método MOIRAE (DORN; BURIOL; LAMB, 2013) é uma tentativa anterior de se extrair características estruturais das proteínas criando intervalos de ângulos de torção usando informações de estruturas de proteínas determinadas experimentalmente. Assim como o método deste trabalho, MOIRAE busca no *Protein Data Bank* (PDB) por segmentos de comprimento 5 de aminoácidos de uma proteína, recupera a informação da estrutura secundária e dos ângulos de torção ϕ e ψ e os organiza em grupos com o algoritmo *k-means*, posteriormente passando essa informação para o treinamento de uma rede neural *multilayer perceptron* (MLP) com *backpropagation*, resultando em um intervalo de busca para cada aminoácido (DORN; BURIOL; LAMB, 2013).

Este trabalho expande e aprimora esta primeira tentativa de criar espaços de busca ao mudar a forma de dividir os segmentos de aminoácidos, de agrupar os pontos, de treinar e utilizar redes neurais para classificação e de construir intervalos de busca de ângulos de torção para aminoácidos. Entre as principais diferenças está a utilização, neste método, de redes neurais evolutivas com o algoritmo *NeuroEvolution of Augmenting Topologies*, que permite a criação de várias redes neurais com diferentes topologias (estruturas).

Já o trabalho da APL (*Angle Probability List*) descreve a frequência normalizada de pares de aminoácidos e estruturas secundárias presentes no PDB (BORGUESAN et al., 2015). Tal trabalho não só discorre sobre a criação e utilização de intervalos de ângulos de torção no problema da predição da estrutura 3D de proteínas como valida a sua vantagem através da implementação de metaheurísticas (BORGUESAN et al., 2015).

Mais especificamente sobre a utilização de NEAT, a sua capacidade de encontrar topologias de redes neurais ótimas mostrou retornar resultados melhores que os de uma rede convencional MLP para o problema da seleção de características (SOHANGIR;

RAHIMI; GUPTA, 2014). No campo da predição da estrutura de proteínas, nos últimos anos, o uso de redes neurais foi expandido. Elas foram utilizadas com bom resultados, por exemplo, para predizer a probabilidade de distribuição de ângulos de torção em resíduos na região de volta de proteínas (HELLES; FONSECA, 2009), e para predizer a estrutura secundária de proteínas (WANG et al., 2015).

As seguintes publicações, artigos e resumos referem-se a pesquisas desenvolvidas antes do método apresentado neste trabalho e que serviram de base e inspiração para esta monografia:

(i) O artigo *"APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction"*, de autoria de Bruno Borguesan, Mariel Barbachan e Silva, Bruno Grisci, Mario Inostroza-Ponta e Márcio Dorn, publicado no jornal *Computational Biology and Chemestry*, versão 58, 2015.

(ii) O artigo *"Using conformational preferences of amino acid residues and meta-heuristics to predict 3-D protein structures"*, de autoria de Bruno Iochins Grisci, Bruno Borguesan, Márcio Dorn e Mario Inostroza-Ponta, apresentado no *Third International Society for Computational Biology Latin America*, versão 1, realizado em Belo Horizonte, Brasil, 2014.

(iii) O resumo *"Desenvolvimento de uma estratégia computacional baseada em algoritmos genéticos para a predição da estrutura tridimensional de proteínas"*, de autoria de Bruno Iochins Grisci e orientação de Márcio Dorn, publicado nos anais do XXVI Salão de Iniciação Científica da UFRGS, 2014.

2 FUNDAMENTOS BIOLÓGICOS

2.1 Proteínas

As proteínas são macromoléculas fundamentais para a existência de todos os seres vivos. Proteínas estruturais formam partes da estrutura de seres vivos, como a camada mais externa da pele, enzimas catalizam as reações do metabolismo e a replicação e transcrição do DNA e anticorpos combatem agentes patogênicos. Proteínas transportadoras controlam a entrada e saída das células, sensoriais atuam na criação e detecção de sinais, regulatórias controlam a transcrição de genes e transdutoras convertem energia química em mecânica (LESK, 2010).

Essa grande gama de funções justifica a importância do estudo e pesquisa de proteínas, pois compreendê-las corretamente abre espaço para entendermos o funcionamento dos organismos dos seres vivos, para buscarmos curas para doenças e para tentarmos desvendar a própria origem e evolução da vida na Terra. O que define qual função uma determinada proteína desempenhará é o formato da sua estrutura tridimensional, mas o que define esta estrutura e quais as regras biológicas em ação?

Proteínas são essencialmente uma sequência unidimensional de moléculas orgânicas chamadas de aminoácidos (TRAMONTANO, 2006; LEHNINGER; NELSON; COX, 2005). Os aminoácidos são formados por um átomo de carbono central chamado Carbono alfa ($C\alpha$) ao qual estão ligados um átomo de hidrogênio ($-H$), um grupo amino ($-NH_3^+$), um grupo carboxílico ($-COO^-$) e uma cadeia lateral. É essa cadeia lateral que determina as propriedades físico-químicas de cada resíduo de aminoácido. Na natureza, existem 20 cadeias laterais, diferenciando os 20 aminoácidos mostrados na Tabela 2.1. Essas cadeias laterais variam em suas propriedades físico-químicas, o que afeta como elas interagirão. Por exemplo, algumas das cadeias são polares e podem participar de ligações de hidrogênio e interações eletrostáticas com outros resíduos ou solventes (LESK, 2010). Cadeias laterais com carga positiva e negativa podem se aproximar e formar uma ponte salina (LESK, 2010). Já cadeias laterais apolares são hidrofóbicas e possuem interação desfavorável com a água (LESK, 2010). A Tabela 2.2 mostra a escala de hidrofobicidade dos 20 resíduos de aminoácidos comuns.

Além dos 20 aminoácidos comuns existem ainda alguns aminoácidos pouco usuais que podem ocorrer em proteínas. Entre eles estão a selenocisteína, a pirrolisina, a hidroxiprolina e a hidroxilisina. Além dos aminoácidos, íons, ligantes orgânicos e moléculas

de água podem se integrar às estruturas de algumas proteínas (LESK, 2010).

Tabela 2.1: Os 20 aminoácidos comuns. Na literatura, os aminoácidos normalmente são abreviados por uma sigla de três letras minúsculas ou uma letra maiúscula.

Aminoácido	*3-letras*	*1-letra*	*Propriedade físico-química*
Glicina	gly, gli	G	Apolar
Alanina	ala	A	Apolar
Leucina	leu	L	Apolar
Valina	val	V	Apolar
Isoleucina	ile	I	Apolar
Prolina	pro	P	Apolar
Fenilalanina	phe ou fen	F	Apolar
Metionina	met	M	Apolar
Serina	ser	S	Polar
Treonina	thr, tre	T	Polar
Cisteina	cys, cis	C	Polar
Tirosina	tyr, tir	Y	Polar
Asparagina	asn	N	Polar
Glutamina	gln	Q	Polar
Triptofano	trp, tri	W	Polar
Aspartato ou Ácido aspártico	asp	D	Polar
Glutamato ou Ácido glutâmico	glu	E	Polar
Arginina	arg	R	Polar
Lisina	lys, lis	K	Polar
Histidina	his	H	Polar

Fonte: (LESK, 2010)

Para a formação da cadeia polipeptídica de uma proteína, os aminoácidos vizinhos precisam se ligar. Essa ligação ocorre através de uma síntese de desidratação e chama-se ligação peptídica. Durante sua formação, o grupo carboxílico de um dos aminoácidos perde um agrupamento hidroxila ($-OH$), liberando uma ligação, enquanto o outro aminoácido perde um hidrogênio ($-H$) em seu grupo amina, também liberando uma ligação. Os aminoácidos são unidos por essas ligações livres e o hidrogênio e agrupamento hidroxila liberados unem-se formando uma molécula de água (H_2O) (AMABIS; MARTHO, 2006).

Cada proteína possui uma sequência de aminoácidos própria e única, mas seria equivocado pensar que qualquer sequência de aminoácidos randômica daria origem a uma proteína viável. São as sequências que geram uma estrutura equilibrada e eficiente que vêm a originar proteínas naturais (LEVINTHAL, 1968). Além disso, seres vivos são capazes de montar proteínas necessárias para seu funcionamento através da codificação do DNA, o que serve de base para o chamado "dogma central da biologia molecular",

Tabela 2.2: Propriedades físico-químicas dos resíduos de aminoácidos.

Aminoácido	*Hidrofobicidade*	*N° de ângulos* χ
Triptofano	2.25	2
Isoleucina	1.80	2
Fenilalanina	1.79	2
Leucina	1.70	2
Cisteina	1.54	1
Metionina	1.23	3
Valina	1.22	1
Tirosina	0.96	2
Prolina	0.72	0
Alanina	0.31	0
Treonina	0.26	1
Histidina	0.13	2
Glicina	0.00	0
Serina	-0.04	1
Glutamina	-0.22	3
Asparagina	-0.60	2
Glutamato	-0.64	3
Aspartato	-0.77	2
Lisina	-0.99	4
Arginina	-1.01	4

Fonte: (FAUCHERE; PLISKA, 1983; CUTELLO; NARZISI; NICOSIA, 2006)

cunhado em 1958 por Francis Crick: "DNA faz RNA faz proteína" (LESK, 2010). Não se pode deixar de apontar, contudo, que o ambiente onde a proteína se encontra também contribui para este processo.

O significado do dogma central é que o DNA contém sequências de código genético que podem ser traduzidas para o RNA mensageiro, que por sua vez é traduzido em sequências de aminoácidos. As interações dos aminoácidos entre si e o meio ao seu redor criam a estrutura tridimensional das proteínas, e esta estrutura determina a função das proteínas (LESK, 2010). Como proteínas são geradas a partir do DNA, mutações no código genético podem provocar pequenas variações nas proteínas finais. Isso deixa as proteínas sujeitas ao processo de seleção natural de Darwin, no qual indivíduos com as características que permitem uma melhor adaptação ao ambiente serão "selecionados" para passar essas características para as gerações futuras (LESK, 2010).

Nesse sentido, é preciso observar que a relação entre a cadeia de aminoácidos e a estrutura tridimensional de uma proteína é robusta, ou seja, ela pode tolerar pequenas variações na cadeia de aminoácidos original. De outro modo, se a estrutura de uma proteína pudesse ser formada apenas através de uma sequência de aminoácidos específica, seria impossível atingi-la na natureza através da evolução, pois qualquer outra sequência

a destruiria. A comparação de proteínas correspondentes em espécies diferentes mostra que, apesar das suas sequências de aminoácidos divergirem, muitas vezes as estruturas tridimensionais e as funções permaneceram as mesmas (LESK, 2010).

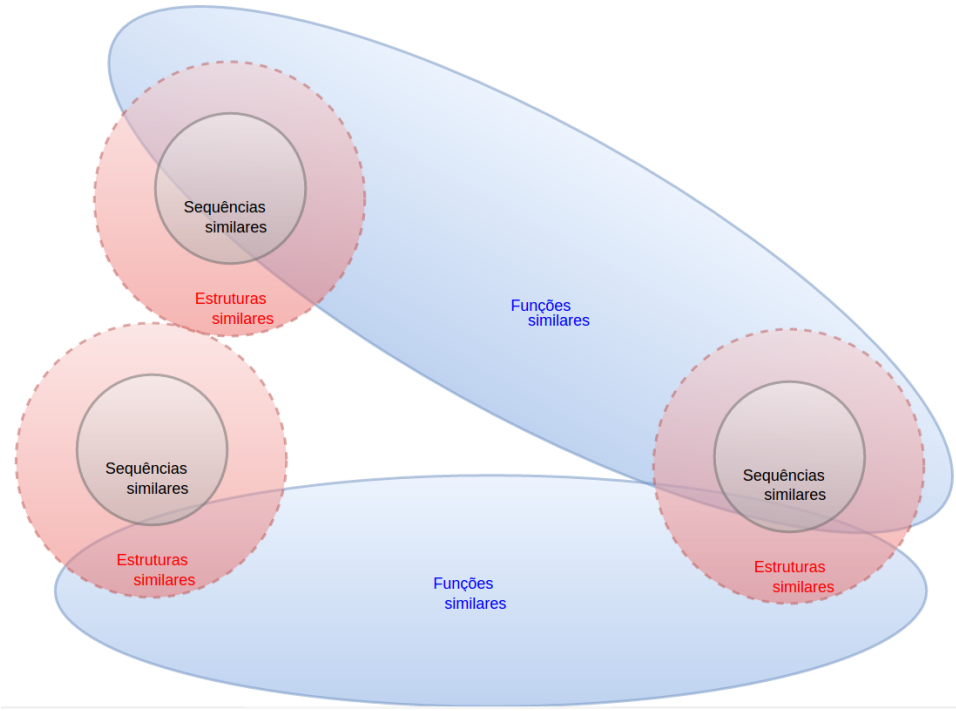
2.1.1 Estrutura das proteínas

Como dito anteriormente, é principalmente a sequência de aminoácidos de uma proteína que determina sua estrutura tridimensional, e é esta estrutura que determina a função. Também foi mencionado que as estruturas são robustas. Pode-se dizer com certa confiança que cadeias de aminoácidos similares originarão estruturas similares. Ao mesmo tempo, estruturas tridimensionais similares podem se originar de sequências de aminoácidos drasticamente diferentes. Sequências de aminoácidos e estruturas tridimensionais semelhantes comumente produzem proteínas com funções semelhantes, mas estruturas tridimensionais distintas também podem possuir funções similares (LESK, 2010). A Figura 2.1 traz um diagrama detalhando as relações esperadas entre sequência, estrutura e função. Essa robustez, contudo, não pode ser vista como um indicativo de que trocas de aminoácidos são inconsequentes. A anemia falciforme ou siclemia, por exemplo, é uma forma de anemia grave hereditária causada pela substituição de um único aminoácido da molécula de hemoglobina (AMABIS; MARTHO, 2006).

É importante observar que a estrutura da grande maioria das proteínas não é fixa, e sim apenas suficientemente estável em condições restritas de solvente e temperatura. Caso os limites não sejam respeitados a proteína se deforma e perde sua função. No caso do corpo humano, a manutenção de temperatura constante, concentração de sal e acidez é essencial para o correto funcionamento das proteínas. Uma mesma proteína pode ainda apresentar mais de uma estrutura tridimensional, pois essa mudança entre estados é essencial para o funcionamento de várias proteínas (LESK, 2010). Esse é o caso das proteínas intrinsecamente desordenadas. Um exemplo é a proteína com código `PDB` 2MM8 (DOBROVOLSKA et al.,), que será vista no Capítulo 5.

Quanto à organização da estrutura das proteínas, ela é normalmente descrita em três diferentes níveis, com um extra para proteínas compostas por mais de uma subunidade. Estes níveis representam diferentes formas para o estudo das proteínas.

Figura 2.1: Relações entre sequência, estrutura e função de uma proteína



Fonte: Adaptado de (LESK, 2010)

2.1.1.1 Estrutura primária

A estrutura primária de uma proteína é representada pela sua sequência de aminoácidos, unidos por ligações peptídicas. As sequências de aminoácidos podem ser diferenciadas pela quantidade de aminoácidos da cadeia, pelos tipos de aminoácidos e pela ordem em que eles aparecem. Duas proteínas com estrutura primária de mesmo comprimento e mesmos tipos de aminoácido nas mesmas quantidades ainda serão diferentes caso os aminoácidos estejam dispostos em ordens diferentes (AMABIS; MARTHO, 2006).

2.1.1.2 Estrutura secundária

Os aminoácidos da estrutura primária interagem entre si e com as moléculas de solvente ao seu redor. Dessa forma, acabam surgindo padrões repetitivos de organização espacial, a chamada estrutura secundária. Existem poucos tipos desses padrões, e a combinação destes pode ser vista como a estrutura tridimensional da proteína (VERLI, 2014).

Um mesmo tipo de estrutura secundária pode surgir a partir de diferentes estruturas primárias, mas mesmo assim elas apresentam semelhanças. Os três principais tipos de

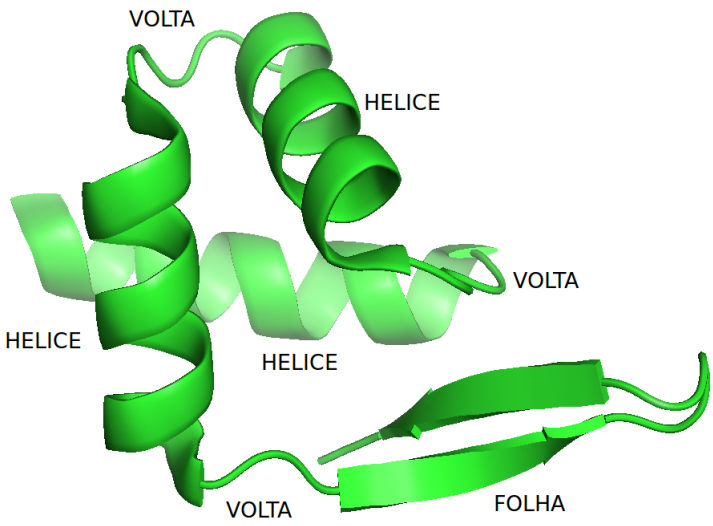
estrutura secundária são alças, folhas e hélices (VERLI, 2014).

As alças, também chamadas de voltas, são o tipo de estrutura secundária com maior flexibilidade, por se tratar da região de conexão entre as hélices e as folhas. Por conta disso possuem uma variedade de orientações maior, com a flexibilidade aumentando quanto maior for a alça. Isso também acaba tornando-as as regiões mais propensas a sofrerem mutações (troca de resíduos de aminoácidos). Diferentemente das folhas e hélices, ainda, as voltas não possuem periodicidade ao longo de suas estruturas, frequentemente sendo consideradas sem estrutura definida ou com estrutura randômica, embora para certos tamanhos e composições possam ter forma mais definida. Além de servir de zona de ligação entre outras estruturas secundárias, as alças podem cobrir sítios ativos, regular o acesso de moduladores ou substratos e se envolver em contatos proteínas-proteínas por estarem frequentemente mais expostas ao solvente (VERLI, 2014).

As folhas β são compostas por duas ou mais sequências de aminoácidos, as quais são chamadas de fitas, extendidas. Essas fitas interagem lado a lado longitudinalmente através de ligações de hidrogênio entre o grupamento $N - H$ de uma das fitas e o grupamento $C = O$ da outra. Nessa organização, os átomos do carbono α orientam-se intercaladamente acima e abaixo do plano da folha, numa organização semelhante a dobraduras. Nas fitas de uma folha, a porção N-terminal da fita pode interagir com a porção N-terminal da fita vizinha, ou a porção N-terminal da fita pode interagir com a porção C-terminal da fita vizinha, dando origem às folhas β paralelas e antiparalelas respectivamente. Nas folhas antiparalelas, as ligações de hidrogênio possuem um ângulo de 90° com as fitas, enquanto nas paralelas esses ângulos são maiores, deixando as interações mais fracas e a estrutura menos estável. As folhas β podem ser puras ou mistas, quando folhas paralelas pareiam com folhas antiparalelas (VERLI, 2014).

A última das estruturas secundárias principais são as hélices. Sua formação decorre da realização de ligações de hidrogênio entre os grupos $N - H$ e $C = O$, assim como nas folhas β, mas dessa vez não entre resíduos de aminoácidos de fitas vizinhas, mas sim entre os mais próximos na sequência, entre as voltas da hélice. Existem diferentes tipos de hélices, a mais comum sendo a hélice α, com $3,6$ resíduos de aminoácidos em média por volta da hélice, cada aminoácido (n) com uma ligação de hidrogênio com o quarto aminoácido seguinte $(n + 4)$. A hélice poli-prolina II também é relativamente comum e pode ser encontrada no colágeno e em proteínas de parede celular de plantas. Nesse caso, não ocorre formação de ligação de hidrogênio durante a organização da hélice porque o átomo de nitrogênio da prolina está ligado a três átomos de carbono. Mais raras

Figura 2.2: Visualização da estrutura terciária da proteína com código PDB 2P5K



Fonte: (GARNETT et al., 2007)

são a hélice π e a hélice 3_{10}, contendo $4, 4$ e 3 resíduos de aminoácidos por volta da hélice respectivamente (VERLI, 2014).

2.1.1.3 Estrutura terciária

A estrutura terciária é representada pela distribuição das estruturas secundárias no espaço tridimensional, que se organizam através de um fenômeno chamado de enovelamento, quando uma combinação de forças promove a adoção de um conformação mais estável pela proteína. Os resíduos de aminoácidos hidrofóbicos se aproximam ao afastarem-se da água, expulsando este solvente do centro da proteína. Ao mesmo tempo os resíduos polares ficam expostos ao solvente e as interações inter-resíduo se estabelecem. O enovelamento é influenciado pelas interações não covalentes entre os aminoácidos de uma proteína e entre estes e o solvente, e também por interações covalentes associadas a mudanças co ou pós-traducionais (VERLI, 2014). A Figura 2.2 ilustra a estrutura terciária da proteína 2P5K (GARNETT et al., 2007) em termos de suas estruturas secundárias.

2.1.1.4 Estrutura quaternária

Várias proteínas contém mais de uma subunidade, e o agrupamento delas é chamado de estrutura quaternária. Essas subunidades podem ser cópias da mesma cadeia de aminoácidos ou combinações de diferentes cadeias (LESK, 2010). Entre as proteínas com

estrutura quaternária destaca-se a hemoglobina.

2.1.2 Conformação de biomoléculas

A conformação é a descrição da forma de uma molécula. Ela não pode ser confundida, entretanto, com a estrutura. Essa última refere-se a uma única forma bem definida e conhecida, enquanto a conformação é uma forma dentre várias possíveis em um dado ambiente molecular. Numa solução existem diversas formas co-existindo simultaneamente, e cada uma delas é uma conformação (VERLI, 2014).

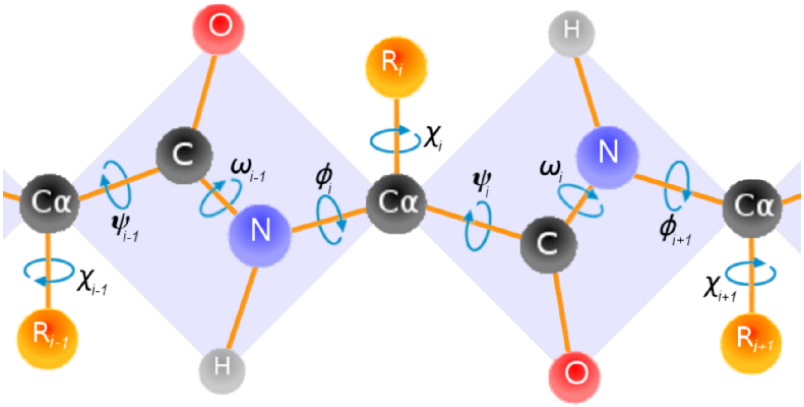
Quando falamos na conformação da cadeia de uma proteína, estamos falando da curvatura que ela assume no espaço. Por ser flexível, a quantidade de formas possíveis é imensa, e quando a proteína está em estado desnaturado ela de fato assume diferentes conformações. No seu estado nativo, contudo, sob condições fisiológicas de solvente e temperatura, todas as moléculas com a mesma sequência de aminoácidos adotam a mesma conformação. Essa limitação da liberdade conformacional é termodinamicamente custosa, o que explica porque sequências randômicas de aminoácidos não assumem estados nativos únicos, enquanto proteínas desenvolveram conjuntos de interações coerentes que estabilizam seus estados nativos (LESK, 2010).

As proteínas em seu estado nativo possuem uma estrutura tridimensional definida, em geral compactas e bem empacotadas, raramente apresentando buracos ou canais. Entre as propriedades que contribuem para a estabilidade desse estado estão a satisfação de ligações de hidrogênio potencial de grupos polares, aprofundamento de átomos não polares e compactação densa dos aminoácidos no interior da proteína (LESK, 2010).

2.1.3 Representação da estrutura tridimensional de proteínas

Existem na literatura duas representações principais da estrutura de proteínas, em especial para a computação (TRAMONTANO, 2006; SCHWARTZ, 2008; LILJAS et al., 2001). O primeiro modelo utiliza as posições cartesianas dos átomos que formam uma proteína para representar a sua estrutura tridimensional. Para cada átomo da proteína se atribui um vetor de coordenadas tridimensionais. Assim sendo, para uma proteína com n átomos, se tem $3n$ graus de liberdade nessa abordagem, e esse número aumenta ainda mais quando o solvente é considerado. Por conta disso, a utilização de todos os átomos na

Figura 2.3: Representação dos ângulos de torção em uma cadeia de aminoácidos. As ligações peptídicas ocorrem entre os átomos de carbono (C) e nitrogênio (N).



Fonte: (BORGUESAN et al., 2015)

representação se mostra computacionalmente custosa (DORN; BURIOL; LAMB, 2013).

A outra maneira de representação é pensarmos na forma tridimensional de uma proteína em termos das suas rotações internas. Nesse caso, a forma de dois aminoácidos vizinhos pode ser descrita pelos ângulos de torção ao redor do átomo de $C\alpha$, chamados de phi (ϕ) e psi (ψ), com valores que podem variar entre $-180°$ e $180°$ (VERLI, 2014). Estes ângulos internos são definidos por conjuntos de quatro átomos sucessivos na cadeia principal da proteína. Para cada resíduo i, o ângulo ϕ_i é definido pelos átomos $C_{i-1} - N - C\alpha - C$, e o ângulo ψ_i é definido pelos átomos $N - C\alpha - C - N_{i+1}$ (LESK, 2010). A posição dos ângulos na cadeia está ilustrada na Figura 2.3.

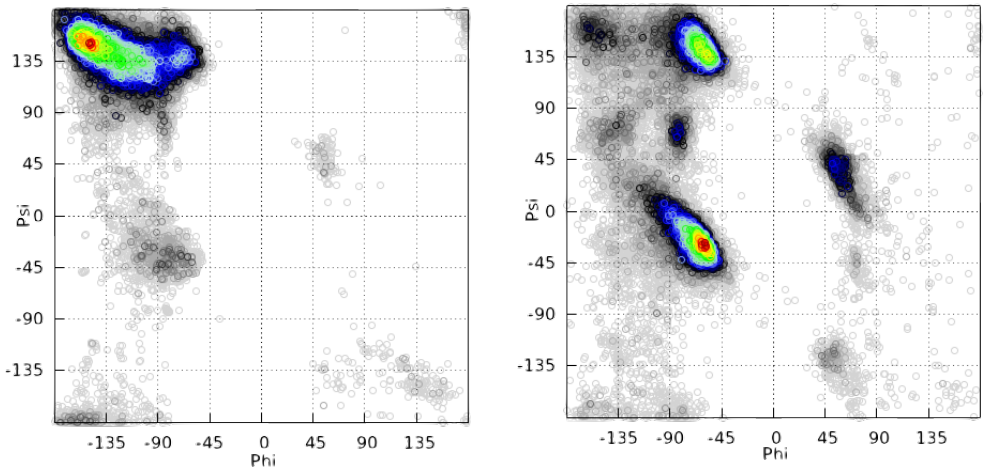
As rotações utilizam a regra da mão direita. O valor de ϕ é igual a $0°$ quando a ligação $C\alpha - C$ é *cis* à $N - C\alpha$. Já o valor de ψ é igual a $0°$ quando a ligação $C - N$ é *cis* à $N - C\alpha$. O ângulo ψ é perdido no último aminoácido de cada cadeia, enquanto ϕ é perdido no primeiro (DORN; BURIOL; LAMB, 2013).

Algumas combinações de ϕ e ψ são proibitivas dadas as restrições estereoquímicas, tornando-as fisicamente impossíveis. V. Sasiekharan, C. Ramakrishnan e G. N. Ramachandran foram os primeiros a representar graficamente as regiões permitidas estereoquimicamente, no que ficou conhecido como "mapa de Ramachandran". As duas principais regiões permitidas para a conformação de aminoácidos são a α e β, correspondendo às estruturas secundárias hélice α e folha β (LESK, 2010). De fato, resíduos de aminoácidos presentes em estruturas secundárias tendem a adotar um conjunto particular de ângulos de torção ϕ e ψ (BORGUESAN et al., 2015), como pode ser visto na Figura 2.4, contendo mapas de Ramachandran para os aminoácidos alanina e tripofano

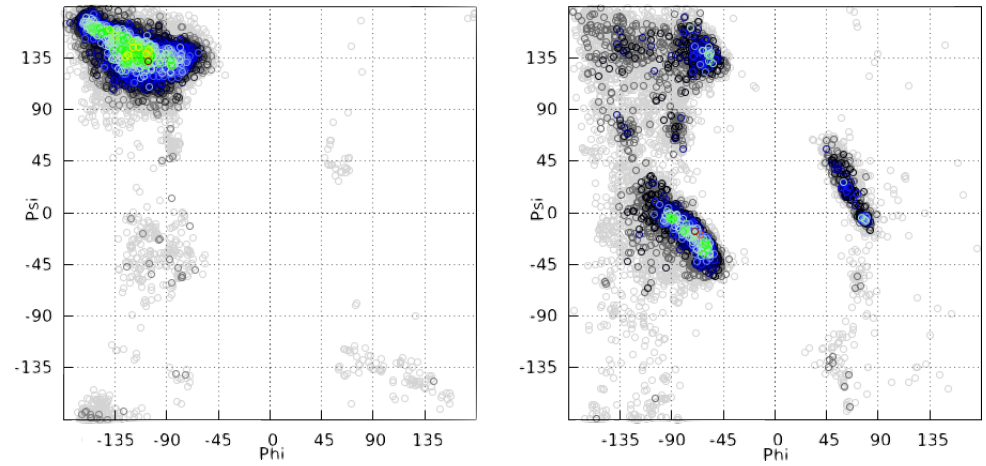
quando em regiões de folha β e volta. Percebe-se que as regiões de volta apresentam irregularidade muito maior que as demais, o que está associado à grande flexibilidade desta estrutura secundária.

Figura 2.4: Preferências conformacionais de aminoácidos de acordo com a suas estruturas secundárias. As áreas em vermelho escuro são as regiões mais densas.

(a) ALA - Folha β (b) ALA - Volta



(c) TRP - Folha β (d) TRP - Volta



Fonte: (NIAS-SERVER, 2016)

Ainda é possível considerar o ângulo de torção ao redor da própria ligação peptídica, definido pelos átomos $C\alpha - C - N_{i+1} - C\alpha$. Este ângulo é chamado de ω e é restrito aos valores aproximados de $180°$ (*trans*), mais comum, e $0°$ (*cis*), raro (LESK, 2010). Por fim, ainda existem os ângulos que descrevem a forma da cadeia lateral. A quantidade de ângulos da cadeia lateral varia de acordo com o aminoácido, a glicina não possui nenhum, enquanto a arginina possui cinco, por exemplo. A Tabela 2.2 contém o número de ângulos da cadeia lateral para os 20 resíduos de aminoácidos básicos. Estes ângulos são denominados χ_1, χ_2 e assim sucessivamente até atingir o número específico do aminoácido em questão (LESK, 2010).

Para a representação da estrutura tridimensional da cadeia principal de resíduos de aminoácidos de uma proteína, bastam os pares de ângulos ϕ e ψ (e idealmente também os ângulos χ). Isso dá a grande vantagem de reduzir os graus de liberdade para $2m$, m sendo o número de resíduos de aminoácidos da cadeia (DORN; BURIOL; LAMB, 2013).

2.2 Protein Data Bank

Existem, na internet, muitos repositórios de informações da área da biologia molecular, como bancos de dados de sequências, estruturas, funções, bibliografias ou coleções específicas (LESK, 2010). O PDB (*Protein Data Bank*, ou Banco de Dados de Proteínas em português) é uma das principais bases de dados para estruturas tridimensionais de proteínas (BERMAN et al., 2000), e consiste em uma coleção pública de estruturas de proteínas, ácidos nucleicos e outras macromoléculas biológicas (LESK, 2010).

As informações das estruturas tridimensionais armazenadas no PDB são determinadas experimentalmente através de cristalografia com raios-X ou espectroscopia com ressonância nuclear magnética. A Tabela 2.3 detalha a quantidade de conteúdo disponível no PDB e a Figura 2.5 mostra o crescimento da base de dados para proteínas ao longo das últimas quatro décadas. Informações adicionais podem ser calculadas a partir dos dados do PDB com programas externos, como os ângulos de torção ϕ e ψ, a estrutura secundária ou a área da superfície exposta ao solvente (DORN; BURIOL; LAMB, 2013). As proteínas são identificadas dentro do PDB por códigos PDB compostos por um dígito seguido por uma combinação de três letras ou dígitos.

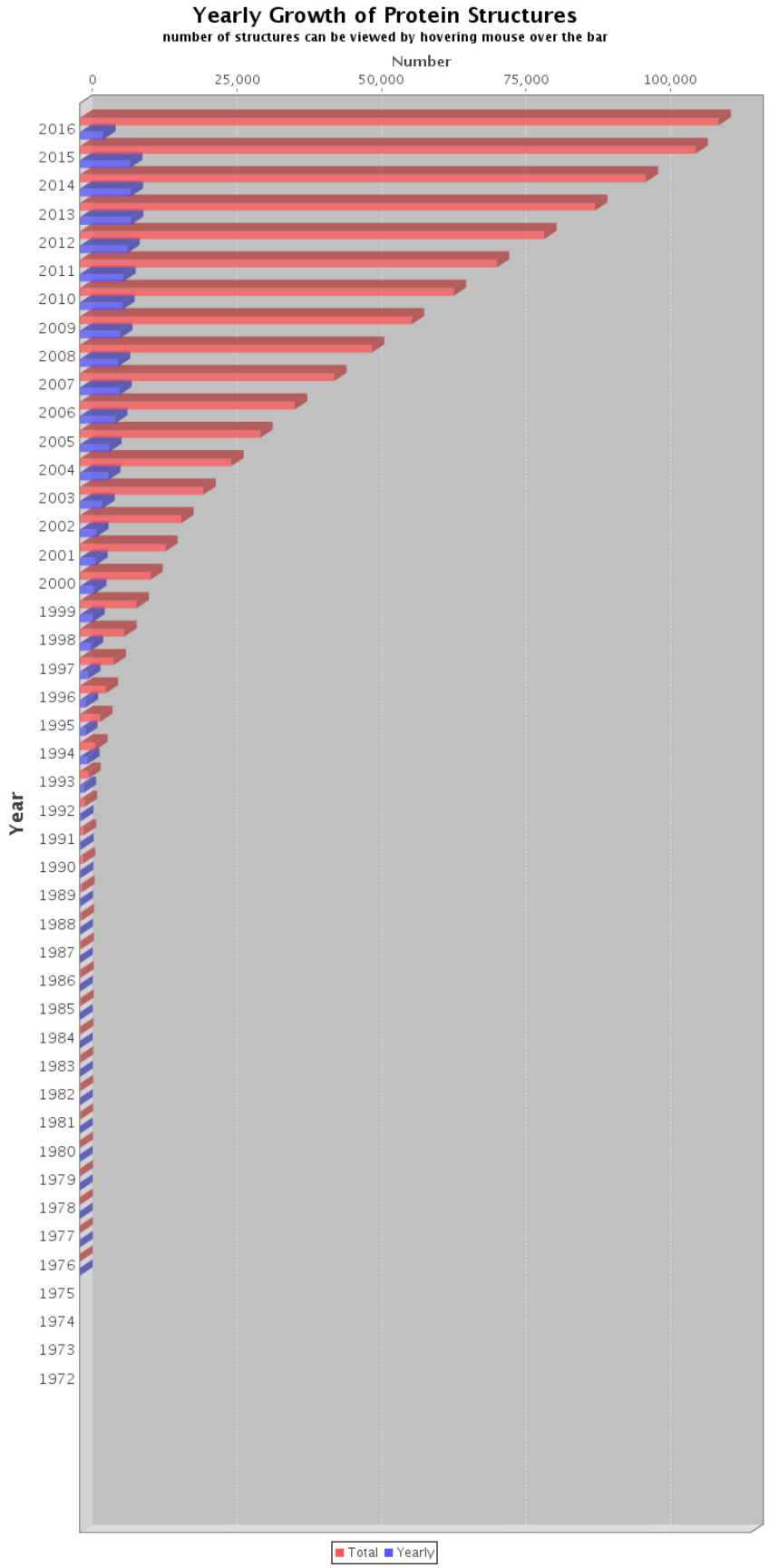
Tabela 2.3: Detalhamento do conteúdo do PDB (como disponível em 23 de maio de 2016)

Método experimental	*Proteínas*	*Ácidos nucleicos*	*Complexos proteína / AN*	*Outros*	*Total*
Raio-X	99437	1729	5041	4	106211
NMR	10004	1141	232	8	11385
Microscopia	745	30	266	0	1041
Híbrido	89	3	2	1	95
Outro	173	4	6	13	196
Total	110448	2907	5547	26	118928

Fonte: (RCSB-PDB, 2016a)

Uma comparação relevante pode ser feita entre os tamanhos do PDB, que até o dia 23 de maio de 2016 continha as estruturas de 110448 proteínas (Tabela 2.3), e bancos de dados de sequências, como o GenBank (BENSON et al., 2009), uma coleção de

Figura 2.5: Crescimento anual da quantidade de estruturas de proteínas disponíveis no PDB até o dia 17 de maio de 2016.

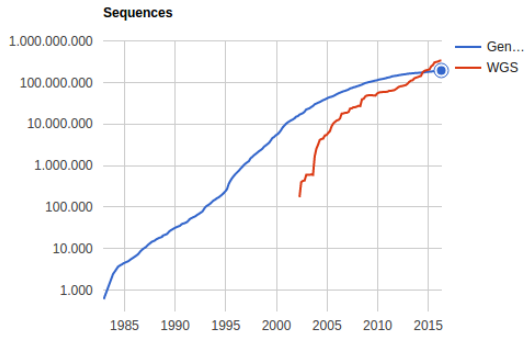


Fonte: (RCSB-PDB, 2016b)

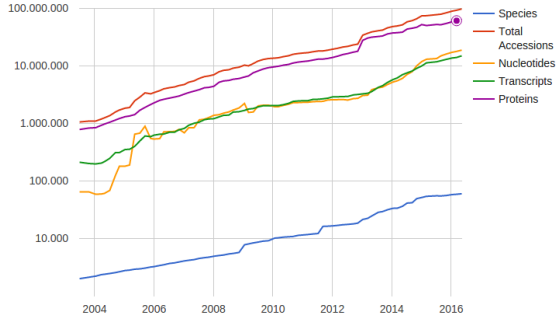
todas as sequências de DNA disponíveis publicamente, e o RefSeq (PRUITT et al., 2013), uma base de dados com sequências de DNA, de transcrições e de proteínas. O primeiro continha em sua base de dados aproximadamente 194 milhões de sequências até o dia 15 de abril de 2016 (Figura 2.6a), enquanto o segundo, considerando apenas as sequências de proteínas, continha quase 64 milhões de entradas até o dia 9 de maio de 2016 (Figura 2.6b). Estes dados apontam uma clara defasagem entre os métodos de obtenção de sequências e os de obtenção de estruturas de proteínas, o que será visto em mais detalhe na Seção 2.3.

Figura 2.6: Crescimento das bases de dados de sequências GenBank e RefSeq como disponível até o dia 1° de junho de 2016.

(a) GenBank



(b) RefSeq



Fonte: (NCBI-GENBANK, 2016; NCBI-REFSEQ, 2016)

2.3 Problema da predição da estrutura 3D de proteínas

Como foi explorado nas seções anteriores, a sequência de aminoácidos da cadeia principal de uma proteína determina a forma 3D que ela assumirá no espaço, e essa forma é a responsável pela função da proteína. Por isso, descobrir a estrutura tridimensional das proteínas é fundamental para compreender sua atuação nos organismos vivos, seu processo de evolução e na sintetização de novas proteínas com objetivos específicos como aplicação na medicina e indústria farmacêutica.

A grande questão é que, enquanto a descoberta da estrutura primária de uma proteína é relativamente simples, e já existem programas capazes de predizer a estrutura secundária com grande acurácia, a predição da estrutura terciária a partir da sequência de aminoácidos ainda é um dos problemas mais desafiadores e importantes da biologia molecular.

Determinar a estrutura 3D de uma proteína experimentalmente é muito caro e de-

morado, devido aos custos e tempo associados aos processos de cristalografia, microscopia eletrônica e ressonância nuclear magnética. A dificuldade em determinar a estrutura 3D criou uma grande discrepância entre o volume de sequências de aminoácidos produzidas pelos projetos genoma e o número de estruturas de proteínas definidas experimentalmente (BORGUESAN et al., 2015), como pode ser visto comparando-se as Figuras 2.5 e 2.6. Portanto, o desenvolvimento de um método capaz de determinar com precisão, de forma rápida e barata, a estrutura 3D de uma proteína seria extremamente útil para as diversas áreas envolvidas. A este problema se convencionou chamar PSP (*Protein Structure Prediction* ou "Predição da Estrutura de Proteína"em português) (DORN; BURIOL; LAMB, 2013)

O PSP é um problema desafiador de otimização matemática (LANDER; WATERMAN, 1999) e, de acordo com a teoria de complexidade computacional, é classificado como NP-completo (CRESCENZI et al., 1998; FRAENKEL, 1993; HART; ISTRAIL, 1997; LEVINTHAL, 1968; NGO; MARKS; KARPLUS, 1997). O enovelamento de proteínas é extremamente complexo e não foi totalmente compreendido. Além disso, outros fatores que não a sequência de aminoácidos acabam por influenciar no formato final da proteína, como o ambiente e local. Soma-se a isso o fato de que a manipulação de estruturas 3D é bem mais complexa que a de sequências unidimensionais (VERLI, 2014).

Diversos métodos computacionais foram propostos para a resolução do PSP, podendo ser divididos em quatro classes distintas de acordo com a estratégia empregada (FLOUDAS et al., 2006; DORN; BURIOL; LAMB, 2013).

(i) Métodos de primeiro princípio sem uso de informação de base de dados (OSGUTHORPE, 2000)

(ii) Métodos de primeiro princípio com uso de informação de base de dados (ROHL et al., 2004; SRINIVASAN; ROSE, 1995)

(iii) Métodos de reconhecimento de enovelamento (FR) (BOWIE; LUTHY; EISENBERG, 1991; JONES; TAYLOR; THORNTON, 1992; BRYANT; ALTSCHUL, 1995)

(iv) Métodos de modelagem comparativa (CM) (MARTÍ-RENOM et al., 2000; SÁNCHEZ; SALI, 1997)

Os métodos da classe (i) não utilizam informações de similaridade de sequências cuja estrutura já é conhecida. Em vez disso, tentam simular no computador os fenômenos

fisicoquímicos responsáveis pelo enovelamento das proteínas e utilizam o conceito de energia livre para encontrar o estado nativo (ANFINSEN et al., 1961; ANFINSEN, 1973).

Os métodos das classes (ii), (iii) e (iv) utilizam bibliotecas de enovelamentos e modelos de estruturas para predições de estruturas 3D de forma rápida e efetiva (KOLINSKI, 2004). Os métodos de primeiro princípio com uso de informação da base de dados extraem regras das estruturas de proteínas de bancos de dados para começar a construir estruturas 3D, os métodos de reconhecimento de enovelamento se aproveitam do fato de que proteínas sem sequências similares podem ter enovelamentos similares e os métodos de modelagem comparativa utilizam relações de evolução de sequência entre a proteína alvo e outras com estrutura 3D conhecida (DORN; BURIOL; LAMB, 2013).

Os principais desafios dos métodos com informações de base de dados são encontrar uma representação computacional para a estrutura da proteína, recuperar a informação dos modelos experimentais, encontrar um critério de diferenciação entre bons e maus candidatos à solução e desenvolver uma estratégia de busca capaz de encontrar a estrutura 3D da proteína. Os avanços mais significativos no CASP (*Critical Assessment of Structure Prediction*), uma competição para métodos dedicados ao PSP, foram alcançados por métodos que fazem uso de bancos de dados (KOOP et al., 2007; COZZETTO et al., 2009; ZHANG, 2008; XU et al., 2011).

2.4 Resumo do capítulo

Neste capítulo foram apresentados conceitos básicos da biologia molecular, as propriedades das proteínas, seus níveis estruturais e como eles se relacionam entre si e com as funções das proteínas. Também foram comentados bancos de dados de informações biológicas e seu conteúdo. Por fim, foi analisado o problema da predição da estrutura tridimensional de proteínas, sua relevância, os métodos já existentes e os seus desafios ainda existentes. O próximo capítulo apresenta uma revisão de métodos computacionais de aprendizado de máquina e inteligência artificial que podem ser utilizados para auxiliar no entendimento dos dados biológicos disponíveis.

3 MÉTODOS COMPUTACIONAIS

3.1 Agrupamento

Técnicas de agrupamento são utilizadas em tarefas de divisão de dados em grupos naturais nos quais não há classes para serem preditas. Grupos devem refletir algum mecanismo que valorize a semelhança entre algumas observações mais do que com outras, ou seja, os dados devem ser agrupados segundo algum grau de semelhança. O critério de semelhança e o cálculo da distância entre os dados normalmente variam de acordo com o problema e são parâmetros dos algoritmos de agrupamento (WITTEN; EIBE; HALL, 2011).

Existem diversos algoritmos diferentes para agrupamento, vários dependendo do conhecimento do número final de grupos. Um dos mais tradicionais, por exemplo, é o *k-means* (MACQUEEN, 1967), que precisa ser inicializado com o parâmetro k que representa quantos grupos estão sendo buscados. Mas nem todos os domínios de problemas permitem saber de antemão o número ideal de grupos. Uma abordagem possível com *k-means*, por exemplo, seria tentar diferentes valores de k e selecionar o que gerou os melhores resultados. Outra opção são algoritmos que recursivamente dividem ou combinam grupos até que o resultado satisfatório seja encontrado, como é o caso do agrupamento hierárquico (WITTEN; EIBE; HALL, 2011).

3.1.1 Agrupamento hierárquico

O algoritmo de agrupamento hierárquico (JOHNSON, 1966) não necessita ser informado do número final de grupos, mas sim de um limiar de distância entre grupos que ao ser atingido interrompa a combinação ou divisão de grupos. Na sua abordagem *bottom-up*, chamada de aglomerativa, basta possuir uma medida de distância entre grupos (WITTEN; EIBE; HALL, 2011).

O método é iniciado considerando cada ponto dos dados como um grupo próprio (linha 2 no Algoritmo 1), e prossegue combinando os grupos mais próximos (linha 5 no Algoritmo 1) até que haja apenas um grupo contendo todos os pontos. Utilizando o limiar de distância é possível definir uma condição de parada, e o método finalizará quando a distância entre os grupos for maior que o limiar (linha 3 no Algoritmo 1), retornando os grupos atuais (linha 14 no Algoritmo 1) (WITTEN; EIBE; HALL, 2011).

Algorithm 1 Pseudo-código para agrupamento hierárquico

Require: Matriz D de pares de distâncias entre pontos, pontos P, limiar de distância t
1: Construa o grafo G atribuindo um vértice a cada grupo;
2: Forme $\|P\|$ grupos com um elemento cada;
3: **while** existe mais de um grupo **and** distância entre grupos $> t$ **do**
4: Selecione os dois grupos mais próximos C_A e C_B;
5: Combine C_A e C_B em um novo grupo C com $\|C_A\|$ + $\|C_B\|$ elementos;
6: Calcule a distância entre C e todos os outros grupos;
7:
8: **if** os grupos são próximos **then**
9: Adicione novo vértice C a G e conecte-o aos vértices C_A e C_B;
10: Remova as linhas e colunas de D correspondendo a C_A e C_B;
11: Adicione uma nova linha e coluna a D correspondendo ao grupo C;
12: **end if**
13: **end while**
14: **return** G;

3.2 Algoritmos genéticos

Métodos estocásticos abrangem uma classe de algoritmos que utiliza de alguma forma a aleatoriedade para encontrar soluções ótimas ou quase ótimas para problemas difíceis (LUKE, 2009). Metaheurísticas são uma subdivisão geral desses algoritmos, aplicadas a uma gama muito grande de diferentes tipos de problemas. Métodos populacionais são uma classe de metaheurísticas que fazem uso de populações de soluções candidatas, isto é, algoritmos que mantém um conjunto de possíveis soluções para um problema, que são alteradas de alguma forma para convergirem para a solução local (LUKE, 2009).

Muitos métodos populacionais são inspirados na biologia, entre eles os algoritmos evolutivos, que fazem uso de conceitos da genética e evolução. Entre os principais algoritmos evolucionários, destacam-se os algoritmos genéticos, inventados na década de 1970 (LUKE, 2009). Um algoritmo genético basicamente opera de forma iterativa, atribuindo valores de *fitness* ou aptidão para as soluções, que são chamadas de "indivíduos", indicando o quão boas elas são, e então passando para uma seleção de indivíduos e cruzamento entre os mesmos, produzindo uma nova população (LUKE, 2009).

Em algoritmos genéticos, os indivíduos de uma população são as soluções candidatas ao problema e eles são representados por um "cromossomo" formado por "genes". Existem diferentes representações possíveis, algumas das mais populares sendo vetores de valores binários ou vetores de valores reais (KUTHAN; LANSKY, 2007).

Duas das principais operações dos algoritmos genéticos que os diferenciam dos demais são o *crossover* e a mutação. A mutação de um indivíduo pode ser realizada de várias formas, mas é a modificação aleatória do seu cromossomo. *Crossover* é a grande

distinção dos algoritmos genéticos. Essa operação promove o cruzamento entre dois indivíduos, chamados de pais, combinando os seus cromossomos de forma a gerar um novo indivíduo, com características de ambos, possivelmente mais capaz de ser uma solução ótima para o problema (GOLDBERG, 1989). A ideia é que os melhores indivíduos sejam selecionados para cruzar, promovendo o avanço das boas soluções e a extinção das más (LUKE, 2009).

3.3 Redes neurais artificiais

Redes neurais artificiais são um grupo de modelos de aprendizado de máquina usados para estimar ou aproximar funções (HORNIK, 1991). O elemento principal das redes neurais é o neurônio, a unidade de processamento. Um neurônio possui três elementos básicos:

(i) Conjunto de sinapses ou conexões, cada uma caracterizada por um peso que multiplica o seu sinal de entrada.

(ii) Somador para somar os sinais de entrada ponderados pelas sinapses.

(iii) Função de ativação que restringe a amplitude de saída do neurônio.

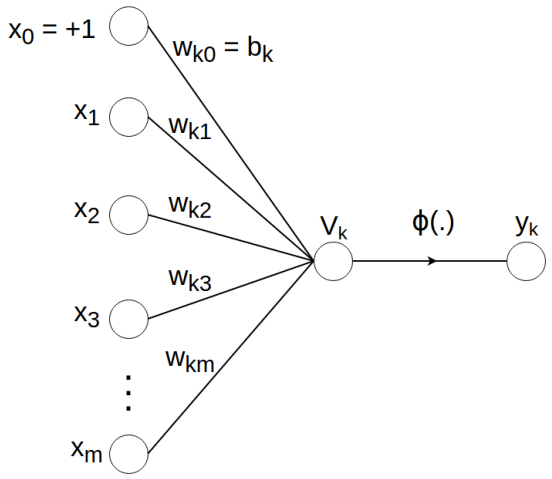
As Equações 3.1 e 3.2 representam matematicamente um neurônio k, onde x_0, x_1, ..., x_m são os sinais de entrada mais um valor constante multiplicado pelo *bias* (b_k), com efeito de aumentar ou diminuir a entrada da função de ativação; w_{k0}, w_{k1}, ..., w_{km} são os pesos das conexões do neurônio k; v_k é a saída do combinador linear; $\Phi(\cdot)$ é a função linear e y_k é o sinal de saída do neurônio (HAYKIN, 1998). A Figura 3.1 representa este modelo de neurônio visualmente. Uma rede neural pode ser vista como um grafo orientado pelo qual flui um sinal dos nós de fonte para os nós computacionais (neurônios).

$$v_k = \sum_{j=0}^{m} w_{kj} x_j \tag{3.1}$$

$$y_k = \Phi(v_k) \tag{3.2}$$

Nos últimos anos, redes neurais mostraram-se capazes de oferecer boas soluções para diferentes problemas de classificação e biologia, como a predição da estrutura de proteínas (WANG et al., 2015) e atracamento molecular (TAYARANI et al., 2013).

Figura 3.1: Modelo de um neurônio artifical



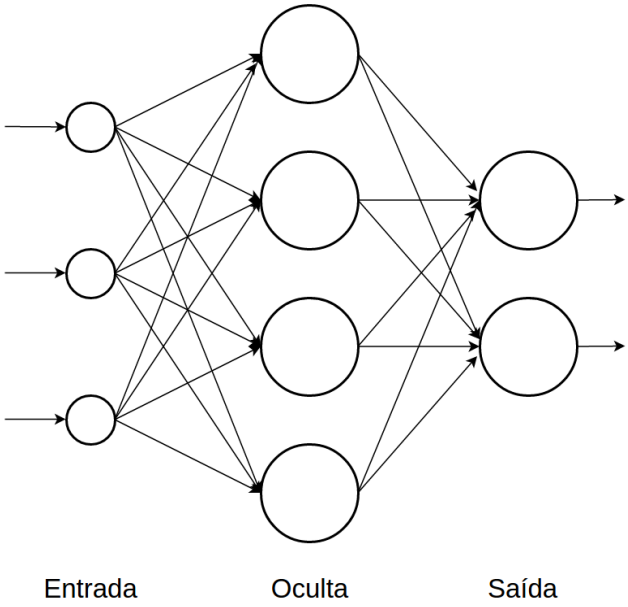
Fonte: Adaptado de (HAYKIN, 1998)

3.3.1 Multilayer perceptron

As redes de múltiplas camadas (MLP ou *multilayer perceptron* em inglês) alimentadas adiante (*feed-forward*) são uma importante classe de redes neurais. Elas são formadas por nós sensoriais da camada de entrada, neurônios computacionais da camada oculta e neurônios computacionais da camada de saída. Entre as características das redes MLP está a alta conectividade entre seus neurônios, a presença de uma ou mais camadas de neurônios ocultos e a inclusão de uma função de ativação não linear no modelo de cada neurônio. A Figura 3.2 representa a estrutura de um MLP com três nós de entrada, quatro neurônios na camada oculta e dois neurônios na camada de saída. Redes MLP já foram bem sucedidas em problemas difíceis através do treinamento de forma supervisionada, no qual dados rotulados são fornecidos como exemplos para o algoritmo, com o popular algoritmo de retro-propagação de erro (LINNAINMAA, 1976) (*error back-propagation*) (HAYKIN, 1998).

O algoritmo de retro-propagação de erro pode ser visto em dois passos através da camada de rede: a propagação (passo para frente) e retro-propagação (passo para trás). No primeiro, o vetor de entrada á aplicado aos neurônios da camada de entrada e seu efeito se propaga camada por camada, percorrendo a rede, até ocorrer a geração de saídas. No segundo, os pesos das conexões são ajustados de acordo com uma regra de correção de erro. A saída da rede é subtraída da resposta desejada, resultando num sinal de erro que é propagado para trás pela rede, e os pesos das conexões são corrigidos para aproximar a

Figura 3.2: Modelo de uma rede MLP



Fonte: O Autor

saída da rede da resposta correta (HAYKIN, 1998; LINNAINMAA, 1976).

Um dos maiores desafios ao se projetar um MLP é definir a sua topologia, ou seja, o número de neurônios, de camadas e as conexões. A criação da estrutura da rede é um dos fatores mais cruciais para o seu sucesso, e geralmente precisa ser planejada caso a caso com o problema e os dados específicos em mente. A criação de uma estrutura de rede sem os cuidados necessários pode produzir resultados improdutivos e ineficientes (CURTEANU; CARTWRIGHT, 2011).

3.3.2 Redes neurais evolutivas

Neuroevolução é o processo de evoluir redes neurais através de algoritmos genéticos, procurando por comportamentos para a rede que produzam bons resultados para a tarefa proposta. Entre as vantagens da neuroevolução estão a maior eficiência e rapidez em relação a métodos de aprendizado por reforço para certos problemas (MORIARTY; MIIKKULAINEN, 1996; MORIARTY, 1997), eficiência para problemas com espaços de estados contínuos e de alta dimensionalidade e fácil representação de memória (GOMEZ; MIIKKULAINEN, 1999; GOMEZ; MIIKKULAINEN, 2002).

Tradicionalmente, métodos neuroevolutivos trabalham sobre redes neurais com

topologia (estrutura) fixa, normalmente com uma única camada oculta totalmente conectada a camada de entrada e de saída, e o objetivo é otimizar os pesos das conexões. Mas a topologia da rede também afeta a sua funcionalidade e pode efetivamente fazer parte do processo de evolução e otimização (ANGELINE; SAUNDERS; POLLACK, 1993; BRANKE, 1995; GRUAU F.; PYEATT, 1996; YAO, 1999).

3.3.3 NeuroEvolution of augmenting topologies

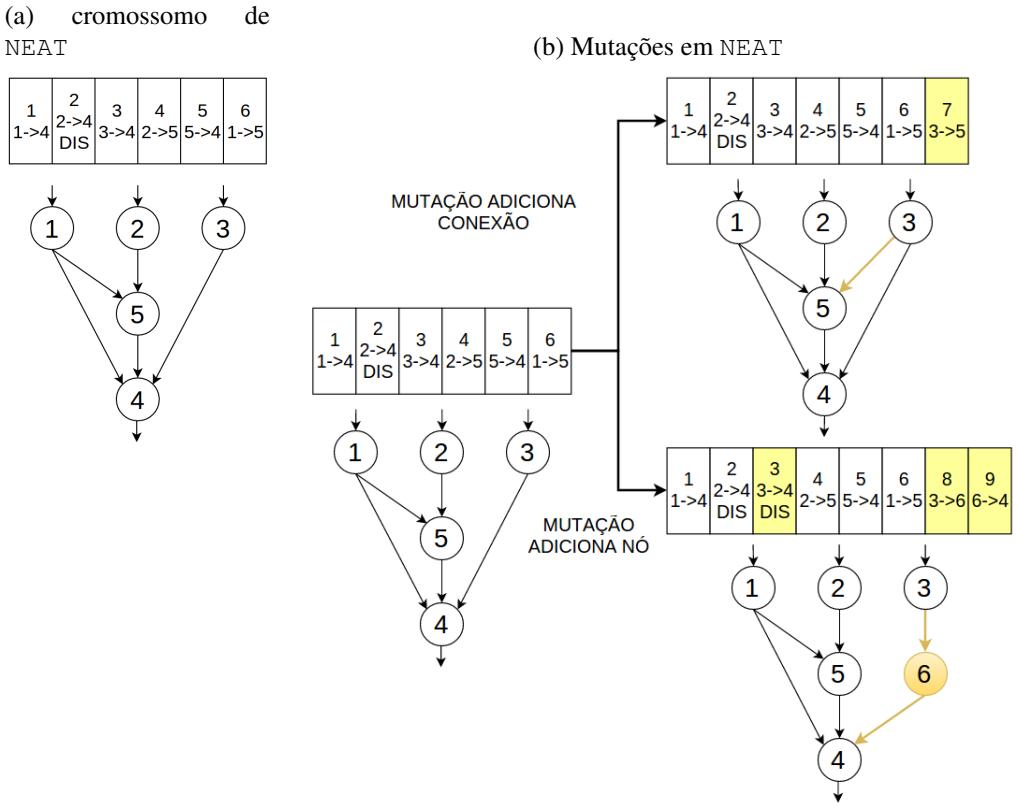
NeuroEvolution of Augmenting Topologies (NEAT), ou NeuroEvolução de Topologias Crescentes, em português, é um algoritmo de construção e treinamento de redes neurais que faz uso de algoritmos genéticos para evoluir a topologia e pesos das ligações de redes neurais. Esse método é adequado para problemas para os quais uma estrutura de rede satisfatória é desconhecida (STANLEY; MIIKKULAINEN, 2002).

NEAT começa com uma população inicial randômica de redes neurais que compartilham a mesma topologia básica, ou seja, nós de entrada, neurônios de saída e conexões entre os neurônios das duas camadas inicializados com pesos randômicos. Essa condição de partida não é por acaso, o minimalismo é importante pois garante que complexidade inútil não será acrescentada às redes neurais, já que apenas as melhorias na topologia da rede que gerarem resultados melhores serão mantidas. Se o método fosse inicializado com topologias aleatórias, neurônios ou conexões desnecessários poderiam estar presentes desde o começo e não seria possível removê-los, o que poderia causar um impacto negativo na evolução. Essa escolha ainda produz resultados mais simples e compactos (STANLEY; MIIKKULAINEN, 2002).

A partir dessa população inicial de redes neurais, novas são criadas iterativamente através dos operadores tradicionais de algoritmos genéticos, especialmente o crossover, que combina dois indivíduos da população atual para criar um novo indivíduo, e mutação, que pode alterar o peso de uma conexão de uma rede neural, adicionar novos neurônios ocultos ou adicionar uma nova conexão entre neurônios. A mutação em NEAT (Figura 3.3) nunca remove um neurônio ou conexão existente pois isso poderia causar inconsistências ao longo da evolução. Em vez disso, há um marcador que pode ser definido para ignorar uma conexão entre neurônios (STANLEY; MIIKKULAINEN, 2002).

O maior desafio dessa estratégia é combinar duas redes neurais durante o crossover sem produzir uma rede defeituosa, já que as topologias dos pais podem não ser diretamente compatíveis para a troca de neurônios e conexões. Devido a isso, NEAT

Figura 3.3: (a) Representação do cromossomo de um indivíduo em uma população de NEAT. O primeiro número em cada gene é o marcador histórico usado para identificar as transformações estruturais. A segunda informação é a conexão entre nós. "DIS" indica que o gene está desativado e é ignorado. (b) Representação dos dois tipos de mutação estrutural de NEAT. A primeira adiciona uma conexão com peso aleatório entre os nós 3 e 5 e recebe o novo marcador histórico. A segunda adiciona o nó 6 entre os nós 3 e 4, criando novas conexões e desativando a conexão entre 3 e 4. A conexão entre 3 e 6 herda o peso da antiga conexão entre 3 e 4, e a conexão entre 6 e 4 recebe um peso aleatório.

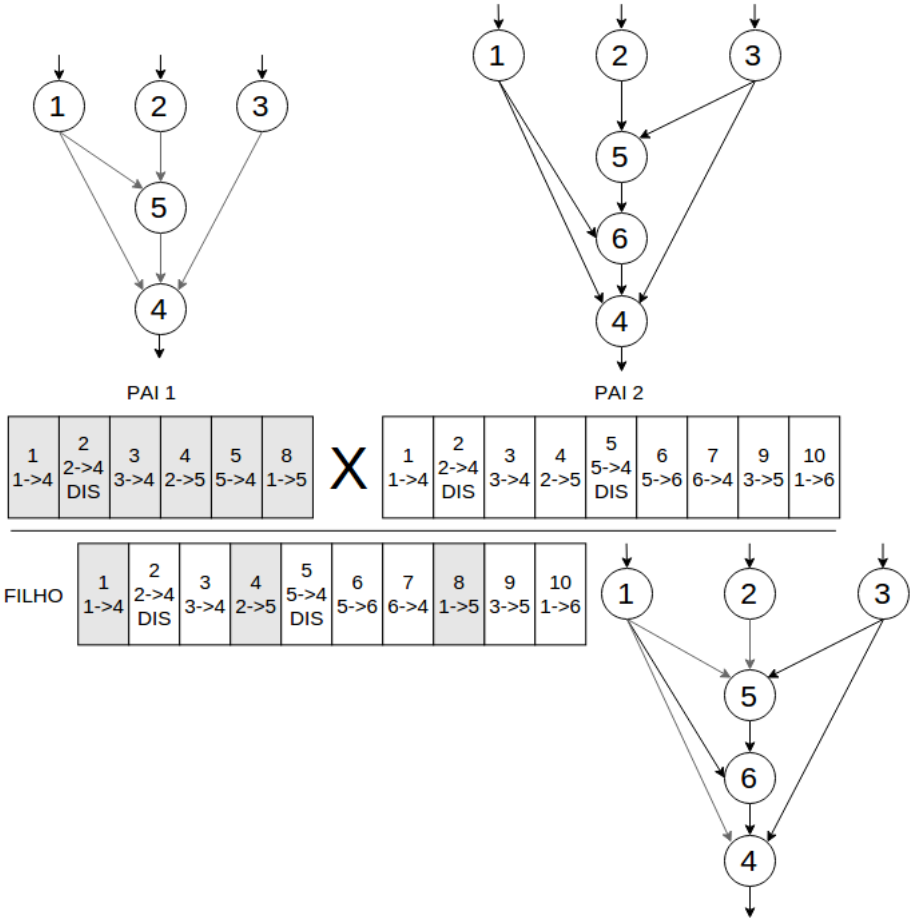


Fonte: (STANLEY; MIIKKULAINEN, 2002)

faz uso de marcadores históricos, um valor numérico atribuído a cada nova estrutura que aparecer ao longo do processo evolutivo que é transferido sem mudanças durante o crossover. Eles permitem ao método alinhar perfeitamente as mesmas porções das topologias de duas redes neurais diferentes, originando uma nova rede neural funcional que respeita a organização de seus pais (Figura 3.4). Os marcadores históricos são atribuídos ao fim de cada geração, de modo que caso duas novas estruturas idênticas apareçam em redes diferentes seus marcadores recebam os mesmos valores (STANLEY; MIIKKULAINEN, 2002).

O último problema da implementação de NEAT é que adicionar novas estruturas às redes neurais já existentes sem mais ajustes normalmente é prejudicial a elas. Nesse caso, redes neurais receberiam novos neurônios e conexões que produziriam resultados

Figura 3.4: Exemplo de crossover em NEAT entre dois indivíduos. Os genes dos pais são alinhados pelo marcador histórico para evitar inconsistências estruturais.



Fonte: (STANLEY; MIIKKULAINEN, 2002)

ruins de imediato e acabariam sendo excluídas da formação de novas populações, mesmo que suas novidades sejam benéficas a longo prazo. Por isso, NEAT utiliza especiação, também conhecido como nicho, uma técnica que agrupa os indivíduos pela sua similaridade estrutural (utilizando os marcadores históricos) e promove a competição entre eles em vez da população geral. Dessa forma, as redes têm tempo para se ajustar, não sendo eliminadas assim que surgem (STANLEY; MIIKKULAINEN, 2002).

NEAT é uma ferramenta poderosa para a evolução artifical de redes neurais e, através de experimentos, demonstrou-se mais eficiente que outros métodos neuroevolutivos e que evoluir a topologia em conjunto com os pesos pode ser uma grande vantagem, otimizando e complexificando soluções simultaneamente (STANLEY; MIIKKULAINEN, 2002).

3.4 Resumo do capítulo

Neste capítulo foram introduzidos diferentes métodos computacionais de propósito geral que podem ser aplicados a problemas da bioinformática. Foram vistos métodos de agrupamento de dados, em detalhe o agrupamento hierárquico, capaz de distribuir os dados em um número não pré-estabelecido de grupos. Foram descritos o funcionamento e operadores básicos de algoritmos genéticos. Com mais enfoque, foram vistas diferentes estratégias de treinamento de redes neurais, destacando-se `MLP` com *error back-propagation* e `NEAT`. O próximo capítulo faz a junção deste capítulo e do anterior ao propor um método que faz uso das técnicas computacionais aqui resumidas, aplicadas aos problemas biológicos apresentados.

4 PREDIÇÃO DA PREFERÊNCIA CONFORMACIONAL DE AMINOÁCIDOS

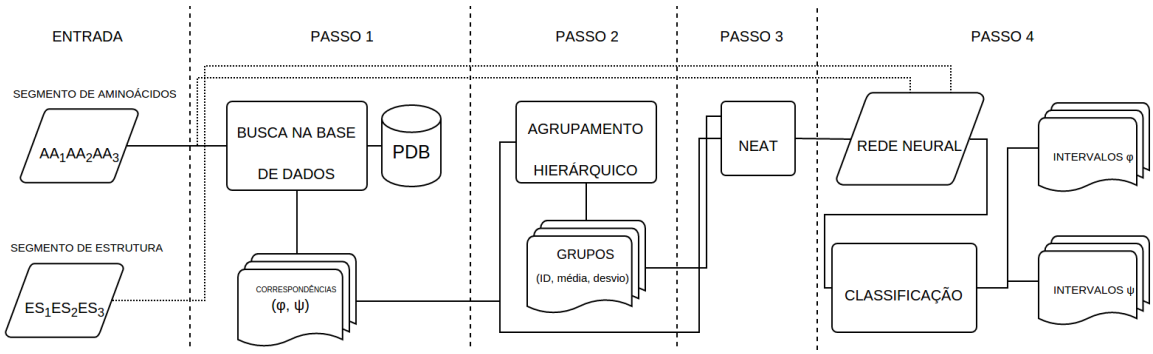
Como foi visto na Seção 2.3, o problema da predição da estrutura 3D de proteínas ainda é relevante e desafiador em diferentes áreas da ciência, como biologia, matemática e ciência da computação. Apesar do desenvolvimento de diferentes técnicas computacionais, não há ainda uma ferramenta com soluções suficientemente satisfatórias. O que se tem visto nas últimas décadas é que os avanços mais significativos foram obtidos por métodos que fazem uso de informações de bases de dados, mas eles ainda contam com desafios como a representação computacional da estrutura de proteínas, recuperação de modelos experimentais e desenvolvimento de estratégias de busca.

Estes métodos podem se beneficiar de estratégias capazes de extrair e organizar informações das bases de dados de proteínas. Como foi descrito na Seção 2.1.2, as proteínas não são rígidas na natureza, alterando entre diferentes conformações. Essa flexibilidade é fundamental no entendimento da função das proteínas (BORNOT; ETCHEBEST; BREVERN, 2011; DOBSON, 2003), mas ela não é descrita em dados experimentais. O método aqui descrito se propõe a extrair informações do `PDB` (Seção 2.2) de forma a predizer a flexibilidade das posições dos resíduos de aminoácidos em uma proteína dada a sua vizinhança. Essa flexibilidade é representada na forma de intervalos de valores possíveis para os ângulos de torção ϕ e ψ e pode ser incorporada em estratégias de busca como forma de otimizar o espaço de busca.

A ideia deste trabalho é inspirada no método MOIRAE, que também tem como objetivo a criação de intervalos de ângulos de torção. Mas enquanto o MOIRAE busca simplesmente restringir o espaço de busca ao tentar criar intervalos que "capturem" o valor dos ângulos dos dados experimentais, o método que será apresentado divide o espaço de busca em diversos intervalos diferentes e atribui probabilidades a eles, de forma que certas regiões sejam favorecidas, mas ainda considerando outras conformações menos comuns. O método proposto neste trabalho ainda tenta implementar melhorias à metodologia do MOIRAE ao utilizar algoritmos como agrupamento hierárquico e `NEAT`, numa tentativa de se adequar aos diferentes e variados padrões que os dados podem assumir.

As próximas seções unem os Capítulos 2 e 3 ao mostrar como o uso de diferentes técnicas computacionais sobre dados de um banco de dados de proteínas pode servir para o aprendizado de características importantes sobre preferências conformacionais de resíduos de aminoácidos.

Figura 4.1: Representação do método proposto para um segmento de três aminoácidos



Fonte: O Autor

4.1 Apresentação do método

O método proposto neste trabalho utiliza informações estruturais do `PDB` para predizer intervalos de valores para os ângulos de torção ϕ e ψ. O objetivo é encontrar intervalos contidos no domínio de $-180°$ a $180°$ para cada aminoácido da sequência alvo a fim de restringir o espaço de busca conformacional em métodos de predição da estrutura 3D de proteínas. Tais intervalos de ângulos representam a flexibilidade conformacional de cada aminoácido analisado, e podem ser utilizados por diferentes métodos de predição baseados em conhecimento, onde técnicas de otimização e metaheurísticas como algoritmos genéticos (BORGUESAN et al., 2015; DORN et al., 2013) ou *simulated annealing* (SAKAE et al., 2011) podem ser desenvolvidas objetivando percorrer de forma mais eficiente o espaço de busca conformacional.

A Figura 4.1 apresenta os quatro passos básicos do método descrito.

1) Busca na base de dados

2) Agrupamento dos dados

3) Treinamento das redes neurais artificiais

4) Criação de intervalos

A entrada do método é uma sequência de aminoácidos e a sua estrutura secundária, que pode ser obtida com acurácia através de programas já existentes como o `Stride` (HEINIG; FRISHMAN, 2004) e `DSSP` (TOUW et al., 2015). A ideia por trás da proposta é predizer os valores esperados de ϕ e ψ de aminoácidos em uma sequência de aminoácidos utilizando dados a respeito das preferências conformacionais de cada aminoácido em

conjunto com sua vizinhança e estrutura secundária procedentes de estruturas de proteínas determinadas experimentalmente.

O método proposto inicia por dividir a sequência de aminoácidos alvo em segmentos consecutivos de comprimento 3 (Figura 4.2) e buscar por correspondências no PDB (Passo 1 na Figura 4.1). Essa segmentação é uma forma de padronizar os dados para os algoritmos mantendo a ideia de vizinhança de aminoácidos (quais aminoácidos aparecem juntos na sequência), o que por si só pode fornecer informações conformacionais úteis como mostra o trabalho da APL (BORGUESAN et al., 2015). Uma correspondência ocorre quando o segmento é encontrado na cadeia de aminoácidos de uma proteína diferente armazenada no PDB. Com a lista de proteínas que contém o segmento em questão, os seus arquivos PDB são baixados e a estrutura secundária delas é atribuída pelo Stride. Os ângulos de torção ϕ e ψ do segmento são calculados. Então, cada segmento possuirá uma lista de pares ϕ e ψ vindos das diferentes proteínas onde ele ocorre. Esses pares são então agrupados (Passo 2 na Figura 4.1) e, com cada ponto atribuído a um grupo, é criado um conjunto de dados composto de padrões representados pelos três aminoácidos do segmento, sua estrutura secundária e o grupo ao qual pertencem.

Esses dados são passados como entrada para o treinamento das redes neurais que utiliza NEAT. As redes neurais são treinadas para aprender como classificar os aminoácidos e sequências secundárias de um segmento nos grupos encontrados anteriormente (Passo 3 na Figura 4.1). Uma vez que o treinamento esteja completo, as redes neurais são capazes de classificar novas entradas com generalização. No último passo, a sequência de aminoácidos e estruturas secundárias alvo originais são submetidas às redes neurais correspondentes, que devolvem como saída a probabilidade delas pertencerem a cada um dos grupos. Neste ponto, temos a flexibilidade do resíduo de aminoácido na cadeia e a probabilidade desta flexibilidade.

Com essa informação são criados intervalos centrados nos valores médios dos grupos e limitados por mais e menos um desvio padrão e meio dos grupos (Passo 4 na Figura 4.1). A saída final do método é um conjunto de intervalos para os ângulos ϕ e ψ dos aminoácidos da sequência alvo. As próximas seções descrevem em detalhe cada um dos passos do método.

Figura 4.2: Construção dos segmentos de aminoácidos de comprimento 3 a partir de uma sequência alvo de comprimento n. Cada segmento é representante do seu aminoácido central, em negrito nesta figura. O primeiro e o último aminoácidos da sequência alvo não são considerados pelo método.

Segmento	$\mathbf{AA}_1$	$\mathbf{AA}_2$	$\mathbf{AA}_3$	$\mathbf{AA}_4$	$\mathbf{AA}_5$	**...**	$\mathbf{AA}_n$
S_1	AA_1	$\mathbf{AA}_2$	AA_3	-	-	...	-
S_2	-	AA_2	$\mathbf{AA}_3$	AA_4	-	...	-
S_3	-	-	AA_3	$\mathbf{AA}_4$	AA_5	...	-
...							
S_{n-2}	-	-	-	-	-	...	AA_n

Fonte: O Autor

4.2 Busca na base de dados

Para encontrar a informação necessária e limitar a extensão dos intervalos finais (ou seja, a flexibilidade do resíduo de aminoácido), a sequência de aminoácidos alvo é dividida em segmentos de comprimento 3 e é feita uma busca no `PDB` por ocorrências deles. Para uma sequência de comprimento n, são criados $n-2$ segmentos de comprimento 3. Esse processo de segmentação da sequência de aminoácidos é mostrado na Figura 4.2. Posteriormente, cada segmento fornecerá informações a respeito de seu aminoácido central, ou seja, não apenas o próprio aminoácido é considerado para a aprendizagem dos ângulos através da base de dados, mas também os seus vizinhos à esquerda e à direita. O primeiro e o último aminoácidos da sequência alvo acabam por não serem representados por faltar uma vizinhança a eles, mas isto não é um grande comprometimento pois os limites de uma sequência de aminoácidos são geralmente áreas de grande instabilidade (DORN; BURIOL; LAMB, 2013).

Para cada segmento é executado o *Protein BLAST* (ALTSCHUL et al., 1990), um programa que faz buscas em bancos de dados de proteínas, no caso o `PDB`, a procura de ocorrências do segmento em proteínas do banco de dados. Ele retorna uma lista de todas as proteínas com correspondência (ou seja, que contém o segmento), e os arquivos `PDB` são baixados e analisados por meio de `Stride`, retornando as estruturas secundárias e ângulos de torção experimentais ϕ e ψ para cada aminoácido. É necessário observar que uma correspondência não necessariamente será exatamente idêntica ao segmento buscado, pois o *Protein BLAST* leva em consideração substituições de aminoácidos que podem ocorrer preservando as propriedades físico-químicas. O método proposto apenas fixa que o aminoácido central do segmento obrigatoriamente deverá ser o mesmo em

suas correspondências. Se uma correspondência ocorreu no começo ou no final de uma proteína, ela é ignorada porque essas regiões são instáveis. Também são ignoradas correspondências encontradas em cadeias de aminoácidos de mesma família da que está sendo testada. Todos os segmentos da sequência alvo original são agrupados a listas formadas de segmentos iguais provenientes das proteínas do PDB, e estes dados são salvos.

4.3 Agrupamento dos dados

Do último passo se obteve, para cada segmento de 3 aminoácidos da sequência de aminoácidos alvo, uma lista de todas as correspondências de proteínas do PDB e respectivas estruturas secundárias e valores experimentais de ϕ e ψ do aminoácido central dos segmentos. O método APL mostra que a utilização do aminoácido central de segmentos permite a extração de informações relevantes da base de dados a respeito das opções conformacionais dos aminoácidos (BORGUESAN et al., 2015). O próximo passo é organizar esses dados pelos valores dos ângulos na forma de grupos, a fim de generalizar a informação mais tarde.

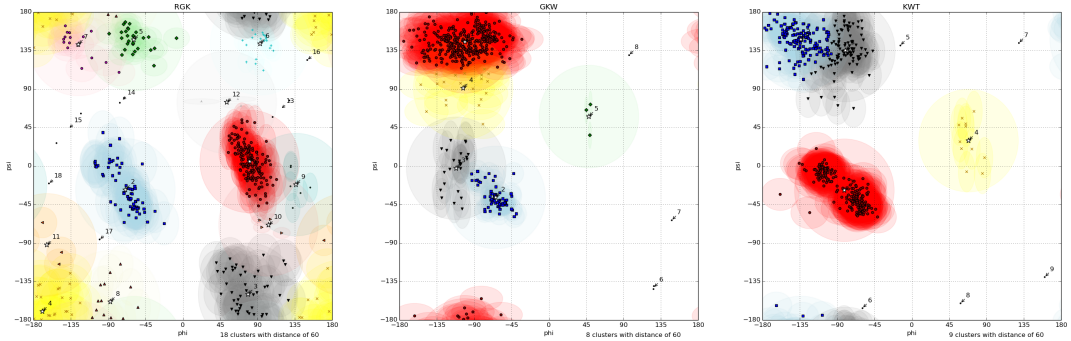
Os pares ϕ e ψ são tratados como pontos bidimensionais com valores de $-180°$ a $180°$ em ambos os eixos, com ϕ sendo o eixo X e ψ sendo o eixo Y. Devido a natureza cíclica dos ângulos, é preciso considerar que os valores das extremidades são os mesmos, ou seja, $-180°$ e $180°$ são duas representações do mesmo ângulo. A partir de agora esta propriedade será considerada em todos os cálculos, incluindo no cálculo da distância entre dois pontos.

Como a busca na base de dados retorna um número desconhecido de pontos distribuídos em formas diferentes para cada aminoácido (BORGUESAN et al., 2015; NIASSERVER, 2016), não é possível adivinhar o número de grupos necessários para melhor encaixar os dados. Por isso, agrupamento hierárquico (descrito na Seção 3.1.1) foi escolhido pela sua capacidade de criar grupos sem a necessidade de se saber a quantidade final. Para este método o limiar de distância entre grupos selecionado foi $60°$, já que este valor normalmente compreende a distância entre diferentes regiões estruturais (HOVMOLLER; OHLSON, 2002) e produziu, em geral, os melhores resultados em inspeção visual. A métrica de distância é a distância euclidiana, modificada para considerar a propriedade cíclica dos valores tratados.

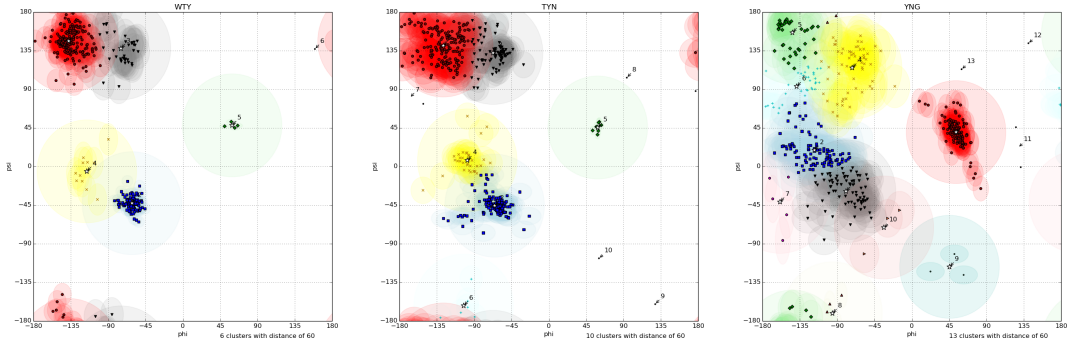
O algoritmo de agrupamento é executado para todos os segmentos da sequência de aminoácidos alvo. Uma sequência de comprimento n e $n-2$ segmentos terá, portanto,

Figura 4.3: Exemplos de grupos hierárquicos para os segmentos de aminoácidos de comprimento 3 da proteína com código PDB 1K43. Cada ponto é um par ϕ, ψ de uma correspondência do PDB. Cada grupo é representado por uma cor diferente, e a sombra mais clara ao redor dos pontos é o desvio padrão do grupo.

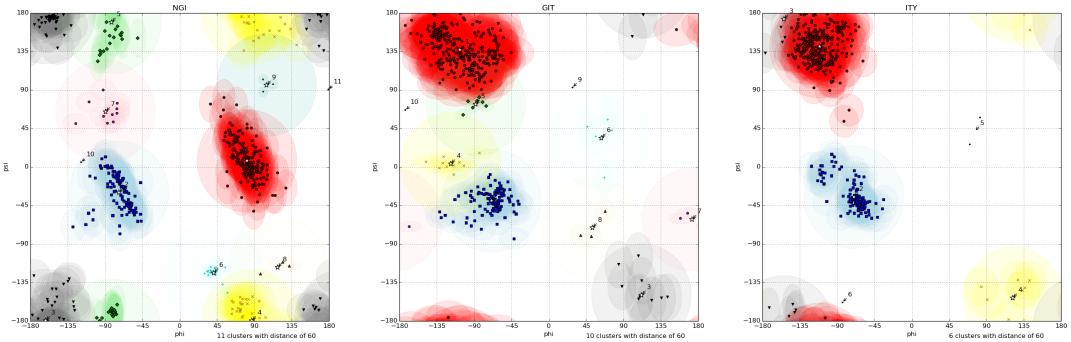
(a) RGK (b) GKW (c) KWT



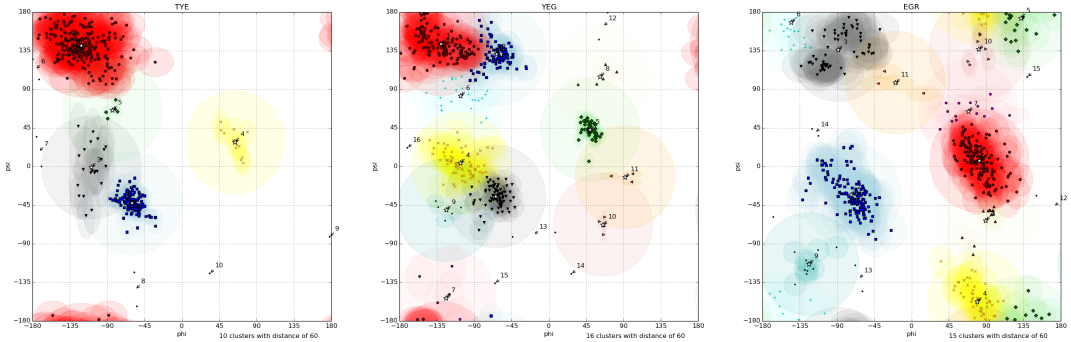
(d) WTY (e) TYN (f) YNG



(g) NGI (h) GIT (i) ITY



(j) TYE (k) YEG (l) EGR



Fonte: O Autor

$n-2$ diferentes conjuntos de grupos. Cada ponto recebe um valor inteiro correspondendo a um grupo. Grupos com muito poucos pontos são considerados ruído e removidos dos próximos passos do método. A Figura 4.3 ilustra todos os conjuntos de grupos obtidos para a sequência de aminoácidos da proteína de código `PDB` 1K43 (PASTOR et al., 2002) (`RGKWTYNGITYEGR`), correspondendo aos segmentos de 3 aminoácidos. Cada ponto dos gráficos é um par ϕ, ψ de uma das correspondências do segmento. Como pode ser percebido ao observar-se as regiões de borda dos gráficos, a ciclicidade se faz presente. Na Figura do segmento `RGK`, por exemplo, o grupo amarelo pode ser visto nos quatro cantos do gráfico.

4.4 Treinamento das redes neurais artificiais

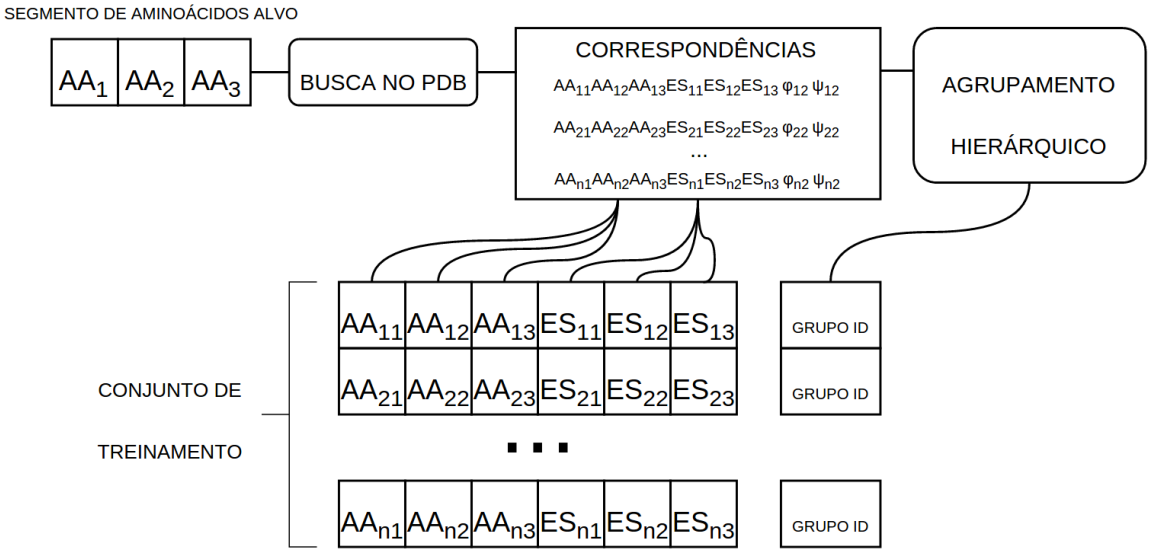
Uma vez que os segmentos da base de dados estão classificados em grupos, já é possível treinar redes neurais para aprender a classificar novas entradas das quais apenas o três aminoácidos e suas estruturas secundárias são conhecidos. Para isso são necessárias redes neurais diferentes para cada segmento, já que eles possuem conjuntos diferentes de pontos para serem analisados.

Um método tradicional seria utilizar redes neurais com um algoritmo de aprendizagem como *back-propagation*, mas desse jeito apenas os pesos das conexões entre os nós da rede seriam modificados, deixando a cargo do desenvolvedor definir uma topologia, ou seja, número de neurônios, camadas e conexões para a rede, o que pode ser um problema nesse caso em que cada segmento de aminoácidos possui um conjunto de treinamento particular. Esta foi a estratégia implementada pelo método MOIRAE, que utiliza uma mesma topologia `MLP` com uma única camada oculta e cinco neurônios ocultos (DORN; BURIOL; LAMB, 2013). Já o método descrito adotou a técnica de evolução de redes neurais `NEAT` (Seção 3.3.3) para contornar essas questões.

Uma sequência de aminoácidos alvo de comprimento n origina $n-2$ segmentos de tamanho 3, cada um correspondendo ao seu aminoácido central e com seu conjunto de grupos produzido no último passo. Então são evoluídas $n-2$ redes neurais capazes de classificar estes segmentos em grupos utilizando informações extraídas da base de dados.

Tais redes neurais possuem sete nós na camada de entrada, que recebem os sinais dos três aminoácidos dos segmentos, os sinais das três estruturas secundárias e um sinal para o *bias* constante em $1,0$. A Figura 4.4 ilustra a criação de um conjunto de treinamento usado para treinar as redes neurais. Para evitar o problema de sobre-ajuste,

Figura 4.4: Diagrama da criação do conjunto de treinamento para evolução das redes neurais



Fonte: O Autor

apenas 80% dos dados do último passo são utilizados para a construção do conjunto de treinamento, e os 20% restantes são usados como conjunto de teste para avaliação das redes neurais (VANHOUCKE; CHAKRABORTY, 2016). O conjunto de treinamento é composto pelos segmentos de aminoácidos e estruturas secundárias associadas ao grupo ao qual foram classificados no passo anterior. As redes neurais devem aprender a predizer como saída o valor correto do grupo de acordo com a entrada.

Os grupos são representados com a codificação *one-hot* (VANHOUCKE; CHAKRABORTY, 2016). Dessa forma, se uma entrada precisa ser classificada entre n grupos diferentes, cada grupo é representado por um vetor de comprimento n com todas as posições valendo 0, exceto a posição no índice correspondente ao identificador numérico do grupo, que recebe o valor 1. Por conta disso a rede neural capaz de calcular essa classificação necessita de n neurônios na camada de saída, cada um correspondente a uma das posições do vetor de grupos (VANHOUCKE; CHAKRABORTY, 2016). Para se comparar o vetor de grupos do conjunto de treinamento com o vetor de números reais obtido como saída da rede neural, estes valores da saída precisam passar por uma função *softmax* (Equação 4.1) que os escala entre $0,0$ e $1,0$. Estes novos valores podem ser vistos como a probabilidade da entrada da rede neural pertencer a cada um dos grupos (VANHOUCKE; CHAKRABORTY, 2016).

$$\mathbf{SOFTMAX}(X) = \mathrm{e}^{X_i} / \sum_{i=1}^{n} \mathrm{e}^{X_i} \tag{4.1}$$

onde X é um vetor de comprimento n.

É importante notar que redes neurais operam apenas com entradas e saídas numéricas, enquanto o método apresentado usa entradas (aminoácidos, estruturas secundárias) e saídas (grupos) simbólicas. Como já foi abordado, os grupos podem ser representados por valores inteiros codificados em *one-hot*, então não provocam um problema de representação. A representação dos aminoácidos e estruturas secundárias, contudo, apresenta vários desafios. A utilização da codificação *one-hot* é inviável pois acabaria por aumentar demais a quantidade de nós de entrada necessários, o que reduziria a eficiência de `NEAT`. A ordem dos valores de entrada precisa também ser preservada para o aprendizado dos padrões pelas vizinhanças. Por fim, simplesmente atribuir um valor numérico arbitrário a cada aminoácido e estrutura secundária poderia afetar o processo de aprendizado ao inserir correlações e ordenamentos que não existem na natureza.

Para lidar com este problema, foram buscadas propriedades dos aminoácidos e estruturas secundárias que permitissem fazer sentido de uma conversão numérica. Para os aminoácidos, foi escolhido o parâmetro hidrofobicidade pi (Tabela 2.2), que tende a manter próximos os aminoácidos que compartilham dos mesmos grupos naturais, e para as estruturas secundárias foram atribuídos valores no intervalo de 0 a 1 que as ordenam de acordo com a similaridade e complexidade estrutural, agrupando os sete tipos de estruturas secundárias possíveis no `Stride`. Resíduos hidrofóbicos com aparição periódica já foram utilizados para a predição da estrutura secundária de proteínas já que a hidrofobicidade afeta a estabilidade da estrutura secundária (HUANG; CHEN, 2013).

$$\mathbf{CROSS-ENTROPY}(A, B) = \sum_{i=1}^{n} (B_i \times \log_{\mathrm{e}} A_i) \tag{4.2}$$

onde A e B são vetores de mesmo comprimento n, A sendo a saída da rede neural aplicada à função *softmax* e B a codificação *one-hot* do conjunto de treinamento.

Como `NEAT` usa algoritmos genéticos (Seção 3.2) para evoluir redes neurais, ele precisa de uma função de aptidão capaz de avaliar e ordenar cada indivíduo por um valor. Para este método, a aptidão é a perda da *cross entropy* (entropia cruzada em português) (Equação 4.2) da saída da rede neural aplicada à função *softmax* e da codificação da classificação original no conjunto de treinamento. O indivíduo com menor perda, ou seja, a rede neural que classifica a maior quantidade de entradas corretamente, é seleci-

onado (VANHOUCKE; CHAKRABORTY, 2016). Para cada segmento da sequência de aminoácidos alvo de comprimento n, um ciclo completo de `NEAT` é computado usando seu próprio conjunto de treinamento, resultando em $n - 2$ redes neurais diferentes e independentes. A Figura 4.5 ilustra as redes obtidas para os grupos da Figura 4.3. Isso resolve o problema original de conhecer previamente a melhor topologia para cada um dos segmentos, já que `NEAT` permite encontrar topologias e pesos de conexões diferentes sem intervenção humana. Por questões de controle e comparação, exatamente o mesmo processo de treinamento foi realizado também com redes neurais `MLP` e o algoritmo de aprendizagem *back-propagation* (Seção 3.3.1).

4.5 Criação dos intervalos

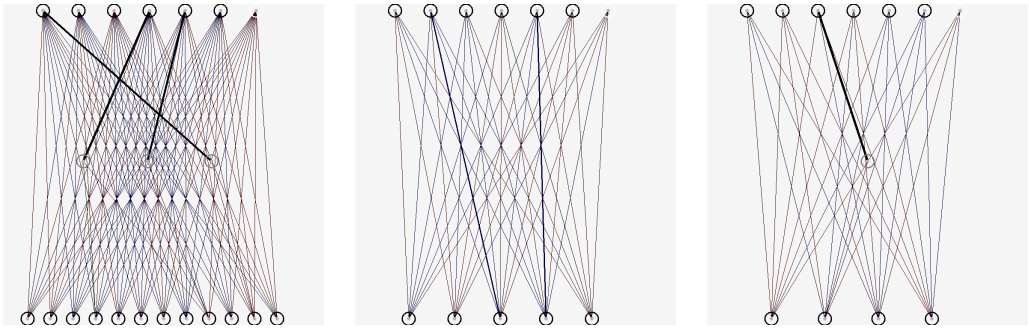
O último passo do método trata da criação de intervalos representando a flexibilidade de um aminoácido. Cada aminoácido passa a ser visto como um conjunto de intervalos de busca para os ângulos de torção ϕ e ψ com uma probabilidade associada. Para tanto, basta fornecer como entrada para as redes neurais já treinadas os aminoácidos e estruturas secundárias de cada um dos segmentos de comprimento 3. Cada segmento possui uma rede neural correspondente treinada no passo anterior, e a saída delas é um vetor com as probabilidades do segmento de entrada pertencer a cada um dos grupos.

Para a criação de intervalos para os ângulos ϕ de um aminoácido em particular, a média dos valores de ϕ é extraída de cada grupo. Esses valores são o centro dos intervalos. Os limites são apenas os valores das médias acrescidos e decrescidos de um desvio padrão e meio do grupo, respeitando a ciclicidade dos ângulos. A probabilidade do ângulo sendo predito estar em um dos intervalos é definida pela probabilidade do grupo que o originou, obtida das saídas das redes neurais. O processo para criação dos intervalos para o ângulo ψ é exatamente o mesmo, mas utilizando a coordenada ψ dos grupos.

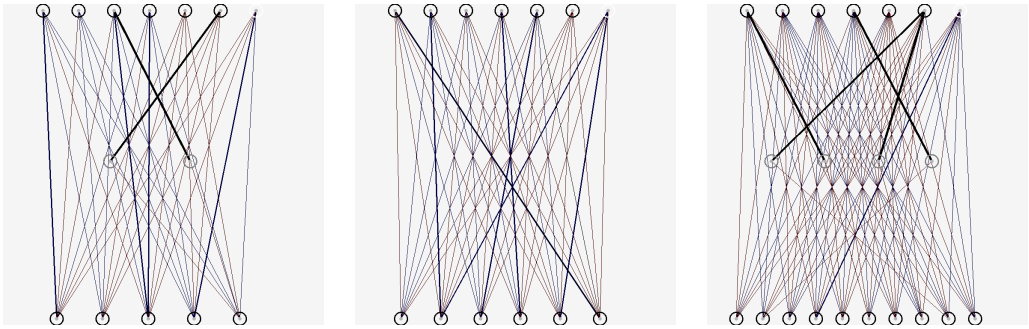
A repetição desses passos para cada um dos segmentos de comprimento 3 da sequência de aminoácidos alvo produzirá $n - 2$ conjuntos de intervalos para o ângulo ϕ e $n - 2$ conjuntos de intervalos para o ângulo ψ, correspondendo a todos os aminoácidos da sequência exceto o primeiro e o último. Isso não é um problema, como já foi abordado, pois as regiões extremas da cadeia são muito instáveis. Esses dois aminoácidos recebem um único intervalo para cada ângulo de torção, de $-180°$ a $180°$, com probabilidade definida como $1, 0$. A Figura 4.6 traz os intervalos finais dos ângulos de torção ϕ e ψ obtidos com este método para a proteína com código `PDB` 1K43, cujos grupos e redes

Figura 4.5: Exemplos de redes neurais evoluídas com NEAT para os segmentos de 3 aminoácidos da Figura 4.3. A orientação é camada de entrada na parte superior e camada de saída na parte inferior. A espessura das linhas é proporcional ao valor absoluto dos pesos das conexões. O nó branco representa o *bias*.

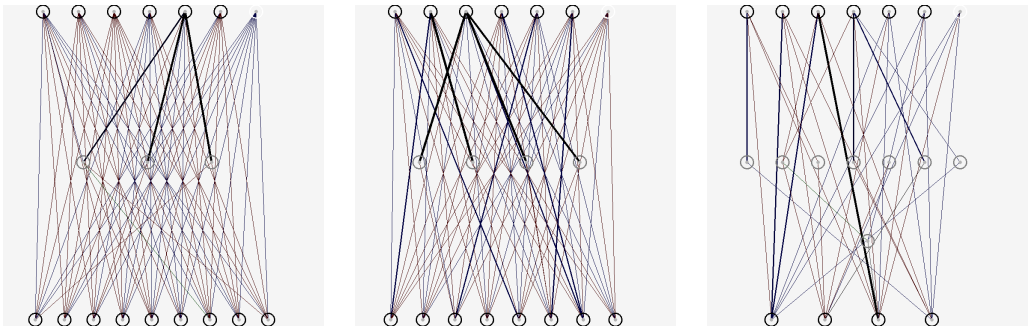
(a) RGK (b) GKW (c) KWT



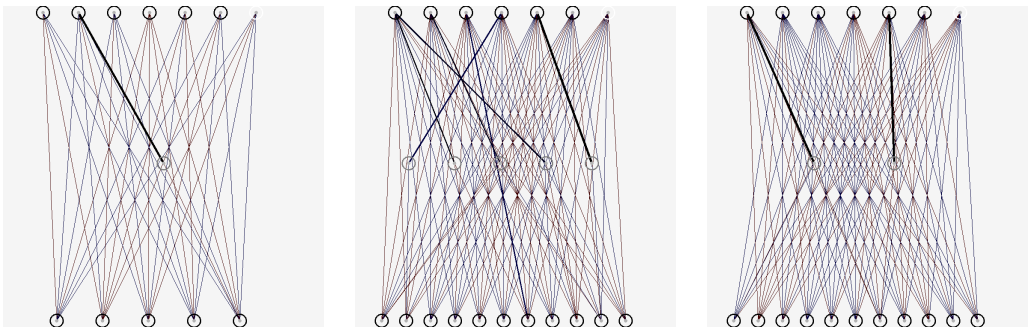
(d) WTY (e) TYN (f) YNG



(g) NGI (h) GIT (i) ITY



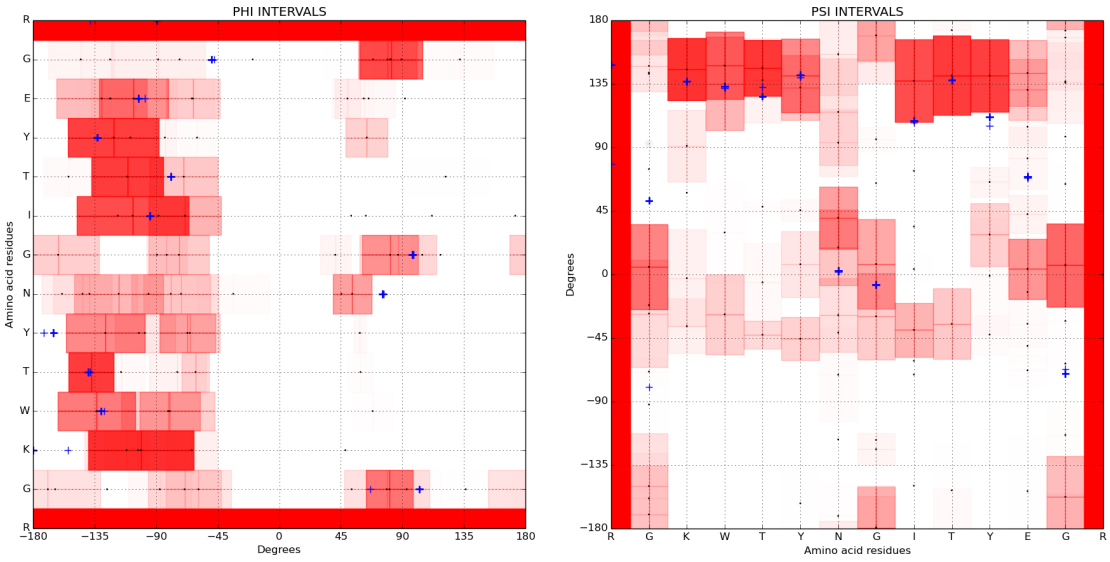
(j) TYE (k) YEG (l) EGR



Fonte: O Autor

Figura 4.6: Representação gráfica de um conjunto de intervalos representando a flexibilidade dos ângulos ϕ e ψ obtidos com os grupos da Figura 4.3 e redes neurais da Figura 4.5 para a proteína com código PDB 1K43. Neste modelo de visualização, quanto mais escura a cor de um intervalo, maior a sua probabilidade associada. Cada linha (para ϕ) ou coluna (para ψ) representa um aminoácido. O domínio dos valores é de $-180°$ até $180°$, preservando a ciclicidade nos limites. Pequenos pontos pretos indicam o centro dos intervalos. Cruzes azuis são os valores experimentais dos ângulos ϕ e ψ da proteína com código PDB 1K43.

(a) Intervalos ϕ (b) Intervalos ψ



Fonte: O Autor

neurais já foram mostrados nas Figuras 4.3 e 4.5 respectivamente.

Os intervalos podem ser vistos como uma estimativa da flexibilidade dos resíduos de aminoácidos ao determinar regiões para a ocorrência dos ângulos ϕ e ψ. É importante lembrar que as proteínas, como visto na Seção 2.1.2, não ficam imóveis, por isso considerar o grau de liberdade da posição dos resíduos de aminoácidos pode auxiliar no problema da predição da estrutura tridimensional de proteínas ao não se buscar apenas um único valor para os pares de ângulos de torção. Ao mesmo tempo, métodos como o MOIRAE ao criar um único intervalo para cada aminoácido correm o risco de eliminar áreas menos favorecidas mas onde ainda pode haver ocorrências dos ângulos (DORN; BURIOL; LAMB, 2013), o que é contornado aqui com o uso de mais de um intervalo com propriedades associadas. Os intervalos podem ser incorporados em métodos de busca usados para a predição da estrutura 3D de proteínas, servindo como guia e limitando consideravelmente o espaço de busca, estratégia já testada pelo trabalho APL (BORGUESAN et al., 2015).

4.6 Resumo do capítulo

Neste capítulo foi apresentado um método para estimar a flexibilidade conformacional de aminoácidos em proteínas a partir de dados extraídos do PDB. Foram detalhados os seus quatro passos básicos, a busca no banco de dados, agrupamento dos dados, treinamento das redes neurais e criação dos intervalos. Também foi comentada a relação com o método MOIRAE. O próximo capítulo explica a implementação do método e analisa os resultados obtidos.

5 EXPERIMENTOS E RESULTADOS

O método apresentado no Capítulo 4 foi implementado e testado a fim de demonstrar a sua funcionalidade. A linguagem de programação adotada foi `Python 2.7` (PYTHON, 2016) devido a existência de várias bibliotecas científicas para aprendizado de máquina e bioinformática e compatibilidade com outros programas desenvolvidos no *Structural Bioinformatics and Computational Biology Lab* do Instituto de Informática da UFRGS com a mesma linguagem. Para suporte matemático foi usada a biblioteca `NumPy` (WALT; COLBERT; VAROQUAUX, 2011). Os experimentos foram executados em um servidor 1 x HP ML350E G8: Intel Xeon E5-2407, 2 CPUs, 2.2Ghz, 8 *cores*, 32G, 2TB.

O método foi testado com a sequência de aminoácidos de 25 proteínas distintas presentes no `PDB`, detalhadas na Tabela 5.1. Foi selecionado um grupo de proteínas com o objetivo de representar diferentes comprimentos de sequências, conteúdos de estrutura secundária e método experimental para determinação da estrutura terciária. São elas 1AB1 (YAMANO; HEO; TEETER, 1997), 1ACW (BLANC et al., 1996), 1DFN (HILL et al., 1991), 1K43 (PASTOR et al., 2002), 1L2Y (NEIDIGH; FESINMEYER; ANDERSEN, 2002), 1ROP (BANNER; KOKKINIDIS; TSERNOGLOU, 1987), 1ZDC (STAROVASNIK M.A., 1997), 2PMR (BONANNO et al.,), 1WQC (CHAGOT et al., 2005), 2MTW (CIFUENTES et al.,), 3P7K, 2P81 (RELIGA et al., 2007), 3V1A (DER B.S., 2012), 1ENH (CLARKE et al., 1994), 2F4K (KUBELKA et al., 2006), 2P6J (SHAH et al., 2007), 1AIL (LIU et al., 1997), 2MR9 (NOWICKA et al., 2015), 2JUC (BONET; RAMIREZ-ESPAIN; MACIAS, 2008), 1D5Q (VITA et al., 1999), 1Q2K (CAI et al., 2004), 2P5K (GARNETT et al., 2007), 1CRN (TEETER, 1984), 1UTG (MORIZE et al., 1987) e 2MM8 (DOBROVOLSKA et al.,).

O primeiro passo, a busca de dados com o *software Protein BLAST*, foi implementada com a biblioteca `BioPython` (COCK et al., 2009) que já incorpora a busca de sequências de aminoácidos no `PDB`. Os parâmetros de busca utilizados nessa etapa estão descritos na Tabela 5.2. As correspondências de cada segmento de comprimento 3 aminoácidos foram salvas em arquivos no formato XML (*eXtensible Markup Language*). Eles são lidos também com a biblioteca `BioPython`, de onde são extraídos os códigos `PDB` das proteínas que contém as correspondências encontradas. Como visto na Seção 4.2, correspondências ocorrendo no começo e no final de proteínas foram descartadas, bem como as que não mantém o aminoácido central do segmento. Também foram desconsi-

Tabela 5.1: Descrição das proteínas utilizadas para avaliação do método

Código da proteína	*Tamanho da sequência*	*Conteúdo da estrutura secundária*	*Método experimental*	*Número de modelos conformacionais*
1AB1	46	Duas hélices e uma folha	Raio-X	1
1ACW	29	Uma hélice e uma folha	NMR	25
1DFN	30	Duas folhas	Raio-X	1
1K43	14	Uma folha	NMR	10
1L2Y	20	Duas hélices	NMR	38
1ROP	63	Duas hélices	Raio-X	1
1ZDC	35	Duas hélices	NMR	24
2PMR	87	Três hélices	Raio-X	1
1WQC	26	Duas hélices	NMR	30
2MTW	20	Uma hélice	NMR	1
3P7K	45	Uma hélice	Raio-X	1
2P81	44	Duas hélices	NMR	25
3V1A	48	Duas hélices	Raio-X	1
1ENH	54	Três hélices	Raio-X	1
2F4K	35	Quatro hélices	Raio-X	1
2P6J	52	Três hélices	NMR	43
1AIL	73	Três hélices	Raio-X	1
2MR9	44	Três hélices	NMR	10
2JUC	59	Quatro hélices	NMR	15
1D5Q	27	Uma hélice e uma folha	NMR	1
1Q2K	31	Uma hélice e uma folha	NMR	21
2P5K	64	Três hélices e uma folha	Raio-X	1
1CRN	46	Uma hélice e uma folha	Raio-X	1
1UTG	70	Cinco hélices	Raio-X	1
2MM8	99	Uma hélice	NMR	5

Fonte: O Autor

deradas as ocorrências em proteínas de sequência idêntica a buscada pelo método, já que seriam a própria proteína alvo.

Com as listas dos códigos `PDB` das correspondências, os arquivos das estruturas das proteínas correspondentes foram baixados do `PDB`. O programa `Stride` foi usado para calcular as estruturas secundárias e os ângulos de torção ϕ e ψ destas proteínas, e a informação foi salva.

Para o segundo passo, o agrupamento de dados, foi usada a implementação de agrupamento hierárquico (Seção 3.1.1) da biblioteca `SciPy` (JONES et al., 2001–). O parâmetro limiar de distância entre os grupos adotado foi de $60°$, como discutido na Seção 4.3. Exemplos de grupos gerados podem ser vistos na Figura 4.3.

Tabela 5.2: Parâmetros de busca por correspondências dos segmentos de aminoácidos de comprimento 3 no *BLAST* utilizados na implementação. A matriz de substituição atribui uma pontuação para o alinhamento de qualquer possível par de resíduos. O custo por lacuna são referentes à matriz e aumentar o seu custo implica e menos alinhamentos com lacunas. O tamanho de palavra determina o quão similares duas sequências devem ser para que se considere compará-las. O valor esperado determina o limiar da significância estatística para reportar correspondências na base de dados.

Parâmetro	*Valor*
Programa	BLAST Protein
Base de dados	`PDB`
Filtro	Nenhum
Matriz de substituição	PAM30
Custo de lacuna	9 1
Valor esperado	200000
Código genético	1
Tamanho da lista de correspondências	500
Limiar	11
Tamanho de palavra	1
Formato de arquivo	XML

Fonte: O Autor

Para a etapa de treinamento de redes neurais (Seção 4.4) foram implementados dois métodos distintos para fins de comparação. O primeiro foi `NEAT`, descrito na Seção 3.3.3, utilizando a biblioteca `MultiNeat` (CHERVENSKI, 2012–). Cada rede tinha 7 nós de entrada e n neurônios de saída, n sendo o número de grupos associado ao segmento em questão. As populações `NEAT` foram compostas por 100 indivíduos cada e foram executadas por 100 gerações evolutivas. Devido à aleatoriedade envolvida no método, foram executados 10 treinamentos independentes para cada rede neural artificial, e o melhor resultado das 10 foi escolhido. Exemplos de redes neurais artificiais criadas podem ser vistos na Figura 4.5. A avaliação das redes foi feita sobre o conjunto de testes a fim de reduzir as chances de haver sobreajuste. Um detalhamento das topologias médias encontradas para cada uma das proteínas usadas como teste pode ser visto nas Tabelas 5.3 e 5.4.

Além de `NEAT`, também foram treinadas redes `MLP` com *error back-propagation*. Para a implementação foi usada a biblioteca `PyBrain` (SCHAUL et al., 2010). Como para `NEAT`, cada rede neural contém 7 nós de entrada e n neurônios de saída, n sendo o número de grupos do segmento. Mas como em `MLP` a estrutura da rede precisa ser definida antes do treinamento, também foi definido que as redes neurais teriam uma única camada oculta com 5 neurônios, e as camadas são totalmente conectadas. Essa topologia foi a escolhida por ser análoga à usada no método MOIRAE. O treinamento das redes

Tabela 5.3: Análise das redes neurais evoluídas com NEAT. Os valores apresentados são as médias de todas as redes criadas para cada proteína testada, com os desvios padrão entre parênteses. A taxa de sucesso é a porcentagem do conjunto de teste classificada corretamente pela rede. Os resultados foram obtidos após a execução do algoritmo NEAT 10 vezes para cada segmento, selecionando o resultado com melhor taxa de sucesso. Cada rodada durou 100 gerações com 100 indivíduos.

Proteína	*N° total de nós*	*N° nós ocultos*	*N° conexões*	*Profundidade (camadas)*	*Taxa de sucesso (prob.)*	*Tempo de evolução (s)*
1AB1	15.13 (3.52)	1.84 (1.59)	46.29 (18.52)	1.86 (0.55)	0.75 (0.16)	303.62 (43.56)
1ACW	14.56 (2.17)	1.3 (1.18)	45.41 (11.47)	1.81 (0.67)	0.72 (0.16)	301.64 (20.72)
1DFN	14.79 (3.26)	1.64 (1.67)	44.93 (17.47)	2.21 (2.7)	0.78 (0.19)	298.92 (37.16)
1K43	17.25 (3.96)	2.75 (2.2)	56.17 (20.86)	2.17 (0.8)	0.73 (0.19)	316.18 (24.66)
1L2Y	14.94 (2.41)	1.56 (1.26)	46.89 (16.93)	4.33 (5.25)	0.78 (0.18)	299.39 (47.68)

Fonte: O Autor

neurais MLP foi executado por 100 épocas de treinamento cada, e as redes neurais foram avaliadas sobre o conjunto de treinamento. Como foi feito para NEAT, cada rede neural foi treinada independentemente 10 vezes e a melhor foi selecionada ao final. A avaliação final das redes neurais MLP obtidas pode ser vista na Tabela 5.5.

Comparando-se os dois modelos percebe-se que ambos atingiram níveis de sucesso semelhantes. Uma taxa de sucesso perfeita na classificação não era esperada para este problema uma vez que as entradas das redes neurais são os aminoácidos e estruturas secundárias dos segmentos, e não há homogeneidade perfeita desses elementos dentro dos grupos. O que se pretende é justamente tentar generalizar e abstrair essa informação na forma de intervalos com diferentes probabilidades, de forma que áreas mais propensas a ocorrência dos valores ângulos de torção sejam privilegiadas sem a eliminação total de áreas menos favorecidas, onde a solução ainda possa estar.

Outra comparação que pode ser feita entre os dados das Tabelas 5.3 e 5.4 e a Tabela 5.5 é topológica. Todas as redes MLP possuem cinco neurônios na camada oculta. Os dados das redes neurais evoluídas com NEAT indicam que com um número menor de nós ainda é possível chegar aos mesmos resultados, mas com uma estrutura mais simples. Os tempos dos dois algoritmos revelam que a evolução de NEAT é mais custosa que o *error back-propagation* de MLP, mas essa desvantagem já era esperada dada a comple-

Tabela 5.4: Continuação da análise das redes neurais evoluídas com NEAT da Tabela 5.3

Proteína	*N° total de nós*	*N° nós ocultos*	*N° conexões*	*Profundidade (camadas)*	*Taxa de sucesso (prob)*	*Tempo de evolução (s)*
1ROP	15.73 (2.52)	2.56 (2.04)	46.15 (10.37)	2.06 (0.6)	0.8 (0.12)	322.89 (44.5)
1ZDC	13.72 (2.91)	1.22 (1.43)	40 (14.79)	1.69 (0.68)	0.81 (0.13)	219.06 (126.15)
2PMR	15.08 (3.16)	1.96 (2.31)	45.37 (14.09)	2.63 (3.28)	0.8 (0.15)	317.84 (61.29)
1WQC	14.75 (2.37)	1.92 (1.63)	43.29 (12.35)	3.71 (4.67)	0.83 (0.1)	305.67 (37.46)
2MTW	16.11 (2.56)	2.17 (1.64)	51.28 (11.39)	2.78 (3.22)	0.73 (0.1)	312.14 (20.7)
3P7K	14.41 (2.17)	1.44 (1.28)	43.69 (12.5)	2.56 (3.17)	0.82 (0.11)	329.53 (32.4)
2P81	15.31 (3.05)	1.9 (2)	47.29 (14.49)	2.17 (2.25)	0.79 (0.13)	308.58 (63.38)
3V1A	15.3 (2.75)	2.22 (1.88)	45.3 (13.74)	2.8 (3.53)	0.79 (0.12)	317.68 (32.17)
1ENH	15.78 (2.67)	2.24 (1.7)	48.67 (13.19)	2.47 (2.79)	0.78 (0.12)	330.68 (52.83)
2F4K	15.42 (2.14)	1.81 (1.4)	48.55 (12.07)	2.74 (3.51)	0.76 (0.12)	325.51 (27.8)
2P6J	15.1 (2.3)	1.52 (1.55)	48.14 (11.56)	2.56 (3.46)	0.79 (0.09)	333.95 (36.24)
1AIL	15.93 (3.06)	2.4 (2.25)	48.76 (14.99)	3.43 (4.31)	0.76 (0.15)	346.9 (36.33)
2MR9	15.5 (3.33)	1.86 (1.96)	48.69 (16.77)	2.62 (3.09)	0.79 (0.13)	336.9 (36.3)
2JUC	15.79 (3.5)	2.58 (2.66)	46.68 (13.75)	2.7 (3.33)	0.76 (0.13)	316.43 (57.58)
1D5Q	16.64 (3.4)	2.52 (2.28)	52.76 (17.63)	2.68 (2.8)	0.68 (0.17)	326.57 (46.27)
1Q2K	15.21 (3)	1.9 (1.63)	46.31 (14.44)	1.9 (0.66)	0.74 (0.16)	316.24 (40.81)
2P5K	15.11 (2.46)	1.79 (1.61)	46.44 (12.11)	2.11 (1.88)	0.78 (0.14)	328.69 (45.3)
1CRN	14.7 (2.95)	1.57 (1.56)	44.89 (17.58)	2.36 (3.04)	0.77 (0.14)	306.17 (42.88)
1UTG	15.15 (2.61)	1.76 (1.64)	46.71 (13.08)	1.84 (0.5)	0.78 (0.11)	318.62 (42.68)
2MM8	16.26 (3.36)	2.05 (1.92)	52.93 (17.21)	2.41 (2.65)	0.74 (0.17)	332.28 (27.29)

Fonte: O Autor

Tabela 5.5: Análise das redes neurais treinadas com MLP e *back-propagation*. Os valores apresentados são as médias de todas as redes criadas para cada proteína testada, com os desvios padrão entre parênteses. A taxa de sucesso é a porcentagem do conjunto de teste classificada corretamente pela rede. Os resultados foram obtidos após a execução do algoritmo MLP 10 vezes para cada segmento, selecionando o resultado com melhor taxa de sucesso. Cada rodada durou 100 épocas.

Proteína	*Taxa de sucesso (prob.)*	*Tempo de treinamento (s)*
1AB1	0.77 (0.13)	34.68 (5.44)
1ACW	0.74 (0.15)	32.76 (2.73)
1DFN	0.79 (0.18)	33.45 (4.78)
1K43	0.72 (0.18)	35 (3.48)
1L2Y	0.8 (0.15)	33.36 (6.11)
1ROP	0.81 (0.11)	36.05 (5.39)
1ZDC	0.81 (0.12)	25.25 (14.97)
2PMR	0.82 (0.13)	37.94 (10.02)
1WQC	0.83 (0.11)	33.86 (4.34)
2MTW	0.73 (0.11)	35.36 (2.82)
3P7K	0.84 (0.12)	39.18 (6.41)
2P81	0.8 (0.12)	38.53 (11.53)
3V1A	0.81 (0.12)	35.36 (6.12)
1ENH	0.79 (0.12)	35.22 (5.89)
2F4K	0.79 (0.11)	36.67 (3.68)
2P6J	0.83 (0.09)	35.15 (4.44)
1AIL	0.77 (0.15)	35.88 (4.32)
2MR9	0.8 (0.14)	36.84 (4.64)
2JUC	0.77 (0.13)	34.53 (6.56)
1D5Q	0.7 (0.16)	36.82 (5.63)
1Q2K	0.77 (0.14)	34.57 (4.79)
2P5K	0.8 (0.13)	34.68 (2.83)
1CRN	0.79 (0.13)	34.5 (5.38)
1UTG	0.8 (0.1)	34.99 (5.02)
2MM8	0.75 (0.16)	37.12 (3.72)

Fonte: O Autor

xidade envolvida em métodos populacionais. Vale ressaltar, contudo, que redes neurais artificiais, uma vez treinadas, podem ser usadas de forma bastante rápida e eficiente.

O último passo de implementação do método é usar as redes neurais treinadas para criar os intervalos para os ângulos de torção ϕ e ψ como descrito na Seção 4.5. Para cada proteína testada se obtém dois conjuntos de intervalos, um para o ângulo ϕ e outro para o ângulo ψ. Como foram testadas 25 proteínas diferentes, tanto com NEAT quanto com MLP, foram obtidos no total 100 conjuntos diferentes de intervalos, que podem ser visualizados nas Figuras 5.1, 5.2, 5.3, 5.4 e 5.5.

Nestes gráficos, os intervalos são representados por faixas vermelhas. A intensidade da cor é proporcional à probabilidade que foi atribuída a eles. O centro de cada

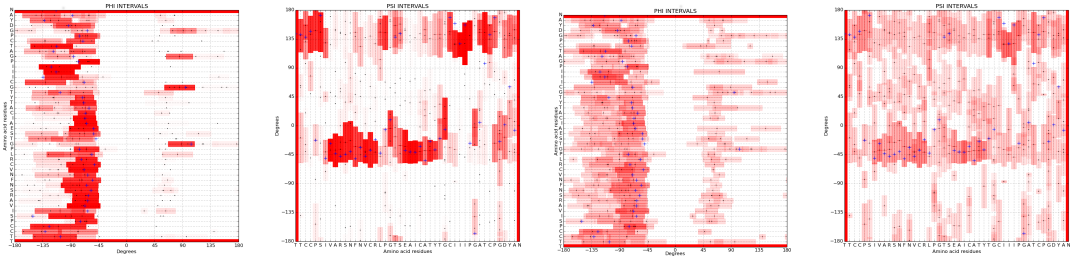
intervalo é marcado por um pequeno ponto preto. Nos gráficos dos intervalos ϕ, os aminoácidos estão distribuídos ao longo do eixo Y, enquanto o eixo X varia de $-180°$ a $180°$. Nos gráficos do ângulo ψ, essa disposição está trocada. As cruzes azuis marcam o valor dos ângulos de torção encontrados pelo método experimental. Proteínas cuja estrutura foi experimentalmente determinada através de NMR (indicadas na Tabela 5.1) podem apresentar mais de um valor de ângulo de torção para o mesmo aminoácido pois elas possuem mais de uma conformação determinada. Esse é mais um indício da vantagem de se usar intervalos em vez de valores únicos na representação, afinal a estrutura de uma proteína não é estática.

Uma análise da cobertura dos intervalos é feita na Tabela 5.6. Nela estão indicadas as porcentagens de ângulos de torção determinados experimentalmente que estão contidos dentro dos intervalos criados, a chamada "cobertura". Idealmente 100% dos ângulos estariam contidos nos intervalos, o que acaba não ocorrendo. Uma possível explicação é que os dados existentes no `PDB` não são suficientes para se ter uma acurácia total na predição dos ângulos, uma vez que o método é treinado descartando a própria proteína buscada. Outra possibilidade é que alguns dos grupos obtidos possuem desvio padrão muito grande, o que acabaria por excluir os ângulos afastados em demasia da média do grupo. Apesar disso, na grande maioria dos casos os intervalos conseguem cobrir corretamente os possíveis valores para os ângulos de torção ϕ e ψ.

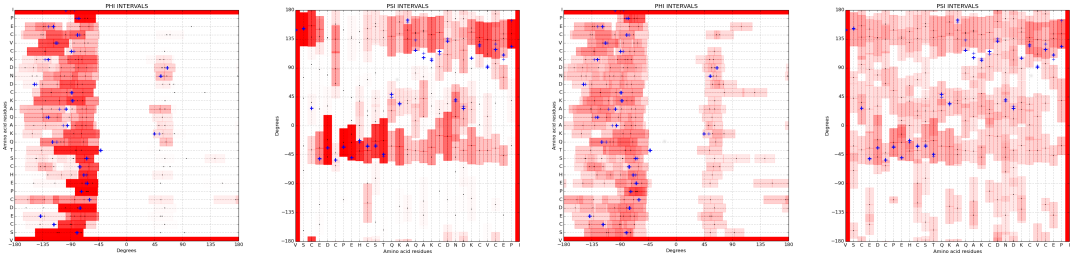
A Tabela 5.6 ainda possibilita uma comparação dos resultados de `NEAT` e `MLP`. Os valores de cobertura de todos os intervalos são os mesmos para os dois métodos, pois a diferença entre eles são as probabilidades atribuídas. Os valores indicados para comparação na Tabela 5.6 são as coberturas dos intervalos de maior probabilidade de cada aminoácido para os dois métodos, para os quais é possível ver que os dois algoritmos obtiveram resultados semelhantes, atribuindo majoritariamente a maior probabilidade aos intervalos que de fato contém os valores dos ângulos de torção experimental. Através de inspeção das Figuras 5.1, 5.2, 5.3, 5.4 e 5.5, contudo, é visível a diferença de atribuição de probabilidades entre os dois tipos de redes neurais. As evoluídas com `NEAT` tenderam a concentrar a probabilidade em uma ou duas regiões, enquanto as `MLP` distribuíram as probabilidades de forma mais uniforme ao longo do intervalo de $-180°$ a $180°$. Uma possível explicação para essa diferença é que as redes `NEAT` tiveram sua topologia evoluída de acordo com os dados apresentados, enquanto as `MLP` possuem topologia fixa, o que criaria uma maior especialização por parte de `NEAT` e por consequência uma maior certeza no momento de selecionar os intervalos.

Figura 5.1: Visualização dos intervalos de ângulos de torção ϕ e ψ gerados com o método usando NEAT e MLP. Cada linha (para ϕ) ou coluna (para psi) representa um aminoácido. A intensidade da cor é proporcional à probabilidade do intervalo. As cruzes azuis representam os valores dos ângulos da estrutura experimental. Estruturas obtidas com NMR podem manifestar mais de um ponto por ângulo de aminoácido.

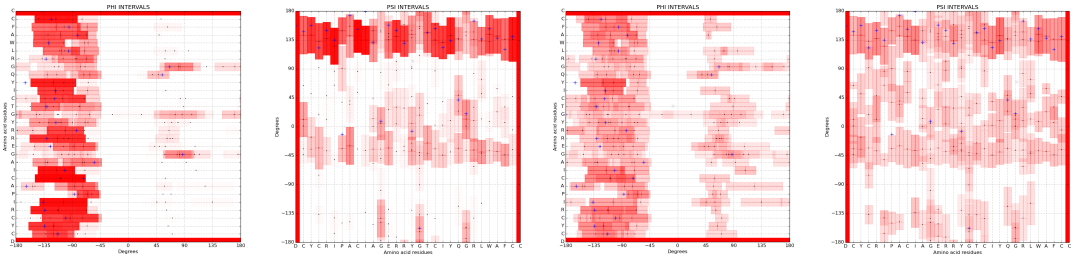
(a) 1AB1 NEAT ϕ (b) 1AB1 NEAT ψ (c) 1AB1 MLP ϕ (d) 1AB1 MLP ψ



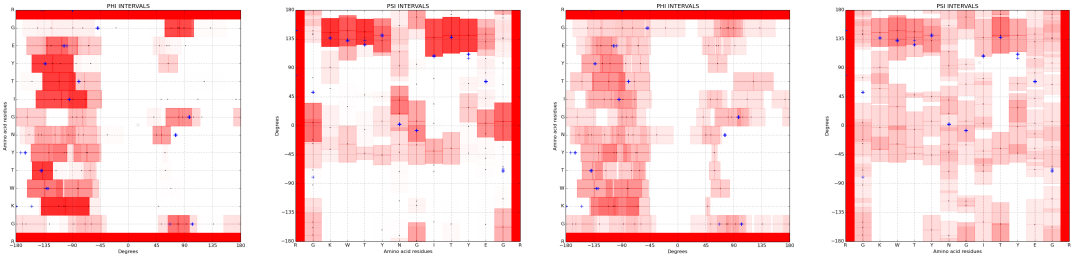
(e) 1ACW NEAT ϕ (f) 1ACW NEAT ψ (g) 1ACW MLP ϕ (h) 1ACW MLP ψ



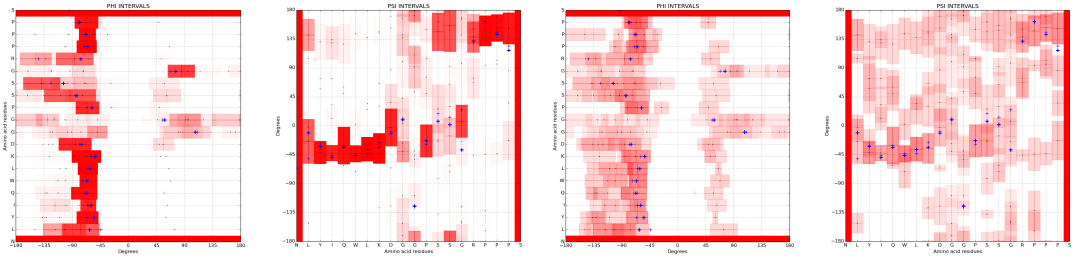
(i) 1DFN NEAT ϕ (j) 1DFN NEAT ψ (k) 1DFN MLP ϕ (l) 1DFN MLP ψ



(m) 1K43 NEAT ϕ (n) 1K43 NEAT ψ (o) 1K43 MLP ϕ (p) 1K43 MLP ψ



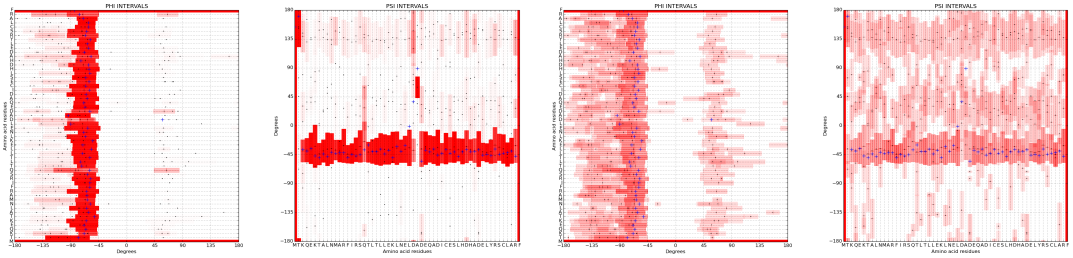
(q) 1L2Y NEAT ϕ (r) 1L2Y NEAT ψ (s) 1L2Y MLP ϕ (t) 1L2Y MLP ψ



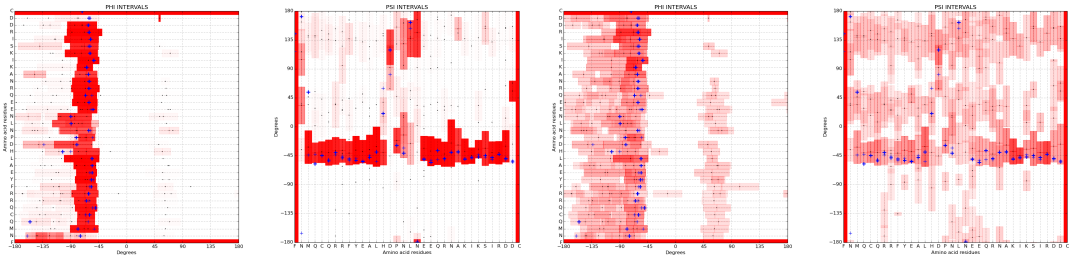
Fonte: O Autor

Figura 5.2: Continuação da Figura 5.1

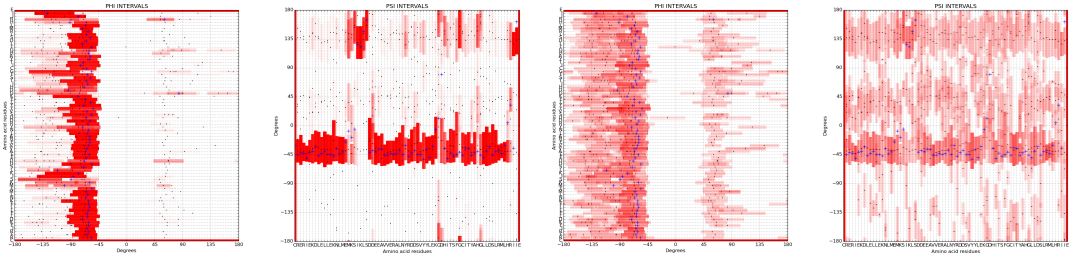
(a) 1ROP NEAT ϕ (b) 1ROP NEAT ψ (c) 1ROP MLP ϕ (d) 1ROP MLP ψ



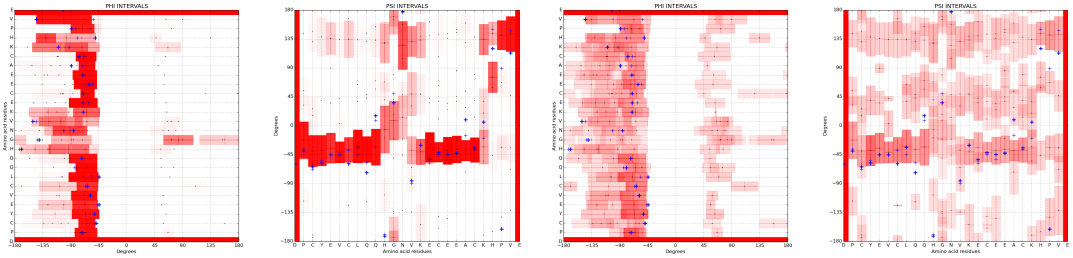
(e) 1ZDC NEAT ϕ (f) 1ZDC NEAT ψ (g) 1ZDC MLP ϕ (h) 1ZDC MLP ψ



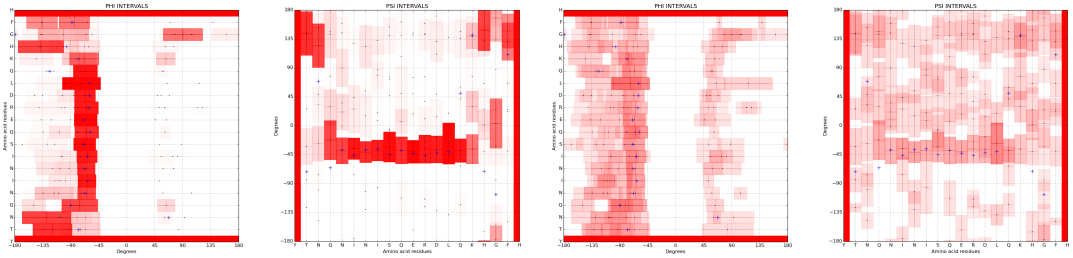
(i) 2PMR NEAT ϕ (j) 2PMR NEAT ψ (k) 2PMR MLP ϕ (l) 2PMR MLP ψ



(m) 1WQC NEAT ϕ (n) 1WQC NEAT ψ (o) 1WQC MLP ϕ (p) 1WQC MLP ψ



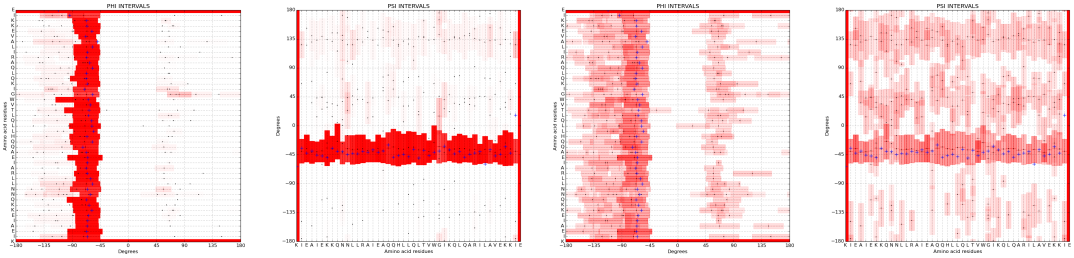
(q) 2MTW NEAT ϕ (r) 2MTW NEAT ψ (s) 2MTW MLP ϕ (t) 2MTW MLP ψ



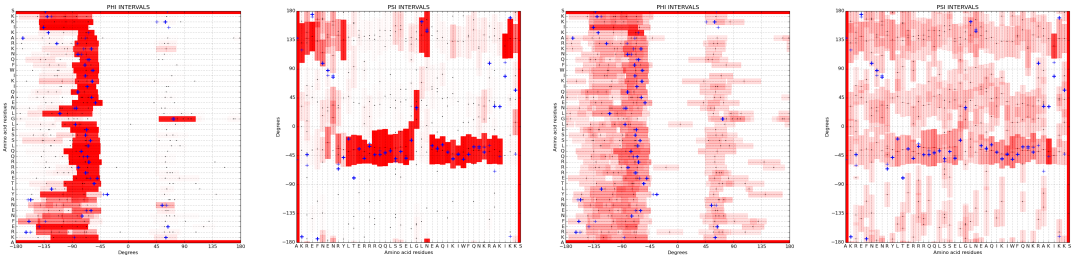
Fonte: O Autor

Figura 5.3: Continuação da Figura 5.2

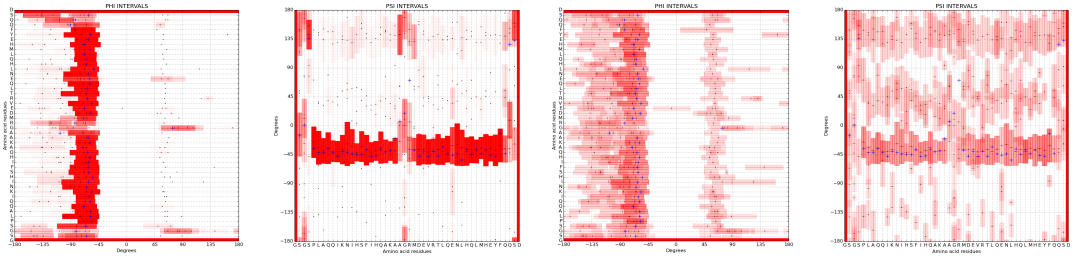
(a) 3P7K NEAT ϕ (b) 3P7K NEAT ψ (c) 3P7K MLP ϕ (d) 3P7K MLP ψ



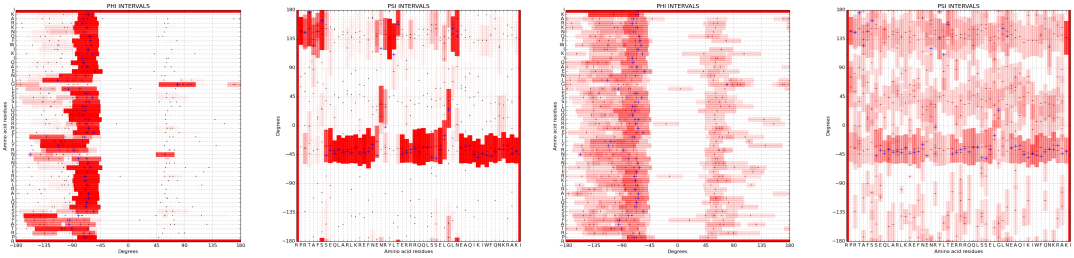
(e) 2P81 NEAT ϕ (f) 2P81 NEAT ψ (g) 2P81 MLP ϕ (h) 2P81 MLP ψ



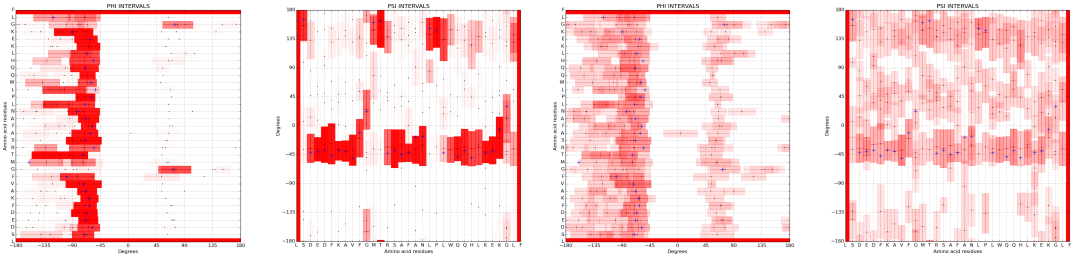
(i) 3V1A NEAT ϕ (j) 3V1A NEAT ψ (k) 3V1A MLP ϕ (l) 3V1A MLP ψ



(m) 1ENH NEAT ϕ (n) 1ENH NEAT ψ (o) 1ENH MLP ϕ (p) 1ENH MLP ψ



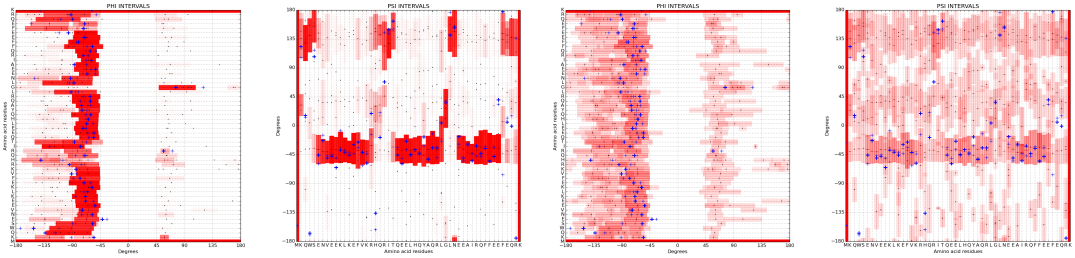
(q) 2F4K NEAT ϕ (r) 2F4K NEAT ψ (s) 2F4K MLP ϕ (t) 2F4K MLP ψ



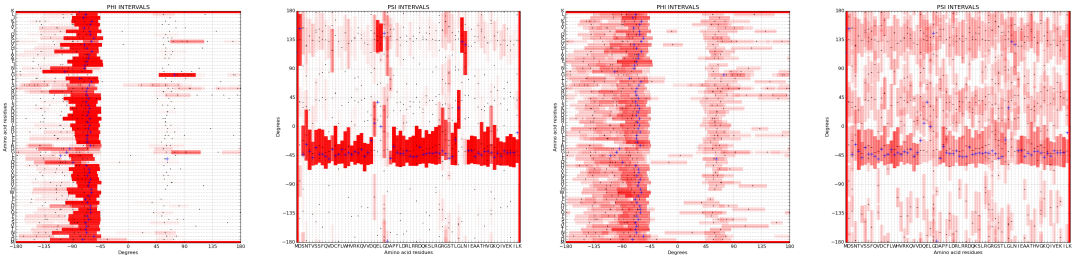
Fonte: O Autor

Figura 5.4: Continuação da Figura 5.3

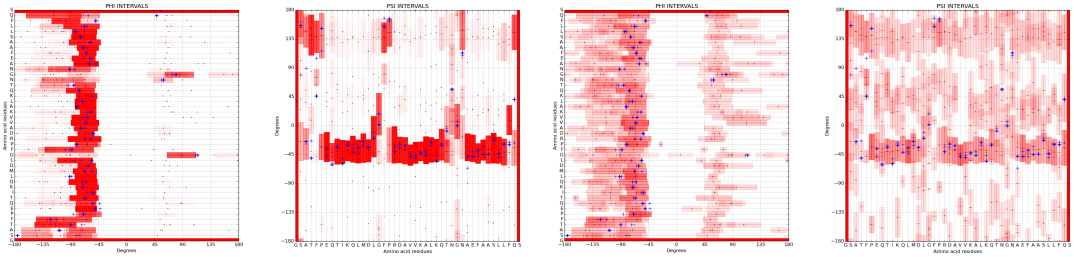
(a) 2P6J NEAT ϕ (b) 2P6J NEAT ψ (c) 2P6J MLP ϕ (d) 2P6J MLP ψ



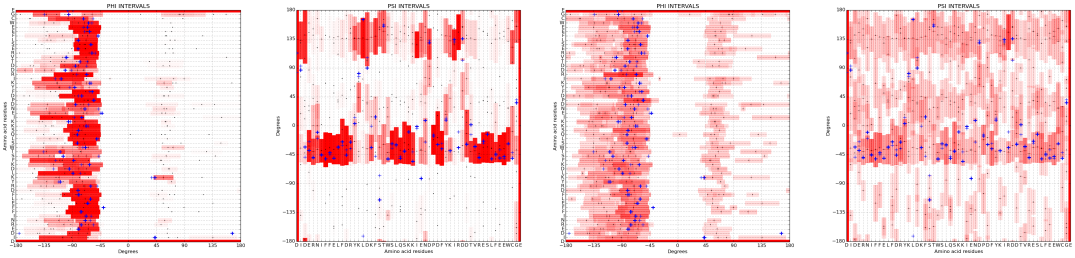
(e) 1AIL NEAT ϕ (f) 1AIL NEAT ψ (g) 1AIL MLP ϕ (h) 1AIL MLP ψ



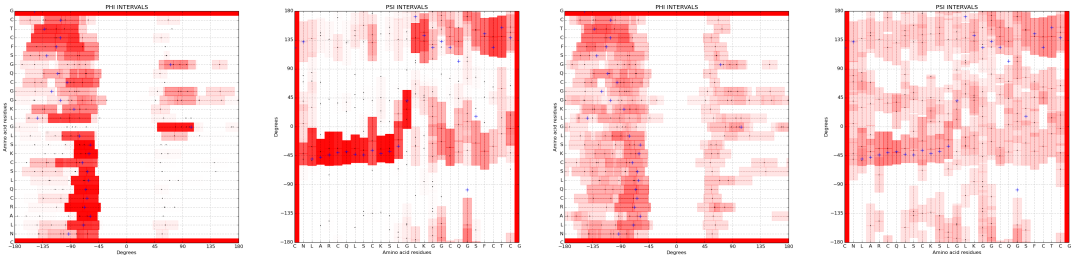
(i) 2MR9 NEAT ϕ (j) 2MR9 NEAT ψ (k) 2MR9 MLP ϕ (l) 2MR9 MLP ψ



(m) 2JUC NEAT ϕ (n) 2JUC NEAT ψ (o) 2JUC MLP ϕ (p) 2JUC MLP ψ



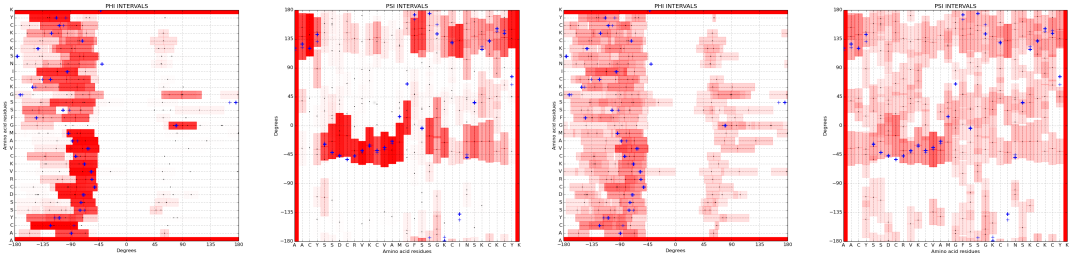
(q) 1D5Q NEAT ϕ (r) 1D5Q NEAT ψ (s) 1D5Q MLP ϕ (t) 1D5Q MLP ψ



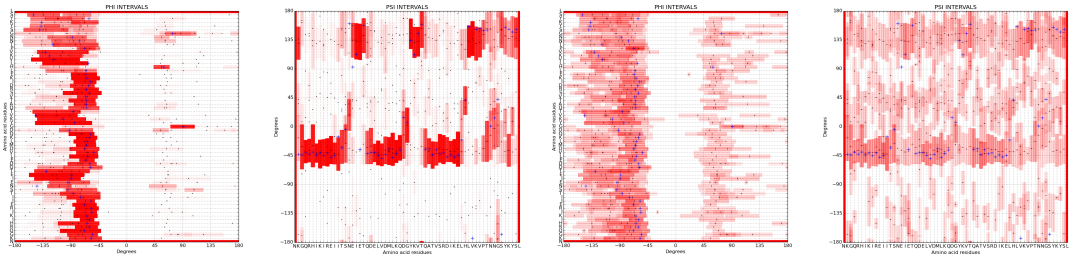
Fonte: O Autor

Figura 5.5: Continuação da Figura 5.4

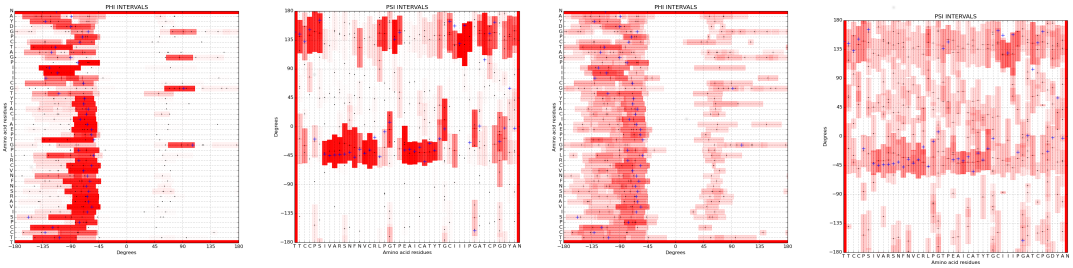
(a) 1Q2K NEAT ϕ (b) 1Q2K NEAT ψ (c) 1Q2K MLP ϕ (d) 1Q2K MLP ψ



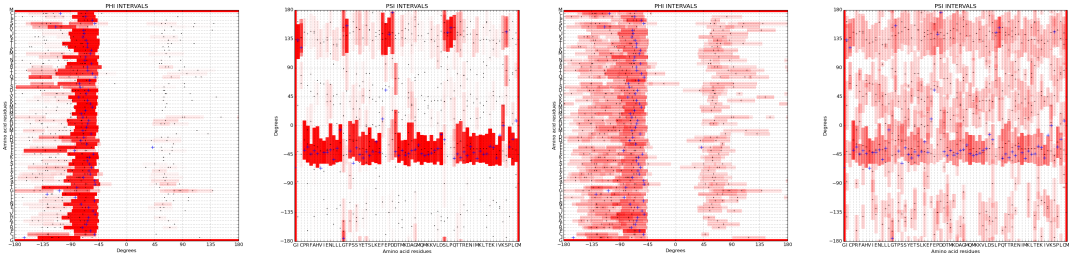
(e) 2P5K NEAT ϕ (f) 2P5K NEAT ψ (g) 2P5K MLP ϕ (h) 2P5K MLP ψ



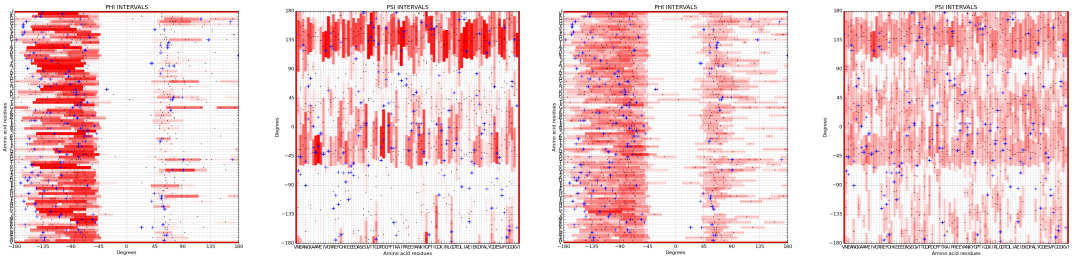
(i) 1CRN NEAT ϕ (j) 1CRN NEAT ψ (k) 1CRN MLP ϕ (l) 1CRN MLP ψ



(m) 1UTG NEAT ϕ (n) 1UTG NEAT ψ (o) 1UTG MLP ϕ (p) 1UTG MLP ψ



(q) 2MM8 NEAT ϕ (r) 2MM8 NEAT ψ (s) 2MM8 MLP ϕ (t) 2MM8 MLP ψ



Fonte: O Autor

Tabela 5.6: Descrição da cobertura dos intervalos gerados, ou seja, a porcentagem de ângulos de torção ϕ e ψ experimentais (NMR ou raio-X) contidos nos intervalos criados com o método apresentado, tanto com redes `NEAT` quanto com `MLP`. O valor representa a cobertura dos intervalos com maior probabilidade associada por aminoácido, com a cobertura considerando todos os intervalos de um aminoácido entre parênteses. Também é apresentada uma comparação com os resultados do método MOIRAE para as proteínas disponíveis.

Proteína	*Cobertura* ϕ `NEAT` / `MLP` *(Todos) %*	*Cobertura* ψ `NEAT` / `MLP` *(Todos) %*	*Cobertura* ϕ *MOIRAE %*	*Cobertura* ψ *MOIRAE %*
1AB1	76.09/71.74(93.48)	69.57/69.57(93.48)	92.85	80.95
1ACW	55.17/65.52(93.1)	51.72/51.72(86.21)	68.00	36.00
1DFN	73.33/76.67(90)	73.33/73.33(90)	80.76	84.61
1K43	57.14/57.14(71.43)	50/50(78.57)	70.00	80.00
1L2Y	75/75(90)	70/70(90)		
1ROP	96.43/94.64(98.21)	92.86/92.86(96.43)	98.07	94.00
1ZDC	85.29/88.24(97.06)	82.35/82.35(91.18)		
2PMR	92.11/90.79(98.68)	84.21/85.53(93.42)		
1WQC	65.38/69.23(88.46)	50/57.69(73.08)	54.54	50.00
2MTW	70/70(90)	60/60(85)		
3P7K	95.56/95.56(97.78)	93.33/93.33(93.33)		
2P81	81.82/79.55(93.18)	68.18/65.91(84.09)		
3V1A	85.42/87.5(97.92)	89.58/91.67(95.83)		
1ENH	90.74/92.59(98.15)	92.59/90.74(98.15)	98.00	96.00
2F4K	84.85/78.79(87.88)	84.85/81.82(87.88)		
2P6J	76.92/75(92.31)	71.15/73.08(88.46)		
1AIL	94.29/92.86(98.57)	91.43/88.57(95.71)	98.48	95.45
2MR9	72.73/75(90.91)	68.18/70.45(86.36)		
2JUC	74.55/74.55(85.45)	67.27/69.09(85.45)		
1D5Q	74.07/81.48(96.3)	66.67/74.07(85.19)		
1Q2K	67.74/74.19(87.1)	58.06/61.29(77.42)	51.85	40.74
2P5K	90.48/90.48(98.41)	82.54/84.13(92.06)		
1CRN	80.43/78.26(97.83)	71.74/67.39(93.48)		
1UTG	84.29/87.14(95.71)	85.71/88.57(94.29)		
2MM8	32.22/33.33(84.44)	24.44/23.33(78.89)		

Fonte: O Autor

A Tabela 5.6 permite uma última comparação entre o método proposto e o MOIRAE. Os dados de cobertura do MOIRAE foram retirados do artigo original do método (DORN; BURIOL; LAMB, 2013), por isso não há informações para todas as proteínas testadas nesse trabalho. O que pode ser percebido é que ambos os métodos tiveram desempenho parecido para proteínas com alta porcentagem de cobertura, mas nas com menor sucesso como as com código `PDB` 1ACW, 1WQC e 1Q2K o método proposto teve melhor desempenho, em especial quando considerados todos os intervalos. Essa diferença provavelmente se deve ao fato de MOIRAE utilizar um único intervalo absoluto

por aminoácido, o que acaba por excluir outras regiões em que os valores dos ângulos de torção podem estar.

Outra análise relevante dos intervalos é o seu tamanho. Quanto menor o intervalo, maior sua capacidade de reduzir o espaço de busca, que seria de $360°$ (de $-180°$ a $180°$). A Tabela 5.7 contém o tamanho médio dos intervalos de maior probabilidade de cada aminoácido das proteínas testadas. Também estão declarados os tamanhos médios dos intervalos do método MOIRAE disponíveis. Como o MOIRAE gera um único intervalo por aminoácido, uma comparação pode ser feita com os intervalos de maior probabilidade de cada aminoácido do método descrito neste trabalho. O MOIRAE apresenta, em geral, intervalos de tamanhos menores, o que era esperado pois diferentemente do que foi adotado neste trabalho, tal método divide a cadeia de aminoácidos em segmentos de comprimento 5 aminoácidos, considerando os dois vizinhos anteriores e seguintes do central (DORN; BURIOL; LAMB, 2013). Essa escolha gera intervalos mais concisos, mas precisa lidar com bem menos dados, o que pode gerar perda de informação. Neste trabalho se preferiu optar por uma maior generalização a partir do banco de dados pois o que se observou na etapa de busca na base de dados e agrupamento foi que segmentos de comprimento 5 retornaram uma quantidade insuficiente de correspondências em algumas das proteínas testadas. Tal resultado acaba por gerar intervalos de ângulos de torção menores e mais precisos, mas ao mesmo tempo prejudicam a generalização do método. Mesmo assim os intervalos gerados são capazes de reduzir o espaço de busca em cerca de 85% quando considerados os de maior probabilidade.

A última análise feita foi a construção de estruturas tridimensionais a partir dos intervalos obtidos. Para tanto, selecionamos os valores centrais dos intervalos de maior probabilidade gerados a partir de `NEAT` de cada aminoácido para obter os valores de ϕ e ψ da estrutura 3D estimada de cada uma das proteínas usadas para teste. O objetivo desta etapa não é de fato predizer a estrutura terciária, e sim mostrar que os intervalos de ângulos gerados compreendem estruturas semelhantes à estrutura nativa da proteína e portanto podem ser usados por um algoritmo de busca e otimização para encontrar soluções para o problema da predição.

Como pode ser visto na Figura 5.6, que sobrepõem as estruturas terciárias das proteínas determinadas experimentalmente e as preditas com os intervalos de ângulos, a utilização dos ângulos de torção correspondentes ao valor central dos intervalos de maior probabilidade de cada aminoácido foi capaz de apresentar um enovelamento compatível com o das estruturas 3D nativas. A Tabela 5.8 faz a comparação do conteúdo das estru-

Tabela 5.7: Descrição dos tamanhos dos intervalos gerados com NEAT. O valor representa a média do tamanho dos intervalos com maior probabilidade associada por aminoácido, com o desvio padrão entre parênteses. Também é informado o número médio de intervalos por aminoácido. Por fim, há uma comparação com os resultados do método MOIRAE para as proteínas disponíveis.

Proteína	*Tamanho médio ϕ (desvio)*	*Tamanho médio ψ (desvio)*	*N° médio de intervalos*	*Tamanho médio MOIRAE ϕ*	*Tamanho médio MOIRAE ψ*
1AB1	51.69 (14.67)	50.25 (11.31)	6.07	24.6	29.63
1ACW	52.03 (17.2)	51.86 (12.8)	5.9	71.33	105.6
1DFN	64.08 (15.77)	54.13 (6.87)	5.8	16.16	17.13
1K43	52.13 (16.28)	51.49 (7.55)	6.57	2.4	6.51
1L2Y	43.11 (14.51)	46.94 (14.26)	5.85		
1ROP	39.49 (10.09)	42.95 (9.48)	6.02	12.22	13.3
1ZDC	40.49 (12.12)	44.9 (12.72)	5.24		
2PMR	47.11 (14.33)	46.25 (11.74)	6		
1WQC	44.85 (17.11)	47.52 (10.97)	5.46	17.77	48.91
2MTW	46.12 (15.56)	49.5 (13.97)	6.35		
3P7K	41.73 (7.96)	42.84 (7.51)	5.69		
2P81	49.96 (17.15)	47.54 (9.75)	6.16		
3V1A	43.27 (10.32)	46.12 (11.45)	5.88		
1ENH	44.53 (13.41)	46.51 (9.43)	6.37	17.03	29.05
2F4K	45.47 (15.62)	50.44 (11.51)	6.27		
2P6J	46.64 (16.03)	46.52 (11)	6.37		
1AIL	46.78 (12.89)	49.9 (12.86)	6.37	12.97	15.6
2MR9	44.55 (15.31)	43.16 (9.29)	6.39		
2JUC	49.76 (15.93)	51.17 (13.09)	6.02		
1D5Q	52.65 (15.03)	54.62 (13.52)	6.67		
1Q2K	55.1 (13.57)	54.55 (9.24)	5.97	45.98	37.05
2P5K	52.04 (18.21)	48.6 (10.73)	6.16		
1CRN	49.35 (15.23)	49.35 (11.06)	5.91		
1UTG	45.1 (11.65)	48.62 (11.06)	6.23		
2MM8	57.55 (16.31)	54.1 (11.76)	7.08		

Fonte: O Autor

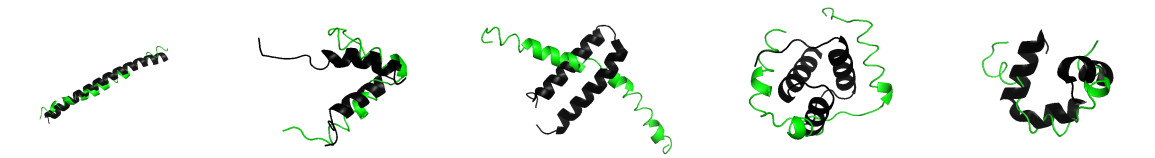
turas secundárias das estruturas nativas e preditas e também revela similaridade entre as duas. A qualidade das estruturas 3D preditas foi avaliada pelo cálculo do desvio médio quadrático (RMSD) (Equação 5.1) entre a posição dos átomos C_α da estrutura predita e a da estrutura experimental. O valor de RMSD calculado permite a comparação de similaridade entre duas estruturas 3D, sendo que duas estruturas idênticas teriam RMSD igual a zero. A Tabela 5.8 traz o RMSD entre as estruturas experimentais das proteínas e as estruturas preditas com os intervalos. A medida do RMSD foi calculada usando a ferramenta PyMOL.

Figura 5.6: Representação das estruturas 3D experimentais (preto) e calculadas usando os valores centrais dos intervalos de maior probabilidade obtidos com NEAT para os ângulos de torção (verde). Os C_α da estrutura experimental e predita foram ajustados. As figuras foram geradas com PyMOL.

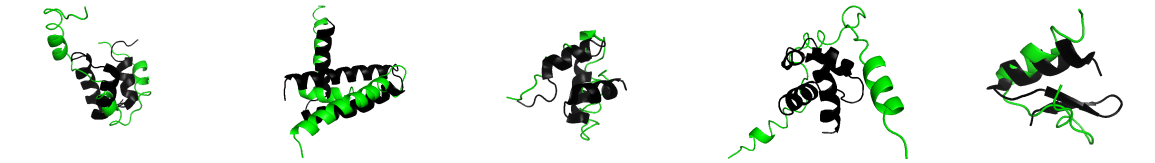
(a) 1AB1 (b) 1ACW (c) 1DFN (d) 1K43 (e) 1L2Y

(f) 1ROP (g) 1ZDC (h) 2PMR (i) 1WQC (j) 2MTW

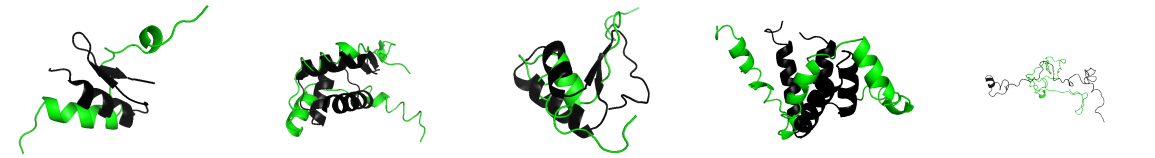
(k) 3P7K (l) 2P81 (m) 3V1A (n) 1ENH (o) 2F4K



(p) 2P6J (q) 1AIL (r) 2MR9 (s) 2JUC (t) 1D5Q



(u) 1Q2K (v) 2P5K (w) 1CRN (x) 1UTG (y) 2MM8



Fonte: O Autor

$$\mathbf{RMSD}(a,b) = \sqrt{\left(\sum_{i=1}^{n} \|r_{ai} - r_{bi}\|^2\right)/n}, \tag{5.1}$$

onde r_{ai} e r_{bi} são vetores representando as posições do mesmo átomo i em cada uma das duas estruturas, a e b respectivamente, e onde as estruturas a e b são otimamente sobrepostas.

É interessante lembrarmos que as proteínas não são totalmente rígidas e podem adotar diferentes estados conformacionais. A Tabela 5.9 é um indicativo disso, calculando o RMSD médio e máximo entre os modelos disponíveis no PDB para as proteínas

Tabela 5.8: Análise estrutural das proteínas preditas com conteúdo da estrutura secundária da proteína predita (P) e experimental (E).

Proteína	*Fita* P (E) %	*Hélice-*α P (E) %	*Hélice-*3_{10} P (E) %	*Outro* P (E) %	*RMSD Å*
1AB1	0.0 (8.7)	41.3 (41.3)	0.0 (0.0)	58.7 (50.0)	7.14
1ACW	0.0 (34.5)	20.7 (24.1)	0.0 (0.0)	79.3 (41.4)	13.27
1DFN	0.0 (60.0)	0.0 (0.0)	0.0 (0.0)	100.0 (40)	15.19
1K43	42.9 (42.9)	0.0 (0.0)	0.0 (0.0)	57.1 (57.1)	2.08
1L2Y	0.0 (0.0)	35.0 (35.0)	0.0 (20.0)	65.0 (45.0)	5.37
1ROP	0.0 (0.0)	92.9 (89.3)	0.0 (0.0)	7.1 (10.7)	19.81
1ZDC	0.0 (0.0)	73.5 (73.5)	0.0 (0.0)	26.5 (26.5)	3.62
2PMR	0.0 (0.0)	80.3 (80.3)	0.0 (0.0)	19.7 (19.7)	21.12
1WQC	0.0 (0.0)	65.4 (65.4)	0.0 (0.0)	34.6 (34.6)	6.58
2MTW	0.0 (0.0)	65.0 (50.0)	0.0 (0.0)	35.0 (50.0)	4.23
3P7K	0.0 (0.0)	95.6 (93.3)	0.0 (0.0)	4.4 (6.7)	3.27
2P81	0.0 (0.0)	61.4 (59.1)	0.0 (0.0)	38.6 (40.9)	5.12
3V1A	0.0 (0.0)	81.2 (77.1)	0.0 (0.0)	18.8 (22.9)	16.98
1ENH	0.0 (0.0)	70.4 (70.4)	0.0 (0.0)	29.6 (29.6)	10.76
2F4K	0.0 (0.0)	66.7 (71.4)	0.0 (0.0)	33.3 (28.6)	6.55
2P6J	0.0 (0.0)	65.4 (59.6)	0.0 (0.0)	34.6 (40.4)	8.61
1AIL	0.0 (0.0)	82.9 (84.3)	0.0 (0.0)	17.1 (15.7)	15.14
2MR9	0.0 (0.0)	68.2 (61.4)	0.0 (0.0)	31.8 (38.6)	4.74
2JUC	0.0 (0.0)	52.7 (52.7)	0.0 (5.5)	47.3 (41.8)	19.46
1D5Q	0.0 (29.6)	44.4 (40.7)	0.0 (0.0)	55.6 (29.6)	5.65
1Q2K	0.0 (19.4)	38.7 (32.3)	6.5 (0.0)	54.8 (48.4)	11.65
2P5K	0.0 (15.9)	55.6 (55.6)	0.0 (0.0)	44.4 (28.6)	6.76
1CRN	0.0 (8.7)	41.3 (41.3)	0.0 (6.5)	58.7 (43.5)	6.04
1UTG	0.0 (0.0)	71.4 (71.4)	0.0 (4.3)	28.6 (24.3)	13.96
2MM8	4.4 (0.0)	4.4 (0.0)	0.0 (3.3)	91.1 (96.7)	22.39

Fonte: O Autor

com estrutura determinada com NMR. Se as proteínas fossem totalmente rígidas o RMSD entre todas deveria ser zero, o que não ocorre. Nos últimos anos, foram também descobertas as proteínas desordenadas, com alta flexibilidade estrutural que permite que adotem diferentes estruturas. Em alguns casos a proteína é totalmente desordenada, mas em outros existem segmentos de aminoácidos desordenados em proteínas que de outra forma são ordenadas e com estrutura terciária bem definida.

Proteínas desordenadas são, sem dúvida, um grande desafio à predição da estrutura terciária por não possuírem uma única estrutura terciária bem definida em seu estado nativo. Um exemplo dessas proteínas é a com código `PDB` 2MM8, que foi usada como caso de teste para este método. Podemos ver na Tabela 5.9 os altos valores de RMSD calculados para ela. Mesmo assim, o método proposto foi capaz de gerar intervalos funcionais para esta proteína. Como pode ser visto na Tabela 5.7, o tamanho médio de intervalo

Tabela 5.9: RMSD médio e máximo calculado entre os diferentes modelos disponíveis no PDB de proteínas da Tabela 5.1 com estrutura terciária determinada experimentalmente através de NMR.

Proteína	*RMSD médio Å (desvio)*	*RMSD máximo Å*
1ACW	0.58 (0.24)	1.37
1K43	0.76 (0.51)	2.52
1L2Y	0.44 (0.13)	0.97
1ZDC	0.64 (0.18)	1.09
1WQC	0.79 (0.34)	2.15
2P81	1.39 (0.32)	2.51
2P6J	1.3 (0.24)	2.26
2MR9	0.5 (0.13)	0.85
2JUC	0.53 (0.13)	0.84
1Q2K	0.48 (0.24)	1.01
2MM8	11.6 (1.57)	14.61

Fonte: O Autor

para a 2MM8 foi de 57,55°, coerente com o encontrado para as outras proteínas testadas. Já conforme a Tabela 5.6, a cobertura do intervalo de maior probabilidade foi de 32% para o ângulo ϕ e 24.44% para o ângulo ψ, bem abaixo das demais. Mas, graças à utilização de mais de um intervalo por aminoácido, chega-se à cobertura total de $84,44\%$ para ϕ e $78,89\%$ para ψ.

6 CONCLUSÃO

O estudo de proteínas e a predição de suas estruturas tridimensionais permanecem como uma das áreas de pesquisa mais relevantes e desafiantes da bioinformática estrutural. A predição da estrutura 3D de proteínas utilizando informações de bases de dados como o PDB, embora promissora, continua em aberto enquanto métodos computacionais ainda não são capazes de solucionar o problema satisfatoriamente. A consequência visível disto é a discrepância entre as sequências de resíduos de aminoácidos disponíveis e a quantidade de estruturas 3D determinadas experimentalmente. Dado que a estrutura 3D da proteína fornece importantes informações sobre seu funcionamento no organismo, tal defasagem é um empecilho às pesquisas em biologia molecular, medicina e farmácia e motiva a busca por métodos eficientes de qualidade.

Neste trabalho, foi proposto um novo método de extração de informação do PDB que permite a criação de intervalos de busca para os ângulos de torção ϕ e ψ baseados em probabilidades de ocorrência de vizinhanças de aminoácidos e estrutura secundária. O método busca no PDB por ocorrências do mesmo segmento de três aminoácidos, realiza o agrupamento hierárquico pelos pares de ângulos ϕ e ψ e treina redes neurais artificiais utilizando NEAT para a criação dos intervalos.

Como foi observado, NEAT foi capaz de aprender padrões estruturais complexos de diferentes proteínas eficientemente e sem a necessidade de conhecimento prévio dos dados para o projeto da topologia das redes neurais. Os intervalos gerados contêm informações estruturais compatíveis com os dados experimentais, como os ângulos de torção e estrutura secundária, mesmo para proteínas não rígidas, como as desordenadas, e podem ser utilizados para a visualização da flexibilidade estrutural de cadeias de aminoácidos e redução do espaço dos dados em estratégias de busca, auxiliando na obtenção de métodos de predição mais eficientes e precisos.

6.1 Trabalhos futuros

O trabalho apresentado abre espaço para pesquisas futuras. Testes podem ser realizados com outras classes de proteínas. Parâmetros como o comprimento do segmento de aminoácidos usado podem ser alterados para comparação de resultados em maior profundidade. Uma base de redes neurais treinadas pode ser criada e disponibilizada para que o projeto torne-se uma ferramenta capaz de auxiliar outras pesquisas, entre outras ideias.

Uma das propostas mais promissoras é a aplicação dos intervalos gerados para os ângulos de torção em um método de busca como algoritmos genéticos, busca tabu, *simulated annealing*, *particle swarm optimization* ou GRASP, de forma a predizer a estrutura 3D de proteínas. Resultados obtidos através da APL (BORGUESAN et al., 2015) sugerem que esta estratégia tem potencial para gerar boas soluções.

6.2 Publicações associadas

1) Uma versão preliminar deste trabalho foi aceita em forma de artigo e apresentação oral, com o nome *"Predicting Protein Structural Features with NeuroEvolution of Augmenting Topologies"*, em co-autoria de Bruno Iochins Grisci e Márcio Dorn, no *IEEE World Congress on Computational Intelligence*, seção *Joint Conference on Neural Networks*, a ser realizado entre 24 e 29 de julho de 2016 em Vancouver, Canadá.

2) Um resumo dete trabalho, com título "Predição da Flexibilidade Conformacional de Resíduos de Aminoácidos através de NeuroEvolução", de autoria de Bruno Iochins Grisci e orientação de Márcio Dorn, foi homologado para participação no XXVIII Salão de Iniciação Científica da UFRGS, a ser realizado de 12 a 16 de setembro de 2016.

3) Um artigo a respeito de toda a pesquisa desenvolvida está sendo finalizado para submissão ao *Journal of Bioinformatics and Computational Biology*.

REFERÊNCIAS

ALTSCHUL, S. et al. Basic local alignment search tool. **J. Mol. Biol.**, v. 215, n. 3, p. 403–410, 1990.

AMABIS, J. M.; MARTHO, G. R. **Fundamentos da Biologia Moderna**. 4. ed. Sao Paulo, Brazil: Moderna, 2006. 128 p.

ANFINSEN, C. Principles that govern the folding of protein chains. **Science**, v. 181, n. 96, p. 223–230, 1973.

ANFINSEN, C. et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proc. Natl. Acad. Sci. U. S. A.**, v. 47, p. 1309–1314, 1961.

ANGELINE, P. J.; SAUNDERS, G. M.; POLLACK, J. B. An evolutionary algorithm that con- structs recurrent neural networks. **IEEE Transactions on Neural Networks**, v. 5, p. 54–65, 1993.

BANNER, D.; KOKKINIDIS, M.; TSERNOGLOU, D. Structure of the cole1 rop protein at 1.7 a resolution. **J. Mol. Biol.**, v. 196, p. 657–675, 1987.

BENSON, D. et al. Genbank. **Nucleic Acids Res.**, v. 36, p. 25–30, 2009.

BERMAN, H. et al. The protein data bank. **Nucleic Acids Res.**, v. 28, n. 1, p. 235–242, 2000.

BLANC, E. et al. Solution structure of p01, a natural scorpion peptide structurally analogous to scorpion toxins specific for apamin-sensitive potassium channel. **Proteins**, v. 24, p. 359–369, 1996.

BONANNO, J. et al. Crystal structure of a protein of unknown function from methanobacterium thermoautotrophicum. **To be published**.

BONET, R.; RAMIREZ-ESPAIN, X.; MACIAS, M. J. 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**, Wiley Online Library, v. 73, n. 4, p. 1001–1009, 2008.

BORGUESAN, B. et al. Apl: an angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. **Comput. Biol. Chem.**, v. 59, n. A, p. 142–157, 2015.

BORNOT, A.; ETCHEBEST, C.; BREVERN, A. e. Predicting protein flexibility through the prediction of local structures. **Proteins**, v. 79, n. 3, p. 839–852, 2011.

BOWIE, J. U.; LUTHY, R.; EISENBERG, D. A method to identify protein sequences that fold into a known three-dimensional structure. **Science**, v. 253, n. 5016, p. 164–170, 1991.

BRANKE, J. Evolutionary algorithms for neural network design and training. In: **Proceedings First Nordic Workshop on Genetic Algorithms and their Applications**. Vaasa, Finland: [s.n.], 1995. p. 145–163.

BRYANT, S. H.; ALTSCHUL, S. Statistics of sequence-structure threading. **Curr. Opin. Struct. Biol.**, v. 5, n. 2, p. 236–244, 1995.

CAI, Z. et al. Solution structure of bmbktx1, a new bkca1 channel blocker from the chinese scorpion buthus martensi karsch. **Biochemistry**, v. 43, p. 3764–3771, 2004.

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

CHERVENSKI, P. **MultiNEAT**. 2012–. Online; accessed 2016-01-28. Available from Internet: <http://www.multineat.com/>.

CIFUENTES, G. et al. Evidence supporting the hypothesis that specifically modifying a malaria peptide to fit hla-dr 1*03 molecules induces antibody production and protection. **To be published**.

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

COCK, P. J. A. et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, 2009.

COZZETTO, D. et al. Evaluation of template-based models in casp8 with standard measures. **Proteins: Struct., Funct., Bioinf.**, v. 77, n. 9, p. 18–28, 2009.

CRESCENZI, P. et al. On the complexity of protein folding. **J. Comput. Biol.**, v. 5, n. 3, p. 423–466, 1998.

CURTEANU, S.; CARTWRIGHT, H. Neural networks applied in chemistry. i. determination of the optimal topology of multilayer perceptron neural networks. **Journal of Chemometrics**, Wiley Online Library, v. 25, n. 10, p. 527–549, 2011.

CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. **J. R. Soc., Interface**, v. 3, n. 6, p. 139–151, 2006.

DER B.S., M. M. M. M. M. J. S. T. K. B. Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. **J.Am.Chem.Soc.**, v. 134, p. 375–385, 2012.

DOBROVOLSKA, O. et al. 1h, 13c, 15n resonance assignments of reduced jaburetox. **To be published**.

DOBSON, C. Protein folding and misfolding. **Nature**, v. 426, p. 884–890, 2003.

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, 2013.

DORN, M. et al. A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: **IEEE Congress on Evolutionary Computation**. Cancun, MX: IEEE, 2013. p. 1233–1240.

FAUCHERE, J.; PLISKA, V. Hydrophobic parameters pi of amino-acid side-chains from the partitioning of n-acetyl-amino-acid amides. **Eur. J. Med. Chem.**, v. 18, p. 369–375, 1983.

FLOUDAS, C. et al. Advances in protein structure prediction and de novo protein design: A review. **Chem. Eng. Sci.**, v. 61, n. 3, p. 966–988, 2006.

FRAENKEL, A. S. Complexity of protein folding. **Bull. Math. Biol.**, v. 55, n. 6, p. 1199–1210, 1993.

GARNETT, J. et al. A high-resolution structure of the dna-binding domain of ahrc, the arginine repressor/activator protein from bacillus subtilis. **Acta Crystallogr.**, v. 63, p. 914–917, 2007.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and Machine Learning**. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201157675.

GOMEZ, F.; MIIKKULAINEN, R. Solving non-markovian control tasks with neuroevolution. **Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence**, p. 1356–1361, 1999.

GOMEZ, F.; MIIKKULAINEN, R. Learning robust nonlinear control with neuroevolution. **Technical Report AI02-292, Department of Computer Sciences, The University of Texas at Austin, Austin, Texas.**, 2002.

GRUAU F., W. D.; PYEATT, L. A comparison between cellular encoding and direct encoding for genetic neural networks. In: **Genetic Programming, Proceedings of the First Annual Conference**. Cambridge, Massachusetts: [s.n.], 1996. p. 81–89.

HART, W.; ISTRAIL, S. Robust proofs of np-hardness for protein folding: general lattices and energy potentials. **J. Comput. Biol.**, v. 4, n. 1, p. 1–22, 1997.

HAYKIN, S. **Neural Networks: A comprehensive foundation**. 2. ed. New York, USA: Prentice Hall Inc., 1998.

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

HELLES, G.; FONSECA, R. Predicting dihedral angle probability distributions for protein coil residues from primary sequence using neural networks. **BMC Bioinformatics**, v. 10, n. 1, p. 338, 2009.

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

HORNIK, K. Approximation capabilities of multilayer feedforward networks. **Neural Networks**, v. 4, n. 2, p. 251—257, 1991.

HOVMOLLER, T.; OHLSON, T. Conformation of amino acids in protein. **Acta Crystallogr.**, v. 58, n. 5, p. 768–776, 2002.

HUANG, Y.-F.; CHEN, S.-Y. Extracting physicochemical features to predict protein secondary structure. **The Scientific World Journal**, v. 2013, 2013.

JOHNSON, S. Hierarchical clustering schemes. **Psychometrika**, v. 32, n. 2, p. 241–254, 1966.

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

JONES, E. et al. **SciPy: Open source scientific tools for Python**. 2001–. Online; accessed 2016-01-28. Available from Internet: <http://www.scipy.org/>.

KOLINSKI, A. Protein modeling and structure prediction with a reduced representation. **Acta Biochim. Pol.**, v. 51, p. 349–371, 2004.

KOOP, S. et al. Assessment of casp7 predictions for template-based modleing targets. **Proteins: Struct., Funct., Bioinf.**, v. 69, n. 8, p. 38–56, 2007.

KUBELKA, J. et al. Sub-microsecond protein folding. **Journal of molecular biology**, Elsevier, v. 359, n. 3, p. 546–553, 2006.

KUTHAN, T.; LANSKY, J. Genetic algorithms in syllable-based text compression. **Dateso**, p. 21–34, 2007.

LANDER, E.; WATERMAN, M. **The secrets of life: a mathematician's introduction to Molecular Biology**. Washington D. C., USA: National Academy Press, 1999. 300 p.

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

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? **J. Chim. Phys. Phys.-Chim. Biol.**, v. 65, n. 1, p. 44–45, 1968.

LILJAS, A. et al. Singapore: World Scientific Printers, 2001. 572 p.

LINNAINMAA, S. Taylor expansion of the accumulated rounding error. **BIT Numerical Mathematics**, v. 16, n. 2, p. 146–160, 1976.

LIU, J. et al. Crystal structure of the unique rna-binding domain of the influenza virus ns1 protein. **Nat. Struct. Biol.**, v. 4, p. 896–899, 1997.

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

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. **Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.**, p. 281—-297, 1967.

MARTÍ-RENOM, M. et al. Comparative protein structure modeling of genes and genomes. **Annu. Rev. Biophys. Biomol. Struct.**, v. 29, n. 16, p. 291–325, 2000.

MORIARTY, D. E. Symbiotic evolution of neural networks in sequential decision tasks. **Ph.D. thesis, Department of Computer Sciences, The University of Texas at Austin. Technical Report UT-AI97-257**, 1997.

MORIARTY, D. E.; MIIKKULAINEN, R. Efficient reinforcement learning through symbiotic evolution. **Machine Learning**, v. 22, p. 11–32, 1996.

MORIZE, I. et al. Refinement of the c222 1 crystal form of oxidized uteroglobin at 1.34 å resolution. **Journal of molecular biology**, Elsevier, v. 194, n. 4, p. 725–739, 1987.

NCBI-GENBANK. **GenBank and WGS Statistics**. 2016. Available from Internet: <http://www.ncbi.nlm.nih.gov/genbank/statistics/>.

NCBI-REFSEQ. **RefSeq Growth Statistics**. 2016. Available from Internet: <http://www.ncbi.nlm.nih.gov/refseq/statistics/>.

NEIDIGH, J.; FESINMEYER, R.; ANDERSEN, N. Designing a 20-residue protein. **Nat.Struct.Biol.**, v. 9, p. 425–430, 2002.

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.

NIAS-SERVER. **Mapas de Ramachandran**. 2016. Available from Internet: <www.sbcb.inf.ufrgs.br/nias>.

NOWICKA, U. et al. Dna-damage-inducible 1 protein (ddi1) contains an uncharacteristic ubiquitin-like domain that binds ubiquitin. **Structure**, Elsevier, v. 23, n. 3, p. 542–557, 2015.

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

PASTOR, M. T. et al. Combinatorial approaches: A new tool to search for highly structured beta-hairpin peptides. **Proc.Natl.Acad.Sci.USA**, v. 99, n. 2, p. 614–619, 2002.

PRUITT, K. et al. Refseq: an update on mammalian reference sequences. **Nucleic Acids Res. PubMed**, 2013.

PYTHON, S. F. **Python Language Reference, version 2.7.** 2016. "http://www.python.org".

RCSB-PDB. **PDB Current Holdings Breakdown**. 2016. Available from Internet: <http://www.rcsb.org/pdb/statistics/holdings.do>.

RCSB-PDB. **Yearly Growth of Protein Structures**. 2016. Available from Internet: <http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=molType-protein&seqid=100>.

RELIGA, T. et al. The helix-turn-helix motif as an ultrafast independently folding domain: The pathway of folding of engrailed homeodomain. **Proc.Natl.Acad.Sci.Usa**, v. 104, p. 9272–9277, 2007.

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

SAKAE, Y. et al. Protein structure predictions by parallel simulated annealing molecular dynamics using genetic crossover. **Journal of Computational Chemistry**, v. 32, n. 7, p. 1353—60, 2011.

SÁNCHEZ, R.; SALI, A. Advances in comparative protein-structure modeling. **Curr. Opin. Struct. Biol.**, v. 7, n. 2, p. 206–214, 1997.

SCHAUL, T. et al. Pybrain. **The Journal of Machine Learning Research**, JMLR. org, v. 11, p. 743–746, 2010.

SCHWARTZ, R. **Biological Modeling and Simulation: a survey of pratical models, algorithms, and numerical methods**. 1. ed. London, UK: MIT Press, 2008. 389 p.

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

SOHANGIR, S.; RAHIMI, S.; GUPTA, B. Neuroevolutionary feature selection using neat. **Journal of Software Engineering and Applications**, v. 7, n. 7, p. 562–570, 2014.

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

STANLEY, K. O.; MIIKKULAINEN, R. Evolving neural networks through augmenting topologies. **The MIT Press Journals Evolutionary Computation**, v. 10, n. 2, p. 99–127, 2002.

STAROVASNIK M.A., B. A. W. J. Structural mimicry of a native protein by a minimized binding domain. **Proc. Natl. Acad. Sci. USA**, v. 94, n. 19, p. 10080–10085, 1997.

TAYARANI, A. et al. Artificial neural networks analysis used to evaluate the molecular interactions between selected drugs and human cyclooxygenase2 receptor. **Iranian Journal of Basic Medical Sciences.**, v. 16, n. 11, p. 1196–1202, 2013.

TEETER, 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**, National Acad Sciences, v. 81, n. 19, p. 6014–6018, 1984.

TOUW, W. G. et al. A series of pdb related databases for everyday needs. **Nucleic Acids Research 2015 January**, v. 43, 2015.

TRAMONTANO, A. **Protein structure prediction: concepts and applications**. 1. ed. Weinheim, Germany: John Wiley and Sons, Inc., 2006. 208 p.

VANHOUCKE, V.; CHAKRABORTY, A. **Deep Learning: Take machine learning to the next level**. 2016. "https://www.udacity.com/course/deep-learning–ud730". Accessed 29/05/2016.

VERLI, H. **Bioinformática: da Biologia à Flexibilidade Molecular**. 1. ed. Sao Paulo, Brazil: SBBq, 2014. 292 p.

VITA, C. et al. Rational engineering of a miniprotein that reproduces the core of the cd4 site interacting with hiv-1 envelope glycoprotein. **Proc.Natl.Acad.Sci.USA**, v. 96, p. 13091–13096, 1999.

WALT, S. v. d.; COLBERT, S. C.; VAROQUAUX, G. The numpy array: A structure for efficient numerical computation. **Computing in Science Engineering**, v. 13, p. 22–30, 2011.

WANG, S. et al. Protein secondary structure prediction using deep convolutional neural fields. **ArXiv e-prints**, 2015.

WITTEN, I. H.; EIBE, F.; HALL, M. A. **Data Mining: Pratical Machine Learning Tools and Techniques**. 3. ed. Burlington, USA: Elsevier, 2011. 629 p.

XU, D. et al. Automated protein structure modeling in casp9 by i-tasser pipeline combined with quark-based ab initio folding and fg-md-based strcuture refinement. **Proteins: Struct., Funct., Bioinf.**, v. 79, n. 10, p. 147–160, 2011.

YAMANO, A.; HEO, N.; TEETER, M. Crystal structure of ser-22/ile-25 form crambin confirms solvent, side chain substate correlations. **J.Biol.Chem.**, v. 272, p. 9597–9600, 1997.

YAO, X. Evolving artificial neural networks. In: **Proceedings of the IEEE**. [S.l.: s.n.], 1999. v. 87, n. 9, p. 1423–1447.

ZHANG, Y. I-tasser server for protein 3d structure prediction. **BMC Bioinf.**, v. 9, n. 40, p. 1–8, 2008.