



UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA

IVAN MAZONI

ANÁLISE DO NANO-AMBIENTE PROPÍCIO PARA
NUCLEAÇÃO E MANUTENÇÃO DOS ELEMENTOS DA
ESTRUTURA SECUNDÁRIA NO CONTEXTO
ESTRUTURAL DAS PROTEÍNAS FUNCIONAIS

CAMPINAS

2018

IVAN MAZONI

**ANÁLISE DO NANO-AMBIENTE PROPÍCIO PARA
NUCLEAÇÃO E MANUTENÇÃO DOS ELEMENTOS DA
ESTRUTURA SECUNDÁRIA NO CONTEXTO
ESTRUTURAL DAS PROTEÍNAS FUNCIONAIS**

*Tese apresentada ao Instituto de
Biologia da Universidade Estadual de
Campinas como parte dos requisitos
exigidos para a obtenção do Título de
Doutor em Genética e Biologia
Molecular, na área de Bioinformática.*

ESTE ARQUIVO DIGITAL CORRESPONDE À
VERSÃO FINAL DA TESE DEFENDIDA PELO
ALUNO IVAN MAZONI E ORIENTADA PELO
PROF. DR. GORAN NESHICH.

Orientador: GORAN NESHICH

CAMPINAS

2018

Agência(s) de fomento e nº(s) de processo(s): Não se aplica.

ORCID: <https://orcid.org/0000-0002-3763-1071>

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

M458a Mazoni, Ivan, 1978-
Análise do nano-ambiente propício para nucleação e manutenção dos elementos da estrutura secundária no contexto estrutural das proteínas funcionais / Ivan Mazoni. – Campinas, SP : [s.n.], 2018.

Orientador: Goran Neshich.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Proteínas - Estrutura. 2. Conformação proteica em alfa-hélice. 3. Conformação proteica em folha beta. 4. Análise multivariada. I. Neshich, Goran. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Analysis of the proper nano environment for nucleating and maintaining the secondary structure elements in the structural context for functional proteins

Palavras-chave em inglês:

Proteins - Structure

Protein conformation, alpha-helical

Protein conformation, beta-strand

Multivariate analysis

Área de concentração: Bioinformática

Titulação: Doutor em Genética e Biologia Molecular

Banca examinadora:

Goran Neshich [Orientador]

Roberto Coiti Togawa

Ljubica Tasic

Ana Carolina de Mattos Zeri

Carla Geovana do Nascimento Macário

Data de defesa: 10-12-2018

Programa de Pós-Graduação: Genética e Biologia Molecular

COMISSÃO EXAMINADORA

Prof. Dr. Goran Neshich

Prof. Dr. Roberto Coiti Togawa

Profa. Dra. Ljubica Tasic

Profa. Dra. Ana Carolina de Mattos Zeri

Profa. Dra. Carla Geovana do Nascimento Macário

Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.

“Se és capaz de manter a tua calma quando
Todo o mundo ao teu redor já a perdeu e te culpa;
De crer em ti quando estão todos duvidando,
E para esses no entanto achar uma desculpa;
Se és capaz de esperar sem te desesperares,
Ou, enganado, não mentir ao mentiroso,
Ou, sendo odiado, sempre ao ódio te esquivares,
E não parecer bom demais, nem pretensioso;

Se és capaz de pensar – sem que a isso só te atires,
De sonhar – sem fazer dos sonhos teus senhores.
Se encontrando a desgraça e o triunfo conseguires
Tratar da mesma forma a esses dois impostores;
Se és capaz de sofrer a dor de ver mudadas
Em armadilhas as verdades que disseste,
E as coisas, por que deste a vida, estraçalhadas,
E refazê-las com o bem pouco que te reste;

Se és capaz de arriscar numa única parada
Tudo quanto ganhaste em toda a tua vida,
E perder e, ao perder, sem nunca dizer nada,
Resignado, tornar ao ponto de partida;
De forçar coração, nervos, músculos, tudo
A dar seja o que for que neles ainda existe,
E a persistir assim quando, exaustos, contudo
Resta a vontade em ti que ainda ordena: “Persiste!”;

Se és capaz de, entre a plebe, não te corromperes
E, entre reis, não perder a naturalidade,
E de amigos, quer bons, quer maus, te defenderes,
Se a todos podes ser de alguma utilidade,
E se és capaz de dar, segundo por segundo,
Ao minuto fatal todo o valor e brilho,
Tua é a terra com tudo o que existe no mundo
E o que mais – tu serás um homem, ó meu filho!”

Rudyard Kipling

Dedicatória

As meus pais (*in memoriam*) Douglas e Idê, e à minha querida esposa Giovana.

Agradecimentos

Agradeço ao amigo e orientador Goran Neshich, pela caminhada ao longo desses anos trabalhando juntos na Embrapa Informática Agropecuária. Entre as dificuldades que enfrentamos e as alegrias conquistadas, acredito que os momentos felizes superaram os tristes.

Agradeço aos meus pais (*in memoriam*), Douglas e Idê, por todo amor, e por sempre insistirem na importância dos estudos.

Agradeço a minha querida esposa, por sua paciência e compreensão quando, querendo estar ao meu lado, eu me ausentei para realizar meu trabalho. Obrigado, Giovana.

Minha gratidão à Chefia da Embrapa Informática Agropecuária, por permitir que esse trabalho de doutorado fosse realizado em serviço.

Obrigado aos colegas de laboratório Luiz, Salim, Fábio, Inácio e Jardine, que de alguma maneira contribuíram para esse trabalho.

Se eu pudesse, agradeceria pessoalmente cada professor que eu tive na vida. Obrigado aos professores da Escola Estadual Professor José Vilagelin Neto, da Escola SENAI Roberto Mange, do Colégio Politécnico Bento Quirino, da FATEC Americana, e aos professores dos diversos institutos e faculdades onde cursei as disciplinas desse doutorado. Como disse Paulo Freire: “*Ensinar não é transferir conhecimento, mas criar as possibilidades para sua própria produção ou a sua construção.*” Sou grato por criarem essas possibilidades.

RESUMO

As proteínas exercem um papel vital na manutenção da vida. Entre as diversas funções que as proteínas têm, destacam-se, por exemplo: proteínas estruturais, de transporte, proteção e defesa, controle e regulação de expressão, catálise, movimento e armazenamento. Para um entendimento melhor da relação entre a sequência de aminoácidos de uma proteína, sua estrutura tridimensional e a função desempenhada por ela, foi proposta a análise do nano-ambiente proteico onde os EES α -hélice, folha- β e *turn* estão inseridos. A hipótese que motivou essa abordagem é a existência de um sinal, ou seja, uma variação nos valores dos descritores físico-químicos e estruturais que distinguem o local específico onde determinado EES está inserido no arcabouço da proteína inteira. Entender como são formados os EES abrirá o caminho para compreendermos como as proteínas assumem sua estrutura final, e consequentemente, sua função. Neste trabalho utilizamos o STING_RDB, uma base de dados única no mundo, que reúne em um único repositório mais de 1500 descritores físico-químicos e estruturais de todos os resíduos de aminoácidos para cada cadeia de todas as estruturas proteicas depositadas no PDB (*Protein Data Bank*). As estruturas armazenadas no STING_RDB foram separadas em diferentes *Datamarts*, que são porções extraídas desta base de dados após uma seleção rígida. As estruturas selecionadas e guardadas nesses *Datamarts* foram então alinhadas posicionalmente pelo respectivo EES, e posteriormente extraíram-se desses alinhamentos os dados referentes aos descritores físico-químicos e estruturais que descrevem o nano-ambiente onde se insere o EES. Esse processo foi usado na busca dos “sinais”. Este trabalho descreve como os dados contidos nesses *Datamarts* foram selecionados, preparados, analisados e interpretados. Baseado nos resultados obtidos, concluímos que o nano-ambiente pode ser descrito não por um descritor, mas por um conjunto de descritores, e que essa descrição varia de acordo com o EES estudado. Isso diferencia o nano-ambiente do restante da proteína, e não apenas entre os diferentes tipos de EES.

ABSTRACT

Proteins play a vital role in maintaining life. Among the various functions that proteins have, one may for example cite those such as: structural proteins, transport, protection and defense, control and regulation of expression, catalysis, movement and storage. For a better understanding of the relationship between the amino acid sequence of a protein, its structure and the function performed by it, it was proposed the analysis of the protein nanoenvironment where α -helix, beta- β and turn secondary structure elements are inserted. The hypothesis that motivated this approach is the existence of a "signal", that is, a variation in the values of the physical-chemical and structural descriptors that distinguish the specific site where the secondary structure element is inserted in the framework of the entire protein. Understanding how the secondary structure elements are formed will open the way to understanding how proteins assume their final structure and hence their function. In this work we use STING_RDB , a unique database, which brings together in a single repository more than 1500 physical-chemical and structural descriptors of all amino acid residues for each chain of all protein structures deposited in the PDB (Protein Data Bank). The structures stored in STING_RDB have been separated into different *Datamarts*, which are portions extracted from the entire database, after a rigid selection. The structures selected and stored in these *Datamarts* were aligned by the respective secondary structure element, and subsequently the data referring to the physical-chemical and structural descriptors that describe the nanoenvironment where the secondary structure element is inserted are extracted from these positional alignments. This process was used in the search for "signs". This Thesis describes how the data contained in these *Datamarts* were selected, prepared, analyzed and interpreted. After this extensive work, we conclude that the nanoenvironment can be described not by a single descriptor, but by a set of descriptors, and that this description varies according to the secondary structure element studied. Also, any of the nanoenvironments suitable for the studied secondary element is different from the rest of the protein.

LISTA DE ABREVIATURAS E SIGLAS

DNA – Ácido Desoxirribonucleico
DSC – *Discrimination of Secondary structure Class*
DSSP – *Define Secondary Structure of Proteins*
ESS – Elemento(s) de Estrutura Secundária
Embrapa – Empresa Brasileira de Pesquisa Agropecuária
EP – Potencial Eletrostático
GOR – *Garnier-Osguthorpe-Robson*
GPBC – Grupo de Pesquisa em Biologia Computacional
HBMM – *Hydrogen Bond Main chain – Main chain*
HBMS – *Hydrogen Bond Main chain – Side chain*
HBMWM – *Hydrogen Bond Main chain – Water – Main chain*
HBMWS – *Hydrogen Bond Main chain – Water – Side chain*
HBMWWM – *Hydrogen Bond Main chain – Water – Water – Main chain*
HBMWWS – *Hydrogen Bond Main chain – Water – Water – Side chain*
HBSS – *Hydrogen Bond Side chain – Side chain*
HBSWS – *Hydrogen Bond Side chain – Water – Side chain*
HBSWWS – *Hydrogen Bond Side chain – Water – Water – Side chain*
HSSP – *Homology-derived Structures of Proteins*
IFR – *Interface Forming Residue*
LHA – Last Heavy Atom
MANOVA – Análise de Variância Multivariada
NMR – *Nuclear Magnetic Resonance Spectroscopy*
PDB – *Protein Data Bank*
PLN – Processamento de Linguagem Natural
PS³A – *Protein Secondary Structure STING Analyzer*
RMN – Ressonância Magnética Nuclear
RNA_m – Ácido Ribonucleico mensageiro
RNA_t – Ácido Ribonucleico transportador
SCOP – *Structural Classification Of Proteins*
SS – Side chain – Side chain
Stride – *Structural Identification*
SVM – *Support Vector Machine*
VMD – Visual Molecular Dynamics
WNA – *Weighted Neighbor Averages* (descritores ponderados pela vizinhança)

SUMÁRIO

1. INTRODUÇÃO	15
2. REVISÃO BIBLIOGRÁFICA.....	18
2.1 Estrutura primária	18
2.1.1 A iniciação da tradução em procariontes.....	18
2.1.2 A iniciação da tradução em eucariontes	19
2.2 Estrutura secundária.....	21
2.2.1 Estruturas helicoidais.....	22
2.2.2 Folhas- β paralelas e antiparalelas	25
2.2.3 <i>Turns</i>	26
2.3 Estrutura terciária.....	29
2.4 Estrutura quaternária	29
2.5 Classificação das proteínas pela presença do EES	30
2.6 Definição do EES usando os algoritmos <i>Define Secondary Structure of Proteins</i> (DSSP) e <i>Structural Identification</i> (Stride)	32
2.6.1 DSSP.....	32
2.6.2 Stride.....	33
2.7 Métodos de determinação das estruturas proteicas	35
2.7.1 Difração de raios X.....	35
2.7.2 Ressonância Magnética Nuclear (RMN)	35
2.7.3 Microscopia eletrônica	37
2.8 Métodos de predição do EES	38
2.8.1 Método de Chou-Fasman	39
2.8.2 Método GOR	40
2.8.3 Redes neurais	40
2.8.4 Máquina de vetor de suporte	41
2.8.5 Comparação entre os métodos de predição do EES	42
2.9 Softwares usados para previsão do EES	42
2.9.1 RaptorX	42
2.9.2 NetSurfP	42
2.9.3 Jpred	43
2.9.4 PredictProtein	43
2.9.5 YASSPP	43

2.9.6	SymPred	43
2.9.7	SSpro	44
2.9.8	DSC	45
2.9.9	PROFphd	46
2.9.10	PSIPRED	46
2.9.11	Predator.....	46
2.9.12	Comparação entre os softwares usados para predição do EES	46
2.10	Métodos de predição das estruturas proteicas.....	46
2.10.1	Modelagem por homologia.....	47
2.10.2	Modelagem <i>ab initio</i>	49
2.10.3	<i>Threading</i>	49
2.11	Plataformas para análise das estruturas proteicas	53
2.11.1	SwissModel	53
2.11.2	Visual Molecular Dynamics (VMD)	53
2.11.3	Jmol	53
2.11.4	Geneious Pro	53
2.11.5	PyMOL	54
2.11.6	RasMol	54
2.11.7	UCSF Chimera	54
2.11.8	Blue Star Sting.....	55
2.12	Bancos de dados com parâmetros estruturais	57
2.12.1	PDB	57
2.12.2	HSSP.....	57
2.12.3	UniProt	58
2.12.4	PROSITE.....	58
2.12.5	STING_RDB	58
2.13	Nano-ambiente dos aminoácidos	58
2.14	Aplicações do conhecimento adquirido pela análise do nano-ambiente onde se insere os EES.....	60
3	MATERIAIS E MÉTODOS	61
3.1	Criação dos <i>Datamarts</i>	61
3.2	Eliminação da redundância	67
3.3	Seleção dos dados	68
3.4	Descritores do STING_RDB	69

3.4.1	Potencial eletrostático.....	69
3.4.2	Acessibilidade.....	70
3.4.3	Espongicidade.....	72
3.4.4	Densidade	72
3.4.5	Space clash	72
3.4.6	Hidrofobicidade	72
3.4.7	Contatos	74
3.4.8	Ordem de Cross Link.....	75
3.4.9	Ordem de Cross Presence	75
3.4.10	Ângulos diedrais	76
3.4.11	Rotâmeros	76
3.4.12	Contatos não usados	76
3.4.13	Fator de temperatura.....	76
3.4.14	Parâmetros ponderados pela vizinhança (WNA).....	77
3.5	Extração, preparação e apresentação dos dados	77
3.6	Testes estatísticos.....	78
3.6.1	Teste de Kolmogorov-Smirnov	79
3.6.2	Teste t de Student	80
3.6.3	Teste de normalidade de Shapiro.....	81
3.6.4	Teste de correlação linear	82
3.6.5	Normalização dos dados	83
3.6.6	Análise multivariada (MANOVA)	84
3.6.7	p-value	85
3.6.8	Sliding Window	85
4	RESULTADOS E DISCUSSÃO	86
4.1	Informações quantitativas dos <i>Datamarts</i>	86
4.2	Eliminação da redundância	89
4.3	Número de EES alinhados	94
4.3.1	Número de cadeias nas proteínas do tipo all- α	94
4.3.2	Número de cadeias nas proteínas do tipo all- β	97
4.3.3	Número de cadeias nas proteínas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$	99
4.3.4	Número de cadeias nas proteínas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$	102
4.3.5	Número de cadeias nas proteínas do tipo desordenado	105

4.4	Comparação entre os sinais de um único descritor	107
4.5	Estudo de caso: comparação entre duas estruturas homólogas.....	108
4.6	Teste estatístico para um descritor selecionado	109
4.7	Análise comparativa entre α -hélices e hélices 3_{10}	113
4.8	Nano-ambiente definido por um conjunto de parâmetros	114
5	CONCLUSÕES.....	123
6	BIBLIOGRAFIA.....	126
	APÊNDICE A – NÚMERO DE CADEIAS COM CONSENSO ENTRE PDB_DSSP	133
	APÊNDICE B – NÚMERO DE CADEIAS COM CONSENSO ENTRE PDB_Stride	141
	APÊNDICE C – NÚMERO DE CADEIAS COM CONSENSO ENTRE DSSP_Stride	151
	APÊNDICE D – DADOS NORMALIZANDOS PELO ICV	160
	APÊNDICE E – DESCRITORES USADOS NO TESTE MANOVA PARA AS ESTRUTURAS ALINHADAS POR TAMANHO.....	167
	APÊNDICE F – DESCRITORES USADOS NO TESTE MANOVA PARA AS ESTRUTURAS ALINHADAS PELO C-TERMINAL	179
	APÊNDICE G – DESCRITORES USADOS NO TESTE MANOVA PARA AS ESTRUTURAS ALINHADAS PELO N-TERMINAL	191
	ANEXO 1 – ARTIGO PUBLICADO	203
	ANEXO 2 - TERMO DE BIOÉTICA/BIOSSEGURANÇA	229
	ANEXO 3 - DECLARAÇÃO DE DIREITOS AUTORAIS	230

1. INTRODUÇÃO

As proteínas são macromoléculas resultantes da combinação de 20 aminoácidos¹ através de ligações peptídicas. Considerando a combinação linear entre os 20 aminoácidos, o número de possíveis variações é de 20^n , onde n é a quantidade de resíduos de aminoácidos da proteína. Por exemplo, para uma proteína com 100 resíduos de aminoácidos, o número de possíveis combinações será igual a $20^{100} = 1,27 \times 10^{130}$. Em comparação, o número total estimado de átomos no Universo é de 9×10^{78} (VILLANUEVA).

Cada organismo, animal ou vegetal, possui milhares de diferentes proteínas. Entre as diversas funções que as proteínas têm, destacam-se: proteínas estruturais, de transporte, proteção e defesa, controle e regulação de expressão, catálise, movimento e armazenamento.

Quando vários resíduos de aminoácidos em série assumem determinadas conformações espaciais, essas conformações são chamadas de elementos de estruturas secundárias das proteínas. Essas conformações são estabilizados por ligações de hidrogênio dentro da cadeia peptídica ou entre duas cadeias vizinhas. Os principais elementos de estruturas secundárias são: α -hélice, folha- β e *turn* (volta).

Neste trabalho, nós analisamos o nano-ambiente onde um EES está presente. Essa região é formada pelos resíduos de aminoácidos que compõem uma α -hélice, folha- β ou *turn*, e os átomos que estão inseridos dentro de uma esfera de 15 Å centrada no último átomo mais pesado para cada um desses resíduos de aminoácidos. Ainda, para verificarmos o comportamento das características físico-químicas, estruturais, geométricas e dos contatos realizados entre os átomos formadores desses resíduos de aminoácidos, ao longo do tempo, incluímos nesse nano-ambiente os 32 resíduos de aminoácidos antes e depois do EES estudado.

Essas características foram extraídas do banco de dados relacional STING_RDB (OLIVEIRA, ALMEIDA, *et al.*, 2007), o que torna essa abordagem inédita. Embora outros grupos estejam trabalhando com a ideia de micro-ambiente (nomenclatura adotada) eles usam apenas a propensão de um resíduo de aminoácido fazer parte de uma α -hélice, folha- β ou *turn* (BAGLEY e ALTMAN, 1995). Nós optamos por usar as informações armazenadas no STING_RDB porque ele tem uma coleção completa e atualizada dos descritores das proteínas,

¹ Em alguns organismos o código genético traduzido pode incluir, por exemplo, a selenocisteína (JOHANSSON, GAFVELIN e AMÉR, 2005) e a pirolisina (GASTON, ZHANG, *et al.*, 2011). Mas por se tratar de casos muito específicos, consideramos apenas os 20 aminoácidos comumente encontrados nas proteínas.

englobando as características físico-químicas, estruturais, geométricas e dos contatos realizados. O STING_RDB é atualizado semanalmente, em sincronia com as novas estruturas disponíveis no PDB², que é uma base de dados pública onde semanalmente entre 150-200 novos arquivos, que descrevem uma estrutura proteica, são depositados e disponibilizados ao público.

A motivação deste trabalho foi analisar o nano-ambiente criado pelos resíduos de aminoácidos que formam o contexto proteico onde se encontram os elementos da estrutura secundária. Zhong (ZHONG e JOHNSON, 1992) e MacDonald (MACDONALD e JOHNSON JR, 2001) demonstraram que as características do nano-ambiente são importantes na determinação das estruturas secundárias de uma proteína. Entretanto, essa conclusão foi obtida usando um conjunto limitado de propriedades do nano-ambiente, conforme será explanado na revisão bibliográfica. Por outro lado, nós usamos um conjunto de dados mais amplo, armazenadas no banco de dados STING_RDB. Através do conhecimento das características do nano-ambiente que são capazes de manter uma determinada sequência dos aminoácidos pertencentes ao elemento da estrutura secundária, pretendemos ter condições de melhor compreender o relacionamento entre sequência, estrutura e função de uma proteína, como também estabelecer uma métrica de qualidade das estruturas que se refere ao nano-ambiente para cada distrito proteico que possui determinadas características.

Esta tese está organizada da seguinte forma: Capítulo 2 Revisão Bibliográfica, onde introduzimos as definições de estrutura primária, secundária, terciária e quaternária, detalhando as estruturas secundárias, que é o tema do trabalho. Esse conhecimento prévio é necessário para se entender os objetivos, metodologia, análise dos dados e conclusões descritas ao longo do texto. Logo após apresentamos os algoritmos usados na definição dos EES, uma vez que essas definições são usadas nas fases de criação e seleção dos dados. Como *background* explicaremos os métodos para determinar as estruturas proteicas, os métodos de previsão dos EES e uma revisão de alguns softwares usados para a predição das estruturas proteicas e uma análise comparativa entre algumas plataformas computacionais usadas na análise dessas estruturas. Introduziremos alguns bancos de dados com parâmetros estruturais, incluindo o STING_RDB, de onde os dados utilizados neste trabalho foram extraídos; Capítulo 3 Materiais e Métodos, há uma descrição sobre como os *Datamarts* empregados neste trabalho foram criados, por que e como eliminamos a redundância entre as estruturas

² <https://www.rcsb.org/>

proteicas extraídas do PDB, como foi feita a seleção dos dados, alguns descritores do STING_RDB e seu significado, e, finalmente, como foi feita a extração, preparação e apresentação dos dados; Capítulo 4 Resultados e Conclusões onde mostraremos o resultado dos testes estatísticos uni e multivariado aplicados nos diferentes *Datamarts*; e finalmente o Capítulo 5 Conclusões, onde explicamos porque concluimos que a hipótese de que existe uma especificidade para cada tipo de nano-ambiente é verdadeira, e como esse trabalho contribuiu no avanço do conhecimento acerca dos nano-ambientes.

2. REVISÃO BIBLIOGRÁFICA

Para serem capazes de desempenhar suas funções biológicas, as proteínas se dobram a partir de uma cadeia de aminoácidos, adquirindo a forma de estrutura secundária e terciária. Finalmente, algumas interagem com várias cadeias originando estruturas quaternárias. Essas dobras são causadas por uma série de interações hidrofóbicas, formação de ligações de hidrogênio e forças de Van der Waals, e ocasionalmente, pela formação das ligações de dissulfeto. As estruturas proteicas variam em tamanho, desde algumas dezenas a milhares de resíduos de aminoácidos. Segundo Alberts, (ALBERTS, JOHNSON e LEWIS, 2002) as estruturas formadas por mais de 50 aminoácidos são consideradas proteínas, e as que possuem menos de 50 aminoácidos são denominadas de peptídeos.

2.1 Estrutura primária

A estrutura primária de uma proteína é formada por uma sequência de unidades estruturais chamadas aminoácidos. Por convenção, a estrutura primária começa no seu terminal amino (N-Terminal) e acaba no seu terminal carboxila (C-Terminal). Porém, é importante observar que isto não é suficiente para que duas sequências representem a mesma proteína, uma vez que cada aminoácido pode estar em posições diferentes em cada cadeia como o exemplo a seguir.



não é o mesmo que



Apesar de serem semelhantes nos aspectos gerais, a síntese de proteínas difere entre células procarióticas e células eucarióticas. A seguir discutimos as principais diferenças.

2.1.1 A iniciação da tradução em procariontes

Nos procariontes, a **sequência de Shine Dalgarno**, ou sítio de ligação do ribossomo, é uma sequência de bases no RNAm que existe próximo ao códon AUG (códon de iniciação). Essa sequência, de tamanho entre 30 e 40 nucleotídeos, é responsável por permitir que a subunidade menor do ribossomo se ligue ao RNAm e posicione o ribossomo

diretamente no códon AUG, dando início à tradução dos nucleotídeos em aminoácidos (Fig. 1). A **sequência de Shine Dalgarno** só existem em procariontes.

305

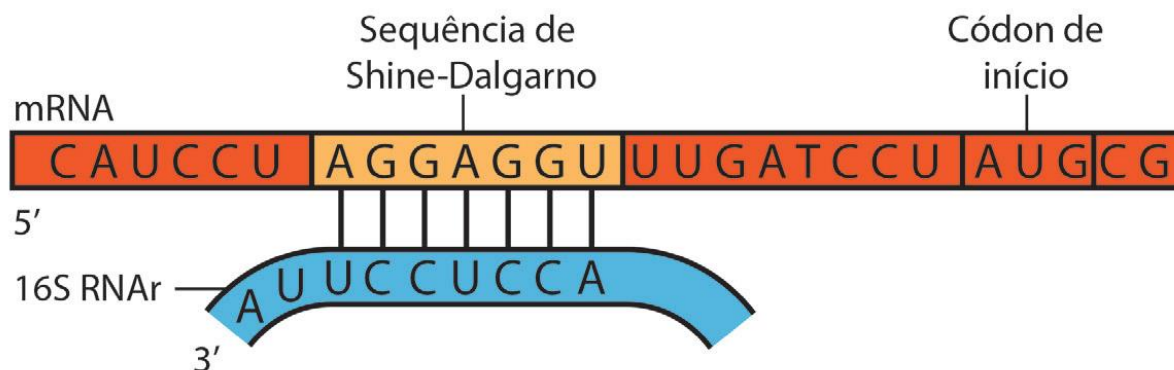


Figura 1. Em bactérias, as bases complementares entre a extremidade 3' do rRNA 16S da subunidade ribossômica menor e a sequência de Shine-Dalgarno do rRNA posicionam o ribossomo para iniciar a tradução do códon de iniciação AUG localizado a seguir. Fonte: https://edisciplinas.usp.br/pluginfile.php/3005466/mod_resource/content/1/BiologiaMolecular_texto07final%20%283%29.pdf consultado em 18 de dezembro de 2018.

2.1.2 A iniciação da tradução em eucariontes

No caso dos eucariontes, a ponta 5' liga-se ao ribossomo com a ajuda de algumas proteínas que reconhecem o **cap 5'**. Então a molécula de rRNA desliza ao longo do ribossomo até o primeiro códon AUG entrar no ribossomo – iniciando a tradução.

A tradução do DNA determina a estrutura primária das proteínas. O código genético presente no DNA é transcrito em uma molécula de rRNA pelo RNA polimerase. No caso dos eucariontes, porções de DNA (*íntrons*) são eliminados e as porções restantes (*exons*) são juntadas dando origem ao rRNA. *Íntrons* são sequências de nucleotídeos que não são traduzidos na síntese proteica, por isso são eliminados. A tradução do código genético encontrado no rRNA produz uma cadeia de aminoácidos que formam, no processo de enovelamento, primeiramente os elementos da estrutura secundária, e posteriormente esses elementos contribuem para a forma final da estrutura 3D da proteína (Fig. 2).

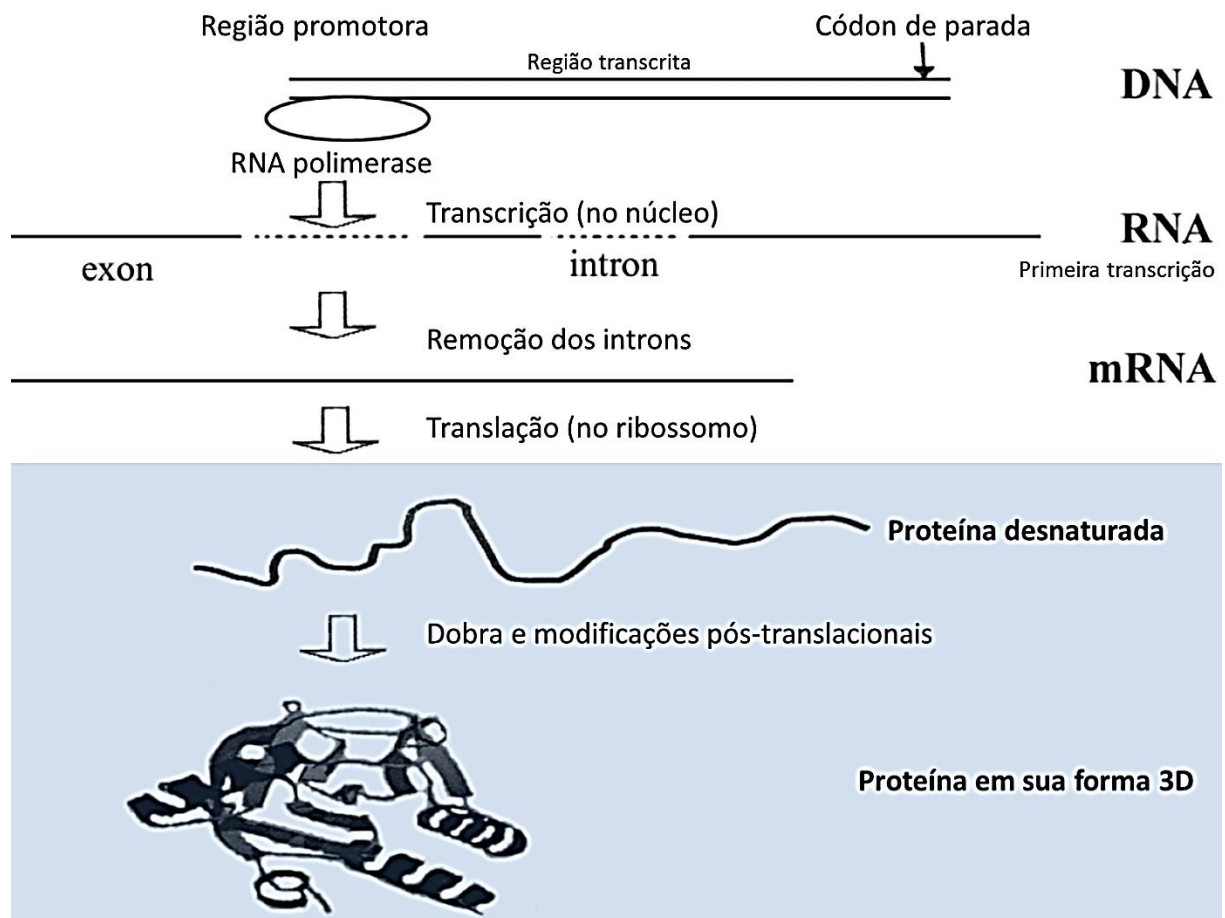


Figura 2. Esquema da síntese proteica em eucariontes. Embora o processo da transcrição DNA → RNA → RNAm → aminoácidos seja bem compreendido, os processos que regem a dobra da proteína a partir da sua sequência primária ainda não são totalmente conhecidos. Fonte: (ORENGO, JONES e THORNTON, 2003)

O processo de síntese proteica costuma ser dividido em tres partes: **iniciação**, **alongamento** e **término**. Na **iniciação** acontece a ligação do ribossomo ao RNAm formando um complexo de iniciação que contem o primeiro aminoacil-RNAt (N-formilmetionina nas bactérias e metionina em eucariontes). Essa etapa precede a união dos primeiros aminoácidos. A fase do **alongamento** compreende a formação da primeira ligação peptídica até a incorporação do último aminoácido do peptídeo. Em bactérias, aproximadamente 15 aminoácidos são adicionados por segundo à cadeia polipeptídica. Isso significa que a síntese de uma proteína com 300 aminoácidos leva cerca de 20 segundos. Nos eucariontes a velocidade é menor. São adicionados aproximadamente dois aminoácidos por segundo, logo, a síntese de uma proteína de mesmo número de aminoácidos leva cerca de 2'30". Finalmente, na fase de **término**, o ribossomo se dissocia do RNAm, liberando o polipeptídeo pronto.

2.2 Estrutura secundária

A estrutura secundária é a forma tridimensional regular que segmentos da estrutura primária assumem. Nas proteínas, a estrutura secundária é definida pelos padrões de ligações de hidrogênio entre os grupos amino da espinha dorsal e carbóximo.

As ligações de hidrogênio podem ser classificadas quanto a sua intensidade. Quando o hidrogênio se une a elementos com volume atômico baixo e eletronegativos, por exemplo, o oxigênio (O), o flúor (F) e o nitrogênio (N) a ligação é **forte**, pois existe uma grande diferença de eletronegatividade entre os elementos. No caso do hidrogênio não interagir diretamente com O, F e N, a força de interação é **intermediária**. Os elétrons estão distribuídos de forma assimétrica e o elemento mais eletronegativo atrai os elétrons para si. Finalmente, quando não há atração elétrica entre as moléculas, diz-se que a intensidade da ligação é **fraca**. Este tipo de interação também é chamado de forças de London ou forças de Van der Waals.

A estrutura secundária também é definida com base nos ângulos diedrais ϕ e ψ da espinha dorsal da cadeia proteica em uma região particular do gráfico de Ramachandran (RAMACHANDRAN, RAMAKRISHNAN e SASISEKHARAN, 1963). Ramachandran definiu os ângulos permitidos onde se mapeiam os EES (Fig. 3). Os EES podem ser agrupados em quatro grandes grupos: estruturas helicoidais (fitas 2.27, α -hélice, hélice 3_{10} , π -hélice), folhas- β (paralelas, antiparalelas, ponte- β), *turns* (tight *turn*, multiple *turn*, hairpins, multiple *turns*, *turns* tipo I, II, VIII, I', II', VIa1, VIa2, VIb, IV) e *random coils*.

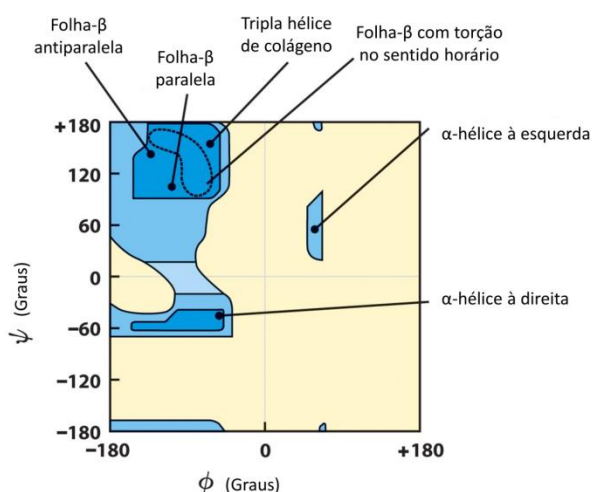


Figura 3. Gráfico de Ramachandran, demonstrando as áreas mais favoráveis à presença de cada EES. Fonte: https://commons.wikimedia.org/wiki/File:Ramachandran%27s_Diagram.jpg consultado e modificado 16 de maio de 2018.

2.2.1 Estruturas helicoidais

Nas estruturas helicoidais, a cadeia principal de uma proteína está “enrolada” ao redor de um eixo imaginário, formando uma hélice. O número de resíduos de aminoácidos envolvidos em uma volta completa determina o tipo da estrutura helicoidal (Fig. 4), sendo quatro tipos possíveis:

1. **α -hélice** (3.6_{13}): estrutura helicoidal mais abundante em proteínas nativas
2. **Fita 2.2₇**: nunca foi observada em proteínas naturais, porém teoricamente possível.
3. **Hélice 3₁₀**: encontrada em aproximadamente 26% das proteínas
4. **π -hélice** (4.4_{16}): encontrada em cerca de 15% do PDB, devido a sua inserção dentro das α -hélices (COOLEY, ARP e KARPLUS, 2010)

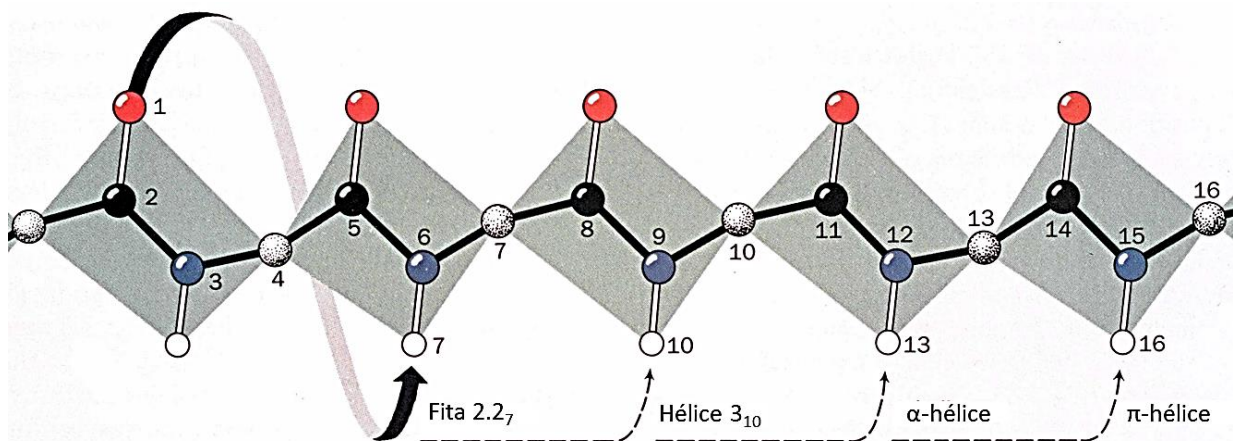


Figura 4. Possíveis tipos de estruturas helicoidais: fitas 2.27, hélice 3₁₀, α -hélice, π -hélice. Fonte: Notas de aula da disciplina NG-110 Tópicos Especiais em Genética: Introdução à estrutura de proteínas, oferecida pelo Instituto de Biologia da Unicamp no 1S/2008.

Segundo Chou e Fasman (CHOU e FASMAN, 1974) os aminoácidos mais presentes nas estruturas helicoidais, em ordem da maior para menor frequência, são:

Ala – Leu – Val – Lys – Glu – Ser – Thr – Gly – Gln – Asp – Ile – Asn – His – Phe – Arg – Tyr – Pro – Trp – Cys – Met

i. α -hélice

A α -hélice foi descrita por Linus Pauling em 1951 (PAULING, COREY e BRANSON, 1951). Os aminoácidos estão arranados em uma estrutura helicoidal com 3,6 resíduos de aminoácidos em cada volta completa. Como a distância vertical entre os resíduos de aminoácidos é de 1,5Å no eixo Y, temos um deslocamento de 5,4Å ($3,6 \times 1,5\text{Å}$) ao longo

do eixo da α -hélice em cada volta. Os ângulos diedrais para a α -hélice são $\phi = -57^\circ$ e $\psi = -47^\circ$. A Fig. 5 apresenta o esquema de uma α -hélice.

ii. *Fita 2.2₇*

Considerando a cadeia principal, na fita 2.2₇ a hélice tem 2,2 resíduos de aminoácidos por volta, e 7 átomos entre o oxigênio da carbonila e o hidrogênio da amida com o qual é estabelecida a ligação de hidrogênio.

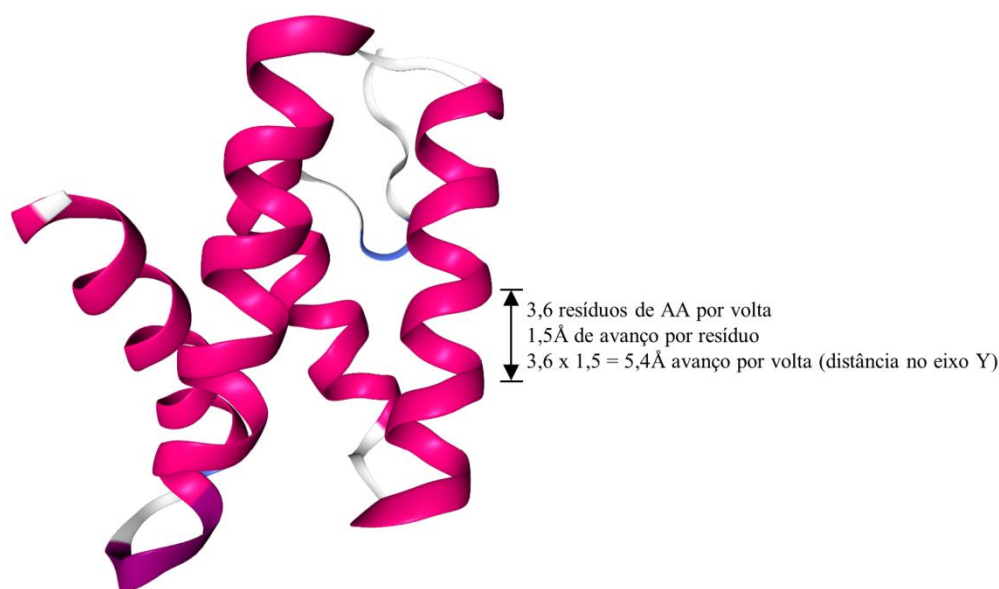


Figura 5. Representação de uma α -hélice. A estrutura abaixo é do PDB 3f2e (“Viral Protein Crystal Structure Of Yellowstone Sirv Coat Protein C-Terminus”). Cada volta tem 3,6 resíduos de aminoácidos com distância vertical de 1,5 Å entre dois resíduos de aminoácidos, o que significa $3,6 \times 1,5 \text{ Å} = 5,4 \text{ Å}$ de deslocamento ao longo do eixo em cada volta. A estabilização se dá pela presença das ligações de hidrogênio entre os grupamentos NH e CO da cadeia principal. O grupamento CO de cada aminoácido forma ponte de hidrogênio com o grupamento NH do aminoácido que está situado a quatro aminoácidos adiante na sequência linear. A imagem foi extraída do site do PDB usando o visualizador NGL.

iii. *Hélice 3₁₀*

Hélice 3₁₀ é um EES onde cada aminoácido encontra-se a 120° na volta da hélice. Isto significa que a hélice tem três resíduos de aminoácidos por volta, e uma translação de 2 Å ao longo do eixo longitudinal. A repetição $i+3 \rightarrow i$ ligações de hidrogênio forma uma hélice 3₁₀, ou seja, ela é formada pelas ligações de hidrogênio entre três resíduos de aminoácidos próximos. Hélices 3₁₀ são encontradas no final de uma α -hélice, compostas de um *turn* e fazendo a transição com o próximo EES. Os valores dos ângulos diedrais ϕ e ψ nas hélices 3₁₀ geralmente são: -49° e 26° respectivamente. A Fig. 6 apresenta um exemplo de estrutura proteica com hélice 3₁₀.



Figura 6. A estrutura 1din.pdb (“Hydrolytic Enzyme Dienelactone Hydrolase At 2.8 Angstroms”) é um exemplo de uma proteína com a presença de hélice 3_{10} (destacada em vermelho). O espaço ocupado por uma volta completa é maior que em uma α -hélice. Imagem produzida pelo software JSmol.

iv. π -hélice

π -hélice é um EES onde cada aminoácido corresponde a 87° na volta da hélice. Isto significa que a hélice tem 4,1 resíduos de aminoácido por volta, e uma translação de 1,15 Å ao longo do eixo Y. A repetição $i+5 \rightarrow i$ ligações de hidrogênio forma uma π -hélice, ou seja, ela é formada pelas ligações de hidrogênio entre resíduos de aminoácidos próximos distanciados por quatro aminoácidos intermitentes na estrutura primária. A Fig. 7 apresenta a representação de uma π -hélice.

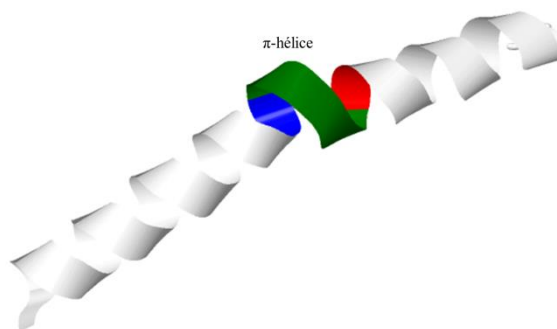


Figura 7. A estrutura 3qhb.pdb (“Metal Transport Crystal Structure Of Oxidized Symerythrin From Cyanophora Paradoxa”) é um exemplo de uma proteína com a presença de π -hélice (resíduos #40-45 da cadeia A. O resíduo 40 está destacado em azul, os resíduos 41-44 em verde e o resíduo 45 em vermelho). O diâmetro da π -hélice é maior que em uma α -hélice. A imagem foi produzida usando o software JSMol.

2.2.2 Folhas- β paralelas e antiparalelas

A primeira descrição teórica da folha- β foi feita na década de 1930. Astbury e Woods (ASTBURY e WOODS, 1934) propuseram a ideia de ligações de hidrogênio entre as ligações peptídicas no sentido paralelo e antiparalelo ao longo das fitas- β , conforme ilustrado na Fig. 8. A folha- β consiste em fitas- β conectadas lateralmente por três ou mais ligações de hidrogênio, formando uma folha geralmente antiparalela. Os ângulos diedrais (ϕ , ψ) são (119° , 113°) na folha- β paralela e (139° , 135°) na folha- β antiparalela. A fita- β tem, tipicamente, entre cinco e dez aminoácidos.

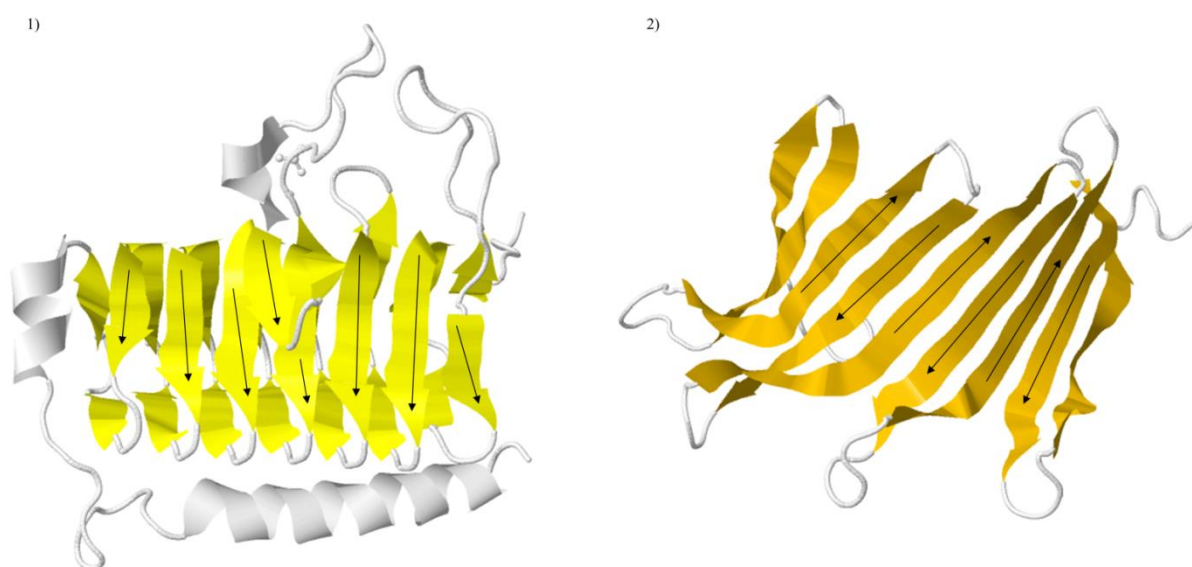


Figura 8. Exemplo de folhas- β paralelas e antiparalelas. Na imagem (1) vemos as fitas- β com a mesma direção (setas no mesmo sentido) formando uma folha- β paralela. Na imagem (2) vemos as fitas- β não paralelas (setas com sentidos opostos em pares umas das outras) formando uma folha- β antiparalela. As imagens foram produzidas pelo software JSmol usando as estruturas: (1) 1qre.pdb: “A Closer Look At The Active Site Of Gamma-Carbonic Anhydrases: High Resolution Crystallographic Studies Of The Carbonic Anhydrase From *Methanosarcina Thermophila*”, e (2) 3msw.pdb: “Crystal Structure Of A Protein With Unknown Function (Bf3112) From *Bacteroides Fragilis* Nctc 9343 At 1.90 Å Resolution”.

Segundo Chou e Fasman (CHOU e FASMAN, 1974) os aminoácidos mais frequentes nas fitas- β por ordem de frequência, da esquerda para a direita, são:

Val – Leu – Ala – Gly – Thr – Ile – Ser – Lys – Tyr – Gln – Phe – Asn – Asp – Arg – Cys – His – Pro – Trp – Met – Glu

As ligações- β são formadas por um único par de ligações de hidrogênio. A Fig. 9 mostra um exemplo de estrutura com a presença da ponte- β .

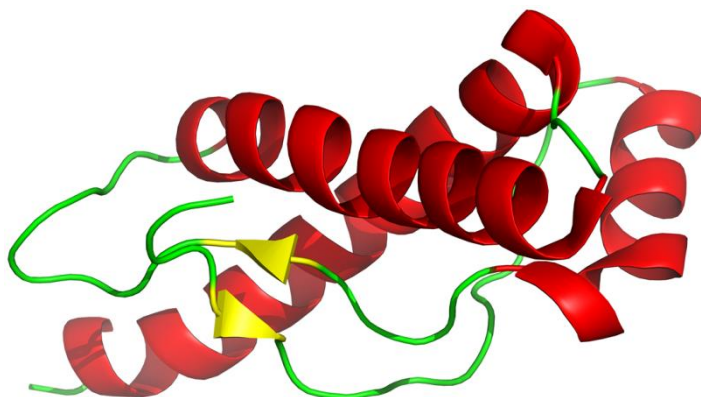


Figura 9. Na estrutura 1hjm.pdb ("Prion Protein Human Prion Protein At Ph 7.0") vê-se uma ponte- β destacada em amarelo. A imagem foi produzida pelo software PyMOL.

2.2.3 Turns

De acordo com a definição de Rose (ROSE, GIERASCH e SMITH, 1985), *turn* é um elemento estrutural em que os átomos C_{α} de dois resíduos de aminoácidos separados por 1 a 5 ligações peptídicas estão em estreita aproximação ($< 7 \text{ \AA}$), enquanto que os resíduos de aminoácidos correspondentes não formam um EES "normal", tais como uma α -hélice ou folha- β . Contrariamente às hélices, os ângulos diedrais da espinha dorsal da cadeia principal não são (aproximadamente) constantes para todos os resíduos de aminoácidos em curva. Os tipos de *turns* são: *tight turn*, *loops*, *turns* múltiplos e β -hairpins.

i. *Tight turns*

Segundo Toniolo (TONIOLO, 1980) *turns* são classificados de acordo com a separação entre os seus dois resíduos de aminoácidos finais em:

- **α -turn:** o resíduo final está separado por quatro ligações peptídicas ($i \rightarrow i \pm 4$)
- **β -turn:** o resíduo final está separado por três ligações peptídicas ($i \rightarrow i \pm 3$)
- **Υ -turn:** o resíduo final está separado por duas ligações peptídicas ($i \rightarrow i \pm 2$)
- **δ -turn:** o resíduo final está separado por uma ligação peptídica ($i \rightarrow i \pm 1$)
- **π -turn:** o resíduo final está separado por cinco ligações peptídicas ($i \rightarrow i \pm 5$)

Os *turns* também podem ser classificados por seus ângulos diedrais. Um *turn* pode ser convertido em um *turn* invertido, mudando o sinal para todos os seus ângulos diedrais. Assim, o Υ -turn tem duas formas: a clássica ($\phi = 75^\circ$, $\psi = -65^\circ$) e a forma inversa ($\phi = -75^\circ$, $\psi = 65^\circ$). A Tabela 1 apresenta os tipos de *turn* e seus respectivos ângulos diedrais.

Tipo	ϕ_{i+1}	Ψ_{i+1}	ϕ_{i+2}	Ψ_{i+2}
I	-60°	-30°	-90°	0°
II	-60°	120°	80°	0°
VIII	-60°	-30°	-120°	120°
I'	60°	30°	90°	0°
II'	60°	-120°	-80°	0°
VIa ₁	-60°	120°	-90°	0° ^(*)
VIa ₂	-120°	120°	-60°	0° ^(*)
Vib	-135°	135°	-75°	160° ^(*)
IV	Turns excluídos de todas as categorias acima			

Tabela 1. Ângulos ideais para diferentes *turns*. ^(*) Os tipos VIa₁, VIa₂ e Vib estão sujeitos à condição adicional que o resíduo (*i*+2) deve ser uma cis-prolina. Fonte: (VENKATACHALAM, 1968)

ii. *Loops*

Loops são segmentos mais longos, desordenados, sem ligações de hidrogênio fixas (Fig. 10).

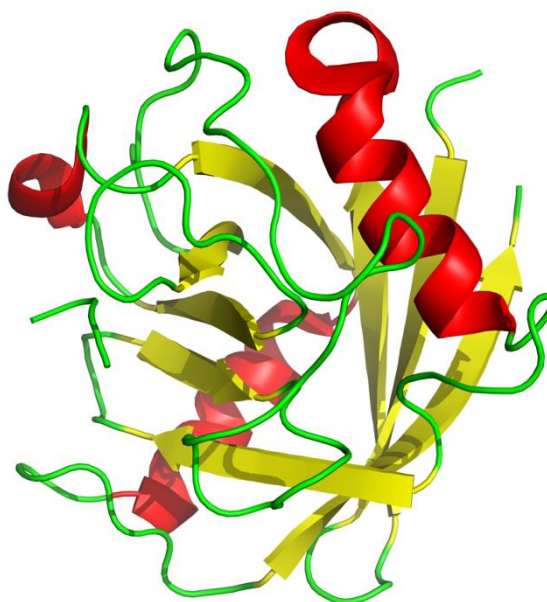


Figura 10. Estrutura do 1rmh.pdb ("Isomerase Substrate Recombinant Cyclophilin A From Human T Cell"). Na imagem os loops estão marcados em verde. Imagem produzida pelo software PyMOL.

iii. Turns múltiplos

Turns múltiplos ocorrem quando um ou mais resíduos de aminoácidos estão envolvidos em dois *turns* parcialmente sobrepostos (Fig. 11). Por exemplo, em uma sequência de cinco resíduos de aminoácidos, os resíduos 1-4 e os resíduos 2-5 formam um *turn*. Neste caso, o EES ($I + I + 1$) é chamando *turn* duplo. *Turns* múltiplos (acima de sete dobras) são mais frequentes que *single turns* (HUTCHINSON e THORNTON, 1994).

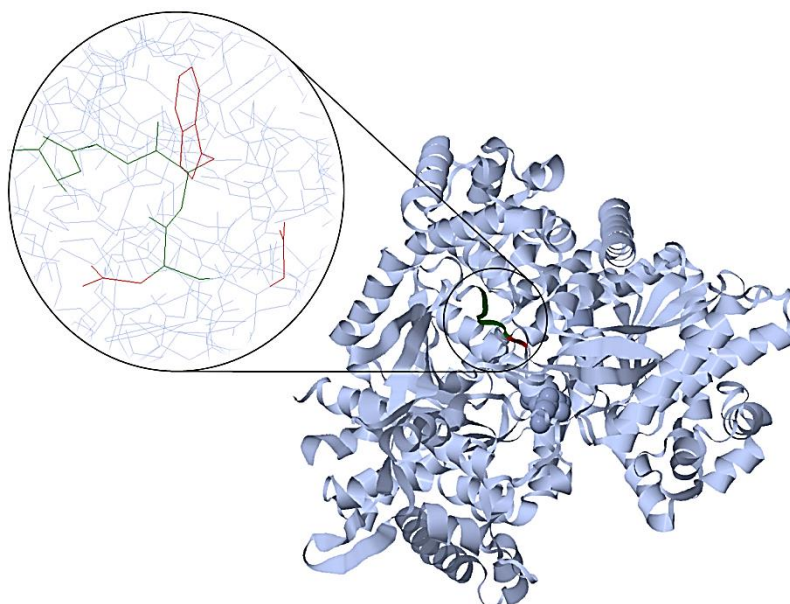


Figura 11. Exemplo de multiple *turn*. Na estrutura 1a8i.pdb ("spirohydantoin Inhibitor Of Glycogen Phosphorylase") os resíduos 180-182 (destacado em **vermelho**) e os resíduos 181-184 (destacados em **verde**) formam um *turn*, de modo que os resíduos 181-182 estão sobrepostos. Na imagem expandida é possível observar a sobreposição desses resíduos de aminoácidos (GURUPRASAD K, 2000). As figuras foram geradas pelo software Blue Star Sting.

iv. β -hairpins

β -hairpin é um tipo de *turn*, formado por quatro resíduos de aminoácidos, que conecta duas fitas- β , promovendo a formação de folhas- β antiparalelas (SIBANDA, BLUNDELL e THORNTON, 1989). A Fig. 12 mostra um exemplo de β -hairpin.

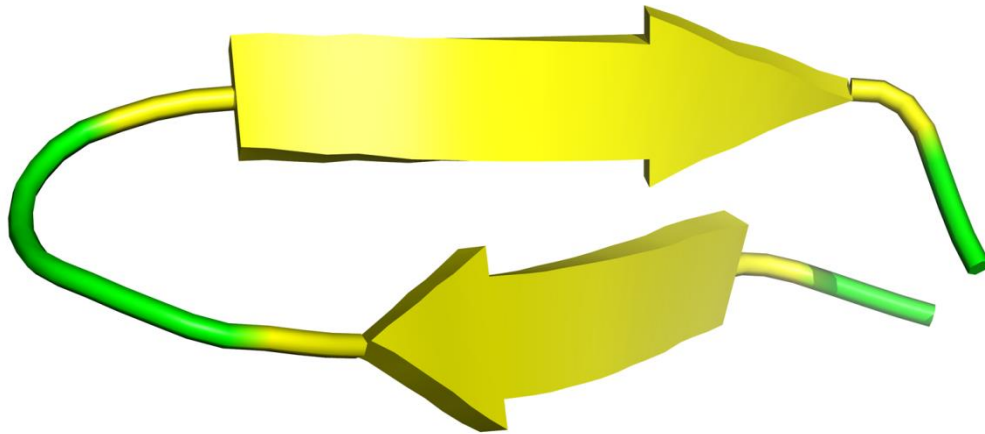


Figura 12. Exemplo de β -hairpin. A estrutura 1le1.pdb ("Nmr Structure Of Tryptophan Zipper 2: A Stable, Monomeric Beta-Hairpin With A Type I' Turn") apresenta um β -hairpin, em verde, ligando as duas fitas- β numa conformação antiparalela. Imagem gerada usando o software PyMOL.

v. *Random coil*

No *random coil* a distância entre os átomos C_α é aproximadamente constante (3,8 Å) e o traçado da cadeia principal pode ser descrito como uma série de laços virtuais que liga os C_α . Para cada resíduo de aminoácido no *random coil* há uma distribuição de ângulos ϕ e ψ , dando origem a um conjunto de conformações (SMITH, FIEBIG, *et al.*, 1996).

2.3 Estrutura terciária

Os EES α -hélices, folhas- β , *turns* e *random coils* são dobrados em uma conformação compacta, formando uma estrutura tridimensional denominada estrutura terciária. Interações hidrofóbicas influenciam na formação da estrutura terciária, que se estabilizam através de suas ligações salinas, ligações de hidrogênio e ligações dissulfeto. A função de uma proteína depende da sua estrutura terciária, e quando ela perde essa estrutura, diz-se que ela está desnaturada e perdeu sua função.

2.4 Estrutura quaternária

A estrutura quaternária é a estrutura tridimensional formada por uma proteína de múltiplas subunidades, ou seja, mais de uma cadeia polipeptídica. A estrutura quaternária é estabilizada pelas mesmas interações não covalentes e pelas ligações dissulfeto que estabilizam a estrutura terciária. Proteínas formadas por duas ou mais subunidades são chamadas multímeros. Especificamente, uma proteína com duas subunidades é chamada dímero, com três subunidades é chamada trímero, e assim sucessivamente (Tabela 2). Multímeros formados por subunidades idênticas são designados pelo prefixo "homo"

(homotetramero, por exemplo), e aqueles formados por subunidades diferentes são designados pelo prefixo “hetero” (*heterotetramero*, por exemplo).

Número de cadeias e nomenclatura	Número de PDBs	% do total de arquivos PDB	Número de cadeias e nomenclatura	Número de PDBs	% do total de arquivos PDB
1 = monômero (101m.pdb)	55601	39,4	13 = tridecamero (1c17.pdb)	83	0,1
2 = dímero (100d.pdb)	43605	30,9	14 = tetradecamero (1avh.pdb)	210	0,1
3 = trímero (10mh.pdb)	8887	6,3	15 = pentadecamero (1a8r.pdb)	137	0,1
4 = tetramero (105d.pdb)	17143	12,2	16 = hexadecamero (1c5f.pdb)	440	0,3
5 = pentâmero (168l.pdb)	1470	1,0	17 = heptadecamero (1slh.pdb)	21	0,0
6 = hexamero (1a2v.pdb)	5410	3,8	18 = octadecamero (1f9e.pdb)	146	0,1
7 = heptamero (1a10.pdb)	377	0,3	19 = nonadecamero (1p6g.pdb)	17	0,0
8 = octamero (190d.pdb)	3191	2,3	20 = eicosamero (1bos.pdb)	152	0,1
9 = nonamero (1c7y.pdb)	294	0,2	21-mero (1aon.pdb)	87	0,1
10 = decamero (1aol.pdb)	836	0,6	22-mero (1bgy.pdb)	51	0,0
11 = undecamero (1be3.pdb)	102	0,1	23-mero (1c9s.pdb)	75	0,1
12 = dodecamero (1a8l.pdb)	1485	1,1	Etc.		

Tabela 2. Nomenclatura das proteínas formadas por uma ou mais cadeias polipeptídicas, e seus respectivos exemplos de PDB. A lista completa pode ser acessada no software PDB Metrics do Blue Star Sting (www.cbi.cnptia.embrapa.br). Em 6 de maio de 2018 o total de PDB era 140390.

2.5 Classificação das proteínas pela presença do EES

Com relação à presença do EES, as proteínas podem ser classificadas em:

- all- α : quando não possuem folhas- β (Fig. 19.A);
- all- β : quando não possuem α -hélices (Fig. 19.B);
- $\alpha+\beta$: quando os EES α -hélice e folhas- β comparecem em ordem aleatória (Fig. 19.C);
- α/β : quando os EES α -hélice e folhas- β comparecem na ordem folha- β – α -hélice – folha- β (Fig. 19.D);
- “desordenadas”: quando não possuem nenhuma α -hélice nem folha- β (Fig. 19.E).

A)

B)

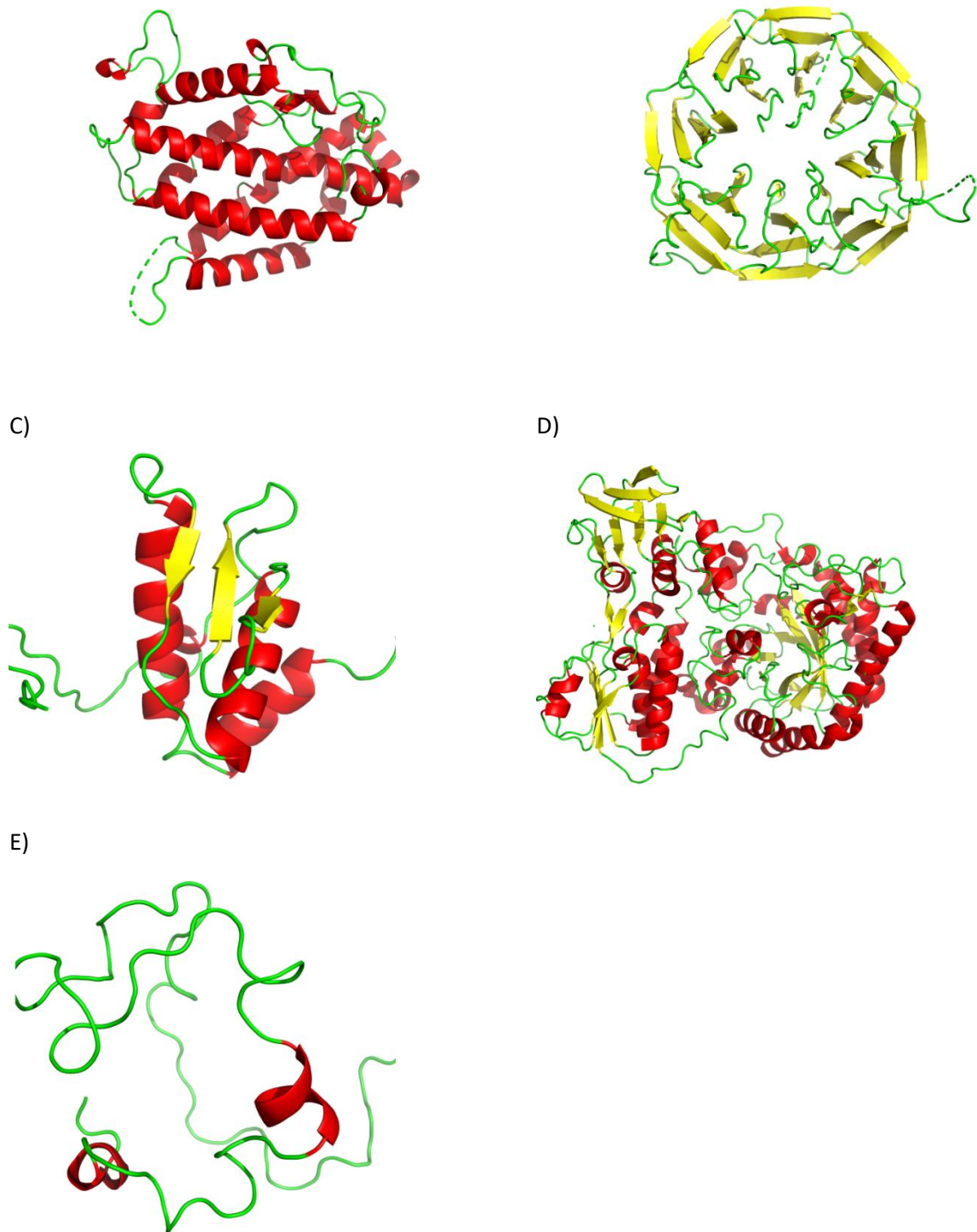


Figura 13. Exemplos de proteínas dos tipos (A) all- α , (B) all- β , (C) ($\alpha+\beta$), (D) (α/β) e (E) “desordenadas”. As estruturas usadas foram: A) 5csm.pdb “Complex Isomerase/Peptide) Yeast Chorismate Mutase, T226s Mutant, Complex With Trp”; B) 3q7m.pdb “Protein Binding The Crystal Structure Of Bamb From The Bam Complex In Spacegroup I222”; C) 1whr.pdb: “structural Genomics, Unknown Function Solution Structure Of The R3h Domain From Human Hypothetical Protein Baa76846”; D) 1ps9.pdb: “Oxidoreductase The Crystal Structure And Reaction Mechanism Of E. Coli 2,4-2 Dienoyl Coa Reductase”; E) 2ju4.pdb: “NMR Structure Of The Gamma Subunit Of Cgmp Phosphodiesterase”. Todas as imagens foram obtidas usando o software PyMOL.

2.6 Definição do EES usando os algoritmos *Define Secondary Structure of Proteins* (DSSP) e *Structural Identification* (Stride)

No arquivo PDB temos a indicação de onde estão os EES na estrutura proteica (Fig. 13). A informação contida nesse campo de um arquivo PDB é atribuída ao seu autor (Proteopedia).

HELIX	1	1	ALA E	55	ALA E	60	5
HELIX	2	2	ASN E	63	VAL E	64	5
HELIX	3	3	PRO E	230	GLN E	243	5MIXED 3/10 + 3.6/13
SHEET	1	B1	7 PRO E	28	ARG E	36	0
SHEET	2	B1	7 GLY E	38	PRO E	49	-1
SHEET	3	B1	7 ASN E	50	SER E	54	-1
SHEET	4	B1	7 ASN E	98	ASN E	109	-1
SHEET	5	B1	7 SER E	74	VAL E	97	-1
SHEET	6	B1	7 VAL E	66	LEU E	73	-1
SHEET	7	B1	7 PRO E	28	ARG E	36	-1
SHEET	1	B2	7 GLY E	133	ARG E	146	0
SHEET	2	B2	7 ASN E	147	VAL E	163	-1
SHEET	3	B2	7 VAL E	181	LEU E	184	-1
SHEET	4	B2	7 ASP E	226	ALA E	229	-1
SHEET	5	B2	7 GLY E	207	PHE E	215	-1
SHEET	6	B2	7 SER E	197	ASN E	202	-1
SHEET	7	B2	7 GLY E	133	ARG E	146	-1

Figura 14. Exemplo da indicação da presença da α -hélice (destacado em azul) e da folha- β (destacado em vermelho) em um arquivo PDB. A imagem é um extrato do arquivo 1ppf.pdb indicando a presença da α -hélice entre os resíduos 55-60, 63-64 e 230-243 (destacado em cinza claro) e da folha- β entre os resíduos 28-36, 38-49, 50-54, 66-73, 74-98, 98-109, 133-146, 147-163, 181-184, 197-202, 207-215, 226-229 (destacado em cinza). As colunas 39-40 indicam o tipo de estrutura helicoidal (1: α -hélices, 3: π -hélices e 5: hélices 3_{10}) (destacado em amarelo) ou o sentido das folhas- β (1: paralela e -1: anti-paralela) (destacado em laranja) (WWPDB).

As α -hélices, folhas- β , *turns* e *random coils* podem ser definidos por vários algoritmos, dentre os quais se destacam os mais usados: *Define Secondary Structure of Proteins* (DSSP) e *Structural Identification* (Stride). Segundo Cuff e Barton (CUFF e BARTON, 1999) a definição do EES dos algoritmos DSSP e Stride concordam em 95% dos casos.

2.6.1 DSSP

O DSSP é um algoritmo que funciona pelo processo de reconhecimento dos padrões das ligações de hidrogênio e características geométricas extraídas das coordenadas espaciais dos átomos que compõem cada aminoácido. Os EES são reconhecidos através dos

padrões de repetições das ligações de hidrogênio chamados *turns* e ligações. A repetição dos *turns* é reconhecida como hélice, enquanto a repetição das ligações é reconhecida como fitas. E a repetição das fitas é reconhecida como uma folha. A estrutura geométrica é definida em termos das torções e curvaturas da geometria diferencial.

As ligações de hidrogênio são descritas através do modelo eletrostático usada pelo algoritmo DSSP. A interação eletrostática entre dois grupos é calculada pela Eq. 1:

$$E = q_1 q_2 \left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right) \times f \quad (1)$$

onde: q_1 = é a carga parcial do oxigênio carbonila, cujo valor é $0.42e$; q_2 = é a carga parcial do hidrogênio amida, cujo valor é $0.20e$; a notação r_{AB} representa a distância interatômica entre A e B, dada em angstroms (Å), sendo O = oxigênio, N = nitrogênio, H = hidrogênio, C = carbono; $f = 332$ (constante dimensional); e E é dado em kcal/mol. Se $E < -0,5 \text{ kcal/mol}$ considera-se a existência da ligação de hidrogênio.

O algoritmo DSSP identifica oito possíveis EES, e os classifica de acordo com o seguinte código: (G) hélice 3_{10} , (H) α -hélice, (I) π -hélice, (B) ponte- β , (E) folhas- β paralelas e antiparalelas, (T) *turn*, (S) *bend*, (C) *coil* (KABSCH e SANDER, 1983).

2.6.2 Stride

O algoritmo Stride usa, além do reconhecimento dos padrões de ligações de hidrogênio, a informação sobre os ângulos diedrais, calculada estatisticamente pela Eq. 2. O algoritmo define que uma α -hélice deve ter, no mínimo, duas ligações de hidrogênio consecutivas entre os resíduos de aminoácidos k e $k+4$ tal que:

$$E_{hb}^{k,k+4} = (1 + W_1^\alpha + W_2^\alpha \times \frac{P_k^\alpha + P_{k+4}^\alpha}{2} < T_1^\alpha) \quad (2)$$

onde: P_k^α e P_{k+4}^α são as probabilidades dos respectivos ângulos diedrais para os resíduos de aminoácidos k ; e W_1^α , W_2^α , T_1^α são pesos e limiares empíricos.

A folha- β antiparalela é definida quando as duas ligações de hidrogênio envolvidas satisfazem as condições expressas na Eq. 3:

$$\begin{aligned} E_{hb1} \left(1 + W_1^\beta + W_2^\beta \times CONF_{Antiparalela} \right) &< T_{Antiparalela}^\beta \\ E_{hb2} \left(1 + W_1^\beta + W_2^\beta \times CONF_{Antiparalela} \right) &< T_{Antiparalela}^\beta \end{aligned} \quad (3)$$

E, para folha- β paralela, as condições expressas na Eq. 4:

$$\begin{aligned} E_{hb1} \left(1 + W_1^\beta + W_2^\beta \times CONF_{Paralela} \right) &< T_{Paralela}^\beta \\ E_{hb2} \left(1 + W_1^\beta + W_2^\beta \times CONF_{Paralela} \right) &< T_{Paralela}^\beta \end{aligned} \quad (4)$$

onde: E_{hb1} e E_{hb2} são as energias da primeira e da segunda ligação de hidrogênio; $CONF = \frac{P_{Int1}^\beta + P_{Int2}^\beta}{2}$ se os resíduos de aminoácidos internos estão presentes em ambos os lados da ponte- β , ou $CONF = P_{Int}^\beta$ se apenas um resíduo é interno na ponte- β ; e W_1^β e W_2^β são pesos e limiares empíricos.

O algoritmo Stride define sete EES, e os classifica usando os códigos: (G) hélice 3_{10} , (H) α -hélice, (I) π -hélice, (B) ponte- β , (E) folhas- β paralelas e antiparalelas, (T) *turn*, (C) *coil* (FRISHMAN e ARGOS, 1995).

A Fig. 14 mostra uma comparação entre os dois algoritmos para a estrutura 1ucs.pdb (“Type Iii Antifreeze Protein Rd1 From An Antarctic Eel Pout”).



2.7 Métodos de determinação das estruturas proteicas

Para compreendermos os mecanismos que regem as funções das proteínas precisamos determinar sua estrutura tridimensional. Esse procedimento faz parte da Biologia Estrutural, que emprega, por exemplo, técnicas de difração de raios X, espectroscopia de ressonância magnética nuclear e microscopia eletrônica para determinar a estrutura das proteínas.

2.7.1 Difração de raios X

A cristalografia de proteínas consiste em isolar, purificar e cristalizar uma proteína, e então medir as direções e intensidades das difrações dos raios X aplicado no cristal obtido. Com isso obtém-se um mapa de densidades eletrônicas dos átomos desse cristal. O modelo da estrutura proteica é feito usando as informações desse mapa de densidades (Fig. 15). É o método mais empregado para decifrar estruturas proteicas tridimensionais. Em 16 de maio de 2018, das 140390 estruturas depositadas no PDB, 125750 foram determinadas através de difração de raios X, ou seja, 89,6% do total.

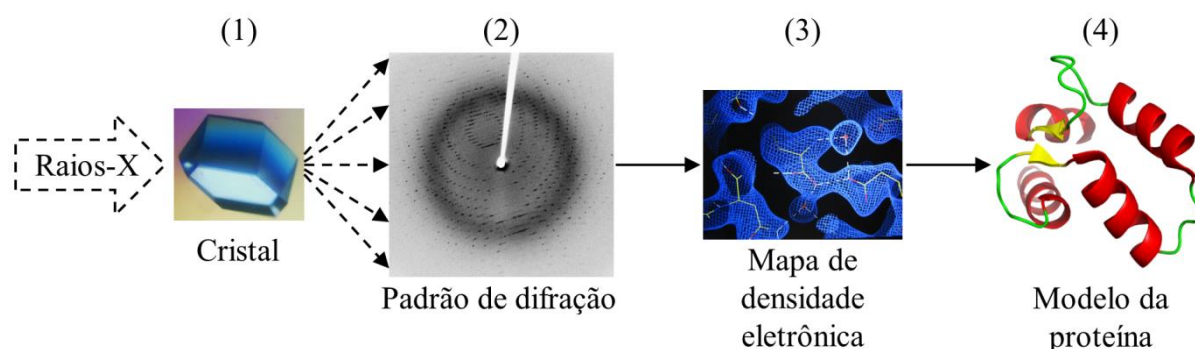


Figura 16. Fases da difração de raios X de proteínas. Após o isolamento, purificação e cristalização, (1) incide-se raios X sobre o cristal, de modo a (2) obter-se um padrão de difração. Com esse padrão é (3) criado um mapa de densidade eletrônica, que é preenchido pelos possíveis aminoácidos presentes, finalmente (4) gerando um modelo da proteína.

2.7.2 Ressonância Magnética Nuclear (RMN)

Também conhecido como NMR (do inglês *Nuclear Magnetic Resonance Spectroscopy*) a espectroscopia de ressonância magnética nuclear é outro método empregado na determinação das estruturas proteicas. É o segundo método mais empregado. Em 16 de maio de 2018, das 140390 estruturas depositadas no PDB, 12144 foram determinadas através de NMR, ou seja, 8,6%.

Diferente da cristalografia de proteínas por raios X, na NMR a proteína não está “imóvel”, ou cristalizada. A amostra é composta por 300-600 microlitros com uma

concentração de 0,1-3 milimolar de proteína, dissolvida em uma solução tampão. Essa solução é colocada no espectrômetro, onde é submetida a ressonâncias magnéticas nucleares. Um algoritmo processa os dados coletados, produzindo vários modelos de igual probabilidade de representar aquela estrutura (Fig. 16 e 17), tudo embasado nas informações que descrevem as distâncias entre os átomos de hidrogênio.

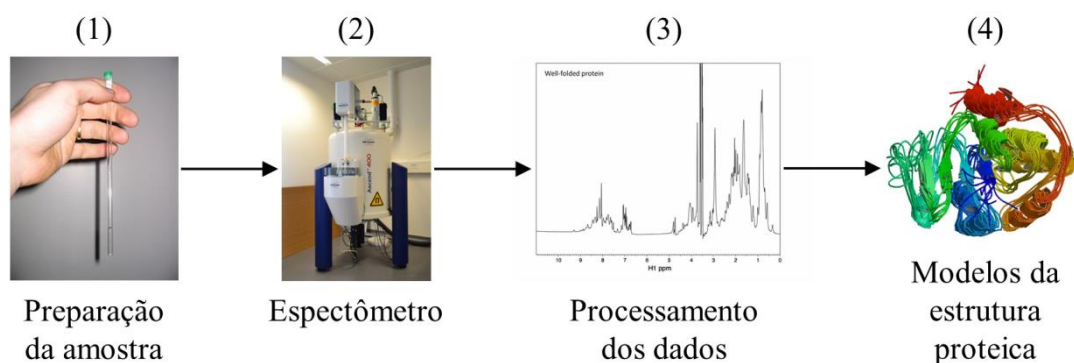


Figura 17. Diagrama do processo de ressonância magnética nuclear. (1) Preparação da amostra; (2) A amostra é submetida as ressonâncias magnéticas nucleares no espectrômetro; (3) coleta de dados do espectrógrafo e medição das distâncias internucleares; (4) determinação da estrutura proteica.

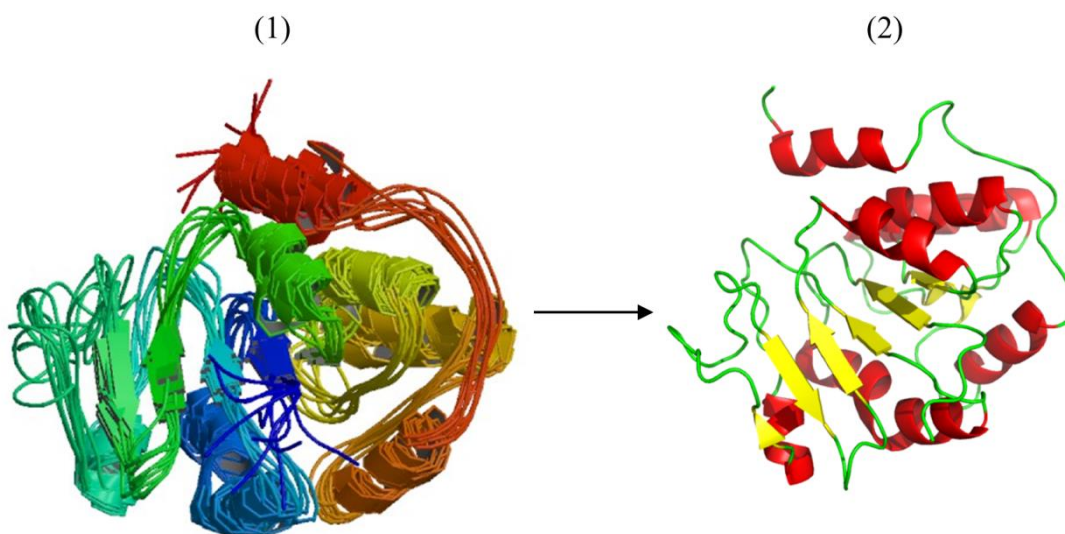


Figura 18. Exemplo de estrutura determinada por NMR. 2jzc.pdb ("Nmr Solution Structure Of Alg13: The Sugar Donor Subunit Of A Yeast N-Acetylglucosamine Transferase. Northeast Structural Genomics Consortium Target Yg1"). Na Fig. 1 veem-se as possíveis estruturas. A imagem foi obtida no site do PDB³. Na Fig. 2 vê-se um dos possíveis modelos dessa estrutura. A imagem foi gerada usando o software PyMOL.

³ www.pdb.org

2.7.3 Microscopia eletrônica

A microscopia eletrônica é outro método usado na determinação das estruturas de proteínas, embora seja bem menos empregado que a difração de raios X e a NMR. Em 16 de maio de 2018, das 140390 estruturas depositadas no PDB, apenas 2139 foram resolvidas por microscopia eletrônica (1,5%). Neste método, um feixe de elétrons é usado para criar uma imagem da estrutura. Se a proteína puder formar um pequeno cristal, a difração de elétrons pode ser usada para gerar um mapa de densidades tridimensional, usando métodos similares aos da difração de raios X. A Fig. 19 apresenta uma estrutura obtida por microscopia eletrônica.

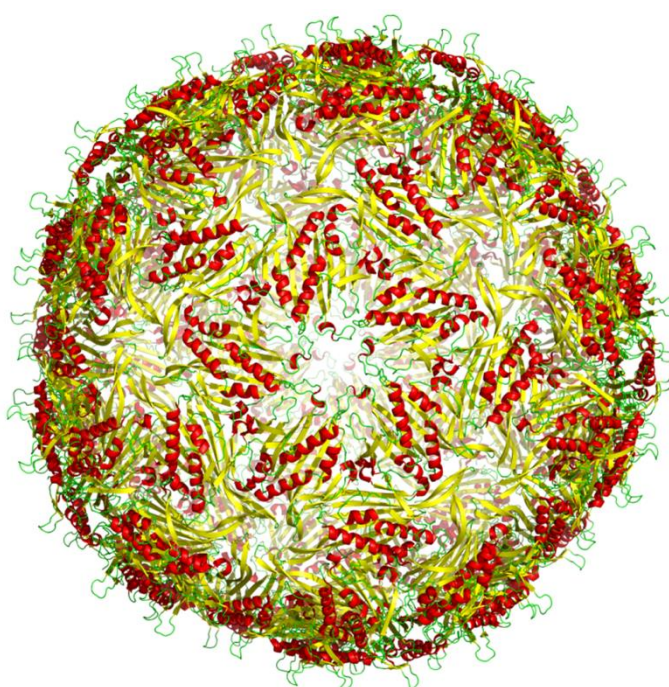


Figura 19. Exemplo de estrutura obtida por microscopia eletrônica. A imagem foi gerada pelo software PyMOL usando a estrutura 4bp4.pdb (“Asymmetric Structure Of A Virus-Receptor Complex”)

Garbuzynskiy (GARBUZYNSKIY, MELNIK, *et al.*, 2005) comparou os métodos mais usados na determinação das estruturas proteicas e concluiu o seguinte:

1. A mesma proteína, determinada por difração de raios X e por NMR, apresenta mais diferenças do que várias estruturas determinadas por difração de raios X para essa proteína, e mais diferenças do que várias estruturas determinadas por NMR para a mesma proteína;
2. Nas estruturas determinadas por NMR, os resíduos de aminoácidos fazem mais contatos entre si que nas mesmas estruturas determinadas por difração de raios X nas

distâncias abaixo de 3,0 Å e entre 4,5-6,5 Å e menos contatos nas distâncias entre 3,0-4,5 Å e entre 6,5-8,0 Å;

3. A diferença no número dos contatos é maior para os resíduos de aminoácidos internos que para os resíduos de aminoácidos presentes na superfície da proteína;
4. A diferença é maior para as proteínas do tipo all- β , α/β e $\alpha+\beta$ que para as proteínas do tipo all- α ;
5. As ligações de hidrogênio na cadeia principal nas estruturas determinadas por NMR presentes no PDB correspondem às ligações de hidrogênio nas estruturas determinadas por difração de raios X em apenas 69% dos casos.

Embora a difração de raios X seja o método mais usado na determinação das estruturas proteicas, ele tem suas limitações. O método funciona melhor para as proteínas rígidas, porque elas formam cristais mais ordenados. Mas a cristalização é um método empírico, e nem sempre se consegue um bom cristal de uma proteína. Por outro lado, a NMR é indicada para determinar a estrutura das proteínas flexíveis. Mas o método tem suas desvantagens. Por exemplo, a proteína precisa estar estável a 30°C por 24 horas, e muitas proteínas se degradam nessas condições. Elas também precisam ter entre 30-40 kDa, e a maioria das proteínas são maiores que isso.

Portanto, cada método tem suas vantagens e desvantagens, sendo que o melhor método é aquele que descreve da melhor maneira a estrutura da proteína estudada.

2.8 Métodos de predição do EES

A predição da estrutura secundária usa o conhecimento da estrutura primária (sequência de aminoácidos) de uma proteína. O processo de predição consiste em atribuir os EES (α -hélice, folha- β , *turn* e *random coil*) a regiões da sequência de aminoácidos. A qualidade da predição é calculada comparando-se seu resultado com os resultados do algoritmo DSSP aplicado a uma proteína cristalizada. Alguns métodos usados são o de Chou-Fasman, método de GOR, uso de redes neurais e máquina de vetor de suporte (SVM). Atualmente, os melhores métodos de predição da estrutura secundária conseguem em torno de 92,1% de acurácia (NANNI, BRAHNAM e LUMINI, 2014).

2.8.1 Método de Chou-Fasman

O método de Chou-Fasman usado para prever os EES baseia-se na propensão, que é a medida da probabilidade de cada aminoácido estar presente em cada tipo de EES, conforme vemos na Tabela 3 (CHOU e FASMAN, 1974).

Propensão		
α -hélice	Folha- β	Turn
ALA	VAL	GLY
LEU	LEU	SER
VAL	ALA	LYS
LYS	GLY	ASN
GLU	THR	THR
SER	ILE	ALA
THR	SER	LEU
GLY	LYS	PRO
GLN	TYR	ASP
ASP	GLN	TYR
ILE	PHE	VAL
ASN	ASN	GLU
HIS	ASP	ARG
PHE	ARG	ILE
ARG	CYS	GLN
TYR	HIS	HIS
PRO	PRO	PHE
TRP	TRP	CYS
CYS	MET	TRP
MET	GLU	MET

Tabela 3. Tabela de Propensão dos resíduos de aminoácidos (CHOU e FASMAN, 1974)

O algoritmo utilizado neste método procura linearmente através da sequência de aminoácidos as regiões de nucleação das α -hélices e folhas- β , e depois estende essa região por uma janela de quatro resíduos de aminoácidos com probabilidade menor que 1. Essa janela considera que de quatro a seis aminoácidos contíguos são suficientes para nuclear uma α -hélice, e de três a cinco aminoácidos são suficientes para nuclear uma folha- β . O limiar da probabilidade é de 1,03 para α -hélice e 1,00 para a folha- β . Como muitos resíduos de aminoácidos que aparecem nas regiões de α -hélice e folha- β também aparecem nas regiões de *turn*, o *turn* é predito apenas se a sua probabilidade for maior que a probabilidade da existência de uma α -hélice ou folha- β , e a probabilidade deste *turn* for maior que um determinado limiar. A probabilidade do *turn* é determinada pela Eq. 5.

$$p(t) = p_t(j) \times p_t(j + 1) \times p_t(j + 2) \times p_t(j + 3) \quad (5)$$

onde j é a posição do resíduo de aminoácido em uma janela de quatro resíduos. Se $p(t)$ exceder o valor arbitrário de $7,5 \times 10^{-3}$, a média de $p_t(j)$ for superior a 1, e $p(t)$ exceder a probabilidade da existência da α -hélice ou folha- β , então ali está predita a existência de um *turn*. O método de Chou-Fasman tem uma acurácia de quase 70% (CHEN, GU e HUANG, 2006).

2.8.2 Método GOR

O método **G**arnier-**O**sguthorpe-**R**obson (GOR) é baseado na teoria da informação para a predição das estruturas secundárias. Semelhante ao método de Chou-Fasman, ele leva em conta a tendência de aminoácidos individuais para formar uma determinada estrutura secundária em particular, e também a probabilidade condicional dos aminoácidos para formarem uma estrutura secundária, dado que seus vizinhos imediatos já formaram essa estrutura (GARNIER, GIBRAT e ROBSON, 1996). O método GOR é, portanto, uma análise Bayseana (GARNIER, OSGUTHORPE e ROBSON, 1978).

O método GOR analisa a sequência de aminoácidos para prever a presença do EES considerando uma janela de 17 resíduos de aminoácidos. Uma matriz 17×20 é usada para pontuar a probabilidade da presença de cada aminoácido em cada uma das 17 posições da sequência. Esse método originalmente apresentou uma acurácia de 73,5% (XIA, DOU, *et al.*, 2011).

2.8.3 Redes neurais

As redes neurais usam um conjunto de dados das estruturas proteicas conhecidas para treinar a rede neural (por esse motivo o método é chamado de aprendizagem de máquina, porque a rede “aprende” a identificar um padrão nos dados de teste). A Fig. 20 mostra a topologia de uma rede neural. Esse método tem uma acurácia em torno de 78% (DONGARDIVE e ABRAHAM, 2016).

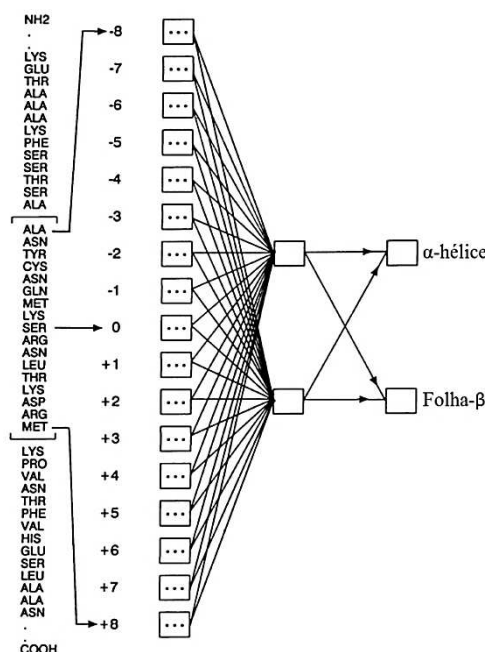


Figura 20. Topologia de uma rede neural. Cada um dos 17 blocos mostrados na camada de entrada representa uma rede de entradas utilizada para codificar o aminoácido na posição correspondente na janela. Cada grupo consiste em 21 entradas, uma para cada possível aminoácido em cada posição, mais uma entrada nula usada quando o movimento da janela se sobrepõe ao final da sequência de aminoácidos. Assim, para uma janela na sequência de aminoácidos, 17 das 357 redes de entrada são definidos como 1 e o restante como 0. A predição é feita para o resíduo de aminoácido central da janela de entrada. Extraído e adaptado de (HOLLEY e KARPLUS, 1989).

2.8.4 Máquina de vetor de suporte

A máquina de vetor de suporte (do inglês “Support Vector Machine”, SVM) é baseada na teoria do aprendizado estatístico. A SVM apresenta bom desempenho e é mais fácil de programar e treinar que as redes neurais. A ideia é usá-la na classificação de um padrão binário (o resíduo de aminoácido está presente, ou não, em um EES). Para isso, faz-se um mapeamento dos vetores de entrada. Esse mapeamento pode ser de forma linear ou não linear. Então se usa uma função *kernel*, que será responsável por dividir esse mapa de vetores em duas classes. Há duas maneiras de se representar uma sequência de aminoácidos como um vetor de entrada:

- Sequência única: cada resíduo de aminoácido de uma proteína é representado por um vetor com 20 posições, com os valores 0 e 1. Cada resíduo de aminoácido da sequência é representado por esse vetor das pontuações de substituição.
- Múltiplas sequências alinhadas: a sequência alvo é primeiramente alinhada com as sequências extraídas de uma base de dados não redundante, a fim de determinar a família proteica a qual a sequência alvo pertence. O alinhamento pode ser expresso por uma matriz de probabilidades estimadas.

Cada sequência proteica é representada por um vetor bidimensional $L \times 20$, onde L é o tamanho da sequência. Esse vetor é usado como entrada da SVM. A SVM tem uma acurácia de 80% (DOR e ZHOU, 2006).

2.8.5 Comparação entre os métodos de predição do EES

Tsilo (TSILO, 2009) fez uma comparação entre as previsões feitas por redes neurais e SVM, concluindo que a acurácia dos métodos está mais ligada aos diferentes modelos de classificadores, que aos métodos em si. Portanto, algumas das variáveis que podem influenciar nos resultados incluem os diferentes sistemas de entrada de dados, tamanho da janela considerada, os métodos de validação cruzada, os parâmetros usados para criar uma tabela de aprendizagem, e atribuições das classes estruturais.

2.9 Softwares usados para previsão do EES

2.9.1 RaptorX

O RaptorX⁴ usa o método de aprendizado estatístico para projetar uma nova função de pontuação, usada para medir a melhor compatibilidade da sequência alvo com uma estrutura molde. O RaptorX usa o algoritmo NEFF para medir a quantidade de informação contida no perfil da sequência de uma proteína. O NEFF pode ser interpretado como o número efetivo de estruturas homólogas não redundantes para uma determinada proteína. Seu valor varia de 1 a 20, o que representa o número de substituições de aminoácidos em cada posição da sequência. Um perfil de sequência esparsa (com um baixo valor de NEFF) geralmente leva a previsões da estrutura secundária menos precisa (PENG e XU, 2011).

2.9.2 NetSurfP

O NetSurfP⁵ usa duas redes neurais para a predição da estrutura secundária. A primeira rede neural é treinada no perfil da sequência e estrutura secundária predita, e tem duas saídas. A saída mais alta define o tipo de EES predito. A segunda rede neural utiliza essas saídas como entrada, junto com o perfil da sequência, e é treinado para prever a superfície relativa exposta de cada resíduo de aminoácido. A taxa de acerto desse método é de 79% (PETERSEN, PETERSEN, *et al.*, 2009).

⁴ <http://raptorx.uchicago.edu/>

⁵ <http://www.cbs.dtu.dk/services/NetSurfP/>

2.9.3 Jpred

Jpred⁶ usa o algoritmo Jnet (CUFF e BARTON, 2000) para fazer a previsão da estrutura secundária através de redes neurais. A acurácia do algoritmo Jnet na predição de α -hélice, folha- β e *random coil* é de 81,5% (COLE, BARBER e BARTON, 2008). Disponível na web, o Jpred faz mais de 1000 previsões por semana para usuários de mais de 50 países.

2.9.4 PredictProtein

O PredictProtein⁷ é um servidor web para análise de sequências e predição de estruturas e função das proteínas. Para a predição da estrutura secundária ele utiliza os algoritmos PHD (ROST, 1996) e PROF (ROST, 2001). O algoritmo PROF tem uma acurácia de 76%, enquanto o algoritmo PHD apresenta uma acurácia de 71% (ROST, YACHDAV e LIU, 2004).

2.9.5 YASSPP

O YASSPP⁸ faz a predição da estrutura secundária através de SVM, usando um modelo baseado em cascata. O primeiro nível do modelo, chamado de modelo sequência-para-estrutura, faz a previsão para cada posição da sequência, considerando a sequência de aminoácidos em torno daquela posição. O segundo nível do modelo, chamado de modelo estrutura-para-estrutura, faz a previsão final dos EES, considerando as previsões do primeiro modelo. Cada modelo é construído usando três conjuntos de classificadores binários que utilizam a abordagem de aprendizagem um-contr-o-restante. O método empregado pelo YASSPP apresentou uma acurácia de 77,83% (KARYPIS, 2006).

2.9.6 SymPred

O SymPred⁹ usa um método baseado em dicionário para predição dos EES. Abordagens baseadas em dicionário são amplamente utilizadas na área de Processamento de Linguagem Natural (PLN). Esse método gera palavras sinônimas a partir da sequência de aminoácidos e sequências similares (Fig. 21). O SymPred apresenta uma acurácia de 81% (LIN, SUNG, *et al.*, 2010).

⁶ <http://www.compbio.dundee.ac.uk/jpred/>

⁷ <https://predictprotein.org/>

⁸ <http://glaros.dtc.umn.edu/yasspp/>

⁹ <http://www.ibi.vu.nl/programs/sympredwww/>

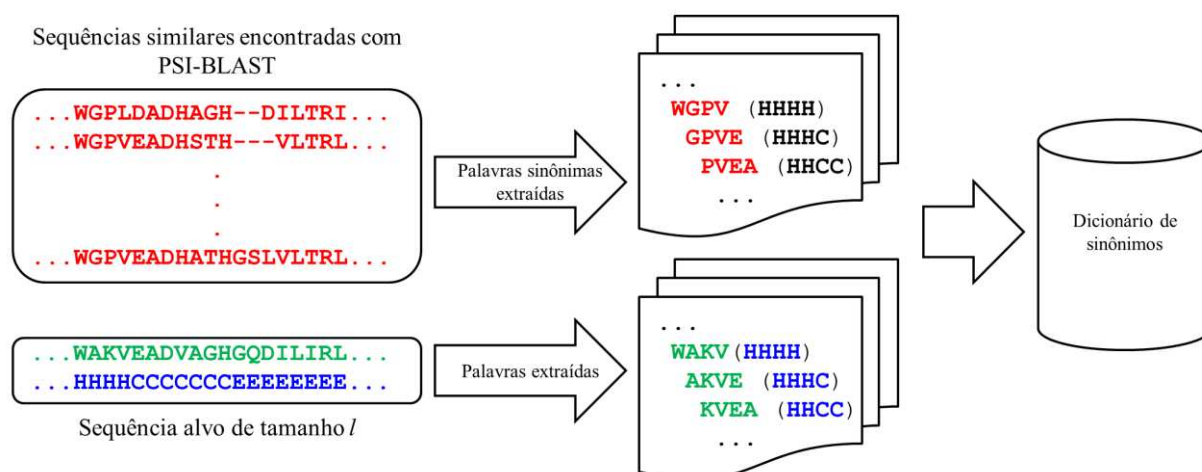


Figura 21. Procedimento usado na extração das palavras e palavras sinônimas para uma proteína alvo. Usa-se uma janela deslizante de tamanho 4 para varrer a sequência alvo e as sequências similares encontradas com o PSI-BLAST e extrair todas as palavras. Cada palavra está associada a uma parte da informação estrutural da região da qual foi extraída. A fonte de todas as palavras extraídas é a proteína alvo, uma vez que toda informação estrutural é obtida a partir dela. Fonte: (LIN, SUNG, *et al.*, 2010)

2.9.7 SSpro

O SSpro¹⁰ usa um conjunto de redes neurais bidirecionais recorrentes para prever a estrutura secundária. Nesta arquitetura de rede neural, a classificação é determinada por três componentes. Em primeiro lugar, existe um componente central associado aos resíduos de aminoácidos, onde ocorre a predição da estrutura secundária para a posição t da sequência. O segundo e o terceiro componente são duas redes neurais recorrentes, que “deslizam” ao longo da sequência de aminoácidos nos sentidos N-Terminal e C-Terminal, até o ponto de predição (Fig. 22). Com essa abordagem, o método apresenta uma acurácia de 78% (POLLASTRI, PRZYBYLSKI, *et al.*, 2002).

¹⁰ <http://download.igb.uci.edu/sspro4.html>

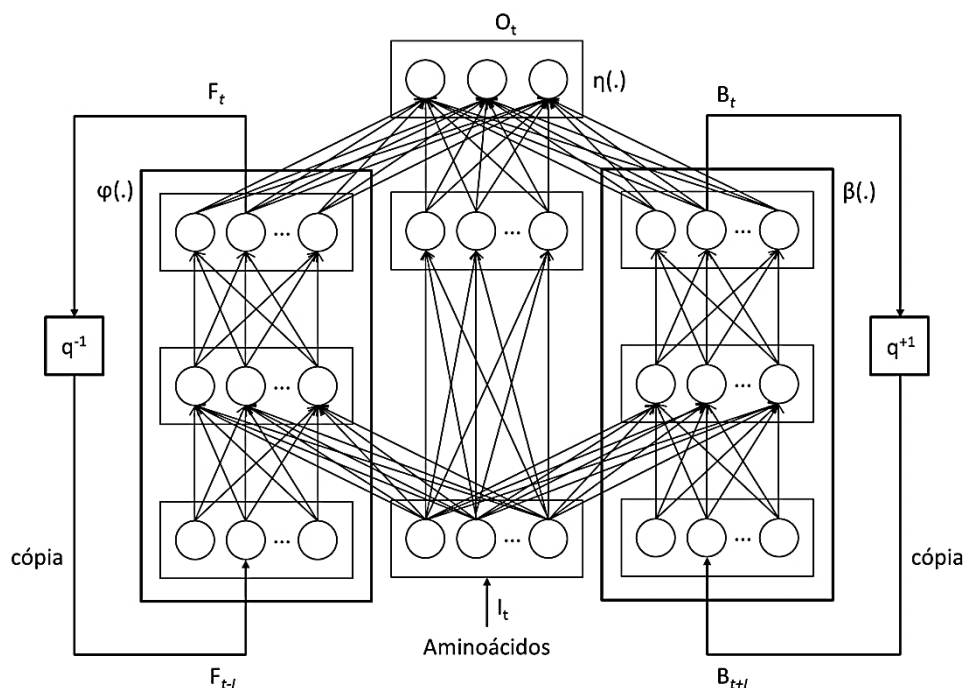


Figura 22. Arquitetura da rede neural bidirecional recorrente. A camada de saída (O_t) tem três unidades exponenciais normalizadas associadas com os membros de cada uma das três classes de EES (α -hélice, folha- β e *turn*) para o resíduo de aminoácido na posição t . As funções β , η e φ são implementadas pela rede neural. Fonte: (POLLASTRI, PRZYBYLSKI, *et al.*, 2002)

2.9.8 DSC

Discrimination of Secondary structure Class (DSC)¹¹ é baseado na decomposição da predição da estrutura secundária em conceitos básicos e, em seguida, utiliza métodos estatísticos para combinar os conceitos de predição. Para cada posição da sequência de aminoácidos é calculada a média do potencial GOR para cada classe de estrutura secundária (α -hélice, folha- β e *turn*), a distância até o final da cadeia, o momento da média da hidrofobicidade para α -hélice e folha- β , a existência de inserções e deleções, e o momento da média da conservação para α -hélice e folha- β . Esses resultados são então suavizados e uma função de discriminação linear é aplicada, para fazer a predição para cada aminoácido da sequência. A quantidade de resíduos de aminoácidos preditos como α -hélice e folha- β é usada para refinar a previsão, usando uma segunda função de discriminação linear. A acurácia desse método é de 70,1% (KING e STERNBERG, 1996).

¹¹ http://www.csb.yale.edu/userguides/seq/dsc/dsc_descrip.html

2.9.9 PROFphd

O software PROFphd¹² usa redes neurais para prever a presença dos EES. Ele combina alguns fatores como parâmetros de entrada, por exemplo, a informação extraída do alinhamento das sequências proteicas, que é quanto um resíduo de aminoácido está presente em cada posição. Ele usa também o número de inserções e deleções para melhorar o desempenho do classificador, e, finalmente, a adição do conteúdo global de aminoácidos também produz uma melhoria, principalmente na predição da classe estrutural. Com essa abordagem, o método consegue uma acurácia de 88% (ROST e SANDER, 1994).

2.9.10 PSIPRED

O PSIPRED¹³ usa uma rede neural de dois estágios para fazer a predição dos EES. A sua acurácia é de 81,4% (BUCHAN, WARD, *et al.*, 2010).

2.9.11 Predator

O método de predição utilizado pelo software Predator¹⁴ consiste em alinhar par-a-par a sequência alvo com todas as sequências relacionadas. Após essa primeira etapa, apenas os fragmentos alinhados com significância são considerados. A propensidade dos EES das sequências auxiliares relacionadas é combinada com aquela da sequência alvo e ponderada de acordo com o grau de similaridade. A acurácia desse método é de 75% (FRISHMAN e ARGOS, 1997).

2.9.12 Comparação entre os softwares usados para predição do EES

Comparando os softwares descritos acima, os melhores resultados foram obtidos pelo algoritmo PROFphd, com 88% de acerto. Mas, conforme explicado por Tsilo (TSILO, 2009) o resultado está mais ligado aos diferentes modelos de classificadores, que aos métodos em si.

2.10 Métodos de predição das estruturas proteicas

Anfinsen (ANFINSEN, 1973) confirmou experimentalmente que a sequência de aminoácidos de uma proteína contém toda a informação necessária que ela necessita para assumir sua estrutura terciária. Ele formulou a “hipótese termodinâmica”, que diz que a estrutura tridimensional de uma proteína em seu meio fisiológico normal (solvente, pH, força iônica, presença de outros componentes como íons de metal, temperatura, etc.) é aquela em

¹² <https://www.predictprotein.org/>

¹³ <http://bioinf.cs.ucl.ac.uk/psipred/>

¹⁴ <http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::predator>

que a energia livre de Gibbs no sistema é a menor. Em outras palavras, a conformação da estrutura proteica é determinada pela totalidade das interações interatômicas e, portanto, pela sequência de aminoácidos. Embora hoje saibamos que este é um caso especial, ele induziu muitas tentativas de prever a estrutura terciária a partir da sequência. Vários métodos foram usados desde então na tentativa de prever a estrutura terciária a partir da sequência de aminoácidos: modelagem por homologia (MARTI-RENOM, STUART, *et al.*, 2000), modelagem *ab-initio* (LEE, WU e ZHANG, 2009), *threading* (JONES, TAYLOR e THORNTON, 1992) etc.

2.10.1 Modelagem por homologia

As proteínas agrupam-se em um limitado número de famílias proteicas. Conhecendo-se pelo menos um representante de uma família proteica, geralmente é possível modelar os demais membros dessa família.

A modelagem por homologia baseia-se em alguns padrões observados em nível molecular:

- A homologia entre sequências de aminoácidos implica em semelhança estrutural e funcional
- Proteínas homólogas apresentam regiões internas conservadas
- As principais diferenças entre proteínas homólogas aparecem nas regiões externas, constituídas principalmente por *turns*, que ligam os EES α -hélice e fita- β .

A Fig. 23 apresenta o fluxograma desse método, que consiste em identificar as estruturas relacionadas com a sequência alvo e selecionar aquelas que serão usadas como modelo. A sequência alvo e os modelos são alinhados, e então constrói-se o modelo para a proteína alvo.

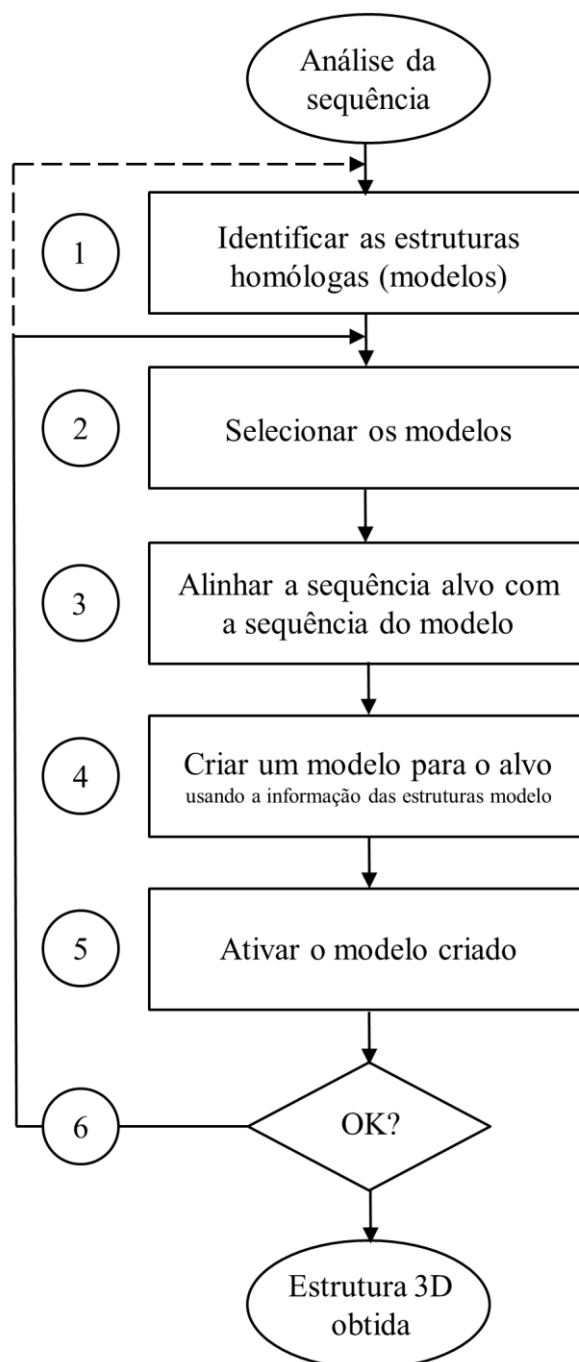


Figura 23. Esquemático da modelagem por homologia. O passo inicial é (1) identificar todas as estruturas proteicas relacionadas com a sequência alvo, e então (2) selecionar aquelas que serão usadas como modelos; (3) uma vez que os modelos foram selecionados, é feito o alinhamento posicional desses modelos com a sequência alvo; (4) após o alinhamento posicional entre a sequência alvo e os modelos selecionados, constrói-se o modelo tridimensional para a proteína alvo. Existe uma variedade de métodos que podem ser usados na construção do modelo tridimensional: modelagem por construção de corpo rígido, modelagem por segmentos correspondentes, modelagem por satisfação de restrições espaciais obtidas a partir do alinhamento posicional. Destes métodos, o mais usado é a construção de corpos rígidos; (5) avalia-se o modelo criado observando se ele possui a dobra correta. Ele terá a dobra correta se a estrutura conhecida corretamente selecionada foi usada como modelo e se a sequência dessa estrutura foi devidamente alinhada com a sequência alvo; (6) se o modelo criado foi bem avaliado, o processo de modelagem por homologia está acabado. Caso contrário, volta-se ao passo 2, ou 3, para a criação de um novo modelo tridimensional e consequente reavaliação. Fonte: (MARTI-RENOM, STUART, *et al.*, 2000)

2.10.2 Modelagem *ab initio*

A modelagem *ab initio* (latim *ab*: do, *initio*: início) considera apenas as propriedades físico-químicas de cada aminoácido para a construção de funções de energia. Estas funções são minimizadas por algoritmos que realizam buscas no espaço de conformações que a proteína de interesse possa assumir. A acurácia da modelagem *ab initio* é baixa, e esse tipo de abordagem se limita a pequenas proteínas (< 100 resíduos de aminoácidos). O sucesso da modelagem *ab initio* depende de três fatores (LEE, WU e ZHANG, 2009):

- Uma função de energia precisa, com a qual a estrutura nativa da proteína corresponda ao estado mais termodinamicamente estável, quando comparado com todas as possíveis estruturas intermediárias;
- Um método de busca eficiente, que identifique rapidamente os estados de baixa energia através da pesquisa conformacional;
- Seleção dos modelos nativos.

A Fig. 24 demonstra o funcionamento do método.

2.10.3 Threading

A modelagem por *threading* é usada quando selecionamos uma proteína alvo com o enovelamento que será o mesmo da estrutura conhecida, mas ela não é homóloga ao alvo. O enovelamento escolhido é aquele que apresenta a menor energia entre vários outros. O método trabalha com o conhecimento estatístico do relacionamento entre as estruturas conhecidas e a sequência da proteína que será modelada. A predição é feita alinhando cada um dos aminoácidos na sequência alvo em uma posição das estruturas molde, e avaliando através da energia total do sistema quão bem o alvo se encaixa no molde. Depois, a energia de cada molde com a sequência alvo é calculada e a menor é escolhida (Fig. 25).

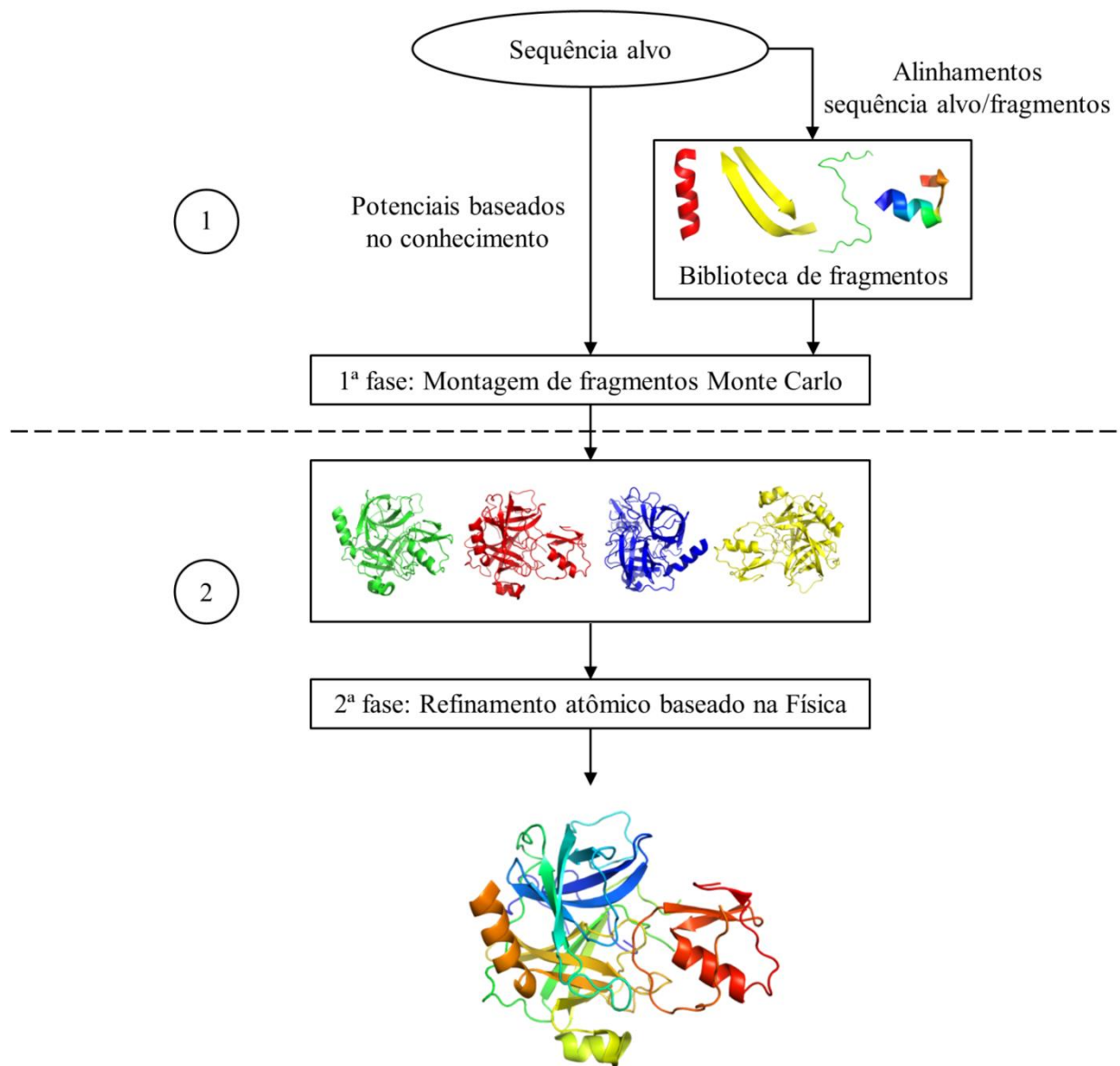


Figura 24. Fluxograma da predição da estrutura de proteína usando a modelagem *ab initio* utilizada pelo software ROSETTA: (1) primeiramente foram gerados modelos em uma forma reduzida com conformações específicas. (2) um conjunto desses modelos de baixa resolução foi refinado usando funções de energia. Imagem extraída de (LEE, WU e ZHANG, 2009) e posteriormente adaptada. As imagens das estruturas foram geradas usando PyMOL.

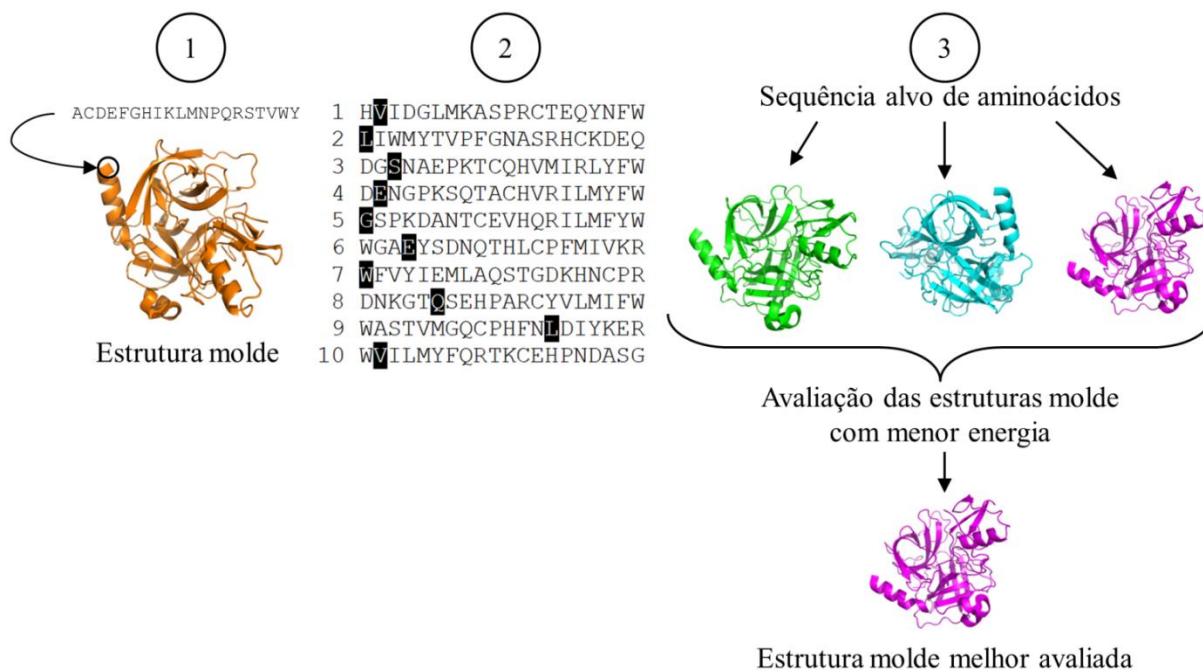


Figura 25. Predição da estrutura pelo método de *threading* (reconhecimento de dobra). (1) Cada um dos 20 aminoácidos é colocado em uma posição da estrutura molde, e então se avalia a sua compatibilidade com o ambiente estrutural. (2) O exemplo da tabela foi feito com a sequência dos primeiros 10 aminoácidos da estrutura 1mbd.pdb (“Oxygen Storage Neutron Diffraction Reveals Oxygen-Histidine Hydrogen Bond In Oxy myoglobin”). Os 20 aminoácidos foram classificados da esquerda para a direita em cada linha, de acordo com seu grau de compatibilidade na estrutura modelo. Os aminoácidos realçados são aqueles energeticamente favoráveis com a estrutura. (3) Visão geral do processo: uma sequência alvo é “encaixada” dentro de diversas estruturas moldes, selecionadas a partir de uma base de dados. A energia de cada molde que recebeu a sequência alvo é calculada e a menor escolhida. As imagens das estruturas acima foram geradas usando o software PyMOL.

Ao compararmos os métodos de predição das estruturas proteicas – modelagem por homologia, modelagem ab initio e threading – concluímos que o que apresenta maior acuraria é a modelagem por homologia, especialmente quando a sequência alvo tem grande similaridade com as sequências homólogas. Entretanto, a modelagem ab initio e threading também têm suas aplicações, conforme explicado na Fig. 26.

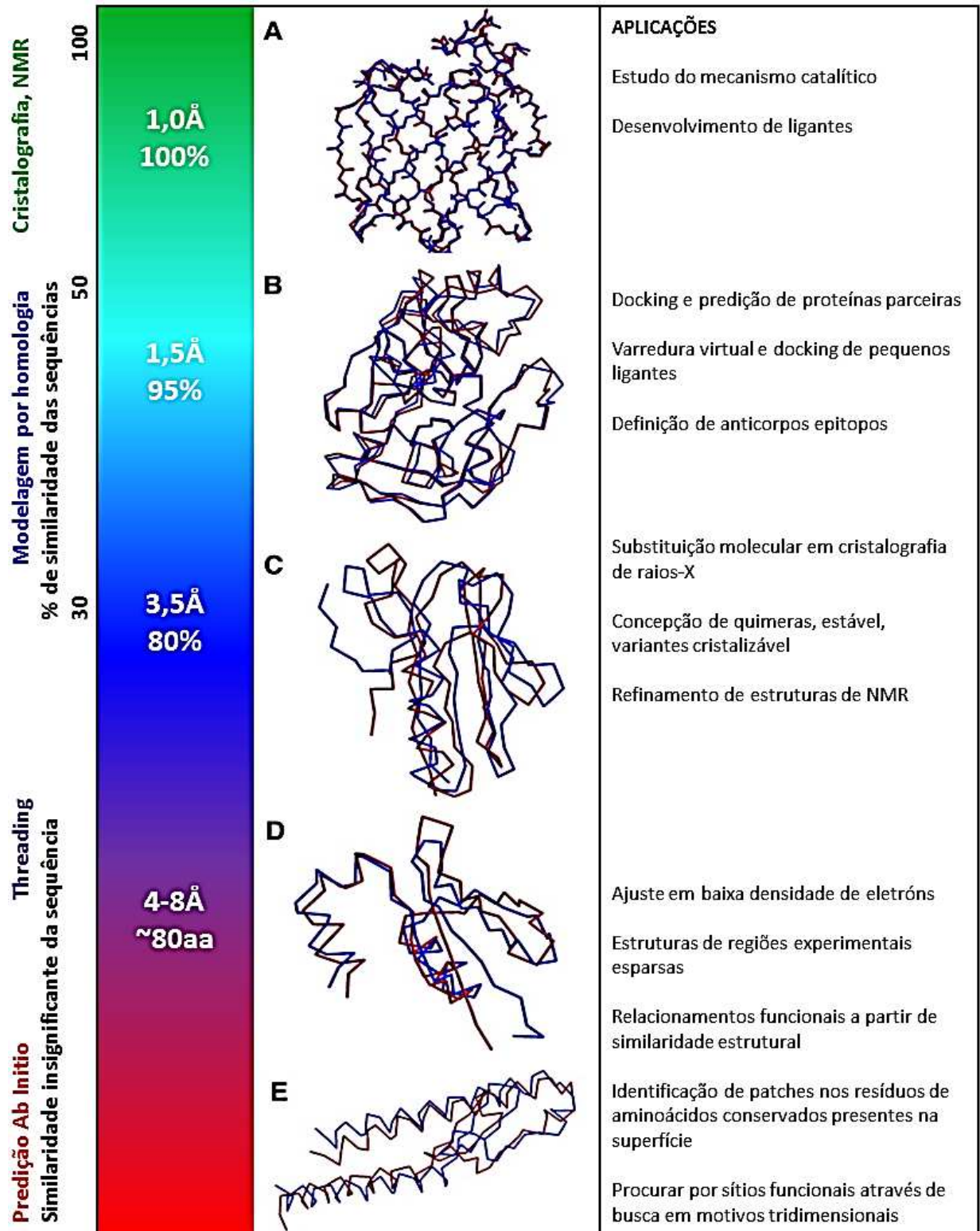


Figura 26. Precisão e aplicação dos modelos de estrutura de proteínas. São mostrados os diferentes intervalos de aplicação de modelagem por homologia, *threading* e modelagem *ab initio*. Modelos comparativos de amostras com base cerca de 60% (A), 40% (B) e 30% (C) de identidade de sequência com a sua estrutura modelo. (D e E) Exemplos de predição *ab initio* usando o software Rosetta para o CASP4. Estruturas previstas estão em vermelho, e as estruturas reais são em azul. A precisão dos modelos diminui significativamente de (A) a (E), mas a estrutura global é ainda aproximadamente correta. Fonte: (BAKER e SALI, 2001)

2.11 Plataformas para análise das estruturas proteicas

Existem vários softwares para visualização e análise das estruturas proteicas. Por exemplo: SwissModel, Visual Molecular Dynamics (VMD), Jmol, Geneious Pro, PyMOL, RasMol, UCSF Chimera e Blue Star Sting.

2.11.1 SwissModel

O SwissModel¹⁵ é um serviço integrado, disponível na Web, que auxilia e orienta o usuário na modelagem por homologia de proteínas com diferentes níveis de complexidade. Ele foi desenvolvido de modo a permitir que vários projetos sejam executados em paralelo. O usuário pode acessar bancos de dados de estruturas de proteínas a partir do SwissModel, usar ferramentas para seleção, construção e avaliação da qualidade do modelo (ARNOLD, BORDOLI, *et al.*, 2006).

2.11.2 Visual Molecular Dynamics (VMD)

O VMD¹⁶ foi desenvolvido para mostrar e analisar estruturas moleculares, em especial proteínas e ácidos nucleicos (HUMPHREY, DALKE e SCHULTEN, 1996). O VMD pode mostrar várias estruturas proteicas simultaneamente, e também permite ao usuário visualizar a dinâmica molecular das proteínas. O software foi desenvolvido na linguagem de programação C++ e é mantido pelo NIH Center for Macromolecular Modelling and Bioinformatics of University of Illinois at Urbana-Champaign. Ele está disponível para download e instalação no computador do usuário para os sistemas operacionais Linux, MacOS, Solaris e Windows.

2.11.3 Jmol

Jmol¹⁷ é um software desenvolvido na linguagem de programação Java, usado para visualização de estruturas moleculares, que permite a interação do usuário via web. O início do seu desenvolvimento foi em 1999, como substituto do programa XMol Rasmol, que era um visualizador de moléculas desenvolvido pelo Minnesota Supercomputer Center (HERRÁE, 2006).

2.11.4 Geneious Pro

O Geneious Pro¹⁸ foi desenvolvido para a organização e análise de dados biológicos, em especial das sequencias moleculares. Ele integra vários algoritmos de análise

¹⁵ <http://swissmodel.expasy.org/>

¹⁶ <http://www.ks.uiuc.edu/Research/vmd/>

¹⁷ <http://jmol.sourceforge.net/>

¹⁸ <http://www.geneious.com/>

de dados biológicos. Por exemplo, o usuário pode executar o BLAST (ALTSCHUL, GISH, *et al.*, 1990) para buscar uma sequência em um repositório online, pode alinhar sequências ou traduzir sequências de DNA em sequências proteicas. Todas essas funcionalidades estão integradas de forma consistente em um único ambiente (KEARSE, MOIR, *et al.*, 2012).

2.11.5 PyMOL

PyMol¹⁹ é um pacote de softwares desenvolvido para a customização de imagens tridimensionais de biomoléculas, com mais de 600 configurações possíveis e 20 tipos de representações. Ele pode interpretar mais de 30 formatos diferentes de arquivos, desde o formato PDB até o formato multi-SDF para mapas de densidade eletrônica. Com o PyMOL também é possível criar animações do movimento molecular (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC).

2.11.6 RasMol

RasMol²⁰ é um programa simples, desenvolvido para visualização de estruturas moleculares. O usuário pode visualizar a estrutura descrita em um arquivo PDB e extrair desse arquivo as seguintes informações: nome e classificação da estrutura, número de cadeias, grupos, átomos e ligações de hidrogênio. É possível medir os ângulos e distâncias entre os átomos. Usando um console, o usuário pode criar seus próprios scripts para, por exemplo, selecionar um conjunto de resíduos de aminoácidos, visualiza-los em diferentes modos e cores. Ele está disponível para download nas versões Windows, Linux e MacOS (SAYLE e MILNER-WHITE, 1995).

2.11.7 UCSF Chimera

O software UCSF Chimera²¹ (ou simplesmente Chimera) é segmentado em um núcleo que fornece serviços básicos de visualização de estruturas moleculares, e possui extensões que fornecem funcionalidades de nível superior. O Chimera oferece as seguintes extensões: Multiscale, que adiciona ao software a capacidade de visualização de grandes moléculas; Collaboratory, que permite aos usuários compartilhar uma sessão do Chimera, mesmo estando em locais separados uns dos outros; Multalign Viewer, que mostram múltiplas sequências alinhadas e suas respectivas estruturas; ViewDock, usado para “screening docked”; Movie, para visualização da trajetória em dinâmica molecular; e Volume

¹⁹ <http://www.pymol.org/>

²⁰ <http://rasmol.org/>

²¹ <http://www.cgl.ucsf.edu/chimera/>

Viewer, para a visualização e análise de dados volumétricos. Chimera está disponível para Windows, Linux, MacOS, SGI Irix e Unix (PETTERSEN, GODDARD, *et al.*, 2004).

2.11.8 Blue Star Sting

O Grupo de Pesquisa em Biologia Computacional (GPBC) da Embrapa Informática Agropecuária desenvolveu um conjunto de dados, ferramentas e técnicas para a análise das estruturas proteicas. Por exemplo, com o intuito de estudar o nano-ambiente proteico foi desenvolvido pesquisas sobre hot spot, interfaces entre proteínas (DE MORAES, 2014), caracterização do sítio catalítico, e este trabalho sobre EES.

O Sequence To and withIN Graphics (STING)²² é um conjunto de programas para análise do relacionamento entre a sequência, estrutura, função e estabilidade das proteínas. Seu acesso é livre para qualquer usuário via web. Mantido pelo GPBC, atualmente o STING encontra-se na versão Blue Star Sting. (NESHICH, KUSER, *et al.*, 2006) Entre as funcionalidades do Blue Star Sting, destaca-se a visualização das estruturas primária e terciária e/ou quaternária simultaneamente, onde relacionamos os resíduos de aminoácidos da sequência com sua posição espacial, buscar padrões na sequência de aminoácidos, visualizar graficamente através de um degrade de cores as características físico-químicas, estruturais e da sequência de uma estrutura.

Um dos módulos do Blue Star Sting é o Multiple Structure Single Parameter 2D Plot (MSSP)²³ (SALIM, 2016). Ele oferece um método para a análise de mutantes ou qualquer conjunto de proteínas estruturalmente similares. O usuário fornece os códigos do PDB que deseja estudar, escolhe quais cadeias serão analisadas e o software faz o alinhamento estrutural. Os dados contidos no STING_RDB são mostrados em um gráfico, onde as diferenças e similaridades entre as estruturas ficam claras.

Para ilustrar o uso do MSSP, vamos comparar duas estruturas proteicas: PDB 1SPD cadeia A e 1N19 cadeia A (estrutura mutada, mas também contendo substituições de seus dois resíduos livres de cisteína: C6A e C111S). Na Fig. 27 observamos que a estrutura mutada tem uma diminuição dramática de EP@Surf em diferentes posições, no entanto, distante do local da alanina mutada (posição número 4). Uma inspeção mais minuciosa dos resíduos de aminoácidos que sofreram uma grande modificação no valor dos seus respectivos

²² <http://www.cbi.cnptia.embrapa.br>

²³ http://www.cbi.cnptia.embrapa.br/SMS/STINGm/MPA/js_mssp.html

potenciais eletrostáticos revela que estão envolvidos e/ou muito próximos dos átomos de ligação ao metal (Fig. 28).

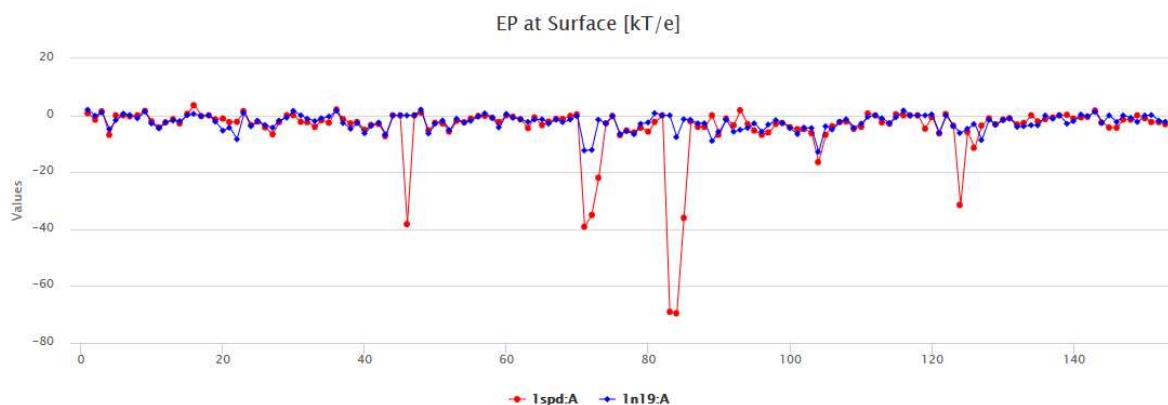


Figura 27. MSSP mostrando potencial eletrostático calculado na superfície do resíduo de aminoácido mais próximo para 1SPD.pdb (tipo selvagem, em vermelho) e 1N19.pdb (estrutura mutada, em azul).

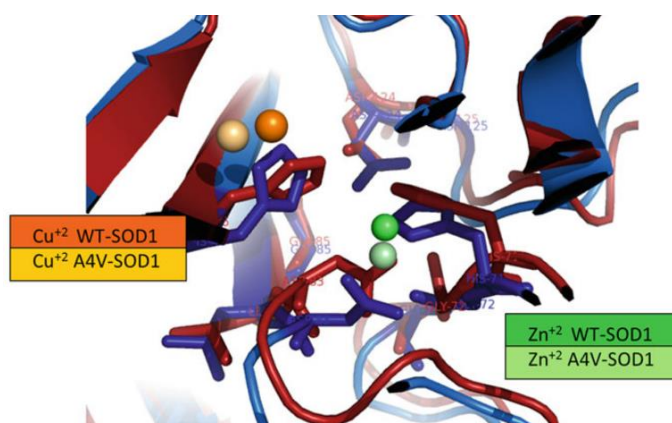


Figura 28. Alinhamento estrutural de 1SPD_A e 1N19_A com ênfase nas posições Cu (superior esquerda) e Zn (inferior direita).

Das plataformas para análise das estruturas proteicas apresentadas acima, o Blue Star Sting destaca-se como a mais completa. Além de oferecer a observação da estrutura tridimensional das estruturas depositadas no PDB, também permite ao usuário observar os modelos estruturais por ele criados. Ele pode calcular alguns dos parâmetros físico-químicos e estruturais oferecidos pelo Blue Star Sting para o modelo e analisa-lo usando o MSSP. Existe uma completa interação entre os módulos do Blue Star Sting, sendo que, a partir de qualquer um deles, é possível interagir com o Jmol integrado à plataforma.

2.12 Bancos de dados com parâmetros estruturais

2.12.1 PDB

Talvez o mais usado banco de dados com parâmetros estruturais seja o **Protein Data Bank (PDB)** (BERNSTEIN, KOETZLE, *et al.*, 1977). Atualmente com mais de 140 mil estruturas armazenadas (147073 em 14 de dezembro de 2018), teve seu início em 1976, com o depósito de 13 estruturas (Fig. 29). Seus dados são de acesso público. Um arquivo PDB tem as seguintes informações: sequência primária da proteína, presença de resíduos que não são padrão nas proteínas, como os heteroátomos, localização dos EES α -hélice e folha- β , dados da geometria dos experimentos cristalográficos e transformação das coordenadas e coordenadas espaciais (eixos X-Y-Z) dos átomos de cada resíduo de aminoácido da estrutura.

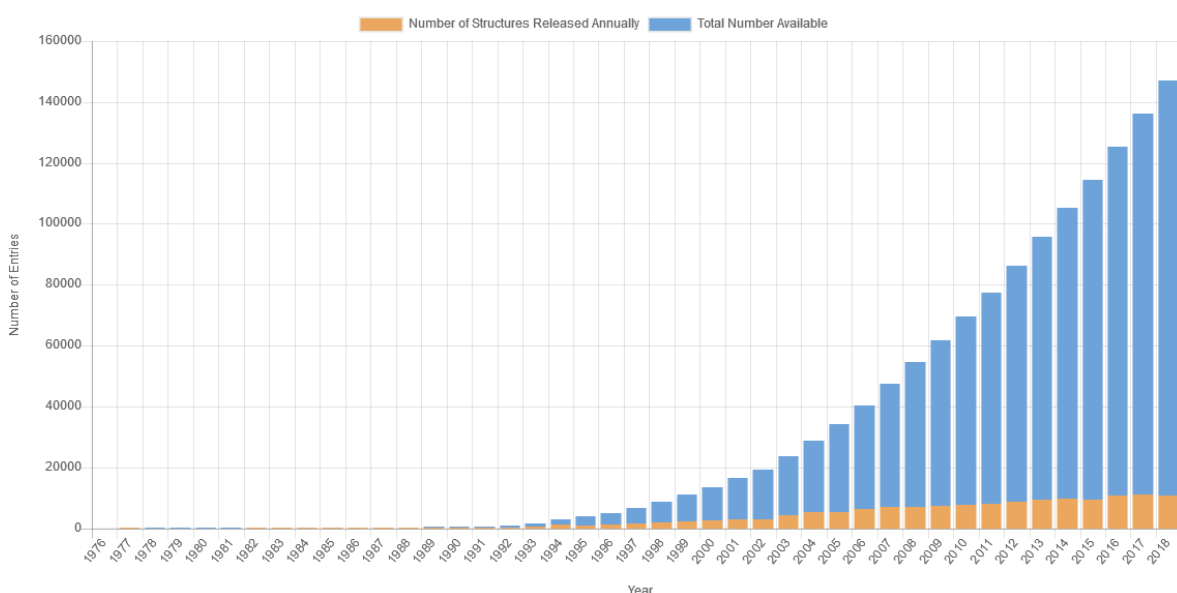


Figura 29. Crescimento anual do número de estruturas depositadas no PDB. Em 14 de dezembro de 2018 tínhamos 147073 entradas. Fonte: <https://www.rcsb.org/stats/growth/overall> pesquisado em 14 de dezembro de 2018.

2.12.2 HSSP

O **Homology-derived Structures of Proteins (HSSP)** (SCHNEIDER e SANDER, 1996) é um banco de dados com informações derivadas da junção entre a estrutura primária e terciária das proteínas. Para cada estrutura proteica conhecida, depositada no PDB, o HSSP contém o alinhamento dessa estrutura com todas as outras estruturas com sequências homólogas. Como resultado, o HSSP não é apenas um banco de dados de alinhamentos sequenciais, mas também um banco de dados com informações das estruturas secundária e terciária. As informações contidas no HSSP são úteis para a análise de conservação dos

resíduos de aminoácidos no contexto estrutural e para definir estruturalmente padrões de sequências significativas.

2.12.3 UniProt

UniProt (LEINONEN, DIEZ, *et al.*, 2004) é o mais abrangente banco de dados de sequências não redundantes. Para garantir que as sequências sejam não redundantes, cada sequência é armazenada uma única vez e recebe um identificador. Quando as proteínas são carregadas na base de dados, são criadas referências cruzadas com as suas sequências. Como resultado, uma busca de sequência no UniProt equivale a uma busca em todos os bancos de dados referenciados por ele.

2.12.4 PROSITE

O PROSITE (SIGRIST, DE CASTRO, *et al.*, 2012) é um banco de dados que descreve as proteínas em termos de seus domínios, famílias e sítios funcionais, seus padrões associados e o perfil associado a elas.

2.12.5 STING_RDB

O GPBC desenvolveu um banco de dados chamado STING_RDB (OLIVEIRA, RODRIGUES, *et al.*, 2006), que armazena em um único lugar as informações extraídas semanalmente do PDB, HSSP, UniProt e Prosite. Além desses dados brutos, vários algoritmos são executados semanalmente, gerando novos dados, por exemplo, os descritores de contatos, físico-químicos, estruturais e geométricos para cada resíduo de aminoácido presente em cada cadeia de todas as estruturas armazenadas no PDB. Essas informações são armazenadas no STING_RDB, fazendo dele o mais completo e robusto banco de dados estruturais. Atualmente o STING_RDB tem 12.043.391.057 registros armazenados em suas 98 tabelas, em um total de 1965 atributos (dados extraídos em 16 de maio de 2018). Considerando o crescimento anual da base de dados do PDB (Fig. 29), e que o STING_RDB cresce em igual proporção, a demanda do parque computacional do grupo é muito grande.

As informações extraídas do STING_RDB foram usadas neste trabalho para caracterizar o nano-ambiente das proteínas funcionais onde ocorre a nucleação (composta por iniciação, propagação e terminação) das estruturas secundárias: α -hélice e folha- β .

2.13 Nano-ambiente dos aminoácidos

Os resíduos de aminoácidos de uma proteína distantes entre si na estrutura primária tornam-se próximos uns dos outros quando ela assume sua forma espacial. Essa

conformação tridimensional origina o nano-ambiente onde o EES está inserido (Fig. 30). Manavalan e Ponnuswamy (MANAVALAN e PONNUSWAMY, 1977) propuseram que cada um dos 20 aminoácidos possíveis de estarem presentes em uma proteína tem seu próprio nano-ambiente, e a cooperação entre esses aminoácidos influencia na determinação da estrutura final da proteína. Zhong e Johnson (ZHONG e JOHNSON, 1992) demonstraram experimentalmente que o nano-ambiente é importante para determinar a estrutura secundária formada pela sequência de aminoácidos, embora a predição da estrutura secundária baseada apenas na sequência de aminoácidos pode nunca ser totalmente bem sucedida. McDonald e Johnson (MACDONALD e JOHNSON JR, 2001) também estudaram as características dos aminoácidos presentes no nano-ambiente de uma estrutura secundária, e concluíram que a acessibilidade ao solvente e a interação de cada resíduo de aminoácido com outros resíduos próximos são importantes na determinação da estrutura secundária.

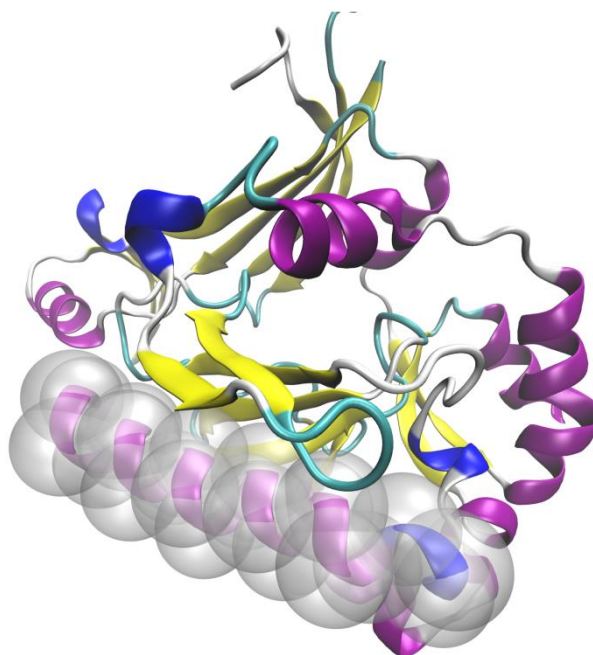


Figura 30. Um exemplo de α -hélice e seu nano-ambiente: o *Synthetic Gene Encoded DcpS bound to inhibitor DG157493* (3bl9.pdb) possui catorze α -hélices e cada uma delas possui seu próprio nano-ambiente. Para ilustrar, vemos uma α -hélice em destaque dentro das esferas transparentes. O nano-ambiente inclui resíduos de aminoácidos da própria α -hélice e os resíduos de aminoácidos ao seu redor que estão ao alcance da esfera de sondagem, cujo raio foi previamente selecionado. Regiões pré e pós-hélice (na extensão arbitrária de 32 resíduos de aminoácidos cada) não são mostrados aqui por causa da clareza da definição básica.

McDonald e Johnson (MACDONALD e JOHNSON JR, 2001) simplificaram o conjunto de características restringindo o nano-ambiente em uma vizinhança de apenas 4 resíduos de aminoácidos. Neste trabalho ampliamos o espaço amostral, adotando-se o número

arbitrário de 32 resíduos de aminoácidos antes e depois do EES como parte do nano-ambiente. Por exemplo, para uma α -hélice formada por nove resíduos de aminoácidos, consideramos como nano-ambiente a região formada por setenta e três resíduos de aminoácidos (32 resíduos antes + α -hélice de tamanho 9 + 32 resíduos depois), dos quais os resíduos de aminoácidos presentes nesta α -hélice fazem parte.

2.14 Aplicações do conhecimento adquirido pela análise do nano-ambiente onde se insere os EES

As possíveis aplicações do conhecimento adquirido pela análise do nano-ambiente propício para a nucleação e a manutenção dos elementos da estrutura secundária no contexto estrutural das proteínas funcionais são as seguintes: a) conhecer o nano-ambiente e suas características; b) melhoria das ferramentas para validação das estruturas obtidas a partir de softwares de modelagem, tais como Modeller²⁴, Swiss-Model²⁵ etc., e c) avaliação dos modelos daquelas estruturas obtidas pelos métodos de difração de raios X e ressonância magnética nuclear.

O entendimento de como o nano-ambiente mantém cada um dos elementos da estrutura secundária abrirá o caminho para melhorar a previsão e a certificação dos modelos relativos a esses elementos. Isso permitirá uma predição mais precisa da estrutura terciária da proteína e, portanto, da sua função.

Compreender a relação existente entre sequência, estrutura e função abre um leque de possíveis aplicações desse conhecimento. No cenário da agricultura brasileira será possível criar, com este conhecimento, novas vacinas, drogas veterinárias, inseticidas etc., com maior eficácia. Podemos obter a sequência proteica a partir da sequência do genoma de um determinado organismo (animal ou planta). Com a sequência proteica poderemos prever os elementos da estrutura secundária, conseqüentemente teremos uma melhor predição da estrutura final tridimensional da proteína, e conseqüentemente a elaboração mais precisa de possíveis inibidores da função destas proteínas que são conhecidos como: agrotóxicos, inseticidas e até vacinas. Isso significa um novo paradigma no design de novos fármacos. A partir da sequência de aminoácidos, será possível modelar e simular reações enzimáticas, iterações proteína-proteína, proteína-substrato, dentre outras, em menor tempo e com menor custo que se realizado *in vivo*.

²⁴ <http://salilab.org/modeller/>

²⁵ <http://swissmodel.expasy.org/>

3 MATERIAIS E MÉTODOS

3.1 Criação dos *Datamarts*

Um banco de dados é um conjunto interligado de arquivos, construído a partir de uma metodologia científica que tem passos bem definidos (SILBERSCHATZ, SUNDARSHAN e KORTH, 2016). O banco de dados STING_RDB foi desenvolvido pelo GPBC usando o modelo relacional, em que os dados são armazenados em tabelas relacionadas entre si. O STING_RDB tem 98 tabelas (Tabela 4). Essas tabelas representam 27 tipos diferentes e independentes de descritores estruturais de proteínas calculados pelo STING, totalizando 1307 variações desses descritores.

Accessibility	Het_Syn	Res_Patche_atoms
Anisou	Hetatm	Residue_Base
Atom	Hetatm_Anisou	Residue_Params
AuditResidue_Base_AUD	Hetatm_Siguij	Resolution
CSA	IFRContact	Revdat
Chain	InternalContact	SW_Density
ChemicalComponent	Jrnl	SW_Energy_Density
Cispep	Link	SW_Entropy_Density
Compnd	Model	SW_Sponge
Conect	Modres	Scale
Conservation	MouthComplexAtom	Seqadv
ContactType	MouthIsolationAtom	Seqres
Contacts	Mouth_Complex	Sheet
ContactsVoronoi	Mouth_Isolation	Sigatm
ContactsWNA	Mtrix	Siguij
CrossOrder	Obsolete	Site
CrossOrderWNA	Origx	SiteRes
Cryst	PDB	Solvation
Dbref	PDBSummary	Source
Density	PLContact	Space_Clash
DensityVoronoi	Physical_Chemical_Param	Sponge
DensityWNA	PhysicoChemicalAndGeometricWNA	SpongeVoronoi
EnergyDensityVoronoi	PhysicoChemicalVoronoi	SpongeWNA
Energy_Density	PocketComplexAtom	Ssbond
Energy_DensityWNA	PocketIsolationAtom	Structural_param
Entropy_Density	Pocket_Complex	Turn
FPocket	Pocket_Isolation	Tvect
FPocketComplex	Predicted_IFR	UnusedContact
Formul	Prosite	UnusedContactWNA
Geometric_Param	ProteinLigandContacts	UnusedContactsVoronoi
Helix	REVINFO	checksum
Het	ReportMessage	mysql_monk
Het_Name	Res_Patche	

Tabela 4. Tabelas do banco de dados STING_RDB. As tabelas destacadas em amarelo referem-se aos descritores do STING. As tabelas destacadas em azul guardam os dados extraídos dos arquivos PDB que identifica o tipo de estrutura helicoidal (α -hélices, π -hélices e hélices 3_{10}), se a folha- β é paralela ou anti-paralela. A tabela Residue_Params destacada em vermelho informa se os resíduos de aminoácidos em cada proteína fazem parte de uma α -hélice, folha- β ou turn. Esse dado foi usado no processo de criação dos *Datamarts*. As demais tabelas armazenam as informações extraídas do arquivo PDB, como aquelas que identificam uma estrutura, ou são auxiliares, por conter dados que complementam as tabelas de descritores.

Ele é atualizado semanalmente, seguindo o fluxograma da Fig. 31.

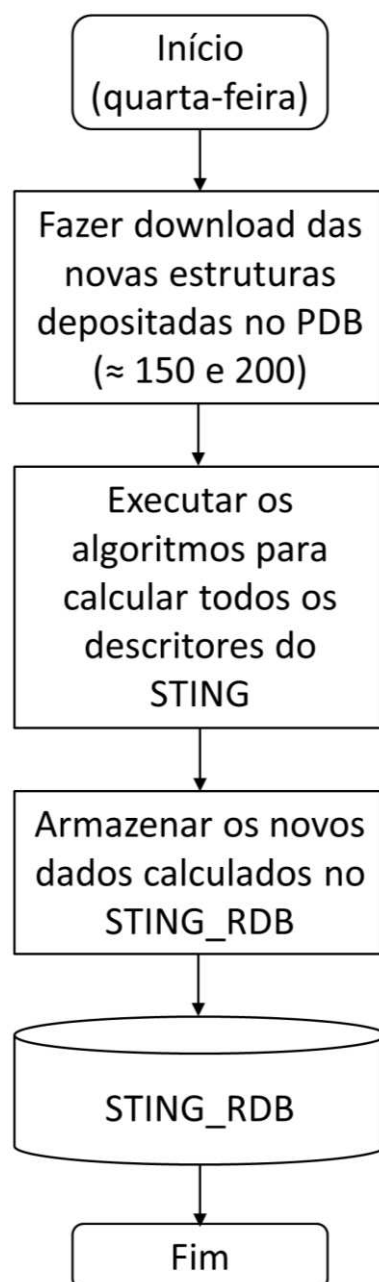


Figura 31. Fluxograma de atualização do STING_RDB. Todas as quartas-feiras o PDB disponibiliza entre 150 e 200 novas estruturas depositadas ao longo da semana. Neste mesmo dia nós fazemos o download dessas novas estruturas. Terminado o download, executamos um conjunto de algoritmos para calcular os descritores do STING. Os novos dados calculados são armazenados no STING_RDB.

Um *Datamart* reflete um subconjunto dos dados armazenados nestas tabelas. Durante o desenvolvimento deste trabalho, nós desenvolvemos um *Datamart* relativo ao nano-ambiente onde as α -hélices, folhas- β e *turns* estão inseridos. O primeiro passo para a sua criação foi selecionar e extrair as estruturas presentes no STING_RDB, agrupando-as em all- α ,

all- β , α -hélice nas $(\alpha+\beta)+(\alpha/\beta)$, folhas- β nas $(\alpha+\beta)+(\alpha/\beta)$ e “desordenadas”. Como explicado no Capítulo 2.7 as estruturas do tipo all- α não possuem folhas- β , enquanto estruturas do tipo all- β não possuem α -hélices. Estruturas do tipo $(\alpha+\beta)+(\alpha/\beta)$ possuem tanto α -hélices como folhas- β . Já as estruturas “desordenadas” não possuem α -hélices e folhas- β . A tabela *Residue_Params* foi usada nessa seleção. Essa tabela diz em qual EES (α -hélice, folha- β ou *turn*) cada resíduo de aminoácido de uma proteína está presente, pelas definições dos algoritmos PDB, DSSP e Stride.

Conforme explicado no Capítulo 2.5, a definição dos EES presentes em um arquivo PDB é atribuída ao autor daquele arquivo. Ele pode ter usado a definição dos algoritmos DSSP, Stride, ou de outro algoritmo. Mas, segundo Cuff (CUFF e BARTON, 1999) o resultado dos algoritmos DSSP e Stride tem consenso de até 95%, optamos por usar neste trabalho os dados desses dois algoritmos, e também a definição encontrada no arquivo PDB, cujo autor pode ter usado, coincidentemente, a definição do DSSP ou do Stride. Neste caso, naturalmente, haverá um consenso entre elas. Caso contrário, usar as três definições aumentará a confiabilidade das análises.

A tabela *Residue_Params* contém os atributos *chain_id*, *pdb_name*, *rb_number*, *SS_DSSP*, *SS_PDB* e *SS_Stride*. Esses atributos apontam se um resíduo de aminoácido de uma cadeia em um arquivo PDB (indicados pelos atributos *rb_number*, *chain_id* e *pdb_name*, respectivamente) está presente em uma α -hélice, folha- β ou *turn*, segundo as definições do PDB, DSSP e Stride. Os atributos *SS_DSSP*, *SS_PDB* e *SS_Stride* usam os seguintes códigos computacionais usados na identificação dos EES: {0, 1, 2, 3} indicam um *turn*, {1000} indica uma α -hélice e {100, 200} indicam uma folha- β . Esses códigos foram adotados arbitrariamente e não tem significado biológico. A Tabela 5 apresenta um extrato da tabela *Residue_Params* extraído do STING_RDB.

chain_id	pdb_name	rb_number	SS_DSSP	SS_PDB	SS_Stride
E	1ppf	55	0	1000	0
E	1ppf	56	1000	1000	1000
E	1ppf	57	1000	1000	1000
E	1ppf	58	1000	1000	1000
E	1ppf	59	2	1000	2
E	1ppf	60	1	1000	1
E	1ppf	61	3	0	1
E	1ppf	62	0	0	3

Tabela 5. Extrato da tabela *Residue_Params* do STING_RDB. Apresentamos os resíduos de aminoácidos 55-62 da cadeia E da estrutura 1ppf.pdb. Nas definições do DSSP e Stride, existe uma α -hélice entre os resíduos 56-58, e na definição do PDB a α -hélice se estende entre os resíduos 55-60. Os demais resíduos de aminoácidos formam *turns*.

Para separar as estruturas de acordo com a pertinência do EES foi usado o seguinte código em linguagem SQL:

```
select pdb_name, chain_id, rb_number, ss_pdb from
STING_RDB.Residue_Params where pdb_name NOT IN(select pdb_name from
STING_RDB.Residue_Params where SS_PDB IN(code)) order by pdb_name;
```

Esse código funciona em duas etapas:

- Seleciona as estruturas proteicas que possuem um EES específico, determinado pelo *code*. É o *select* interno, escrito em vermelho no quadro acima;
- Seleciona todas as estruturas que não fazem parte do primeiro subconjunto.

Portanto, pelo processo de seleção e exclusão, as estruturas proteicas foram selecionadas e agrupadas pela pertinência do EES em: all- α , all- β , α -hélice em $(\alpha+\beta)+(\alpha/\beta)$, folha- β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada”.

O próximo passo foi agrupar esses conjuntos pelo consenso que elas tem entre si. Isso significa que o EES deve começar no mesmo resíduo de aminoácido e ter o mesmo tamanho (Fig. 32). O consenso mais restritivo é aquele entre o PDB-DSSP-Stride. Isso significa que o EES começa no mesmo resíduo de aminoácido e tem o mesmo tamanho definido pelos três algoritmos. As outras possíveis variações são: PDB-DSSP, PDB-Stride e DSSP-Stride.

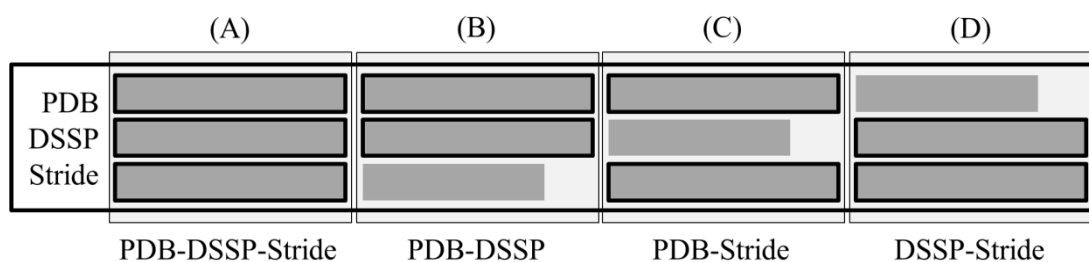


Figura 32. Agrupamento segundo o consenso entre PDB, DSSP e Stride. Temos quatro possíveis consensos: (A) PDB-DSSP-Stride quando o EES começa e termina no mesmo resíduo, e, portanto, tem o mesmo tamanho, nas definições do PDB, DSSP e Stride; (B) PDB-DSSP quando o EES começa e termina no mesmo resíduo, e, portanto, tem o mesmo tamanho, nas definições do PDB e DSSP; (C) PDB-Stride quando o EES começa e termina no mesmo resíduo, e, portanto, tem o mesmo tamanho, nas definições do PDB e Stride; e (D) DSSP-Stride quando o EES começa e termina no mesmo resíduo, e, portanto, tem o mesmo tamanho, nas definições do DSSP e Stride.

Cada variação do consenso originou uma tabela do *Datamart*. As tabelas de consenso forneceram a informação de quais proteínas são do tipo all- α , all- β , α -hélice em $(\alpha+\beta)+(\alpha/\beta)$, folha- β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada”, com maior ou menor rigidez em suas

definições. Após criarmos as tabelas de consenso, fizemos o alinhamento por tamanho das estruturas presentes nessas tabelas (Fig. 33). Usamos o número arbitrário de 32 resíduos de aminoácidos antes e depois do EES alinhada como universo amostral do nano-ambiente. Quando não havia 32 resíduos de aminoácidos disponíveis, completamos com *gaps*.

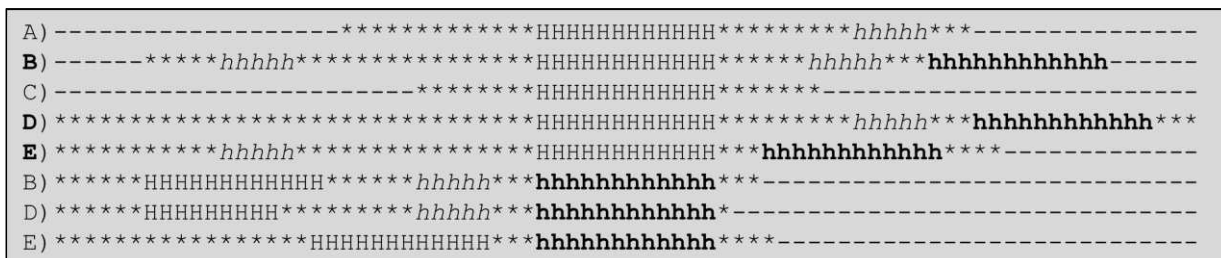


Figura 33. Alinhamento por tamanho. No exemplo abaixo há cinco estruturas do tipo all- α alinhadas pelo pela α -hélice de tamanho 12 ("H"). Dessas cinco estruturas, três (B, D, E) possuem uma segunda α -hélice de tamanho 12 ("h"), que também é alinhada com a primeira ocorrência de um EES deste mesmo tamanho.

A diferença no número de estruturas alinhadas influencia perceptivelmente no tamanho do ruído gerado nas análises estatísticas. Quanto maior foi o número de estruturas alinhadas, menor será o ruído obtido (Fig. 34). Para diminuir o ruído, nós alinhamos os EES pelo seu C-Terminal e N-Terminal, independente do seu tamanho (Fig. 35 e 36). Isso aumenta o número de EES alinhadas, melhorando a razão sinal/ruído e a análise estatística do sinal encontrado. Embora com esta abordagem não seja possível observar o que acontece durante a presença do EES na sua extensão integral, a relação sinal/ruído fica muito mais favorável no seu início (N-Terminal) e fim (C-Terminal).

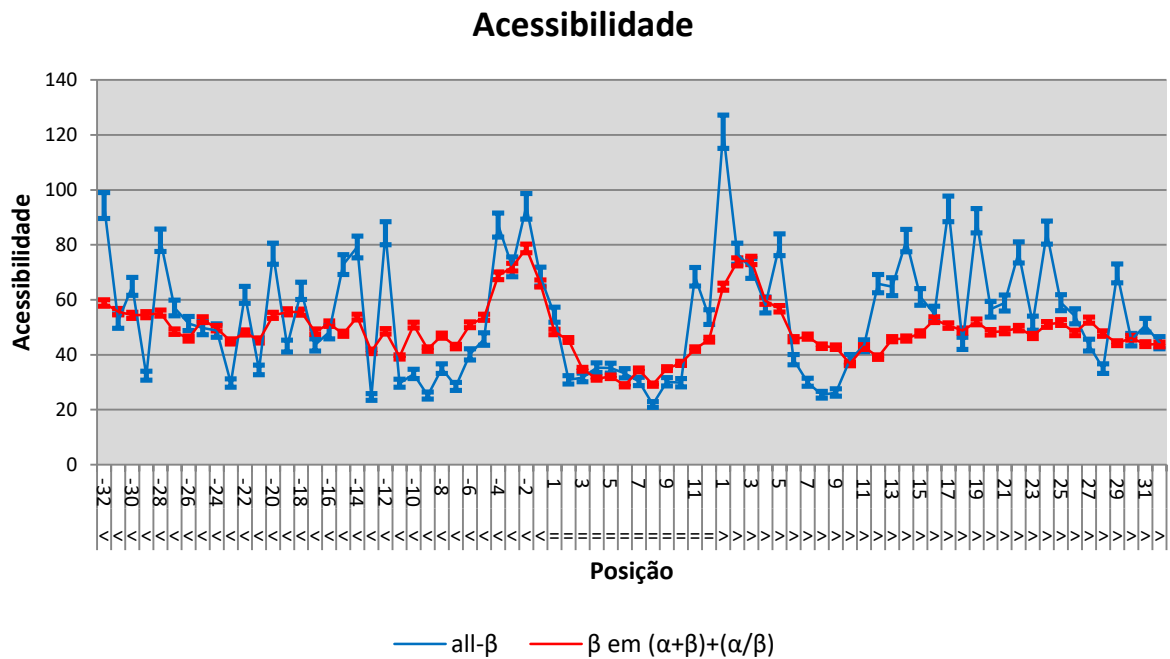


Figura 34. Influência do número de estruturas alinhadas na obtenção do sinal com menos ruído. No exemplo, alhamos 47 folhas- β do tipo all- β (azul) e 6022 folhas- β do tipo $(\alpha+\beta)+(\alpha/\beta)$ (vermelho). Em ambos os casos o tamanho da folha- β é 12 resíduos de aminoácidos. A diferença entre o número de estruturas alinhadas é de ≈ 128 vezes mais estruturas do tipo $(\alpha+\beta)+(\alpha/\beta)$. Nota-se que o desvio padrão é maior nas proteínas do tipo all- β , devido o relativamente baixo número de amostras.



Figura 35. Alinhamento posicional pelo C-Terminal. No exemplo abaixo temos oito estruturas proteicas do tipo all- α . Todas as α -hélices foram alinhadas pelo seu C-Terminal. Neste caso, não importa o tamanho do EES. Isso aumenta a quantidade de estruturas alinhadas, possibilitando a extração de um sinal mais forte, com maior razão sinal/ruído perto do C-Terminal do EES.

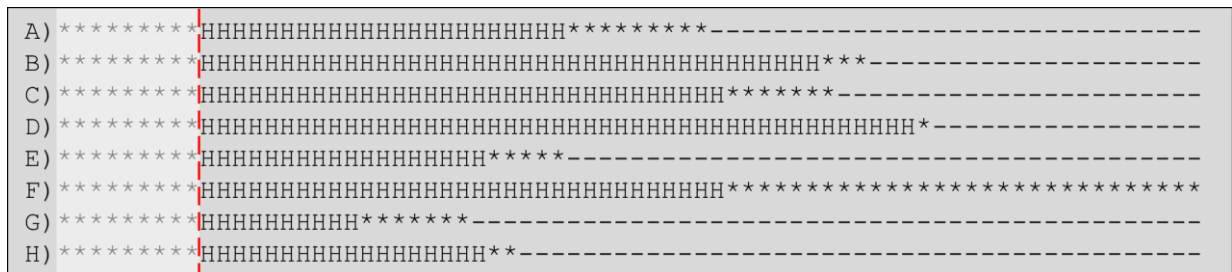


Figura 36. Alinhamento posicional pelo N-Terminal. No exemplo abaixo temos oito estruturas proteicas do tipo all- α . Todas as α -hélices foram alinhadas pelo seu N-Terminal. Neste caso, não importa o tamanho do EES. Isso aumenta a quantidade de estruturas alinhadas, possibilitando a extração de um sinal mais forte, com maior razão sinal/ruído perto do N-Terminal do EES.

Depois disso, para cada tipo de consenso e alinhamento, criamos os *Datamarts* dos descritores, separados pelos tipos de estrutura. Por exemplo, para as proteínas do tipo all- α com consenso entre os algoritmos PDB-DSSP-Stride nós temos os descritores de contatos, físico-químicos, geométricos, estruturais, etc. e assim sucessivamente (Fig. 37). Essa organização do *Datamart* limita e organiza o espaço amostral, diminuindo o tempo de busca no banco de dados.

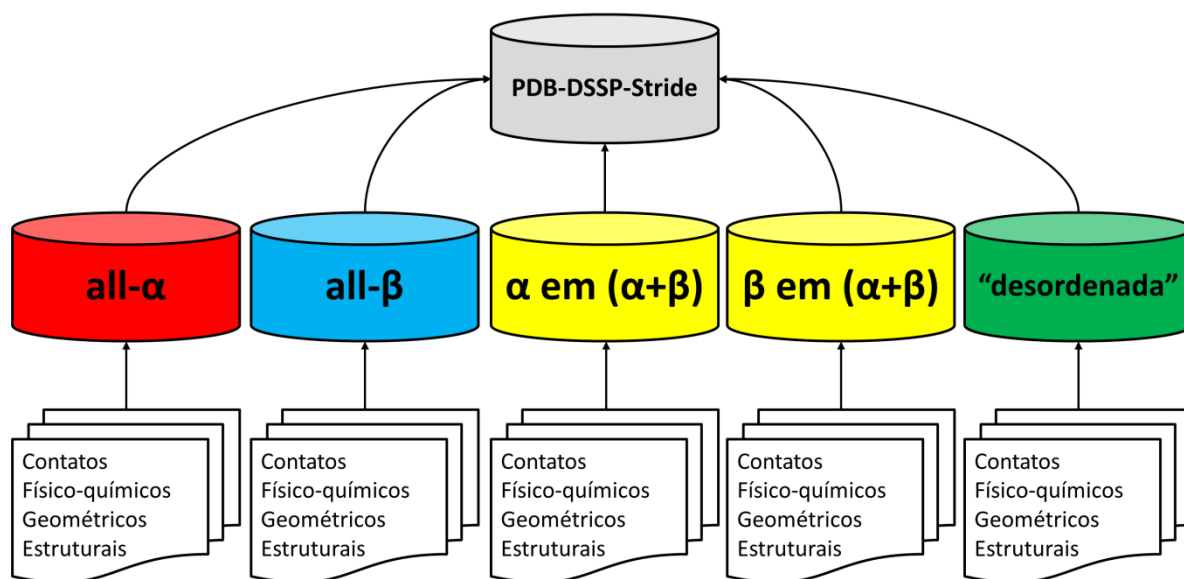


Figura 37. Esquema de criação dos *Datamarts*. Cada variante do consenso entre os algoritmos PDB, DSSP e Stride origina cinco tabelas: no exemplo abaixo, temos as tabelas all- α , all- β , α -hélice em $(\alpha+\beta)+(\alpha/\beta)$, folha- β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada” para o consenso PDB-DSSP-Stride. E para cada tipo de proteína foi criada uma tabela para cada tipo de descritor: de contatos, físico-químico, geométrico, estrutural, etc. e assim sucessivamente para cada variação possível.

3.2 Eliminação da redundância

Muitas das estruturas presentes no STING_RDB são homólogas entre si. Embora essa redundância aumente o número de EES alinhados, ela introduz um erro na compreensão dos dados, porque poderia artificialmente aumentar, ou ampliar, o sinal obtido. A fim de garantir a melhor qualidade dos dados, eliminamos a redundância no nível das sequências usando do software *cd-hit*²⁶ (LI e GODZIK, 2006).

Para tirar a redundância usando o software *cd-hit* usamos o seguinte comando:

```
$ cdhit -i pdb_seqres.txt -o db50 -c 0.50 -n 3
```

²⁶ <http://weizhongli-lab.org/cd-hit/>

onde:

- i arquivo de entrada chamado “pdb_seqres.txt” esse arquivo pode ser baixado do site PDB
- o arquivo de saída chamado “db50” no exemplo acima
- c nível de redundância eliminado
- n tamanho da palavra

O arquivo “pdb_seqres.txt” contém as estruturas primárias das estruturas depositadas no PDB no formato FASTA (Fig. 38).

```
>1PPF:E|PDBID|CHAIN|SEQUENCE
IVGRRARPHAWPFMVSLQLRGGHFCGATLIAPNFVMSAAHCVANVNVRAVRVVLGAHNLSRREPTRQVFVAVQRIFENGY
DPVNLNDIVILQLNGSATINANVQVAQLPAQGRRRLGNGVQCLAMGWLLGRNRGIASVLQELNVTVVVTSLCRRSNVCTL
VRGRQAGVCFGDSGSPVLCNGLIHGIASFVRGGCASGLYPDAFAPVAQFVNWIDSIIQ
>1PPF:I|PDBID|CHAIN|SEQUENCE
LAAVSDCSEYPKPACTLEYRPLCGSDNKTYGNKCNFCNAVVESNGTLTLSHFGKC
```

Figura 38. Formato FASTA. Sequência de aminoácidos do PDB 1ppf em formato FASTA.

Eliminamos as sequências redundantes nos níveis de 50%, 70% e 95%. Depois que o arquivo de saída foi gerado, nós o processamos da seguinte forma:

```
$ fasta2pdblista.pl db50
```

O script *fasta2pdblista.pl*, escrito na linguagem de programação Perl²⁷, lê o arquivo “db50”, extrai a informação do código PDB e cadeia, e grava essas informações em um arquivo de saída com o nome db50list.txt com o seguinte formato {*pdb_name*, *chain_id*, *percentage*} por linha. Depois, esse arquivo processado foi usado para carregar a tabela de dados não redundantes no *Datamart*. Com este procedimento garantimos a “diversidade” do nano-ambiente no contexto da estrutura proteica como toda até para o EES com idêntico tamanho e sequência.

3.3 Seleção dos dados

Utilizando diferentes parametrizações, o STING calcula um total de 1307 descritores estruturais de proteínas. Nós selecionamos os 69 parâmetros mais representativos do STING_RDB para descrever o nano-ambiente onde ocorre a nucleação e manutenção do EES. Eles foram agrupados em sete classes: estrutural, geométrico, contatos, contatos não

²⁷ <https://www.perl.org/>

usados, físico-químicos, ponderados pela vizinhança e outros (Tabela 6). Na sequência explicamos o significado de cada parâmetro.

Estrutural (I)		18	HB-MWS	38, 39	HB-MS
1	Temperature_Factor_CA	19	HB-MWWS	40, 41	HB-MWS
2	Dihedral_Angle_PHI (ϕ)	20	HB-SS	42, 43	HB-MWWS
3	Dihedral_Angle_PSI (ψ)	21	HB-SWS	44, 45	HB-SS
4	Dihedral_Chi1	22	HB-SWWS	46, 47	HB-SWS
5	Dihedral_Chi2	23	Hydrophobic	48, 49	HB-SWWS
6	Dihedral_Chi3	24	Aromatic	50, 51	Hydrophobic
7	Dihedral_Chi4	25	Ch_attractive	52, 53	Aromatic
8	Density IFR	26	Ch_repulsive	54, 55	Ch_attractive
9	Density Internal	27	Disulfide	56, 57	Ch_repulsive
10	Space Clash number of clashes	Unused Contacts (IV)		58, 59	Disulfide
11	Space Clash percent	28	Number_Unused_Contact	60, 61	Number_Unused_Contact
Geométrico (II)		Físico-químico (V)		62, 63	Electrostatic_Potential_at_CA
12	Cross_Link_Order_CA	29	Electrostatic_Potential_at_CA	64, 65	Electrostatic_Potential_Average
13	Cross_Pres_Order_CA	30	Electrostatic_Potential_Average	66, 67	Electrostatic_Potential_at_LHA
Contatos (III)		31	Electrostatic_Potential_at_LHA	Outros (VII)	
14	HB-MM	WNA ^(*) by Distance and at Surface (VI)		68	Accessible_in_Isolation
15	HB-MWM	32, 33	HB-MM	69	Hydrophobicity_KDI
16	HB-MWWM	34, 35	HB-MWM		
17	HB-MS	36, 37	HB-MWWM		

Tabela 6. Parâmetros do STING_RDB, onde HB: hydrogen bond (ponte de hidrogênio), M: Main chain (cadeia principal), S: Side chain (cadeia secundária), W: water (água), CA: carbono- α , LHA: *Last Heavy Atom* (último átomo pesado em cadeia lateral de aminoácido), KDI: refere-se à escala de hidrofobicidade de Kyte Doolittle (KYTE e DOOLITTLE, 1982). Os ângulos diedrais ϕ e ψ foram incluídos na lista dos parâmetros selecionados a fim de servirem como gabarito para as análises dos demais descritores, uma vez que essas informações são intrínsecas para cada aminoácido. Os descritores WNA são contados duas vezes, porque foram usados dados de WNA Distance e WNA Surface.

3.4 Descritores do STING_RDB

3.4.1 Potencial eletrostático

Os resíduos de aminoácidos são formados por átomos, e esses átomos podem apresentar cargas elétricas. O Blue Star Sting calcula o potencial eletrostático usando o

software DelPhi (NESHICH e ROCCHIA, 2007) através da equação de Poisson-Boltzmann (Eq. 6).

$$\nabla \cdot [\epsilon_r(r) \nabla \varphi(r)] = \frac{1}{\epsilon_0} [\rho^{cargasfixas}(r) - \frac{\epsilon_{solv} k^2(r)}{4\pi} \varphi(r)] \quad (6)$$

onde: $\epsilon_r(r)$ é a constante dielétrica relativa local, ϵ_{solv} é a constante dielétrica do solvente, $\rho^{cargasfixas}(r)$ é a distribuição de carga sobre o soluto, e k^2 é o parâmetro de Debye, definido pela Eq. 7:

$$k^2 = \frac{8\pi e^2 C}{\epsilon_{solv} k_B T} [kT/e] \quad (7)$$

onde: C é a concentração em massa do sal, k_B é a constante de Boltzmann e T é a temperatura absoluta.

O STING_RDB armazena os valores de potencial eletrostático calculado para o carbono- α , para o último átomo pesado da cadeia lateral (LHA, do inglês “Last Heavy Atom”), para a média de todos os átomos do resíduo de aminoácido e para a superfície, criada por cada aminoácido analisado se o resíduo faz parte da superfície proteica.

3.4.2 Acessibilidade

O Blue Star Sting calcula a acessibilidade usando o programa *SurfV* (SRIDHARAN, NICHOLLS e HONIG, 1992). São calculados três valores: acessibilidade da cadeia isolada, em um complexo com outra cadeia, e a acessibilidade relativa. A acessibilidade relativa é calculada usando os valores apresentados na Tabela 7 (NESHICH, HIGA, *et al.*, 2002). Neste trabalho usamos apenas a acessibilidade da cadeia isolada.

Quatro arquivos PDBs foram selecionados com um dos 20 aminoácidos presente no C-Terminal de cada um desses arquivos. Após isso, foi calculada a área acessível na superfície da estrutura deste aminoácido sozinho (do inglês “Accessible Surface Area” (ASA)).

Para cada um desses quatro PDBs selecionamos os valores de ASA máxima e mínima. A diferença entre os dois valores foi adicionada ao maior valor. Esses quatro valores são apresentados na tabela 6 (ASA_1, ASA_2, ASA_3, ASA_4). O Blue Star Sting usa os

números na última coluna (destacada em azul) como valores de ASA para cada aminoácido dentro de uma cadeia.

Residue	ASA_1	ASA_2	ASA_3	ASA_4	Difference =(max-min)	%difference	[%diff]	Mult.fact.	BLUE STAR STING ASA VALUE USED = MAX EXP. VALUE X MULT.FACT.
ALA	231,680	232,768	233,585	234,345	2,665	1,14	2	1,02	239,0
ARG	362,479	366,063	371,165	372,958	10,479	2,81	3	1,03	384,1
ASN	277,159	278,898	282,891	284,687	7,528	2,64	3	1,03	293,2
ASP	271,343	273,849	276,055	276,055	5,484	1,98	2	1,02	282,4
CYS	255,831	256,499	260,818	262,350	6,519	2,48	3	1,03	270,2
GLN	302,331	303,265	309,626	310,915	8,584	2,76	3	1,03	320,2
GLU	292,397	303,215	304,521	312,052	19,655	6,30	7	1,07	333,9
GLY	202,875	205,294	205,443	205,471	2,596	1,26	2	1,02	209,6
HIS	303,173	305,122	307,905	317,645	14,472	4,56	5	1,05	333,5
ILE	302,749	302,944	305,736	306,417	3,668	1,20	2	1,02	312,5
LEU	306,523	310,613	311,242	311,438	4,915	1,58	2	1,02	317,7
LYS	339,108	340,594	343,988	347,742	8,634	2,48	3	1,03	358,2
MET	317,302	317,461	321,639	325,357	8,055	2,48	3	1,03	335,1
PHE	325,050	327,259	340,984	342,137	17,087	4,99	5	1,05	359,2
PRO	265,778	265,784	266,094	266,453	675	0,25	1	1,01	269,1
SER	241,856	242,445	243,840	245,213	3,357	1,37	2	1,02	250,1
THR	265,434	266,091	267,756	269,504	4,070	1,51	2	1,02	274,9
TRP	363,036	368,853	385,297	387,278	24,242	6,26	7	1,07	414,4
TYR	352,526	355,254	359,056	360,794	8,268	2,29	3	1,03	371,6
VAL	276,272	278,396	279,453	280,703	4,431	1,58	2	1,02	286,3

Tabela 7. Valores da área de superfície acessível para os aminoácidos isolados usados pelo Blue Star Sting. Fonte: http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/solvent_accessible_area.html

A área acessível de cada resíduo foi calculada utilizando uma sonda esférica com raio 1,4 Å (raio da molécula de água) que percorre a superfície de Van der Waals da estrutura proteica. A superfície produzida pelo centro geométrico da sonda esférica é a superfície acessível ao solvente (Fig. 40) (LEE e RICHARDS, 1971).

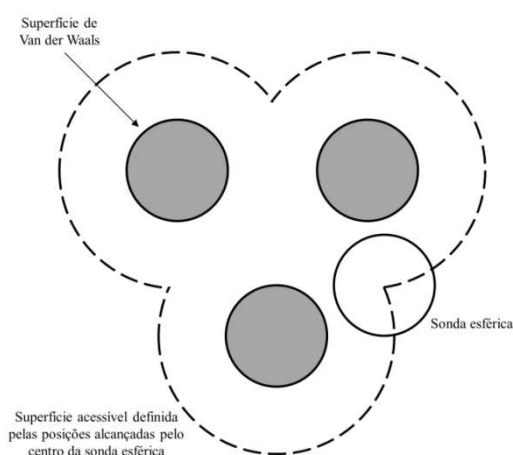


Figura 39. Superfície acessível de acordo com a superfície de Van der Waals de cada molécula. Fonte: (NESHICH, HIGA, et al., 2002)

3.4.3 Espongicidade

A espongicidade é calculada pela soma do volume de Van der Waals de todos os átomos encontrados dentro de uma esfera, com raio igual a 3, 4, 5, 6 e 7 angstroms [Å], centrada no carbono- α ou no último átomo pesado em cadeia lateral de aminoácido, dividido pelo volume dessa esfera (Eq. 8).

$$Espongicidade = \frac{\sum volume_{vdw}}{Volume_{esfera}} \quad (8)$$

3.4.4 Densidade

A densidade é calculada pela soma das massas de todos os átomos encontrados dentro de uma esfera com raio igual a 3, 4, 5, 6 e 7 angstroms [Å], centrada no carbono- α ou no último átomo pesado em cadeia lateral de aminoácido, dividido pelo volume dessa esfera (Eq. 9).

$$Densidade = \frac{\sum massa_{atomos}}{Volume_{esfera}} \quad (9)$$

3.4.5 Space clash

Space clash é a medida do choque estérico que ocorre entre os aminoácidos. Muitas vezes, a medida para o choque estérico é um fator de sobreposição, que é calculado como a razão entre as distâncias entre os dois centros dos átomos para a soma dos seus raios de Van der Waals. Esse fator de sobreposição será menor que 1 se as duas esferas se penetram mutuamente. Isso acontece, por exemplo, em um arranjo de ligações de hidrogênio. Fatores de sobreposição com valores entre 0,8 e 0,75 são comuns nas estruturas obtidas por difração de raios X. Fatores de sobreposição igual a 0,7 seria um choque estérico relativamente menor, enquanto valores iguais ou menores que 0,65 seriam mais graves, indicando possível erro no modelo estrutural.

3.4.6 Hidrofobicidade

A *hidrofobicidade* é a propriedade que mede quão hidrofóbico ou hidrofílico é um aminoácido, ou seja, sua capacidade de estarem, ou não, expostos ao solvente. É uma propriedade fundamental para estabilidade das proteínas. Atribui-se a característica de *hidrofóbicos* aos aminoácidos que não possuem átomos de oxigênio e nitrogênio em suas cadeias laterais, de modo que não formam ligações de hidrogênio com as moléculas de água.

Os aminoácidos capazes de formar ligações de hidrogênio com as moléculas de água são chamados de *hidrofílicos*.

O Blue Star Sting calcula a hidrofobicidade usando as escalas de Kyte Doolittle (KYTE e DOOLITTLE, 1982) e de Radzicka (RADZICKA e WOLFENDEN, 1988) conforme as equações 10 e 11 apresentadas abaixo:

$$Hidrofobicidade_i = \frac{Acessibilidade_i}{AcessibilidadeMaxima_i} \times kyteDoolittle_i \quad (10)$$

$$Hidrofobicidade_i = \frac{Acessibilidade_i}{AcessibilidadeMaxima_i} \times Radzicka_i \quad (11)$$

onde $KyteDoolittle_i$ é o valor da escala Kyte Doolittle de hidrofocidade de cada resíduo (Tabela 8), $Radzicka_i$ o valor da escala Radzicka de hidrofocidade de cada resíduo (Tabela 9), $Access_i$ é o valor de acessibilidade do resíduo i , e $AccessMax_i$ é o valor máximo de acessibilidade para o resíduo i (Tabela 7).

Aminoácido	Polaridade da cadeia lateral	Carga da cadeia lateral pH 7,4	Constante de hidrofobicidade KyteDoolittle
(Arg)	Polar	Positiva	-4,5
(Lys)	Polar	Positiva	-3,9
(Asp)	Polar	Negativa	-3,5
(Glu)	Polar	Negativa	-3,5
(Asn)	Polar	Neutra	-3,5
(Gln)	Polar	Neutra	-3,5
(His)	Polar	Neutra / positiva	-3,2
(Pro)	Apolar	Neutra	-1,6
(Tyr)	Polar	Neutra	-1,3
(Trp)	Apolar	Neutra	-0,9
(Ser)	Polar	Neutra	-0,8
(Thr)	Polar	Neutra	-0,7
(Gly)	Apolar	Neutra	-0,4
(Ala)	Apolar	Neutra	1,8
(Met)	Apolar	Neutra	1,9
(Cys)	Apolar	Neutra	2,5
(Phe)	Apolar	Neutra	2,8
(Leu)	Apolar	Neutra	3,8
(Val)	Apolar	Neutra	4,2
(Ile)	Apolar	Neutra	4,5

Tabela 8. Escala de hidrofobicidade de Kyte Doolittle. Fonte: (KYTE e DOOLITTLE, 1982)

Aminoácido	Polaridade da cadeia lateral	Carga da cadeia lateral pH 7,4	Constante de hidrofobicidade Radzicka
(Arg)	Polar	Positiva	-14,92
(Asp)	Polar	Negativa	-8,72
(Glu)	Polar	Negativa	-6,81
(Asn)	Polar	Neutra	-6,64
(Lys)	Polar	Positiva	-5,55
(Gln)	Polar	Neutra	-5,54
(His)	Polar	Neutra / positiva	-4,66
(Ser)	Polar	Neutra	-3,4
(Thr)	Polar	Neutra	-2,57
(Tyr)	Polar	Neutra	-0,14
(Gly)	Apolar	Neutra	0,94
(Cys)	Apolar	Neutra	1,28
(Ala)	Apolar	Neutra	1,81
(Trp)	Apolar	Neutra	2,33
(Met)	Apolar	Neutra	2,35
(Phe)	Apolar	Neutra	2,98
(Pro)	Apolar	Neutra	3,5
(Val)	Apolar	Neutra	4,04
(Ile)	Apolar	Neutra	4,92
(Leu)	Apolar	Neutra	4,92

Tabela 9. Escala de hidrofobicidade de Radzicka. Fonte: (RADZICKA e WOLFENDEN, 1988)

3.4.7 Contatos

As interações entre os aminoácidos de uma proteína são indispensáveis para manter sua estabilidade. O Blue Star Sting calcula os contatos entre os átomos dos resíduos de aminoácidos com base em suas distâncias relativas. O método usado na identificação dos contatos internos e externos a uma cadeia de aminoácidos consiste em: a) classificar os átomos em grupos, de acordo com seu comportamento eletrostático e posição na cadeia lateral e/ou principal do aminoácido; e, b) os átomos são selecionados com base no tipo de contato que eles podem fazer e nas restrições de distância definidas experimentalmente para cada tipo de contato. A Tabela 10 apresenta os 5 tipos de contato existentes no STING_RDB e suas respectivas energias (MANCINI, 2004).

Tipo de contato	Energia de contato (kcal/mol)
Interações hidrofóbicas	0,6
Aromático	1,5
Ligações de hidrogênio ^(*)	2,6
Pontes salinas ^(†)	10,0
Ligações de dissulfeto	85,0

Tabela 10. Tipos de contato e suas energias de ligação. ^(*) Existem 9 tipos de ligações de hidrogênio: cadeia principal-cadeia principal, cadeia principal-(1 H₂O)-cadeia principal, cadeia principal-(2 H₂O)-cadeia principal, cadeia principal-cadeia lateral, cadeia principal-(1 H₂O)-cadeia lateral, cadeia principal-(2 H₂O)-cadeia lateral, cadeia lateral-cadeia lateral, cadeia lateral-(1 H₂O)-cadeia lateral, cadeia lateral-(2 H₂O)-cadeia lateral. ^(†) As pontes salinas podem ser dos tipos atrativo ou repulsivo. Fonte: http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/energy_contacts_table.html (MANCINI, 2004).

3.4.8 Ordem de Cross Link

Resíduos de aminoácidos distantes entre si na estrutura primária (sequência de aminoácidos) podem fazer contatos entre si na estrutura tridimensional da proteína. O Blue Star Sting denomina essa característica de *cross link*. Ela é calculada da seguinte forma: resíduos de aminoácidos distantes entre si 15, 20 ou 30 posições na estrutura primária, e que na estrutura tridimensional estejam dentro da mesma sonda esférica com raio de 3,5 Å, 5 Å e 8,5 Å podem fazer contatos entre si. A “ordem” é a quantidade de contatos exercidos por um determinado resíduo (Fig. 40.A).

3.4.9 Ordem de Cross Presence

Semelhante ao parâmetro *cross link*, *cross presence* conta todos os resíduos de aminoácidos que estejam dentro da sonda esférica de raio 3,5 Å, 5 Å e 8,5 Å mesmo que esses resíduos de aminoácidos não façam contato entre si. A “ordem” é a quantidade de resíduos de aminoácidos dentro dessa esfera (Fig. 40.B).

Se dois resíduos de aminoácidos fazem contato (*cross link*) ou apenas estão presentes dentro da sonda esférica centrada no resíduo de aminoácido base, e eles não estão distanciados em pelo menos 5 posições, a ordem não é incrementada.

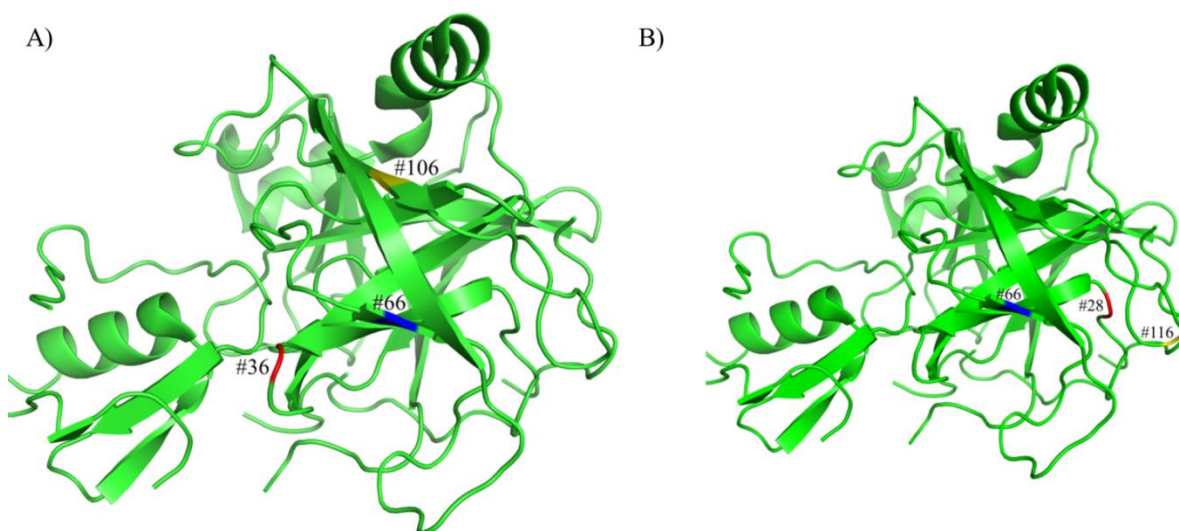


Figura 40. Exemplo de (A) *Cross Link*: 1cho.pdb, cadeia F. O resíduo 36 (vermelho) faz contato com os resíduos de aminoácidos 66 (azul) e 106 (amarelo). Embora distantes na sequência primária, eles estão dentro da sonda esférica de 3,5 Å e fazem contato entre si. (B) *Cross Presence*: 1cho.pdb, cadeia F. Embora não façam contato entre si, os resíduos de aminoácidos 28 (vermelho), 66 (azul) e 116 (verde) estão dentro da sonda esférica, mesmo que distantes na sequência primária. Imagens produzidas pelo software PyMOL.

3.4.10 Ângulos diedrais

Os ângulos PHI (φ) e PSI (ψ) são ângulos assumidos, por esta razão apresentam preferências de configuração de cadeia principal devido a eliminação de possíveis choques estéricos entre os resíduos de aminoácidos. O Blue Star Sting calcula esses ângulos através do método de Ramachandran (RAMACHANDRAN, RAMAKRISHNAN e SASISEKHARAN, 1963) usando a Eq. 12.

$$\varphi = tg^{-1}[2(v,u)] \times 57,29578 \quad (12)$$

onde: $tg^{-1}[2(v,u)]$ é a função que retorna o arco-tangente de v/u em radianos; v e u são autovetores com o menor autovalor correspondente; e 57,29578 é a constante utilizada para converter radianos em graus.

3.4.11 Rotâmeros

Enquanto os átomos dos resíduos de aminoácidos da cadeia principal formam os ângulos PHI (φ) e PSI (ψ), os átomos da cadeia lateral podem formar até cinco ângulos (χ -1, χ -2, χ -3, χ -4, χ -5) dependendo do tamanho da cadeia lateral. Entretanto, como a cadeia lateral dos resíduos de aminoácidos possui tamanhos diferentes, nem todos os aminoácidos possuem a mesma quantidade de ângulos χ . Por exemplo, o ângulo χ -5 é calculado somente para os resíduos de arginina, porque ela tem seis átomos na sua cadeia lateral.

Assim como os ângulos diedrais PHI (φ) e PSI (ψ) apresentam configuração estatisticamente mais aceitável devido a possíveis choques estéricos entre os resíduos de aminoácidos, os ângulos da cadeia lateral também apresentam valores estatisticamente mais aceitáveis (SCHRAUBE, EISENHABER e ARGOS, 1993).

3.4.12 Contatos não usados

O parâmetro *contatos não usados* é calculado subtraindo, para cada resíduo de aminoácido, o número máximo de contatos de um determinado tipo encontrado no inteiro STING_RDB, dos contatos realizados por esse resíduo. Esse parâmetro estabelece um possível potencial de contatos não usados, definindo quais os tipos de contatos os átomos presentes em determinada estrutura proteica não realizam, quando poderiam fazê-lo.

3.4.13 Fator de temperatura

O fator de temperatura é um parâmetro extraído diretamente do arquivo PDB, que indica o nível de mobilidade dos átomos dentro do cristal. Fator de temperatura com valor menor que 30 Å² significa que os átomos não se movem muito e está na mesma posição das

demais moléculas do cristal. Valores maiores que 60 \AA^2 indicam que o átomo se move tanto que mal pode ser visto através da difração de raios X (PROXYCHEM). Este é frequentemente o caso de átomos na superfície das proteínas onde as longas cadeias laterais presentes nos loops estão mais livres para oscilar no espaço disponível.

3.4.14 Parâmetros ponderados pela vizinhança (WNA)

Parâmetros ponderados pela vizinhança, do inglês *Weighted Neighbour Averages* (WNA), é o nome dado ao descritor D ponderado pelos valores dos seus vizinhos, cujo valor resultante é associado ao resíduo de aminoácido de interesse, segundo as Eq. 13 e 14.

$$D_{WNA}^{Surf} = \sum_{i=0}^n D_i Ac_{relativa} \quad (13)$$

$$D_{WNA}^{Dist} = D_0 + \sum_{i=1}^n \frac{D_i}{d_i} \quad (14)$$

onde: D_i são os valores do mesmo descritor para os n vizinhos espaciais; $Ac_{relativa}$ é a área relativa acessível ao solvente; D_0 é o valor do descritor D para o resíduo de aminoácido de interesse; e d_i é a distância do i -ésimo vizinho ao resíduo de aminoácido de interesse. A Eq. 13 indica que os resíduos de aminoácidos com maior porcentagem de área acessível tem influência maior sobre o resíduo de aminoácido de interesse, e a Eq. 14 diz que os resíduos de aminoácidos mais distantes espacialmente influenciam menos o resíduo de interesse, independente da acessibilidade de cada aminoácido.

O algoritmo que calcula os valores de WNA usa a definição de Porollo e Meller (POROLLO e MELLER, 2007) ao estabelecer uma sonda esférica de raio 15 \AA , centrada no carbono- α de resíduo de aminoácido de cada estrutura armazenada no STING_RDB . Segundo Silveira (SILVEIRA, PIRES, *et al.*, 2009), após 15 \AA a influência dos vizinhos praticamente cai para zero.

3.5 Extração, preparação e apresentação dos dados

Após alinharmos os EES por tamanho, pelo C-Terminal e N-Terminal, nós selecionamos um descritor e extraímos do *Datamart* a média, o desvio padrão e calculamos o erro quadrático da média para cada posição no nano-ambiente (32 resíduos de aminoácidos antes do EES, os resíduos durante o EES e 32 resíduos de aminoácidos após o EES). A

estatística descritos abaixo, usando o RStudio³², um amigável ambiente de programação em R³³ para Windows.

3.6.1 Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov é baseado na diferença entre a função de distribuição cumulativa $F_0(x)$ e a função de distribuição empírica da amostra $S_n(x)$. A função de distribuição cumulativa $F_0(x)$ é dada pela Eq. 15:

$$F_0(x) = \frac{1}{n} \sum_{i=1}^n I_{[-\infty, x]}(X_i) \quad (15)$$

onde $I_{[-\infty, x]}(X_i)$ é a função indicador, sendo $\begin{cases} 1, & X_i \leq x \\ 0, & X_i > x \end{cases}$

A função de distribuição empírica da amostra é definida como a proporção das observações da amostra que são menores ou iguais a x para todos os valores reais x . Pelo teorema de Glivenko-Cantelli³⁴ pode-se afirmar que $S_n(x)$ aproxima-se da distribuição teórica. Portanto, quanto maior o valor de n , o desvio entre as duas distribuições $|S_n(x) - F_x(x)|$ se torna cada vez menor para todos os valores de x , de modo que a estatística de Kolmogorov-Smirnov é calculada pela Eq. 16:

$$D_n = \sup |F_n(x) - F(x)| \quad (16)$$

A probabilidade D_n não depende de $F_n(x)$ desde que F_n seja contínua.

Em R usamos a seguinte função para o teste de Kolmogorov-Smirnov³⁵:

```
ks.test(x, y, alternative = c("two.sided", "less", "greater"))
```

³² <https://www.rstudio.com/>

³³ <https://www.r-project.org/>

³⁴ Teorema de Glivenko-Cantelli: $S_n(x)$ converge uniformemente para $F_x(x)$ com a probabilidade 1, que é

$$P \left[\lim_{n \rightarrow \infty} \sup_{-\infty < x < \infty} |S_n(x) - F_x(x)| = 0 \right] = 1$$

³⁵ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ks.test.html>

onde:

x	um vetor numérico de valores de dados
y	um vetor numérico de valores de dados ou uma cadeia de caracteres nomeando uma função de distribuição cumulativa ou uma função de distribuição cumulativa real, como pnorm. Somente CDFs contínuos são válidos.
alternative	os valores possíveis "two.sided", "less" e "greater" de alternativa especificam a hipótese nula de que a função de distribuição real de x é igual a, não menor que ou não maior que a função de distribuição hipotética (caso de uma amostra) ou a função de distribuição de y (caso de duas amostras), respectivamente.

3.6.2 Teste t de Student

O teste t de Student, ou simplesmente teste t, trabalha com duas hipóteses: a nula (H_0) e a alternativa (H_1). Usando a Eq. 17 calcula-se o valor de t e esse valor é aplicado à função densidade de probabilidade da distribuição t de Student, medindo o tamanho da área abaixo dessa função para valores maiores ou iguais a t. Essa área representa a probabilidade da média amostral apresentar os valores observados ou algo mais extremo. Essa probabilidade é dada pelo p-value. Por exemplo, se o p-value usado for igual a 0,05 (5%) e a área abaixo da função densidade de probabilidade da distribuição t de Student for menor do que 5% rejeitamos H_0 com um nível de confiança de 95%.

$$t = \frac{\bar{x} - \mu_0}{\left(\frac{S}{\sqrt{n}}\right)} \quad (17)$$

onde \bar{x} é a média da amostra, μ_0 é um valor fixo usado para comparação com a média da amostra, S é o desvio padrão da amostra e n é o tamanho da amostra. Quanto maior o valor t, podemos afirmar com mais confiança que $\bar{x} \leq \mu_0$ não é verdade, e, portanto, rejeitamos H_0 .

Em R usamos a seguinte função para o teste t de Student³⁶:

```
t.test(x, y, alternative = c("two.sided", "less", "greater"))
```

onde:

x	um vetor numérico (não vazio) de valores de dados.
y	um vetor numérico opcional (não vazio) de valores de dados.
alternative	"greater" é a alternativa que x tem uma média maior que y.

³⁶ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/t.test.html>

3.6.3 Teste de normalidade de Shapiro

O teste de normalidade de Shapiro é usado para determinar se uma população tem distribuição normal (SHAPIRO e WILK, 1965). Uma condição indispensável para o teste MANOVA é a normalidade dos dados. Neste caso, antes da análise multivariada, aplicamos o teste de normalidade de Shapiro, para garantir que apenas os dados que atendam o requisito sejam testados.

O teste segue os seguintes passos:

- Formulação da hipótese:
 - ✓ H_0 : A amostra provém de uma população com distribuição normal;
 - ✓ H_1 : A amostra não provém de uma população com distribuição normal;
- Estabelecer o nível de significância do teste (α);
- Calcular a estatística de teste:
 - ✓ Ordenar as observações da amostra: $x(1), x(2), x(3), \dots, x(n)$;
 - ✓ Calcular $\sum_{i=1}^n (x_i - \bar{x})^2$;
 - ✓ Calcular b ;
 - ✓ Calcular W ;
- Tomar a decisão de rejeitar H_0 ao nível de significância α se $W_{\text{calculado}} < W_{\alpha}$.

onde b e W são dados pelas Eq. 18 e 19, x_i são os valores da amostra ordenados, e a_{n-i+1} são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição normal.

$$b = \begin{cases} \sum_{i=1}^{n/2} a_{n-i+1} \times (x_{(n-i+1)} - x_{(i)}) & \text{se } n \text{ é par} \\ \sum_{i=1}^{(n+1)/2} a_{n-i+1} \times (x_{(n-i+1)} - x_{(i)}) & \text{se } n \text{ é ímpar} \end{cases} \quad (18)$$

$$W = \frac{b^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2} \quad (19)$$

Em R usamos a seguinte função para o teste de normalidade de Shapiro³⁷:

```
shapiro.test(x)
```

onde [x] é um vetor numérico de valores de dados. Valores ausentes são permitidos, mas o número de valores não ausentes deve estar entre 3 e 5000.

3.6.4 Teste de correlação linear

Em estatística, duas ou mais variáveis podem estar linearmente correlacionadas. Isso significa que a mudança em uma das variáveis provoca mudanças nas outras. A palavra “correlacionada” indica que essa relação acontece nos dois sentidos (co + relação), e a correlação linear, portanto, designa a força que mantém essas variáveis unidas.

Ao trabalharmos com análise multivariada, antes de executarmos o teste MANOVA, nós eliminamos as variáveis correlacionadas. Embora a presença delas aumente o universo amostral, elas introduzem um ruído branco na análise estatística, porque “puxam” os dados de maneira tendenciosa para um lado.

O coeficiente de correlação linear entre duas variáveis é dado pela Eq. 20.

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \quad (20)$$

onde S_x e S_y são os desvios padrões da amostra e S_{xy} é a covariância da amostra. Similarmente, o coeficiente de correlação para uma população é dado pela Eq. 21.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (21)$$

onde σ_{xy} é a covariância da população e σ_x e σ_y são os desvios padrões da população.

A interpretação é que, quanto mais próximos a 1 forem os coeficientes de correlação, as variáveis estão mais linearmente correlacionadas (BUSSAB e MORETTIN, 2010). Em R usamos a seguinte função para o teste de correlação linear³⁸:

³⁷ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html>

³⁸ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.html>

```
cor(x, y = NULL, method = c("pearson", "kendall", "spearman"))
```

onde:

x um vetor numérico, matriz ou quadro de dados.
y NULL (padrão) ou um vetor, matriz ou quadro de dados com dimensões compatíveis para x. O padrão é equivalente a $y = x$ (mas mais eficiente).
method Indica qual coeficiente de correlação (ou covariância) deve ser calculado: "pearson" (padrão), "kendall" ou "spearman".

3.6.5 Normalização dos dados

Quando trabalhamos com dados de naturezas diferentes, eles precisam ser normalizados para minimizar os problemas naturais do uso de unidades e dispersões distintas entre si. Fazendo isso, conseguimos colocar dados dentro de um intervalo numérico, por exemplo, $[0, 1]$. Quando conhecemos os valores máximo e mínimo, podemos normalizar os dados usando o método linear (Eq. 22).

$$f(x) = \frac{X - \min}{\max - \min} \quad (22)$$

Ou podemos normalizar pelo valor máximo (Eq. 23).

$$f(x) = \frac{X}{\max} \quad (23)$$

Quando não temos os valores de máximo e mínimo, podemos normalizar os dados usando o coeficiente de variância (CV), que é a relação entre o desvio padrão e a média de uma população (Eq. 24).

$$CV = \frac{\sigma}{\mu} \quad (24)$$

Também é possível trabalhar com o inverso do coeficiente de variação (ICV) (MAZONI, 2018), que é a relação entre a média e o desvio padrão (Eq. 25).

$$ICV = \frac{\mu}{\sigma} \quad (25)$$

3.6.6 Análise multivariada (MANOVA)

Quando temos duas ou mais variáveis que precisam ser analisadas em conjunto, usamos a análise multivariada (MANOVA) (HAIR, 1998). A MANOVA usa a covariância entre as variáveis dependentes para testar a significância estatística das diferenças entre as médias de cada variável. A MANOVA aplica quatro diferentes testes no conjunto de dados: Pillai (Eq. 26), Wilks (Eq. 27), Lawley-Hotelling (Eq. 28) e Roy (Eq. 29).

$$\Lambda_{Pillai} = \sum_{1 \dots p} \left(\frac{\lambda_p}{1 + \lambda_p} \right) \quad (26)$$

$$\Lambda_{Wilks} = \prod_{1 \dots p} \left(\frac{1}{1 + \lambda_p} \right) \quad (27)$$

$$\Lambda_{LH} = \sum_{1 \dots p} (\lambda_p) \quad (28)$$

$$\Lambda_{Roy} = \max_p (\lambda_p) \quad (29)$$

onde λ_p é a proporção da variância entre as variáveis dependentes.

Em R usamos a seguinte função para o teste MANOVA³⁹:

```
summary(object, test = c("Pillai", "Wilks", "Hotelling-Lawley", "Roy"), tol = 1e-7)
```

onde:

object um objeto da classe "manova" com múltiplas respostas.
test nome da estatística de teste a ser usada, pode ser "Pillai", "Wilks", "Hotelling-Lawley", "Roy".
tol tolerância tol é aplicada à decomposição QR da matriz de correlação residual (a menos que alguma resposta tenha essencialmente residuais nulos, quando está sem escala). Assim, o valor padrão protege contra respostas altamente correlacionadas: ele pode ser reduzido, mas isso permitirá resultados bastante imprecisos e normalmente será melhor transformar as respostas para remover a alta correlação.

³⁹ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/summary.manova.html>

3.6.7 p-value

Dada duas hipóteses, a hipótese nula (H_0) e a hipótese alternativa (H_1), um nível de significância (α) é estabelecido, e se o p-value for menor que α então rejeitamos H_0 e aceitamos H_1 como verdadeira (WASSERSTEIN e LAZAR, 2016). Neste trabalho, H_1 é a hipótese que afirma existir um sinal que distingue o EES dentro do seu nano-ambiente. Quando testamos os dados obtidos dentro do EES contra os dados obtidos fora dele, se p-value $< \alpha$ significa que, estatisticamente, o EES se diferencia do restante do seu nano-ambiente.

Em R a sintaxe para obtermos o valor do p-value é a seguinte:

```
var$p.value
```

onde [var] é o nome da variável que recebeu o resultado do teste estatístico aplicado.

3.6.8 Sliding Window

O teste de Sliding Window (“janela deslizante”) consiste em criar uma janela e separar a população em duas amostras: dentro e fora da janela. A cada rodada do teste, movemos a janela em uma posição, mudando o conjunto de dados de cada amostra (Fig. 42-A). Para cada conjunto de amostras aplicamos um teste estatístico. Esse método de análise é usado para demonstrar como os resultados dos testes estatísticos variam ao longo do tempo (Fig. 42-B).

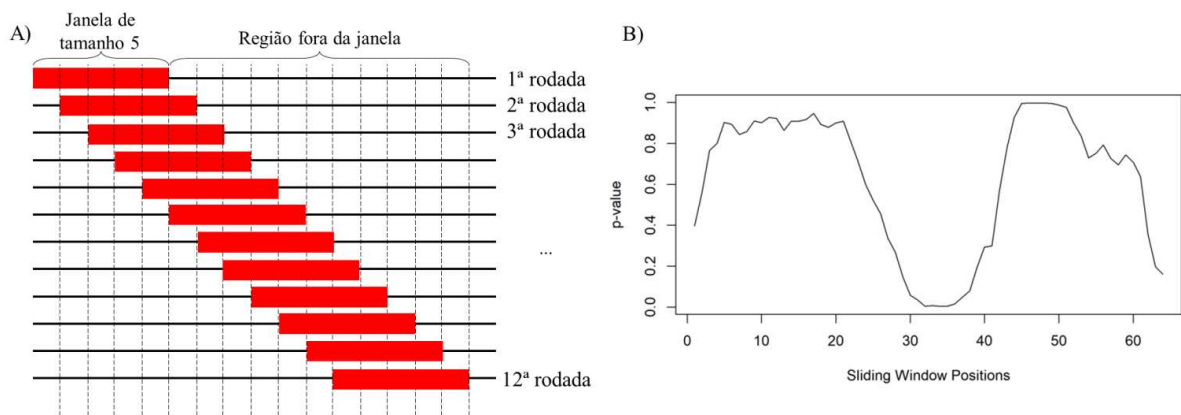


Figura 42. Teste de Sliding Window. Na figura A) vemos como a posição da janela muda ao longo do tempo. A cada rodada do teste, a janela desliza uma posição da esquerda para a direita, até cobrir toda a extensão dos dados. Na figura B) vemos como determinado teste estatístico se comporta ao longo da extensão dos dados. No exemplo, fica claro que a região central do gráfico se distingue em relação às outras regiões.

Essa metodologia pode ser aplicada nos testes descritos acima, usando um simples script escrito em R que varie a posição da janela e a faça deslizar ao longo do tempo.

4 RESULTADOS E DISCUSSÃO

4.1 Informações quantitativas dos *Datamarts*

Os algoritmos DSSP e Stride classificam as estruturas helicoidais em hélice 3_{10} , α -hélice e π -hélice. Embora o STING_RDB, de onde os *Datamarts* foram extraídos, não faça essa distinção, é importante mencionar que o PDB faz. A Tabela 11 introduz o número de estruturas helicoidais nas proteínas do tipo all- α e α em $(\alpha+\beta)+(\alpha/\beta)$.

	all- α	α em $(\alpha+\beta)+(\alpha/\beta)$
α -hélice	8366	104885
π -hélice	0	84
hélice 3_{10}	1075	32348
Total	9441	137241

Tabela 11. Número de estruturas helicoidais presentes no PDB. Para as estruturas do tipo all- α , temos 88,6% de α -hélices e 11,4% de hélice 3_{10} . Para as estruturas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$ temos 76,4% de α -hélice e 23,6% de hélice 3_{10} .

As estruturas proteicas foram agrupadas de acordo com a presença do EES em all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$ e β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenadas”, usando as definições do PDB, DSSP e Stride. A Tabela 12 apresenta a quantidade de cadeias proteicas em cada agrupamento.

Tipo de estrutura proteica	Número de cadeias		
	PDB	DSSP	Stride
all- α	25933	10317	24158
all- β	4553	1882	4237
α em $(\alpha+\beta)+(\alpha/\beta)$	279805	114656	274498
β em $(\alpha+\beta)+(\alpha/\beta)$	251275	103288	248273
“desordenada”	206	165140	369

Tabela 12. Número de cadeias para as proteínas do tipo all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$, β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada”, segundo as definições do PDB, DSSP e Stride, armazenados no STING_RDB em 16 de maio de 2018.

Concluimos a partir desses dados que o algoritmo DSSP é o mais restritivo em relação ao algoritmo Stride nas definições de α -hélice e folha- β . Por esse motivo, o número de estruturas definidas como “desordenadas” por algoritmo DSSP é maior. Isso pode ser observado na Fig. 43, que mostra essa distribuição para cada algoritmo. Uma explicação do por que o algoritmo DSSP é mais restritivo é que ele usa apenas o reconhecimento dos padrões de ligações das ligações de hidrogênio na determinação do EES, o que pode restringir os resultados. O algoritmo Stride se torna mais abrangente ao usar não apenas o reconhecimento dos padrões das ligações de hidrogênio, como também a informação sobre os ângulos diedrais.

Para melhorar a confiabilidade da existência dos EES ao longo da estrutura proteica, combinamos os algoritmos PDB, DSSP e Stride entre si (veja Fig. 32). Usando o consenso entre os algoritmos, mostramos na Tabela 13 os números de cadeias proteicas para cada tipo estrutural. Em linguagem matemática, apresentamos os quatro consensos possíveis em um diagrama de Venn (RUSKEY e WESTON, 1997) para cada agrupamento estrutural na Fig. 44.

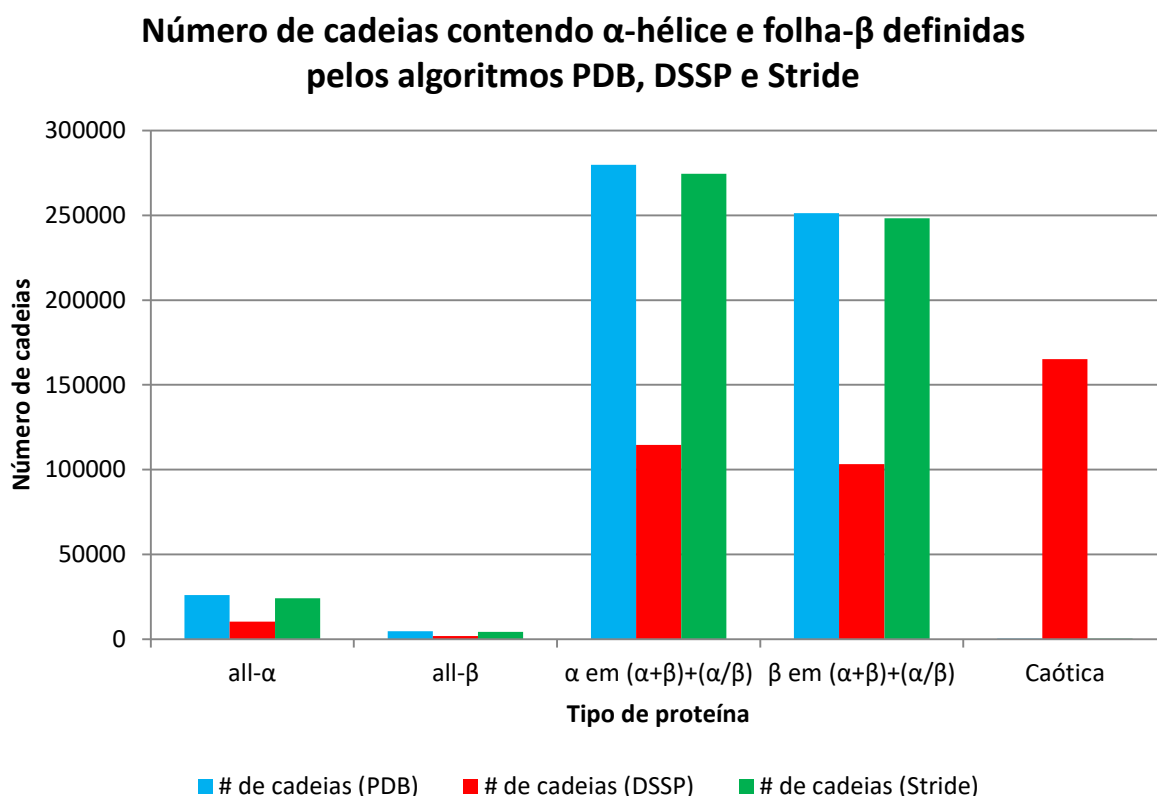


Figura 43. Número de cadeias proteicas no diferentes tipos de estruturas: all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$, β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada”. O gráfico indica que o algoritmo DSSP é o mais restritivo nas definições das α -hélices e folhas- β , e consequentemente, é o algoritmo com o maior número de estruturas “desordenadas”.

	PDB-DSSP-Stride	PDB-DSSP	PDB-Stride	DSSP-Stride
all- α	991	1498	8555	7937
all- β	1144	1394	2659	1248
α em $(\alpha+\beta)+(\alpha/\beta)$	8942	11453	109985	86869
β em $(\alpha+\beta)+(\alpha/\beta)$	85524	88856	211862	88008
“desordenada”	1386	1466	1386	1403

Tabela 13. Número de cadeias com α -hélices e folhas- β dos tipos all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$, β em $(\alpha+\beta)+(\alpha/\beta)$ e também de estruturas “desordenada”, usando o consenso entre os algoritmos PDB, DSSP e Stride. Novamente observamos que o algoritmo DSSP é o mais restritivo ao definir a presença da α -hélice ou folha- β . Por esse motivo, o número de cadeias é sempre maior no consenso entre os algoritmos PDB e Stride.

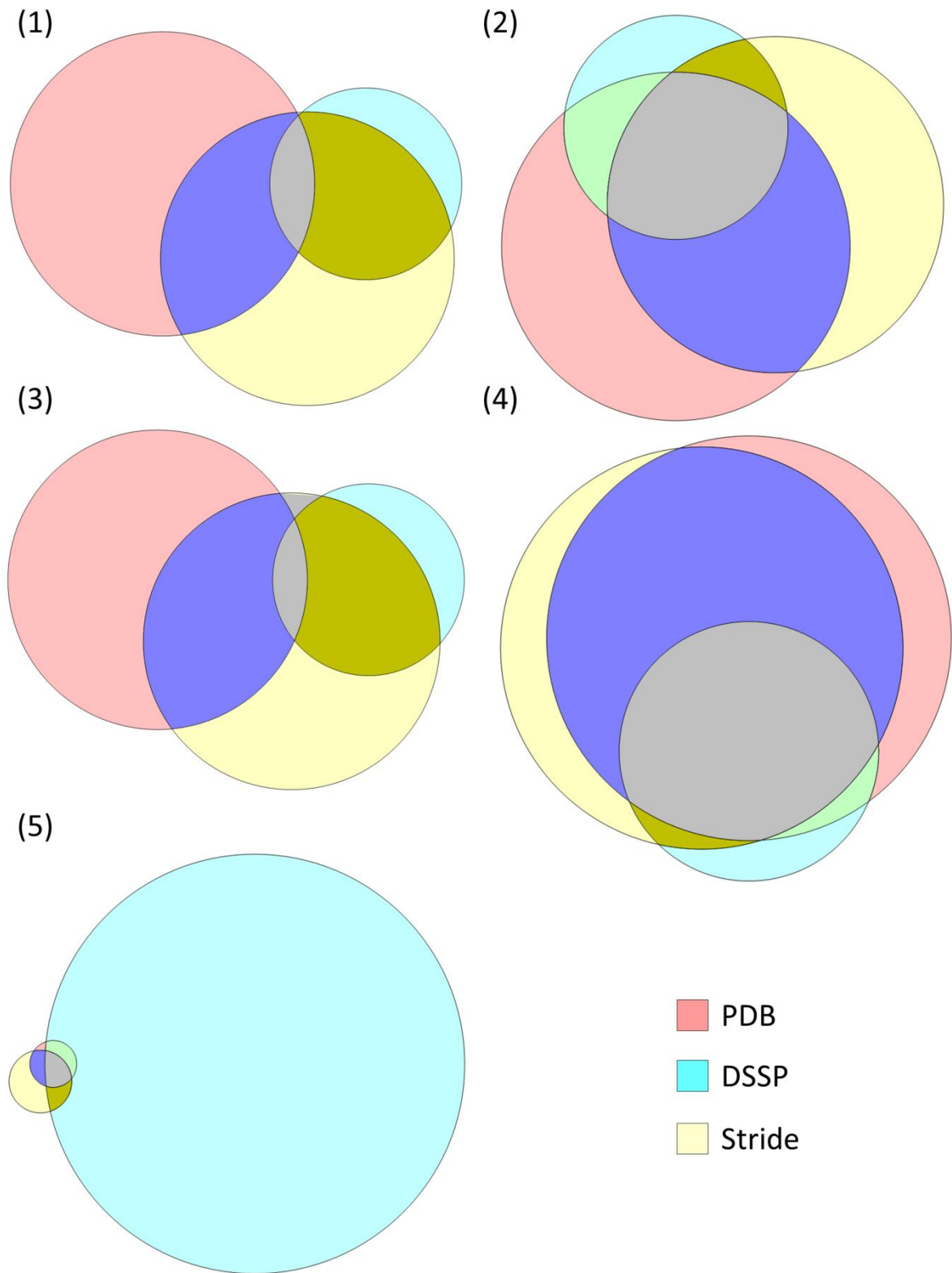


Figura 44. Diagramas de Venn para os consensos entre as definições de α -hélice, folha- β do PDB, DSSP e Stride. Em: (1) all- α , (2) all- β , (3) α em $(\alpha+\beta)+(\alpha/\beta)$ e (4) β em $(\alpha+\beta)+(\alpha/\beta)$ fica claro que o algoritmo DSSP é mais restritivo, e por isso, o círculo em vermelho é menor. Como a estrutura “desordenada” por definição não apresenta α -hélice nem folha- β , o diagrama em: (5) tem o maior número desse tipo de estrutura na definição do DSSP. Observamos ainda que nas estruturas “puras” – α -hélice e folha- β – a intersecção entre os três conjuntos, que representa o consenso mais restritivo, é menor que no caso das estruturas tipo $(\alpha+\beta)+(\alpha/\beta)$.

4.2 Eliminação da redundância

Eliminamos as estruturas redundantes nos níveis de 95%, 70% e 50%, usando o software CD-HIT (LI e GODZIK, 2006). A Tabela 14 e a Fig. 45 apresentam o número de cadeias das proteínas do tipo all- α que satisfazem os possíveis consensos após a eliminação da redundância.

all- α					
Redundância (%)		PDB-DSSP-Stride	PDB-DSSP	PDB-Stride	DSSP-Stride
	100%	991	1498	8555	7937
	95%	353	540	1739	2392
	70%	276	424	1378	1882
	50%	218	349	1153	1549

Tabela 14. Número de cadeias das proteínas do tipo all- α que contém pelo menos uma α -hélice que coincida nas definições dos algoritmos PDB, DSSP e Stride.

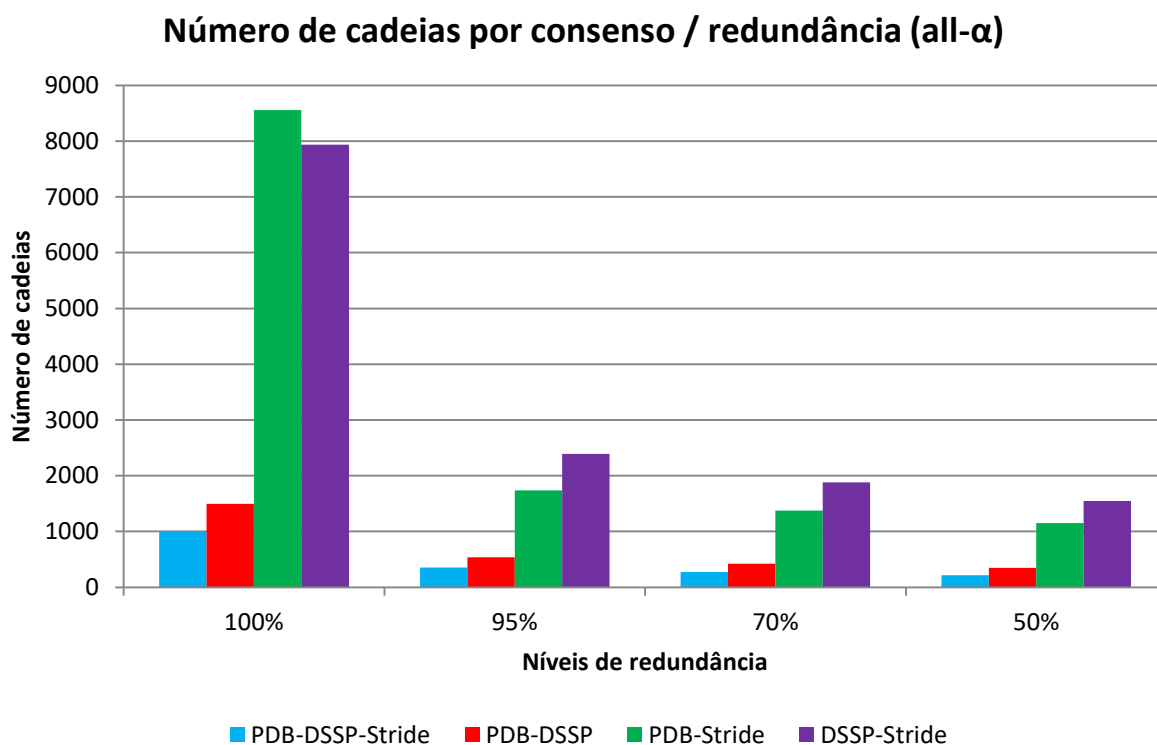


Figura 45. Número de cadeias das proteínas do tipo all- α que contém pelo menos uma α -hélice que coincida nas definições dos algoritmos PDB, DSSP e Stride. É interessante observar que para 100% – que significa que nenhuma estrutura foi eliminada – o consenso com maior número de cadeias é o PDB-Stride, justamente por serem menos restritivos que o algoritmo DSSP. Mas, a partir dos 95% de redundância, o consenso com maior número de cadeias é o DSSP-Stride.

Observamos no gráfico da Fig. 45 que o consenso entre os algoritmos DSSP e Stride é menos restritivo ao definir as α -hélices depois de retirada a redundância que os demais possíveis consensos.

A Tabela 15 e a Fig. 46 mostram o número de cadeias das proteínas do tipo all- β que satisfazem o consenso entre as definições de folha- β pelos algoritmos PDB, DSSP e Stride, após eliminada as redundâncias.

all- β					
Redundância (%)		PDB-DSSP-Stride	PDB-DSSP	PDB-Stride	DSSP-Stride
	100%	1144	1394	2659	1248
	95%	449	566	667	473
	70%	343	427	501	361
	50%	258	318	387	269

Tabela 15. Número de cadeias das proteínas do tipo all- β que contém pelo menos uma folha- β que coincida nas definições dos algoritmos PDB, DSSP e Stride.

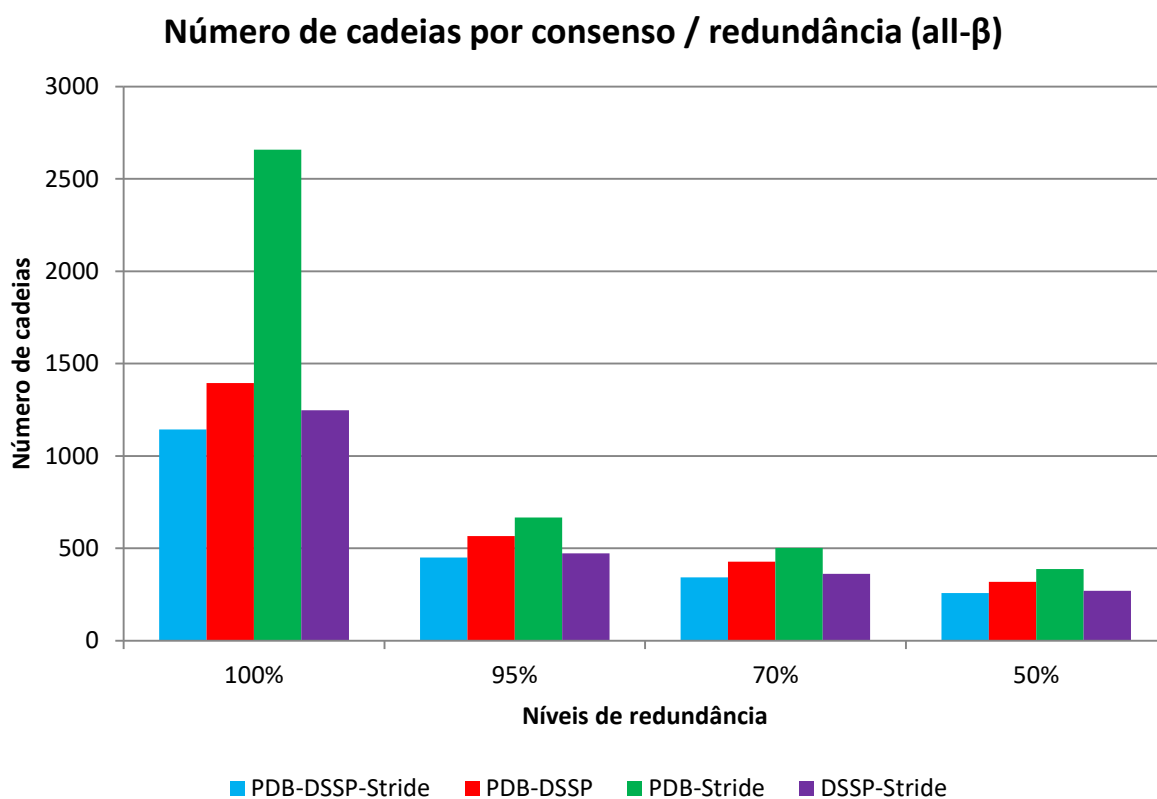


Figura 46. Número de cadeias das proteínas do tipo all- β que contém pelo menos uma folha- β que coincida nas definições dos algoritmos PDB, DSSP e Stride. Aqui o algoritmo DSSP é o mais restritivo em todos os casos, e o comportamento geral é o mesmo para todos os níveis de redundância. Por ordem de restrição, começando pelo mais restritivo consenso, temos PDB-DSSP-Stride, DSSP-Stride, PDB-DSSP e PDB-Stride.

Diferente do que foi observado na Fig. 45, onde o consenso entre DSSP-Stride é o menos restritivo para as proteínas do tipo all- α , aqui o consenso menos restritivo é aquele entre os algoritmos PDB-Stride. Porém, para as proteínas do tipo all- β , a diferença entre o número de cadeias com pelo menos uma folha- β em cada consenso é menor quando comparamos com as proteínas do tipo all- α .

A Tabela 16 e a Fig. 47 apresentam o número de cadeias das proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com pelo menos uma α -hélice que tenha consenso entre os algoritmos PDB, DSSP e Stride.

α em $(\alpha+\beta)+(\alpha/\beta)$					
Redundância (%)		PDB-DSSP-Stride	PDB-DSSP	PDB-Stride	DSSP-Stride
	100%	8942	11453	109985	86869
	95%	1909	2578	14797	17553
	70%	1493	2029	12307	14562
	50%	1162	1595	10279	12001

Tabela 16. Número de cadeias das proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com pelo menos uma α -hélice que coincida nas definições dos algoritmos PDB, DSSP e Stride.

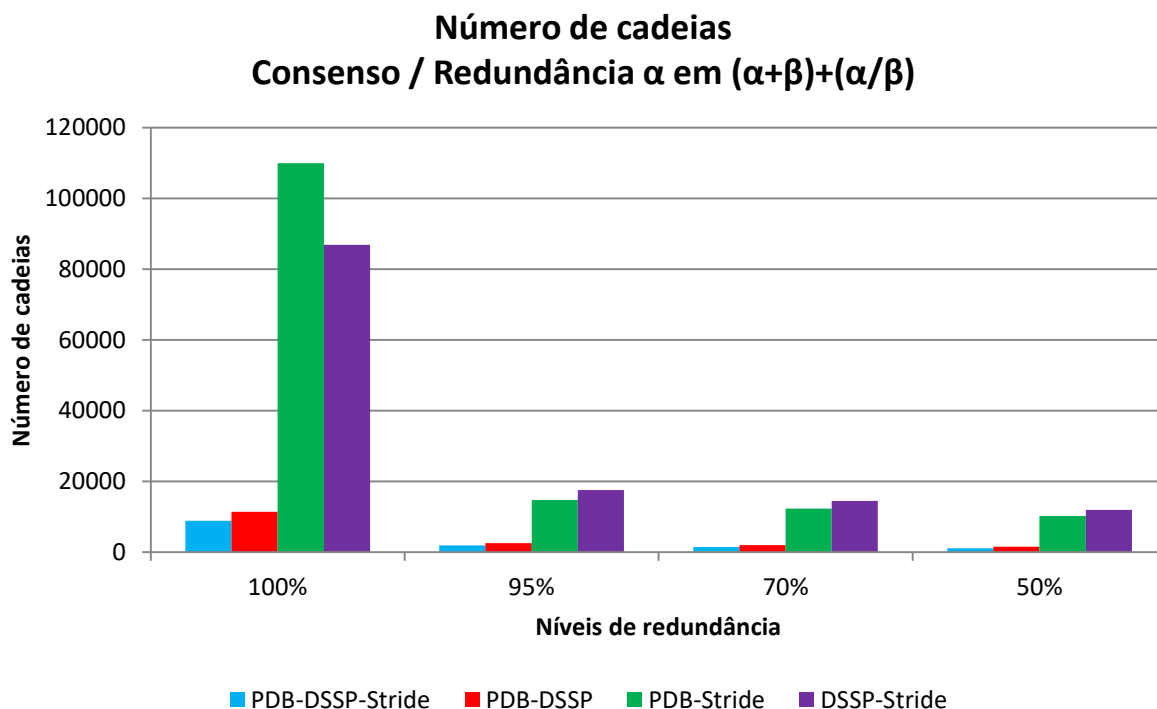


Figura 47. Número de cadeias das proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com pelo menos uma α -hélice que coincida nas definições dos algoritmos PDB, DSSP e Stride. O comportamento é similar às proteínas do tipo all- α : para 100% o consenso com maior número de cadeias é o PDB-Stride, justamente por serem menos restritivos que o algoritmo DSSP. Entretanto, a partir dos 95% de redundância, o consenso com maior número de cadeias é o DSSP-Stride.

Os gráficos da Fig. 47 são semelhantes aos gráficos da Fig. 45. Novamente temos uma restrição menor no consenso entre os algoritmos PDB-Stride quando não removemos a redundância. Mas, quando eliminamos a redundância, o consenso menos restritivo é o DSSP-Stride, similar ao que observamos para as proteínas do tipo all- α .

A Tabela 17 e a Fig. 48 apresentam o número de cadeias das proteínas do $(\alpha+\beta)+(\alpha/\beta)$ com pelo menos uma β -folha que tenha consenso entre os PDB, DSSP e Stride.

β em $(\alpha+\beta)+(\alpha/\beta)$					
Redundância (%)		PDB-DSSP-Stride	PDB-DSSP	PDB-Stride	DSSP-Stride
	100%	85524	88856	211862	88008
	95%	16897	17575	28058	17229
	70%	13764	14303	22499	13992
	50%	11317	11756	18552	11481

Tabela 17. Número de cadeias das proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ que contém pelo menos uma folha- β coincidindo nas definições dos algoritmos PDB, DSSP e Stride.

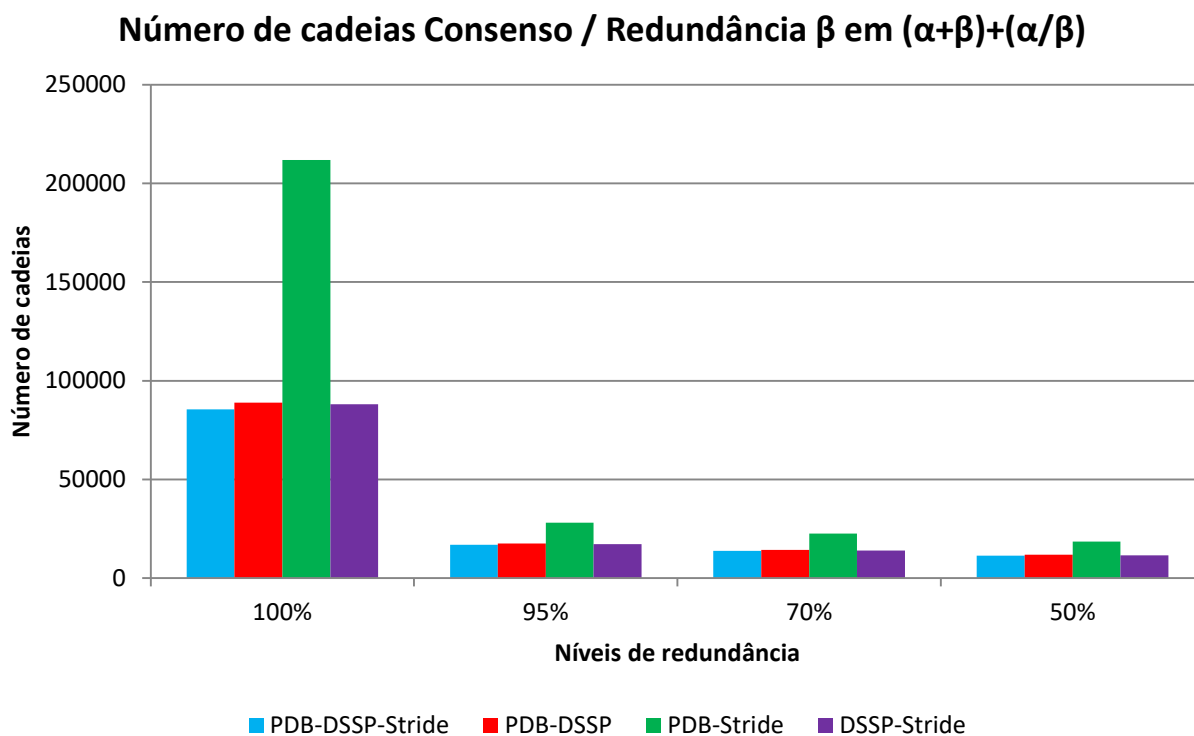


Figura 48. Número de cadeias das proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ que contém pelo menos uma folha- β coincidindo nas definições dos algoritmos PDB, DSSP e Stride. O algoritmo DSSP é o mais restritivo.

No caso das proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ o algoritmo DSSP é o mais restritivo ao definir a presença das folhas- β .

A Tabela 18 e a Fig. 49 mostram o número de cadeias das proteínas do tipo desordenado. Esse tipo de proteína não contém α -hélice nem folha- β na sua estrutura.

“desordenada”					
Redundância (%)		PDB-DSSP-Stride	PDB-DSSP	PDB-Stride	DSSP-Stride
	100%	1386	1466	1386	1403
	95%	102	115	102	108
	70%	64	76	64	70
	50%	48	59	48	52

Tabela 18. Número de cadeias das proteínas do tipo desordenado. Proteína desse tipo não contém α -hélice nem folha- β nas definições dos algoritmos PDB, DSSP e Stride, mas coincidem em número de resíduos de aminoácidos entre si.

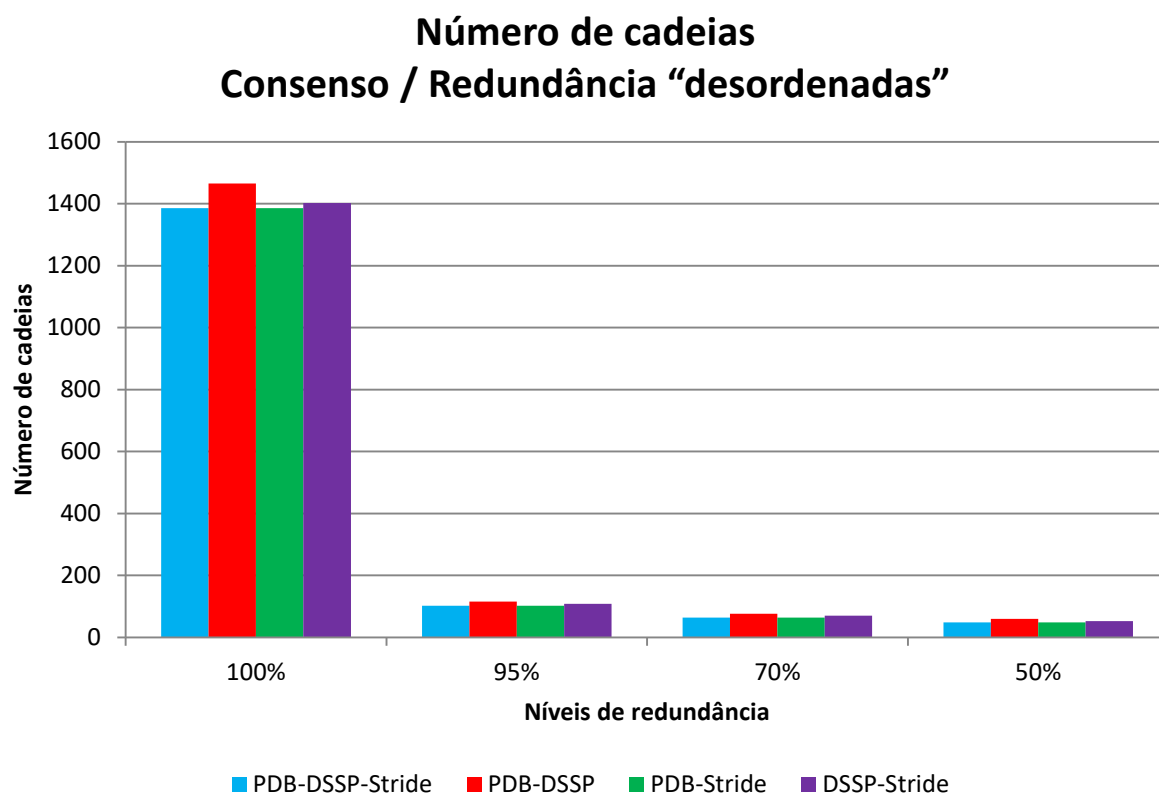


Figura 49. Número de cadeias das proteínas do tipo desordenado que não contém α -hélice nem folha- β nas definições dos algoritmos PDB, DSSP e Stride, mas coincidem em número de resíduos de aminoácidos entre si.

O alto número de cadeias antes de remover qualquer redundância pode ser explicado pelo alto número de mutantes nas proteínas do tipo desordenado e também pelo alto número de complexos para cada proteína. Por exemplo, a estrutura 1t9e.pdb (“sunflower Trypsin Inhibitor, Disulfide Mutant, Hydrolase-Hydrolase Inhibitor Complex”) e a estrutura 3j6u.pdb (“Dengue Virus, Human Antibody, Neutralization, Virus-Immune System

Complex”) são exemplos de estruturas mutantes e/ou com complexos que foram eliminados nos níveis de 70% e 50%.

4.3 Número de EES alinhados

Após removermos as estruturas redundantes, nós agrupamos as α -hélices e folhas- β por seu tamanho (medido em número de resíduo de aminoácidos), tipo (all- α , all- β , ($\alpha+\beta$)+(α/β) e desordenado) e consenso entre as definições do PDB, DSSP e Stride. Nos apêndices A, B e C apresentamos o número de cadeias para os demais consensos: PDB-DSSP, PDB-Stride e DSSP-Stride.

4.3.1 Número de cadeias nas proteínas do tipo all- α

A Tabela 19 e as Figs. 50 e 51 apresentam o número de cadeias proteicas para cada tamanho de α -hélice encontrada nas proteínas do tipo all- α com consenso entre as definições do PDB-DSSP-Stride, antes e depois de removida a redundância.

Tamanho da hélice	Redundância			
	100%	95%	70%	50%
5	82	29	24	22
6	185	54	41	31
7	99	31	25	19
8	137	39	33	26
9	62	27	23	20
10	100	39	29	20
11	120	46	33	28
12	86	30	23	20
13	128	42	32	25
14	96	28	26	18
15	132	38	29	20
16	50	17	14	10
17	32	13	10	8
18	51	17	11	8
19	53	14	11	8
20	20	6	5	4
21	24	13	10	8
22	24	12	12	12
23	12	4	3	2
24	18	11	7	6
25	41	5	4	4
26	4	2	2	2
27	14	5	5	4
28	20	6	5	2
29	10	4	1	1

30	3	1	0	0
31	10	2	1	0
32	18	2	2	2
33	12	3	2	2
34	2	0	0	0
35	4	2	2	2
36	2	1	1	1
37	2	1	0	0
38	1	0	0	0
39	1	0	0	0
40	2	0	0	0
42	1	1	1	1
43	1	1	1	1
44	1	0	0	0
45	2	0	0	0
48	1	1	1	1
50	2	1	1	1
51	7	0	0	0
55	1	1	1	1
62	2	1	1	1
67	1	1	1	1
108	1	1	1	1

Tabela 19. Número de α -hélices nas proteínas do tipo all- α com consenso entre os algoritmos PDB-DSSP-Stride.

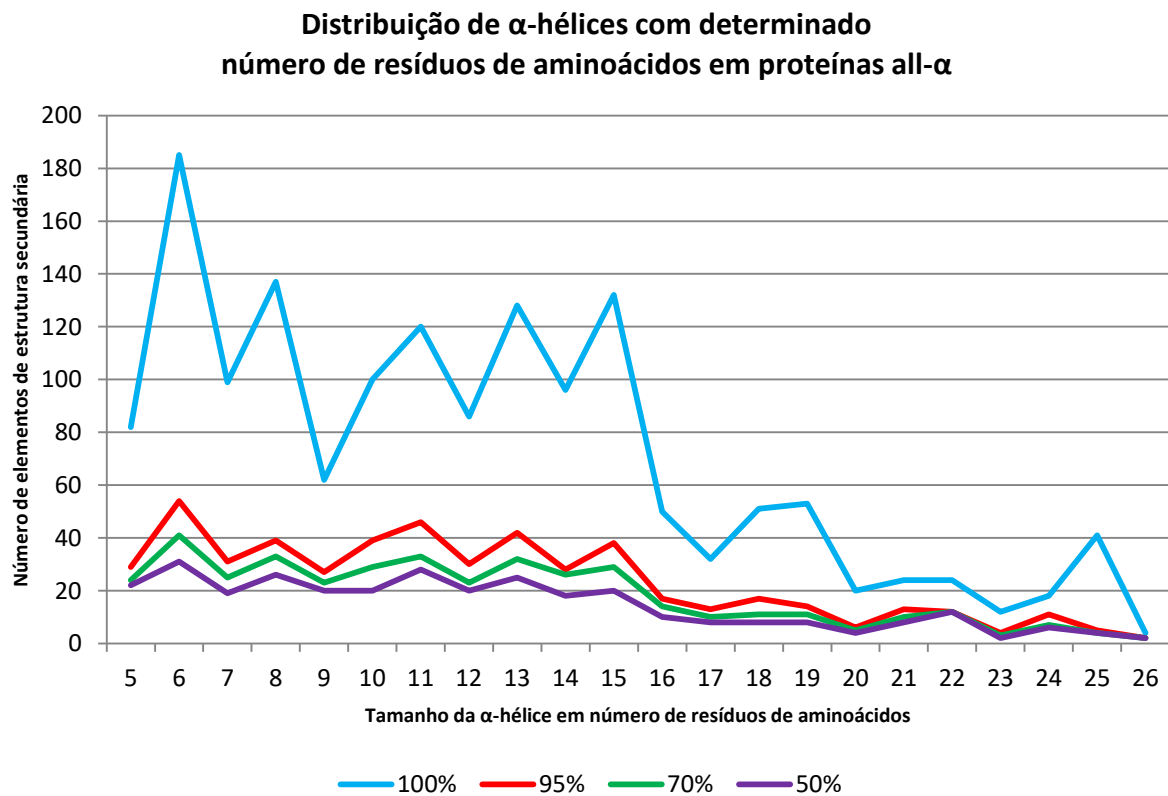


Figura 50. Número de α -hélices nas proteínas do tipo all- α com consenso entre PDB-DSSP-Stride agrupados por tamanho.

A maioria das α -hélices presentes nas proteínas do tipo all- α é formada por 6, 8 ou 15 resíduos de aminoácidos, quando olhamos para a curva de 100% de redundância (linha azul). A partir de 15 resíduos de aminoácidos, o número de α -hélices cai consideravelmente, chegando a apenas uma ocorrência de α -hélices formadas por 26 resíduos de aminoácidos. Entretanto, existem α -hélices com mais de 26 resíduos de aminoácidos, conforme demonstrado no gráfico da Fig. 51. A maior α -hélice encontrada está na estrutura 1sfc.pdb (“Neuronal Synaptic Fusion Complex”). Ela possui 62 resíduos de aminoácidos (Fig. 52).

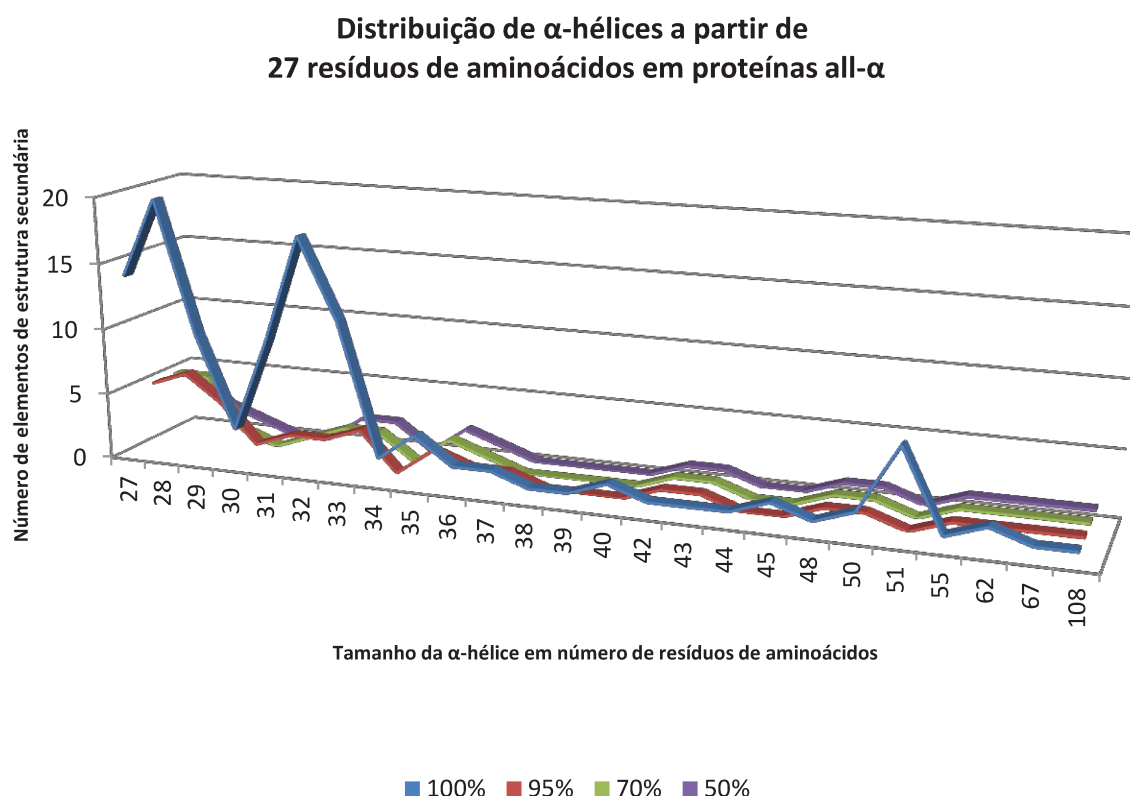


Figura 51. Número de α -hélices de tamanho ≥ 15 resíduos de aminoácidos presente nas proteínas do tipo all- α com consenso entre PDB-DSSP-Stride.

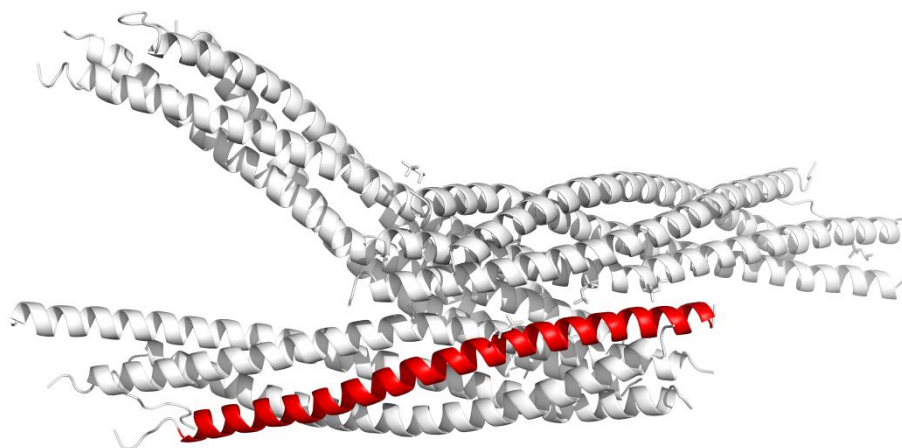


Figura 52. Exemplo de α -hélice longa. A estrutura 1sfc.pdb (“Neuronal Synaptic Fusion Complex”) tem uma α -hélice de 62 resíduos de aminoácidos em sua cadeia I (em vermelho). A imagem foi gerada usando o software PyMOL.

4.3.2 Número de cadeias nas proteínas do tipo all- β

A Tabela 20 e a Fig. 53 mostram o número de cadeias proteicas para cada tamanho de folhas- β encontrada nas proteínas do tipo all- β com consenso entre as definições do PDB-DSSP-Stride.

Tamanho da folha	Redundância			
	100%	95%	70%	50%
5	542	247	187	138
6	637	235	180	140
7	477	169	118	95
8	387	130	103	83
9	355	114	86	67
10	235	71	50	40
11	174	38	26	18
12	46	19	14	10
13	14	8	6	6
14	7	4	3	3
15	7	3	3	3
16	5	1	1	1
17	9	2	2	2
18	7	1	1	1
19	1	1	0	0
20	17	0	0	0
22	9	1	1	0

Tabela 20. Número de folhas- β nas proteínas do tipo all- β com consenso entre os algoritmos PDB-DSSP-Stride.

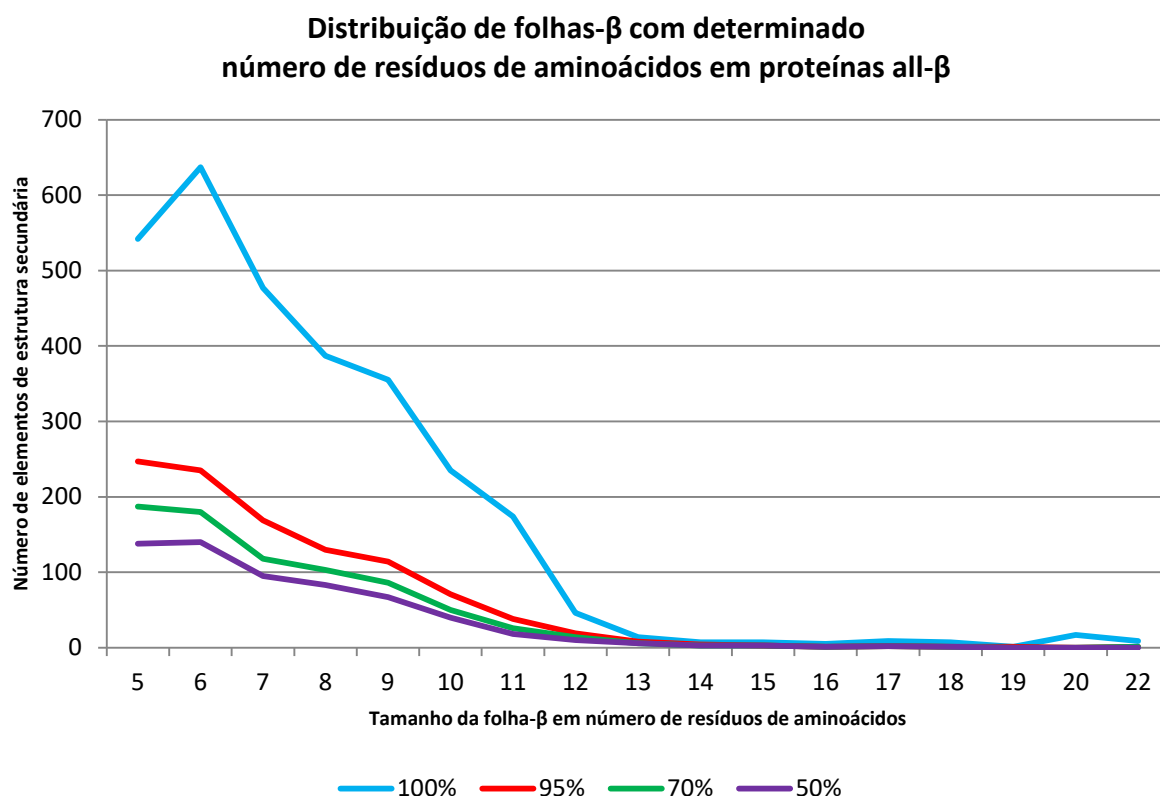


Figura 53. Número de folhas- β nas proteínas do tipo all- β com consenso entre os algoritmos PDB-DSSP-Stride.

A maioria das folhas- β são formadas por seis resíduos de aminoácidos. Sem eliminarmos a redundância (linha azul) existem 637 folhas- β de seis resíduos de aminoácidos. Isso representa 17,5% a mais que o número cadeias com folhas- β formadas por cinco resíduos, que é de 542 cadeias. Entretanto, quando removemos a redundância nos níveis de 95%, 70% e 50%, o número de cadeias com folhas- β formadas por cinco ou seis resíduos de aminoácidos se tornam mais próximos. Por exemplo, considerando 70% de redundância (linha verde do gráfico), temos 187 cadeias com folhas- β de cinco resíduos de aminoácidos, e 180 cadeias com folhas- β de seis resíduos de aminoácidos, uma diferença de 4%. É interessante notar que quando removemos a redundância nos níveis de 95% e 70%, o maior número de folhas- β é para o tamanho cinco. Ainda, da observação do gráfico, concluímos que são raras as folhas- β maiores que 13 resíduos de aminoácidos.

A folha- β mais longa encontrada é formada por 22 resíduos de aminoácidos. A Fig. 54 mostra como exemplo a estrutura com código no PDB 1gr3 (“structure Of The Human Collagen X Nc1 Trimer”).

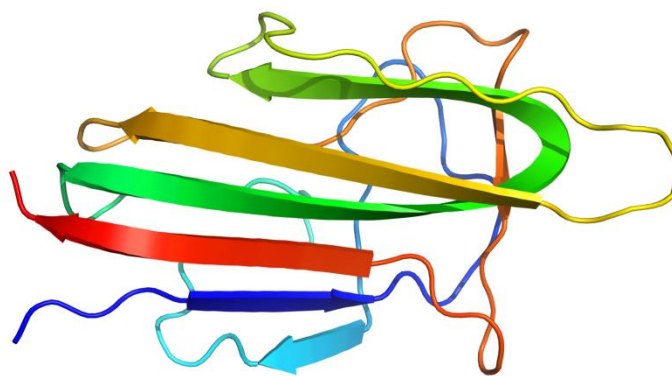


Figura 54. Exemplo de uma das mais longas folhas- β . A estrutura 1gr3.pdb ("structure Of The Human Collagen X Nc1 Trimer") tem uma folha- β de 22 resíduos de aminoácidos (em verde). A imagem foi gerada usando o software PyMOL.

4.3.3 Número de cadeias nas proteínas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$

A Tabela 21 e as Figs. 55 e 56 mostram o número de cadeias proteicas para cada tamanho de α -hélice encontrada nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre as definições do PDB-DSSP-Stride, antes e depois de removida a redundância.

Tamanho da hélice	Redundância			
	100%	95%	70%	50%
5	1503	271	212	171
6	1816	379	291	225
7	1581	319	278	207
8	1317	277	216	172
9	1269	248	189	147
10	1656	323	246	177
11	1764	361	269	202
12	1734	258	180	138
13	1261	206	167	123
14	1269	258	204	148
15	906	178	129	95
16	752	119	89	67
17	609	102	82	56
18	588	101	76	59
19	397	83	60	45
20	332	50	40	30
21	433	45	37	28
22	195	41	33	28
23	100	21	15	9
24	123	25	16	13
25	136	17	14	13
26	70	17	14	9
27	108	14	12	9

28	57	14	12	6
29	83	11	6	5
30	41	7	4	2
31	47	7	5	3
32	61	9	9	8
33	37	9	4	3
34	20	4	3	1
35	7	4	4	4
36	4	1	1	1
37	6	2	0	0
38	8	1	0	0
39	1	0	0	0
40	4	0	0	0
41	1	1	1	1
42	1	1	1	1
43	7	1	1	1
44	1	0	0	0
45	3	1	1	1
48	1	1	1	1
50	2	1	1	1
51	7	0	0	0
54	1	1	1	1
55	1	1	1	1
60	1	0	0	0
62	2	1	1	1
66	1	0	0	0
67	1	1	1	1
71	1	0	0	0
107	1	1	1	1
108	1	1	1	1
109	1	1	1	1

Tabela 21. Número de α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre os algoritmos PDB-DSSP-Stride.

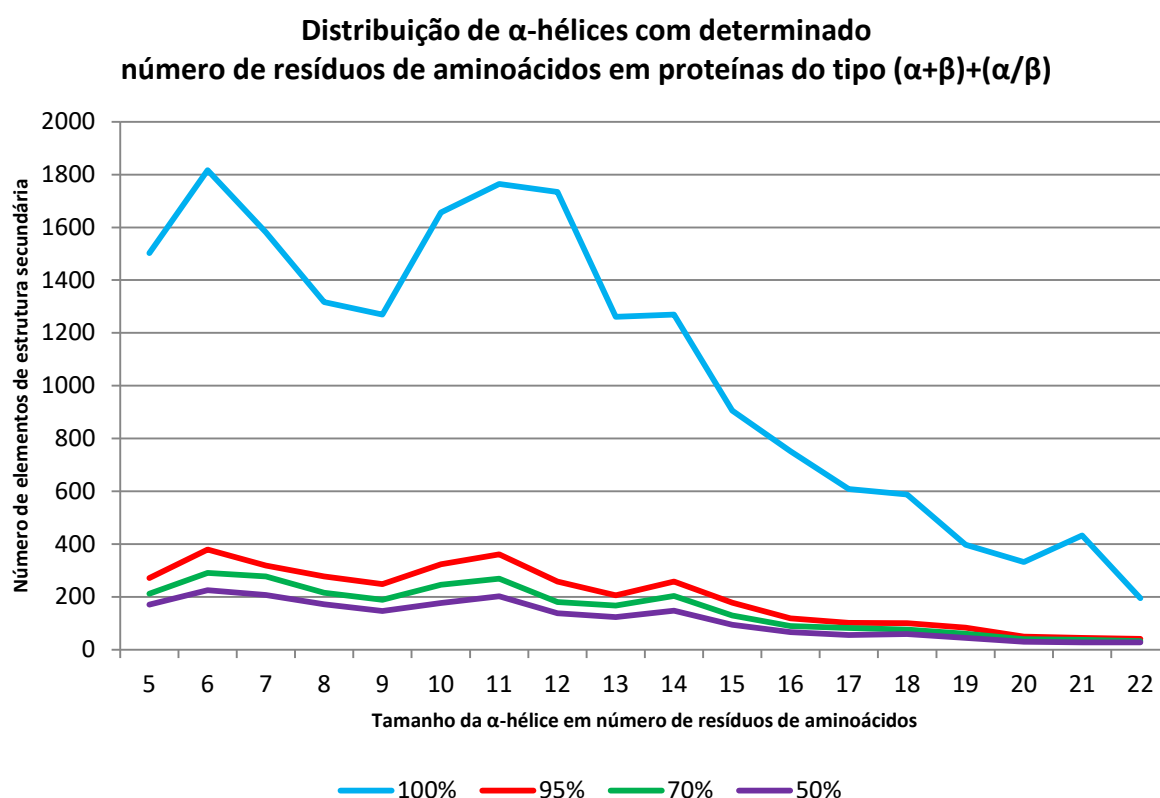


Figura 55. Número de α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre os algoritmos PDB-DSSP-Stride.

O maior número de α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ são formadas por 6, 10, 11 e 12 resíduos de aminoácidos. Mesmo após eliminarmos a redundância esse padrão se mantém, ainda que menos acentuado. Se tomarmos a linha 400 como limite superior ao número de α -hélices depois de removida a redundância e a linha 1200 como limite inferior ao número de α -hélices sem remover a redundância, concluímos que a existência de proteínas mutantes e/ou diversos complexos para cada proteína é da ordem de 300% a mais quando não removemos a redundância.

A partir de 23 resíduos de aminoácidos, o número das α -hélices diminui, principalmente após eliminarmos a redundância. A Fig. 56 mostra esse decréscimo. Embora vejamos no gráfico alguns picos, trata-se de valores baixos comparados aos do gráfico da Fig. 55. Por exemplo, para apenas quatro tamanhos diferentes de α -hélices (23, 24, 25 e 27) temos mais de 100 α -hélices. A α -hélice mais longa tem 109 resíduos de aminoácidos, que é a representada pelo código PDB 1ciib (“Transmembrane Protein Colicin Ia”) e é mostrada na Fig. 58.

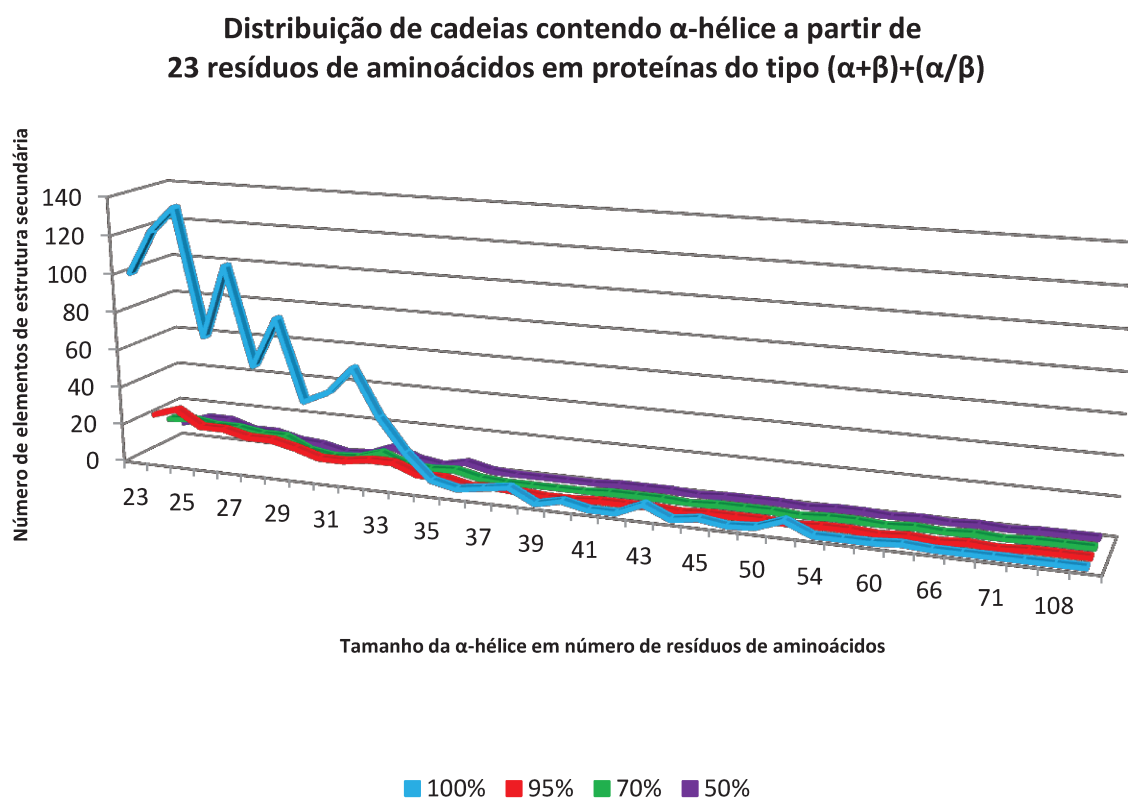


Figura 56. Número de α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre os algoritmos PDB-DSSP-Stride.

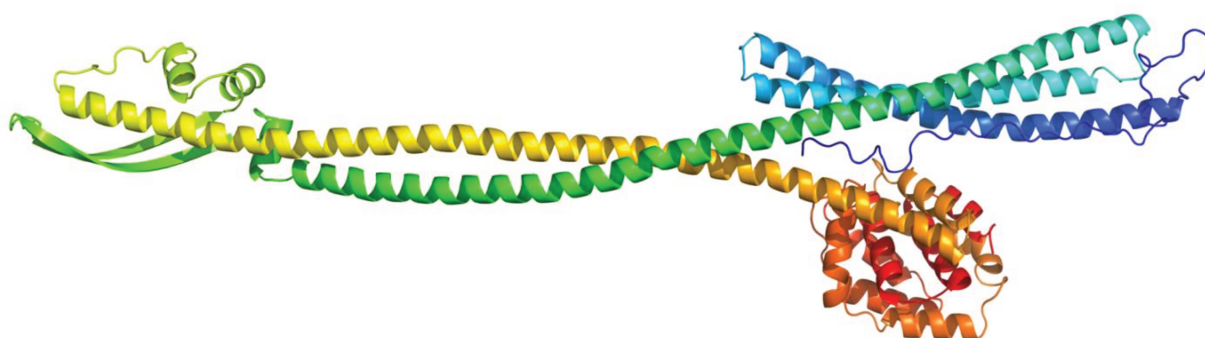


Figura 57. Exemplo de uma das mais longas α -hélices. A estrutura 1cii.pdb ("Transmembrane Protein Colicin Ia") tem duas α -hélices de 109 resíduos de aminoácidos (em amarelo e verde). A imagem foi gerada usando o software PyMOL.

4.3.4 Número de cadeias nas proteínas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$

A Tabela 22 e a Fig. 58 apresentam o número de cadeias proteicas para cada tamanho de folha- β encontrada nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre as definições do PDB-DSSP-Stride.

Tamanho da folha	Redundância			
	100%	95%	70%	50%
5	55531	11131	9133	7517
6	54078	10730	8627	7065
7	42836	8288	6649	5455
8	32573	6213	4963	4068
9	23588	4272	3473	2849
10	18428	3202	2467	2015
11	11767	2129	1572	1311
12	5890	1204	1010	857
13	3640	619	506	443
14	2158	431	366	315
15	1672	286	242	212
16	910	160	137	118
17	551	116	103	88
18	316	70	60	50
19	309	52	43	36
20	153	30	29	28
21	87	20	18	19
22	88	17	15	12
23	56	17	16	14
24	48	7	6	6
25	12	6	6	6
26	14	5	5	5
27	7	3	2	2
28	2	1	1	1
30	1	0	0	0
31	1	0	0	0
33	20	0	0	0
36	2	1	1	1

Tabela 22. Número de folhas- β nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre os algoritmos PDB-DSSP-Stride.

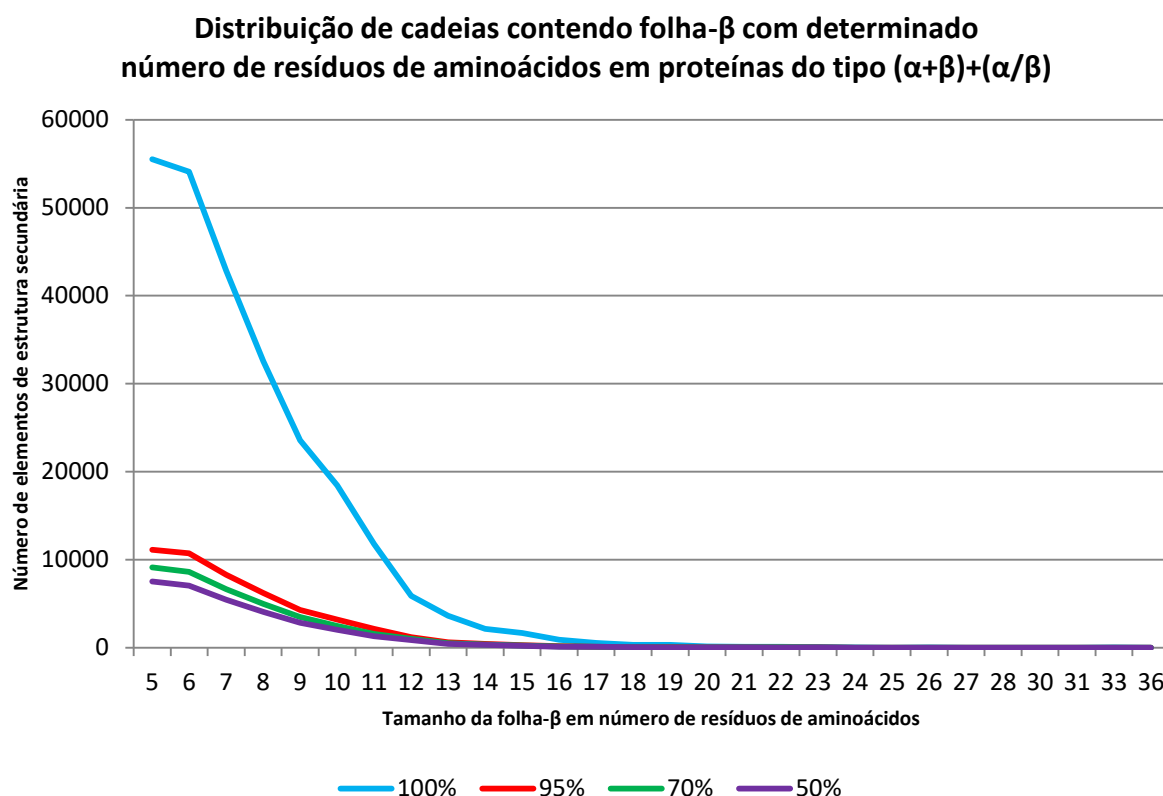


Figura 58. . Número de folhas- β nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre os algoritmos PDB-DSSP-Stride.

A maioria das folhas- β nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ são formadas por cinco resíduos de aminoácidos, e a partir daí o número de folhas- β diminui à medida que o seu tamanho aumenta. A mais longa folha- β tem 36 resíduos de aminoácidos. Ela está presente nas cadeias A e B da estrutura 2ztb.pdb (“Crystal Structure Of The Parasporin-2 Bacillus Thuringiensis Toxin That Recognizes Cancer Cells”), como pode ser visto na Fig. 59.

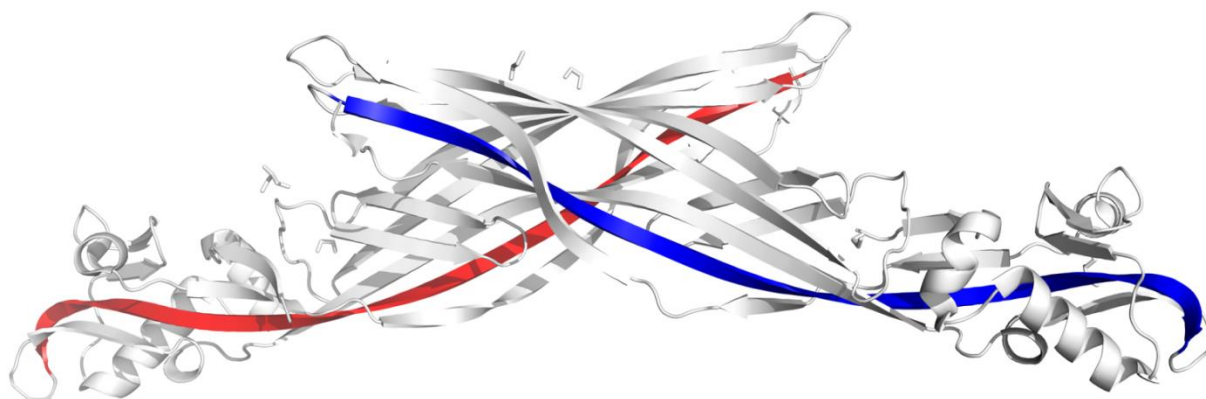


Figura 59. Exemplo de uma das mais longas folhas- β . A estrutura 2ztb.pdb (“Crystal Structure Of The Parasporin-2 Bacillus Thuringiensis Toxin That Recognizes Cancer Cells”) tem uma folha- β de 36 resíduos de aminoácidos na cadeia A (em azul) e uma folha- β de 36 resíduos de aminoácidos na cadeia B (em vermelho). A imagem foi gerada usando o software PyMOL.

4.3.5 Número de cadeias nas proteínas do tipo desordenado

As estruturas do tipo desordenado não tem α -hélice nem folha- β . A Tabela 23 e a Fig. 60 mostram a variação do número de cadeias por tamanho. Consideramos apenas as estruturas formadas por 50, ou mais, resíduos de aminoácidos (ALBERTS, JOHNSON e LEWIS, 2002) e, para efeito de visualização, usamos até o tamanho 99 na construção do gráfico.

Tamanho da folha	Redundância			
	100%	95%	70%	50%
50	17	7	4	0
51	6	1	0	0
52	32	4	3	1
53	18	6	4	1
54	11	1	0	0
55	25	8	6	1
56	27	4	1	0
57	20	5	3	0
58	42	8	6	2
59	12	4	1	0
60	34	3	0	0
61	17	4	2	0
62	50	5	1	0
63	27	4	2	1
64	35	7	2	0
65	29	3	2	1
66	29	5	2	0
67	49	5	2	0
68	53	8	7	0
69	31	7	4	0
70	63	6	5	2
71	43	12	5	0
72	69	12	5	0
73	56	7	8	1
74	86	11	8	0
75	90	7	1	0
76	198	6	2	0
77	69	7	3	2
78	14	1	0	0
79	15	2	2	1
80	28	4	4	2
81	30	2	0	0
82	38	3	1	0
83	20	5	5	2

84	21	1	0	0
85	12	5	4	0
86	28	6	5	0
87	20	1	1	0
88	20	2	2	0
89	19	3	2	1
90	11	3	2	0
91	8	2	2	0
92	15	3	2	1
93	18	3	2	0
94	33	4	2	0
95	8	1	0	0
96	13	3	3	0
97	14	2	1	1
98	12	3	2	0
99	21	3	2	1

Tabela 23 Número de cadeias nas proteínas do tipo "desordenado".

Número de cadeias proteicas do tipo "caótico" agrupadas por tamanho

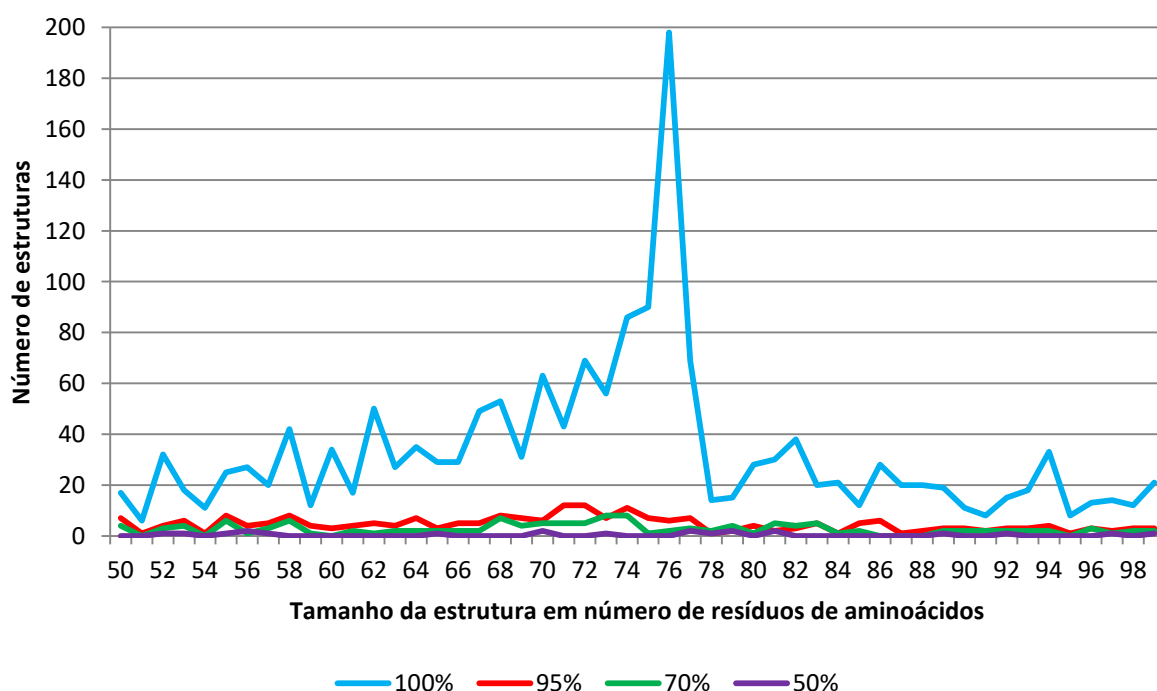


Figura 60. Número de cadeias proteicas do tipo desordenado agrupadas por tamanho e consenso entre PDB-DSSP-Stride.

O número de cadeias proteicas do tipo desordenado não passa de 200 por tamanho, e a maior quantidade dessas estruturas tem entre 70 e 80 resíduos de aminoácidos. A Fig. 61 mostra a maior estrutura "desordenada" encontrada, com 1630 resíduos de aminoácidos.

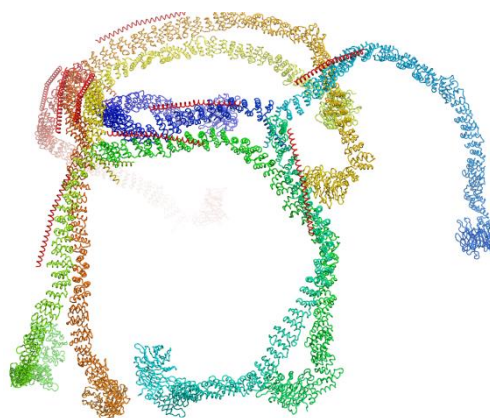


Figura 61. Exemplo de estrutura “desordenada”. A estrutura 3iyv.pdb (Clathrin D6 Coat As Full-Length Triskelions) tem 1630 resíduos de aminoácidos na cadeia A (em azul) e nenhum EES definido. A imagem foi gerada usando o PyMOL.

Para aumentar o número de estruturas alinhadas e, conseqüentemente, diminuir o ruído na análise estatística das médias dos valores de cada descritor, os EES foram alinhados pelo C-Terminal e N-Terminal. A Tabela 24 apresenta o número de EES alinhados para as estruturas do tipo all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$ e folha- β em $(\alpha+\beta)+(\alpha/\beta)$.

Tipo de proteína	Número de elementos de estruturas secundárias alinhadas
all- α	864
α em $(\alpha+\beta)+(\alpha/\beta)$	8525
all- β	1144
β em $(\alpha+\beta)+(\alpha/\beta)$	79173

Tabela 24. Número de cadeias contendo EES que foram alinhados pelo C-Terminal e N-Terminal

4.4 Comparação entre os sinais de um único descritor

Como prova de conceito, comparamos os valores de um único descritor para as α -hélices e folhas- β de mesmo tamanho. Se a hipótese de que existe um sinal que caracteriza determinado EES em seu nano-ambiente for verdadeira, então os mesmos princípios são válidos para as α -hélices e folhas- β , embora o sinal seja diferente em forma, conteúdo (tipo de descritor usado) e intensidade. A comparação entre os conjuntos de dados correspondentes para a região helicoidal contra a região das folhas- β demonstrou que existe uma diferença convincente entre os dois “sinais” que descrevem seus nano-ambientes. A Fig. 62 mostra essa diferença. Os descritores usados foram: A) EP@Ca e B) o número de contatos do tipo HBMM_WNASurf e o tamanho do elemento de estrutura escolhido para a prova de conceito foi 12. Foram selecionadas 1803 α -hélices e 6930 folhas- β , em ambos os casos das proteínas

do tipo $(\alpha+\beta)+(\alpha/\beta)$. A comparação entre os valores do descritor EP@C α mostra que, enquanto o sinal se assemelha a uma letra "N" no caso das α -hélices (linha cheia), o sinal para folhas- β parece mais uma letra "U" (linha tracejada). Para o descritor HBMM_WNASurf os valores são maiores para α -hélices do que para folhas- β , embora tenham um comportamento similar.

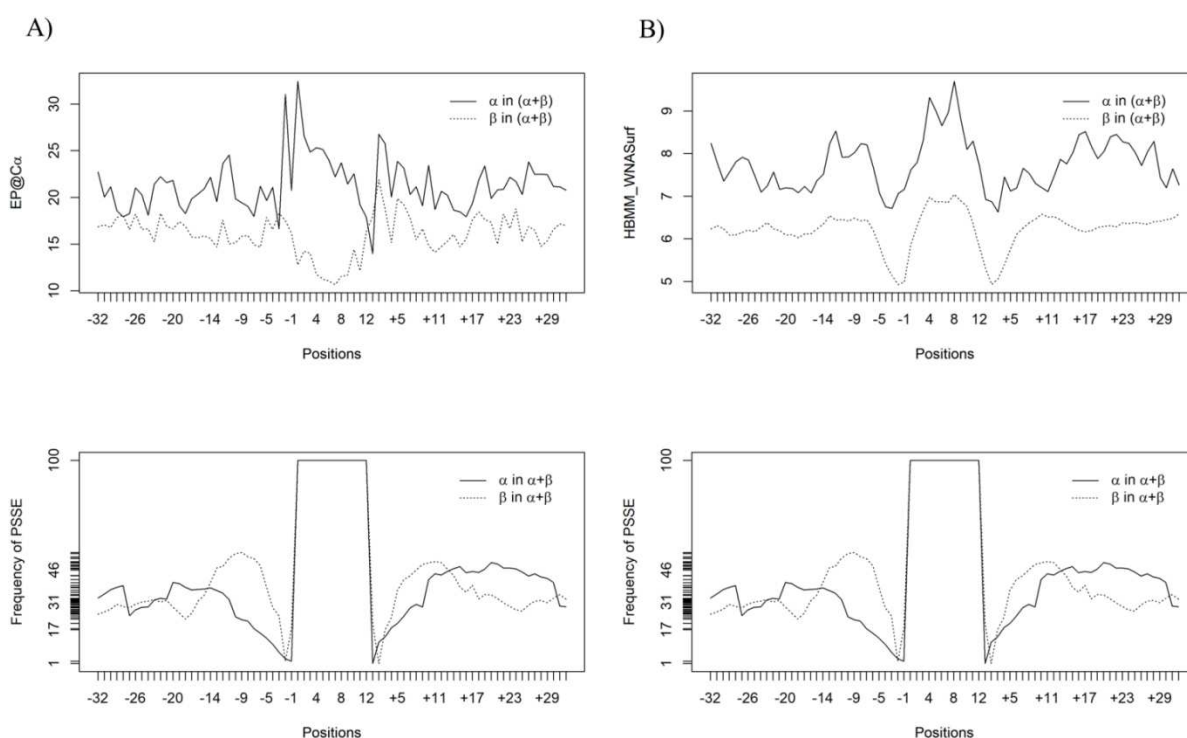


Figura 62 Diferenças no comportamento de dois descritores selecionados em torno de α -hélices (linhas contínuas) e folhas- β (linhas pontilhadas). Na parte superior dos gráficos, as diferenças são claras. Em: A) os valores do descritor EP@C α para α em $(\alpha+\beta)+(\alpha/\beta)$ se assemelha a uma letra "N" na presença das α -hélices e para β em $(\alpha+\beta)+(\alpha/\beta)$ o sinal parece mais uma letra "U"; B) Os gráficos para HBMM_WNASurf em α in $(\alpha+\beta) + (\alpha/\beta)$ e β em $(\alpha+\beta)+(\alpha/\beta)$ são semelhantes, mas os valores encontrados nas α -hélices são maiores. Na parte inferior dos gráficos temos a frequência com que as α -hélices (linhas contínuas) e folhas- β (linhas pontilhadas) estão presentes ao longo das estruturas alinhadas por tamanho.

Essa prova de conceito mostrou que existe um sinal para cada EES. Entretanto, o aprofundamento nos testes estatísticos demonstrou que esse sinal não é composto por um único descritor.

4.5 Estudo de caso: comparação entre duas estruturas homólogas

A metodologia usada para analisar o comportamento de um único descritor pode ser usada como métrica de comparação entre duas estruturas idênticas na estrutura primária, sendo uma delas resolvida com boa resolução e outra com uma resolução ruim.

No exemplo da Fig. 63, nós comparamos os valores do EP@Ca de uma estrutura bem resolvida (1fw4.pdb, com resolução de 1.7 Å, representada em vermelho) com a estrutura idêntica de menor resolução (1trc.pdb, com resolução de 3.6 Å, representada em azul) atualmente obsoleta. No arquivo 1fw4.pdb encontramos uma α -hélice entre os resíduos de aminoácidos 117-127, mas isso não acontece no arquivo 1trc.pdb. Entretanto, usando o software MSSP⁴⁰ nós alinhamos estruturalmente as duas estruturas e observamos a presença da α -hélice em ambas (Fig. 63.1). Na Fig. 63.2 vemos a comparação entre os valores do EP@Ca entre elas e o *template* (linha preta). Esse *template* foi elaborado alinhando as α -hélices de tamanho 11 extraídas do nosso *Datamart* e calculando a média para o descritor EP@Ca em cada posição da α -hélice. Comparando as estruturas 1fw4 (linha vermelha) e 1trc (linha azul), observamos que elas distoam em formato e valores. Mas quando as comparamos com o *template* (linha preta) notamos uma semelhança em formato entre ele e a estrutura 1fw4 (linha vermelha), indicando que a estrutura com melhor resolução (1.7 Å) se aproxima mais daquilo que é esperado para o conjunto das α -hélices de tamanho 11.

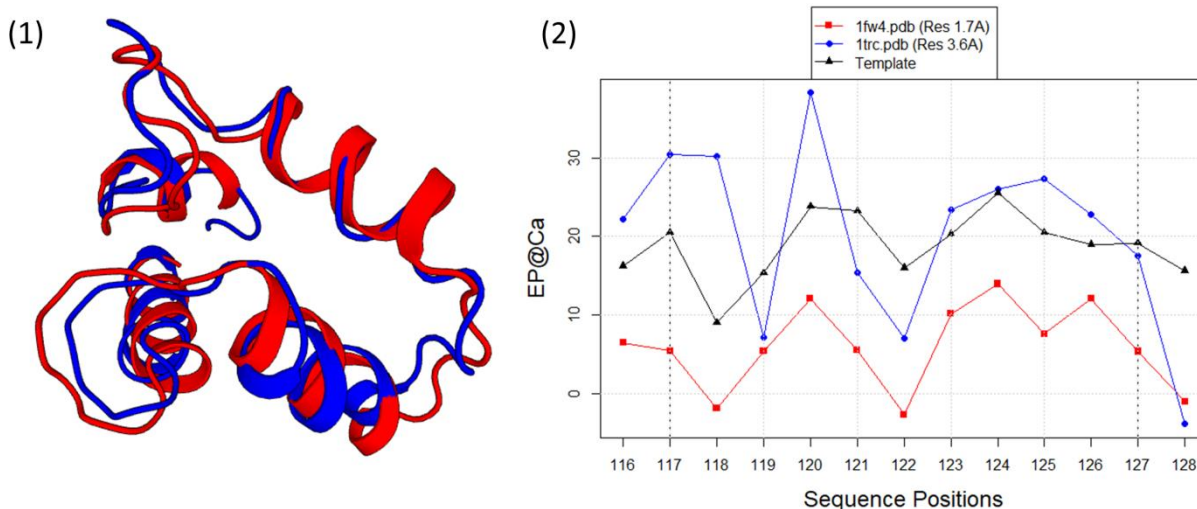


Figura 63. Comparação entre duas estruturas idênticas. A estrutura representada em vermelho tem boa resolução (1.7 Å) e o seu comportamento é semelhante ao comportamento do *template*, representando pela linha preta. A estrutura representada em azul não tem boa resolução (3.6 Å) e o seu comportamento distoa em vários pontos do da estrutura com melhor resolução e também do *template*. Essa comparação estabelece uma metodologia para comparar duas estruturas idênticas, que diferem em resolução, e também para avaliar uma nova estrutura resolvida ou modelo estrutural criado.

4.6 Teste estatístico para um descritor selecionado

Nos gráficos da Fig. 62 vimos um comportamento claramente distinto para as α -hélices e as folhas- β quando comparadas entre si e comparadas com o seu nano-ambiente.

⁴⁰ https://www.cbi.cnptia.embrapa.br/SMS/STINGm/MPA/js_mssp.html

Entretanto, a análise visual tem um caráter subjetivo, porque depende do olhar de quem o analisa. Para confirmar estatisticamente o que foi observado visualmente, aplicamos o teste de Kolmogorov-Smirnov ao inteiro conjunto de dados. Neste teste univariado, trabalhamos com as proteínas agrupadas em all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$ e β em $(\alpha+\beta)+(\alpha/\beta)$. Para cada um desses conjuntos, testamos os 69 descritores (um de cada vez - univariado) apresentados na Tabela 4 para cada tamanho do EES existente. A Tabela 25 mostra os resultados desses testes.

<i>Datamarts</i>	Total de testes	p-value $\leq 10^{-6}$	$10^{-6} > \text{p-value} \leq 10^{-3}$	p-value $> 10^{-3}$
all- α	3165	125 (3,9%)	301 (9,5%)	2739 (86,6%)
all- β	1173	28 (2,4%)	115 (9,8%)	1030 (87,8%)
α em $\alpha+\beta$	3704	298 (8,0%)	468 (12,6%)	2938 (79,4%)
β em $\alpha+\beta$	1860	207 (11,1%)	315 (16,9%)	1338 (72,0%)

Tabela 25. Avaliação do p-value para o teste de Kolmogorov-Smirnov aplicado ao conjunto de descritores para cada tipo de EES, alinhados por tamanho.

Começando pelas estruturas do tipo all- α , existem 46 tamanhos diferentes de α -hélices (Tabela 18). Usando os 69 descritores apresentados na Tabela 5, foram feitos 3165 testes ($46 \times 69 = 3174$; 9 testes foram excluídos porque apresentaram inconsistência nos dados de entrada). Destes, 125 testes apresentaram p-valor igual ou inferior a 10^{-6} (3,9%), e 301 testes tiveram p-value entre 10^{-6} e 10^{-3} (9,5%). Isso significa que apenas 13,4% dos testes analisados mostram valores de p-value compatíveis com a conclusão de que a região helicoidal (em termos de nano-ambiente) é estatisticamente diferente das regiões fora da hélice, considerando um nível de significância de 0,1%.

No caso das folhas- β nas estruturas do tipo all- β , foram realizados 1173 testes (17 tamanhos diferentes \times 69 descritores = 1173). Apenas 28 testes (2,4%) apresentaram p-value menor que 10^{-6} , e 115 testes (9,8%) tiveram p-value entre 10^{-6} e 10^{-3} . Isso significa que apenas 12,2% dos testes indicam que a região das folhas- β é estatisticamente diferente da região ao seu redor, considerando um nível de significância de 0,1%. É um resultado ainda mais baixo do que aquele obtido para as α -hélices do tipo all- α .

Para as α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$, nós encontramos 54 tamanhos diferentes, totalizando 3704 testes (54 tamanhos \times 69 descritores = 3726; 22 testes não foram realizados com sucesso porque apresentaram problema com os dados de entrada). Destes testes, 298 (8,0%) apresentam valor de p-value menor que 10^{-6} , e 468 (12,6%) apresentaram valor de p-value entre 10^{-6} e 10^{-3} . Assim, em 20,6% dos casos a região helicoidal (em termos

de nano-ambiente) é estatisticamente diferente das regiões fora da hélice, considerando um nível de significância de 0,1%.

Finalmente, para as folhas- β proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ nós encontramos 28 tamanhos diferentes, totalizando 1932 testes (28 tamanhos x 69 descritores = 1932; 72 testes não foram realizados com sucesso porque apresentaram problema com os dados de entrada). Destes testes, 207 (11,1%) apresentam valor de p-value menor que 10^{-6} , e 315 (16,9%) apresentaram valor de p-value entre 10^{-6} e 10^{-3} . Assim, em 28,0% dos casos a região das folhas- β é estatisticamente diferente das regiões ao seu redor, no mesmo nano-ambiente, considerando um nível de significância de 0,1%.

Essas análises usando um único descritor em cada teste e agrupando os EES por tamanho, resultaram, em média, em uma taxa de p-value menor que 10^{-3} de aproximadamente 20%. Na tentativa de melhorar esse valor, aplicamos o mesmo teste, mas dessa vez alinhamos as estruturas pelo seu C-Terminal e N-Terminal, ao invés de alinha-las pelo tamanho do EES. Desse modo, para cada grupo nós fizemos um único teste utilizando cada descritor, totalizando 69 testes. Os resultados são apresentados nas Tabelas 26 e 27.

<i>Datamarts</i>	Total de testes	p-value $\leq 10^{-6}$	$10^{-6} > \text{p-value} \leq 10^{-3}$	p-value $> 10^{-3}$
all- α	69	49 (71,0%)	14 (20,3%)	6 (8,7%)
all- β	69	15 (21,7%)	15 (21,7%)	39 (56,6%)
α em $\alpha+\beta$	69	54 (78,2%)	12 (17,4%)	3 (4,4%)
β em $\alpha+\beta$	69	31 (44,9%)	33 (47,8%)	5 (7,2%)

Tabela 26. Avaliação do p-value para o teste de Kolmogorov-Smirnov aplicado ao conjunto de descritores para cada tipo de EES, alinhados pelo C-Terminal das estruturas.

O resultado dos testes usando as estruturas alinhadas pelo C-Terminal foi melhor do que aqueles realizados usando as estruturas alinhadas pelo tamanho das α -hélices ou folhas- β . Por exemplo, para as proteínas do tipo all- α a taxa de testes com p-value menor que 10^{-3} foi 91,3%; para as proteínas do tipo all- β a taxa de testes com p-value menor que 10^{-3} foi 43,4%; para as α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ a taxa de testes com p-value menor que 10^{-3} foi 95,6%; para as folhas- β nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ a taxa de testes com p-value menor que 10^{-3} foi 92,8%.

A Tabela 27 mostrar o resultado dos testes aplicados nas estruturas alinhadas pelo N-Terminal. No caso das proteínas do tipo all- α a taxa de testes com p-value menor que 10^{-3} foi 91,3%; para as proteínas do tipo all- β a taxa de testes com p-value menor que 10^{-3} foi

36,2%; para as α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ a taxa de testes com p-value menor que 10^{-3} foi 95,6%; para as folhas- β nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ a taxa de testes com p-value menor que 10^{-3} foi 76,8%.

<i>Datamarts</i>	Total de testes	p-value $\leq 10^{-6}$	$10^{-6} > \text{p-value} \leq 10^{-3}$	p-value $> 10^{-3}$
all- α	69	40 (58,0%)	23 (33,3%)	6 (8,7%)
all- β	69	12 (17,4%)	13 (18,8%)	44 (63,8%)
α em $\alpha+\beta$	69	55 (79,7%)	11 (15,9%)	3 (4,4%)
β em $\alpha+\beta$	69	22 (31,9%)	31 (44,9%)	16 (23,2%)

Tabela 27. Avaliação do p-value para o teste de Kolmogorov-Smirnov aplicado ao conjunto de descritores para cada tipo de EES, alinhados pelo N-Terminal das estruturas.

As Tabelas 26 e 27 demonstraram que os resultados obtidos alinhando as estruturas pelo C-Terminal e N-Terminal foram melhores que os resultados obtidos pelo alinhamento por tamanho. Comparando as duas tabelas, concluímos o seguinte: para as proteínas do tipo all- α e α em $(\alpha+\beta)+(\alpha/\beta)$ a taxa de testes com p-value menor que 10^{-3} foi igual (91,3% e 95,6% respectivamente) para C-Terminal e N-Terminal, embora quando dividimos esse limiar em p-value $\leq 10^{-6}$ e $10^{-6} > \text{p-value} \leq 10^{-3}$ cada teste resultou em taxas diferentes. No caso das proteínas do tipo all- β , os testes aplicados nas estruturas alinhadas pelo C-Terminal teve uma taxa de p-value menor que 10^{-3} de 43,4% enquanto os testes aplicados nas estruturas alinhadas pelo N-Terminal tiveram uma taxa de p-value menor que 10^{-3} de 36,2%. Os testes aplicados nas estruturas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$ alinhadas pelo C-Terminal teve uma taxa de 92,8% com p-value menor que 10^{-3} , enquanto os testes aplicados nas estruturas alinhadas pelo N-Terminal tiveram uma taxa de p-value menor que 10^{-3} de 76,8%. A Fig. 64 apresenta um gráfico comparativo entre os resultados obtidos com os alinhamentos por tamanho, C-Terminal e N-Terminal.

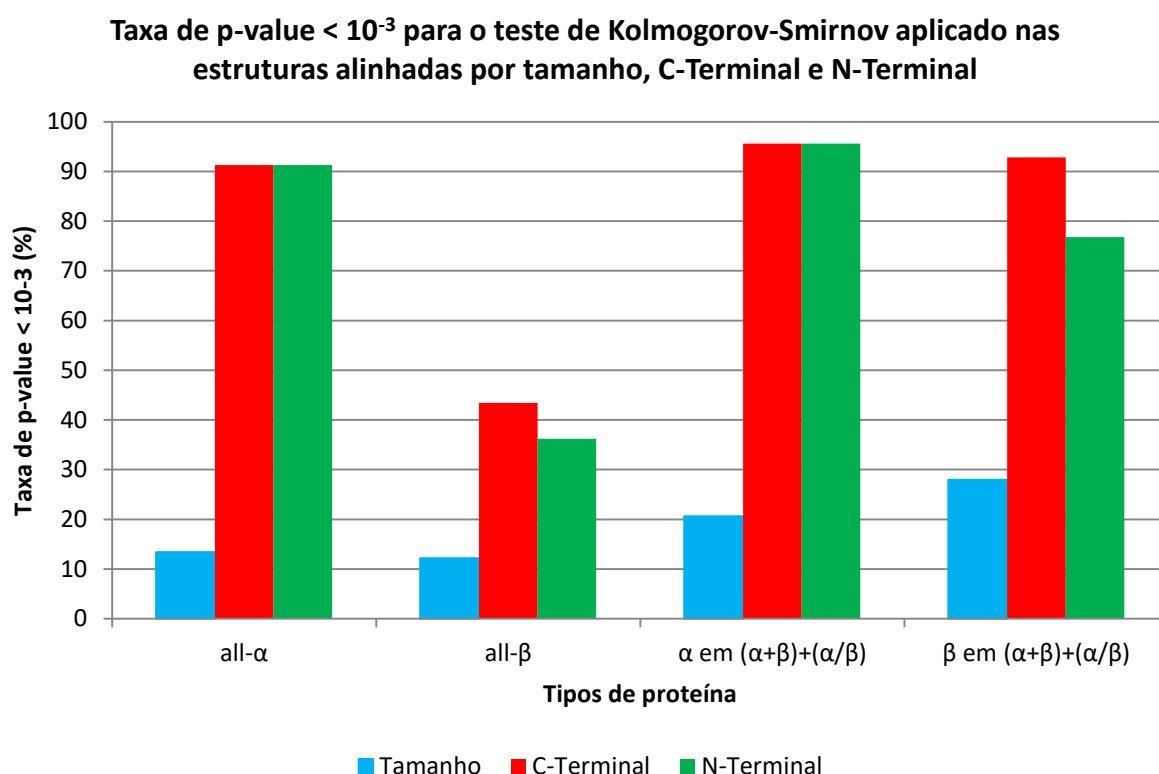


Figura 64. Taxa de p-value < 10^{-3} para a análise univariada (um descritor de cada vez) usando o teste de Kolmogorov-Smirnov aplicado nas proteínas dos tipos all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$ e β em $(\alpha+\beta)+(\alpha/\beta)$ alinhados por tamanho, C-Terminal e N-Terminal.

4.7 Análise comparativa entre α -hélices e hélices 3_{10}

Os *Datamarts* foram extraídos do STING_RDB, que não diferencia as estruturas helicoidais em α -hélices, π -hélices e hélices 3_{10} . Como visto anteriormente na Tabela 10, a quantidade de π -hélices é irrisória (menor que 0,001%), mas a quantidade de hélices 3_{10} é significativa (11,4% no caso das proteínas do tipo all- α e 23,6% nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$). Nós dividimos o *Datamart* das estruturas helicoidais em α -hélices e hélices 3_{10} e aplicamos o teste multivariado MANOVA nos dois sub-*Datamarts* e também no conjunto total de estruturas helicoidais, e apresentamos os resultados na Tabela 28.

Tipo de estrutura helicoidal	p-value < 10^{-3}	
	all- α	α em $(\alpha+\beta)+(\alpha/\beta)$
α -hélices	97,8%	93,9%
hélices 3_{10}	94,3%	93,0%
Estruturas helicoidais (α -hélices + hélices 3_{10})	95,5%	94,2%

Tabela 28. Análise comparativa entre α -hélices e hélices 3_{10} . A taxa de p-value < 10^{-3} é superior a 90% em todos os casos.

Todos os testes tiveram resultados significativos, com taxas de p-value $< 10^{-3}$ sempre maiores que 90%. Isso é relevante, porque a introdução das hélices 3_{10} no *Datamart* das estruturas helicoidais não interferiu negativamente nos testes. Assim, para diminuir o ruído, aumentamos o número de estruturas analisadas, incluindo no mesmo *Datamart* todos os três tipos de estruturas helicoidais. A partir daí passamos à definição do nano-ambiente usando um conjunto de parâmetros.

4.8 Nano-ambiente definido por um conjunto de parâmetros

Os resultados dos testes Kolmogorov-Smirnov aplicados nas estruturas alinhadas por tamanho demonstraram que um único parâmetro pode até ser satisfatório para descrever o nano-ambiente onde uma α -hélice ou uma folha- β está presente. Porém, é intuitivo pensar que um nano-ambiente seria mais bem descrito por um conjunto de parâmetros. Para determinar se isso é verdade, e quais parâmetros usar em cada caso, nós seguimos o seguinte procedimento:

1. Extraímos a média e o desvio padrão para cada descritor, dentro e fora do EES;
2. Normalizamos os dados pelo ICV;
3. Calculamos a diferença entre os dados normalizados, encontrados dentro e fora do EES;
4. Selecionamos apenas aqueles descritores com diferença maior que 0,1 entre seus valores normalizados.

No Apêndice A nós explicamos porque normalizamos os dados (item 2) usando o inverso do coeficiente de variância (ICV). As Fig. 65-68 exibem, na forma de um gráfico de radar, quais descritores foram selecionados em cada caso, e as diferenças entre os valores normalizados encontrados dentro e fora do respectivo EES.

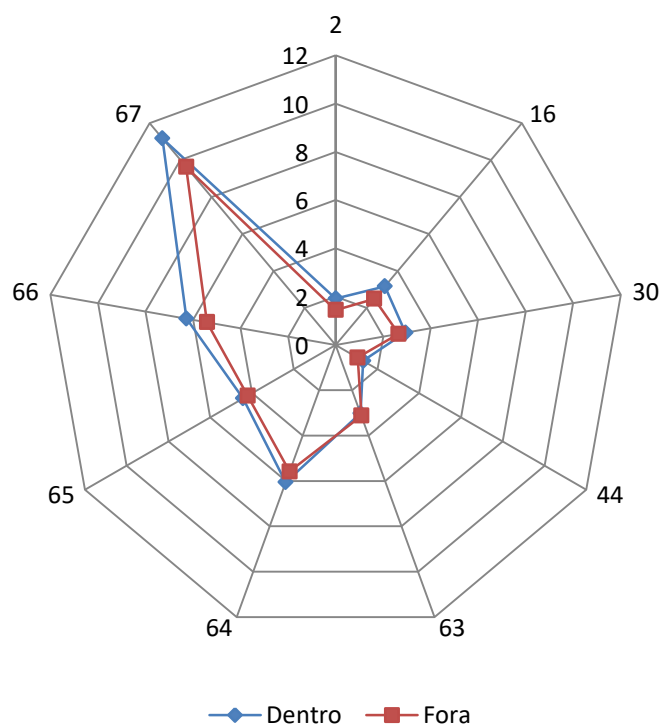


Figura 65. Descriptores selecionados para as proteínas do tipo all-α: 2. hbmm, 16. hbmm_WNADist, 30. hbmm_WNASurf, 44. Electrostatic_Potential_at_CA, 63. Number_Unused_Contact, 64. Number_Unused_Contact_WNADist, 65. Number_Unused_Contact_WNASurf, 66. IFR_CA_3, 67. Internal_CA_3 .

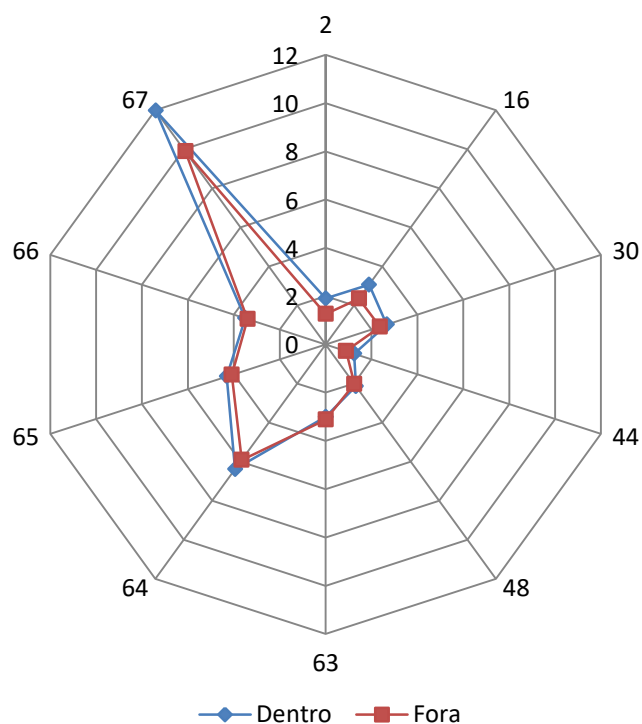


Figura 66. Descriptores selecionados para as proteínas do tipo α em (α+β)+(α/β): 2. hbmm, 16. hbmm_WNADist, 30. hbmm_WNASurf, 44. Electrostatic_Potential_at_CA, 48. Electrostatic_Potential_at_CA_WNADist, 63. Number_Unused_Contact, 64. Number_Unused_Contact_WNADist, 65. Number_Unused_Contact_WNASurf, 66. IFR_CA_3, 67. Internal_CA_3

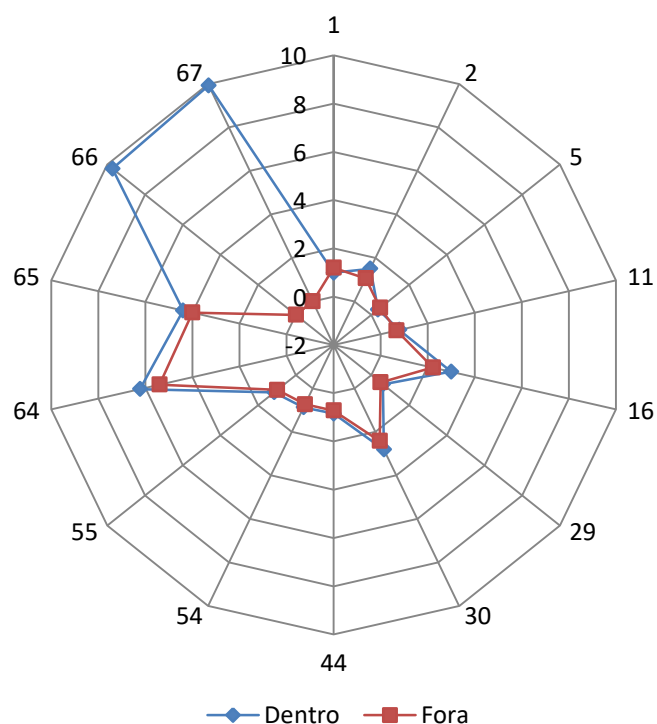


Figura 67. Descriptores selecionados para as proteínas do tipo all-β: 1. Accessible_Surface_in_Isolation, 2. hbmm, 5. hbms, 11. hydrophobic, 16. hbmm_WNADist, 29. ch_repulsive_WNADist, 30. hbmm_WNASurf, 44. Electrostatic_Potential_at_CA, 54. Cross_Link_Order_CA, 55. Cross_Pres_Order_CA, 64. Number_Used>Contact_WNADist, 65. Number_Used>Contact_WNASurf, 66. IFR_CA_3, 67. Internal_CA_3

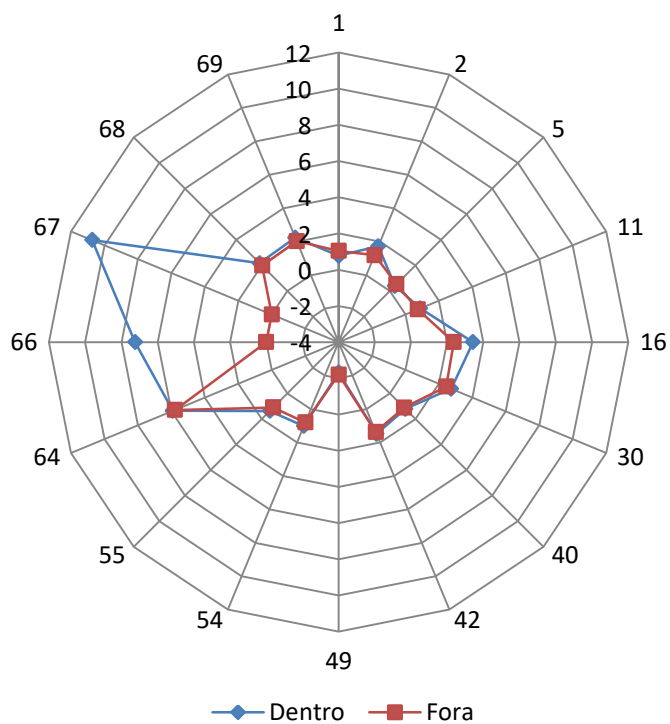


Figura 68. Descriptores selecionados para as proteínas do tipo β em (α+β)+(α/β): 1. Accessible_Surface_in_Isolation, 2. hbmm, 5. hbms, 11. hydrophobic, 16. hbmm_WNADist, 30. hbmm_WNASurf, 40. aromatic_WNASurf, 42. ch_attractive_WNASurf, 49. Electrostatic_Potential_at_LHA_WNADist, 54. Cross_Link_Order_CA, 55. Cross_Pres_Order_CA, 64. Number_Used>Contact_WNADist, 66. IFR_CA_3, 67. Internal_CA_3, 68. Clash, 69. Percent

Após essa análise, descobrimos que os descritores comuns aos quatro diferentes tipos de proteínas são os seguintes:

- hbmm
- hbmm_WNADist
- hbmm_WNASurf
- Number_Unused_Contact_WNADist
- IFR_CA_3
- Internal_CA_3

Também observamos que os seguintes descritores são comuns para as folhas- β :

- Accessible_Surface_in_Isolation
- Cross_Link_Order_CA
- Cross_Pres_Order_CA

Enfim, encontramos um único descritor que demonstrou diferença apenas para as α -hélices e seus nano-ambientes, e que não aparece no caso das folhas- β :

- Number_Unused_Contact

Fizemos essa análise e seleção de descritores para fornecer os dados para a análise de variância multivariada (MANOVA). Assim, trabalhamos na análise multivariada apenas com os descritores que apresentaram diferença estatisticamente significativa entre os valores encontrados dentro e fora do EES. Segundo Johnson e Wichern (JOHNSON e WICHERN, 1999), a análise MANOVA assume que os dados tem uma distribuição normal multivariada. Portanto, o primeiro passo foi remover os descritores que não tem uma distribuição normal. Fizemos isso aplicando o teste de normalidade de Shapiro. A seguir, os descritores linearmente correlacionados usando um limite padrão 0,9 (LEE B, 1971) também foram eliminados. Como resultado dessa filtragem, o número de descritores utilizados na MANOVA foi menor que o número de descritores alimentados na entrada de teste. A Tabela 29 apresenta os resultados do teste MANOVA para as proteínas dos tipos all- α , all- β , α em $(\alpha+\beta)+(\alpha/\beta)$ e β em $(\alpha+\beta)+(\alpha/\beta)$. Em todos os casos, o menor número de descritores usados foi 2, o que significa que os demais não tem distribuição normal, ou são linearmente correlacionados acima do ponto de corte. Os testes foram aplicados usando dois níveis de significância, onde buscamos valores de p-value $< 10^{-6}$ e valores de p-values $\geq 10^{-6}$ e $< 10^{-3}$.

		Tipos de estruturas proteicas			
		all- α	all- β	$\alpha (\alpha+\beta)+(\alpha/\beta)$	$\beta (\alpha+\beta)+(\alpha/\beta)$
Total de testes executados		38 / 46 (83%)	17 / 17 (100%)	44 / 54 (81%)	21 / 28 (75%)
Total de descritores usados		9	14	10	16
Menor n° de descritores usados		2 (45)	2 (22)	2 (109)	2 (36)
Média de descritores usados		4	5	4	5
Pillai	$< 1e^{-6}$	36,8	41,2	70,5	66,7
	$\geq 1e^{-6}$ e $< 1e^{-3}$	26,3	29,4	25,0	23,8
Wilks	$< 1e^{-6}$	42,1	47,1	72,7	76,2
	$\geq 1e^{-6}$ e $< 1e^{-3}$	26,3	23,5	22,7	23,8
Hotelling Lawley	$< 1e^{-6}$	47,4	47,1	75,0	76,2
	$\geq 1e^{-6}$ e $< 1e^{-3}$	21,1	23,5	20,5	14,3
Roy	$< 1e^{-6}$	47,4	47,1	75,0	71,4
	$\geq 1e^{-6}$ e $< 1e^{-3}$	28,9	29,4	22,7	23,8

Tabela 29. Resultados da análise multivariada para as estruturas alinhadas pelo tamanho do EES. Na linha *Total de testes executados* vemos dois valores: (total de testes concluídos com êxito / número total de testes) onde o número total de testes é a quantidade de diferentes tamanhos do EES. A linha *Total de descritores usados* apresenta o número de descritores com diferença $> 0,1$ entre os valores encontrados dentro e fora do EES, após terem sido normalizados pelo ICV. A linha *Menor n° de descritores usados* apresenta dois valores: menor número de descritores usados e (tamanho do EES que usou a menor quantidade de descritores). As demais linhas são autoexplicativas.

Ao compararmos os resultados do teste MANOVA com os resultados do teste univariado para as estruturas alinhadas por tamanho, fica evidente que um conjunto de descritores descreve melhor um nano-ambiente que um único descritor. Nos testes MANOVA, a taxa de resultados com p-value menor que 10^{-3} é maior que nos testes uni variados. A Fig. 69 apresenta uma comparação entre os resultados dos testes uni e multivariado Pillai.

Taxa de p-value $< 10^{-3}$ para testes Uni e Multivariados nas estruturas alinhadas por tamanho do PSSE

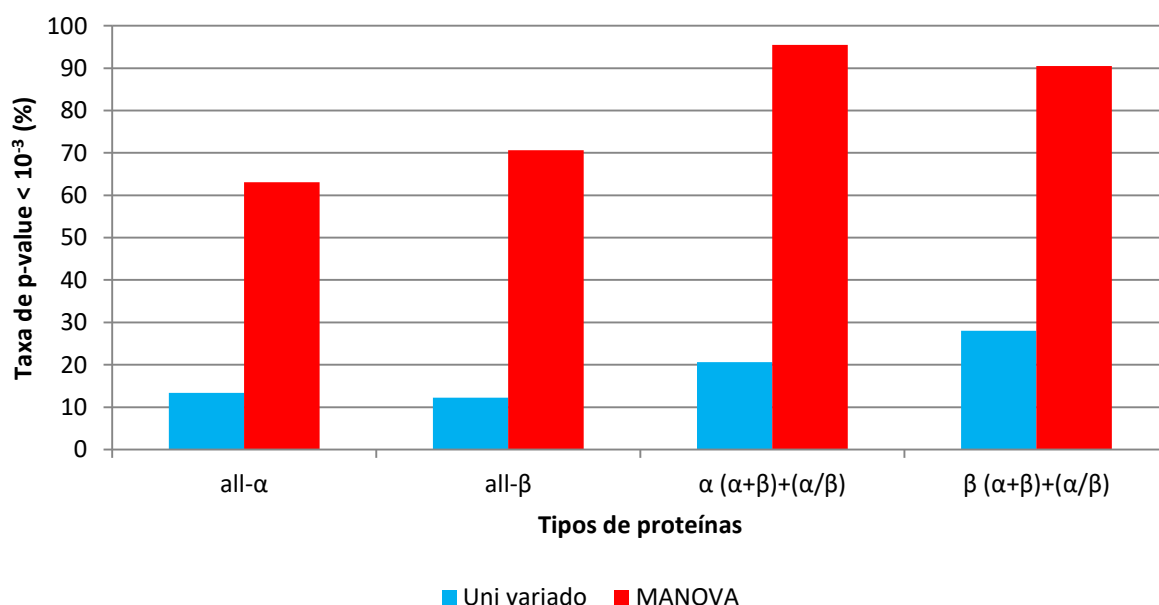


Figura 69. Comparação entre os testes uni e multivariados aplicados nas estruturas alinhadas por tamanho. Enquanto os testes univariados não alcançaram 30% de p-value $< 10^{-3}$, e na média essa taxa foi de 18,5%, para os testes multivariados as taxas passaram de 90%, e na média foram de 79,9%. Isso representa um ganho de 61,4%.

No caso das proteínas do tipo all- α , o teste MANOVA demonstrou que o seu nano-ambiente é descrito pelo número de contatos não realizados, número de contatos do tipo ponte de hidrogênio, pelo seu potencial eletrostático no carbono- α e pela densidade na superfície. Em um total de 46 testes, cada um dos descritores selecionados para o teste MANOVA foram usados o seguinte número de vezes:

1. Number_Unused_Contact (26; 56,5%)
2. Number_Unused_Contact_WNADist (24; 52,2%)
3. hbmm_WNASurf (21; 45,6%)
4. hbmm_WNADist (19; 41,3%)
5. Internal_CA_3 (18; 39,1%)
6. Number_Unused_Contact_WNASurf (18; 39,1%)
7. hbmm (14; 30,4%)
8. Electrostatic_Potential_at_CA (10; 21,7%)
9. IFR_CA_3 (10; 21,7%)

No caso das proteínas do tipo all- β , a análise multivariada mostrou que o seu nano-ambiente é descrito pela acessibilidade dos resíduos de aminoácidos em sua superfície, e basicamente por seus contatos, em especial os do tipo ligações de hidrogênio. Os descritores de Cross Link e Cross Presence também foram selecionados, o que era esperado, uma vez que as fitas- β que compõe uma folha- β se estabilizam pelos contatos entre si. O potencial eletrostático no carbono- α e a densidade na superfície também são importantes na formação do nano-ambiente para as folha- β . Em um total de 17 testes, cada um dos descritores selecionados para o teste MANOVA foram usados o seguinte número de vezes:

1. Accessible_Surface_in_Isolation (9; 52,9%)
2. hbmm_WNADist (9; 52,9%)
3. Number_Unused_Contact_WNASurf (9; 52,9%)
4. hbmm_WNASurf (8; 47,1%)
5. Number_Unused_Contact_WNADist (8; 47,1%)
6. Internal_CA_3 (7; 41,2%)
7. Cross_Link_Order_CA (5; 29,4%)
8. Cross_Pres_Order_CA (5; 29,4%)
9. hbmm (5; 29,4%)
10. IFR_CA_3 (5; 29,4%)
11. hbms (4; 23,5%)
12. hydrophobic (4; 23,5%)
13. Electrostatic_Potential_at_CA (2; 11,8%)

No caso α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$, o teste multivariado selecionou os parâmetros abaixo como importantes para a descrição do nano-ambiente. Em

um total de 54 testes, cada descritor selecionados para o teste MANOVA foram usados o seguinte número de vezes:

1. Number_Unused_Contact_WNADist (33; 61,1%)
2. Electrostatic_Potential_at_CA_WNADist (30; 55,5%)
3. Number_Unused_Contact (24; 44,4%)
4. Number_Unused_Contact_WNASurf (21; 38,9%)
5. hbmm_WNASurf (16; 29,6%)
6. IFR_CA_3 (16; 29,6%)
7. Internal_CA_3 (13; 24,1%)
8. hbmm (11; 20,4%)
9. hbmm_WNADist (11; 20,4%)
10. Electrostatic_Potential_at_CA (5; 9,3%)

O nano-ambiente das α -hélices nas proteínas do tipo $(\alpha+\beta)+(\alpha/\beta)$ pode ser descrito pelo número de contatos não realizados, pelos contatos do tipo ponte de hidrogênio, pelo potencial eletrostático no carbono- α e pela densidade na superfície.

No caso das folhas- β nas proteínas do $(\alpha+\beta)+(\alpha/\beta)$, o nano-ambiente é melhor definido pelo potencial eletrostático no último átomo mais pesado ponderado pela vizinhança, pelo número de contatos não feitos ponderados pela vizinhança, e naturalmente pela capacidade dos resíduos de aminoácidos presentes na folha- β fazerem contatos entre si, ou estarem próximos uns dos outros quando os colocamos dentro de uma esfera de prova (Cross Link e Cross Presence). A acessibilidade dos resíduos de aminoácidos na superfície também é importante na determinação desse nano-ambiente. Em um total de 28 testes, cada descritor selecionados para o teste MANOVA foram usados o seguinte número de vezes:

1. Electrostatic_Potential_at_LHA_WNADist (14; 50%)
2. Number_Unused_Contact_WNADist (10; 35,7%)
3. Cross_Link_Order_CA (9; 32,1%)
4. Cross_Pres_Order_CA (9; 32,1%)
5. hbmm_WNADist (9; 32,1%)
6. hbmm_WNASurf (8; 28,6%)
7. Accessible_Surface_in_Isolation (6; 21,4%)
8. ch_attractive_WNASurf (6; 21,4%)
9. hbmm (6; 21,4%)
10. hbms (6; 21,4%)
11. Internal_CA_3 (6; 21,4%)
12. Percent (5; 17,8%)
13. hydrophobic (4; 14,4%)
14. IFR_CA_3 (4; 14,4%)
15. Clash (3; 10,7%)
16. aromatic_WNASurf (2; 7,1%)

Como visto anteriormente, quando aplicamos o teste univariado nas estruturas alinhadas pelo C-Terminal e N-Terminal, nós melhoramos as taxas de p-value menores que 10^{-3} , porque aumentamos o espaço amostral e, conseqüentemente, diminuimos o ruído nos testes. Nós tentamos a mesma abordagem para o teste MANOVA, mas encontramos alguns problemas inerentes ao teste e a natureza dos dados.

O teste multivariado pressupõe que os dados tenham uma distribuição normal. Nós usamos o teste Shapiro para verificar a normalidade dos dados, o que funciona muito bem para as estruturas alinhadas por tamanho, onde temos 32 resíduos de aminoácidos antes e 32 resíduos de aminoácidos depois delas, definindo seu nano-ambiente. Mas, quando alinharmos as α -hélices e folhas- β pelo C-Terminal e N-Terminal, mudamos a natureza dos dados, porque desconsideramos o que acontece antes ou depois do seu início ou fim. Nestes casos, o teste Shapiro classificou a maioria dos descritores como não tendo uma distribuição normal, e isso impede a realização do teste MANOVA, quando apenas um descritor tem tal distribuição.

Nos testes multivariados aplicados nas estruturas alinhadas por tamanho, tínhamos n testes, onde n representa o número de diferentes tamanhos de α -hélices e folhas- β . Assim, foi possível definir uma métrica: em 100% dos testes aplicados, calculamos a taxa de p-value $< 10^{-3}$. No caso das estruturas alinhadas pelo C-Terminal e N-Terminal não foi possível usar essa métrica, porque para cada tipo de estrutura executamos apenas um teste. O que temos, quando conseguimos aplicá-lo, é um único valor de p-value. As Tabelas 30 e 31 apresentam os valores de p-value para esses testes.

Região analisada	Tipos de estruturas proteicas			
	all- α	all- β	$\alpha (\alpha+\beta)+(\alpha/\beta)$	$\beta (\alpha+\beta)+(\alpha/\beta)$
Início (res. 1-5)	0.1295717	0.0001933965	NA	NA
Toda extensão	2.675539e-24	8.033081e-09	NA	NA

Tabela 30. Resultados do teste MANOVA aplicada nas estruturas alinhadas pelo N-Terminal. Apenas as estruturas dos tipos all- α e all- β puderam ser testadas pelo MANOVA, nos demais casos o teste de normalidade Shapiro não retornou dados suficientes para a análise multivariada. A análise do início dos EES resultou em valores de p-value mais próximos de 1 que a análise de toda a sua extensão. No caso das α -hélices isso pode ser explicado porque seu início se assemelha a um turn, o que “confundi” o teste.

Os testes foram aplicados em duas regiões das α -hélices e folhas- β alinhadas pelo N-Terminal e C-Terminal: primeiro consideramos apenas os cinco primeiros resíduos de aminoácidos (início) para o alinhamento pelo N-Terminal e os últimos cinco resíduos de

aminoácidos (final) para o alinhamento pelo C-Terminal, contra todo o restante dos resíduos de aminoácidos. No segundo teste, consideramos a inteira extensão de cada EES, contra os 32 resíduos de aminoácidos fora.

Região analisada	Tipos de estruturas proteicas			
	all- α	all- β	α ($\alpha+\beta$)+(α/β)	β ($\alpha+\beta$)+(α/β)
Últimos 5 res.	NA	0.03850456	NA	NA
Toda extensão	NA	1.153829e-06	NA	NA

Tabela 31. Resultados do teste MANOVA aplicada nas estruturas alinhadas pelo C-Terminal. Apenas as estruturas do tipo all- β pôde ser testado pelo MANOVA, nos demais casos o teste de normalidade Shapiro não retornou dados suficientes para a análise multivariada. A análise do início das folhas- β resultou em um valor de p-value mais próximo de 1 que a análise de toda a sua extensão.

5 CONCLUSÕES

Durante este trabalho, desenvolvemos o software PS³A como prova de conceito, demonstrando que é possível ver um sinal na região onde o EES está presente. Os gráficos da Fig. 62 demonstram um comportamento dentro das α -hélices, que não é o mesmo visto nas folhas- β . Porém, essa diferença, além de ser subjetiva porque depende da visão de quem a analisa, é insuficiente para se fazer uma classificação robusta de um EES em seu nano-ambiente. Assim, o próximo passo foi aplicar testes estatísticos no conjunto de dados, para comprovar estatisticamente o que foi observado visualmente.

Os primeiros testes estatísticos aplicados foram univariados, onde nós usamos apenas um descritor em cada teste. Primeiramente aplicamos os testes nas estruturas alinhadas por tamanho, e neste caso, tivemos uma taxa de sucesso de aproximadamente 20%, onde definimos taxa de sucesso o $p\text{-value} < 10^{-3}$. Entretanto, o alinhamento por tamanho está sujeito à introdução de ruído, uma vez que o número de amostras decresce nas regiões periféricas à estrutura alinhada. Para diminuir o ruído e melhorar o resultado, alinhamos as estruturas pelo seu C-Terminal e N-Terminal. Os resultados obtidos foram melhores. No caso das proteínas do tipo all- β , a taxa de sucesso foi um pouco acima de 70%, e nos demais casos a taxa de sucesso foi superior a 90%.

Aprofundando a investigação, nós submetemos os dados à análise MANOVA. Os resultados da MANOVA aplicada às estruturas alinhadas por tamanho foi sensivelmente melhor que àqueles obtidos nos testes univariados. Enquanto para os testes univariados a taxa de sucesso ficou em 20%, no teste MANOVA a taxa de sucesso foi em média de 80%.

Portanto, o uso da análise multivariada mostrou-se mais eficaz que os testes aplicados em um único descritor. Isso demonstra que o nano-ambiente funciona como um sistema chave-fechadura, onde todas as ranhuras da chave precisam estar devidamente ajustadas, caso contrário a fechadura não irá abrir. Em nosso caso, todos os descritores de um conjunto selecionado pelo teste MANOVA devem ter seus valores ajustados para aquele nano-ambiente, caso contrário, o EES não se manterá.

Ao compararmos os descritores mais usados pelo teste MANOVA e que foram comuns para os diversos nano-ambientes, concluímos que o número de ligações de hidrogênio entre cadeias principais é maior nas α -hélices que nas folhas- β , o potencial eletrostático no carbono- α é maior nas α -hélices que nas folhas- β , o número de contatos não realizados é maior nas α -hélices que nas folhas- β , a capacidade de fazer contatos cruzados (Cross_Link) é maior nas

folhas- β que nas α -hélices, a acessibilidade à superfície em isolamento é maior nas α -hélices que nas folhas- β , a densidade na superfície e no interior da proteína é maior nas α -hélices que nas folhas- β . Essas especificidades são como as ranhuras da chave, que a fazem específica para a sua fechadura.

Três categorias de descritores foram identificados entre as mais frequentemente usados para identificação das α -hélices e também das folhas- β : número de contatos do tipo ligações de hidrogênio entre cadeias principais (em valores absoluto, ponderado pela distância e ponderado pela superfície), o número de contatos não realizados, densidade na superfície e densidade interna.

Alguns descritores foram selecionados pelo teste MANOVA apenas para a identificação de um EES específico. A acessibilidade na superfície e a capacidade de fazer contatos cruzados foram usados na identificação das folhas- β , e o número de contatos não realizados foi usado na identificação das α -hélices. Concluimos que cada EES tem um nano-ambiente exclusivo, determinado por um conjunto de descritores específico.

Porém, o teste MANOVA não pôde ser aplicado em todos os casos de estruturas alinhadas pelo N-Terminal e C-Terminal, porque os dados não apresentaram uma distribuição normal, condição *sine qua non* para esse tipo de análise. Entretanto, isso abre espaço para uma discussão mais aprofundada sobre o que acontece no início e no final das α -hélices e das folhas- β , o que certamente será tema de trabalho futuro do GPBC. A perspectiva atual do GPBC é reunir todos os diferentes nano-ambientes estudados – *hot spots*, interface proteína-proteína, sítio catalítico, interface proteína-ligante, EES – para que sejam abordados, classificados e descritos em um único trabalho, gerando uma compilação de dados que poderá ser descrita como um dicionário proteico.

Conforme demonstrado no estudo de caso apresentado, onde comparamos duas estruturas idênticas, uma delas com resolução de 1.7 Å e a outra com resolução de 3.6 Å, a metodologia desenvolvida neste trabalho pode ser usada como uma nova métrica para avaliação de estruturas resolvidas ou modelos de estruturas produzidos pelos mais diversos métodos.

A integração de diversas metodologias computacionais, entre elas, o alinhamento sequencial com homólogos distantes, modelagem por homologia, predição de regiões de interação entre proteínas, DNA e outras macromoléculas, *docking*, dinâmica molecular etc., é tida como o objetivo final para um melhor entendimento dos processos que acontecem dentro

e fora da célula. Especificamente, esperamos que o conhecimento adquirido durante este trabalho ajude na importante tarefa de prever regiões contendo um EES e também de avaliar os modelos gerados, o que será útil para o desenho racional de drogas e agroquímicos e estudos bioquímicos sobre os resíduos de aminoácido importantes para o estabelecimento da função proteica.

Esse trabalho focou no estudo dos nano-ambientes onde as α -hélices e as folhas- β estão inseridas. Mas, a abordagem usada abriu o caminho para um estudo mais aprofundado do tema. Por exemplo, embora os algoritmos DSSP e Stride separem as estruturas helicoidais em α -hélices, π -hélices e hélices 3_{10} , neste trabalho nós as colocamos em um único grupo. Uma abordagem mais aprofundada, em um próximo trabalho, será estudar cada um desses tipos individualmente. O mesmo vale para as folhas- β , que poderão ser sub-divididas em folhas- β paralelas e anti-paralelas e estudadas separadamente. Finalmente, os *turns* poderão ser alvo de um novo trabalho investigativo.

Nós publicamos um artigo científico na revista PlosOne⁴¹ sobre o nano-ambiente onde as α -hélices estão presentes (anexo 1). Queremos publicar um segundo artigo, contendo as análises, resultados e conclusões sobre o nano-ambiente onde as folhas- β estão inseridas.

⁴¹ <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0200018>

6 BIBLIOGRAFIA

- ALBERTS, B.; JOHNSON, A.; LEWIS, J. **Molecular Biology of the Cell. 4th edition.** New York: Garland Science, 2002.
- ALTSCHUL, S. et al. Basic local alignment search tool. **Journal of Molecular Biology**, 1990. 403–410.
- ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, 1973. 223–230.
- ARNOLD, K. et al. The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. **Bioinformatics**, 2006. 195-201.
- ASTBURY, W. T.; WOODS, H. J. X-ray studies of the structures of hair, wool and related fibres. II. The molecular structure and elastic properties of hair keratin. **Trans. R. Soc. Lond.**, 1934. 333-394.
- BAGLEY, S. C.; ALTMAN, R. B. Characterizing the microenvironment surrounding protein sites. **Protein Science**, 4, n. 4, 1995. 622-635.
- BAKER, D.; SALI, A. Protein structure prediction and structural genomics. **Science.**, 2001. 93-6.
- BERNSTEIN, F. C. et al. The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. **Journal of Molecular Biology**, 1977. 535.
- BUCHAN, D. W. et al. Protein annotation and modelling servers at University College London. **Nucl. Acids Res.**, 2010. W563-W568.
- BUSSAB, W. D. O.; MORETTIN, P. A. **Estatística Básica 6ª ed.** 6ª. ed. [S.l.]: Saraiva, 2010. 540 p.
- CHEN, H.; GU, F.; HUANG, Z. Improved Chou-Fasman method for protein secondary structure prediction. **BMC Bioinformatics**, v. Volume 7, n. Issue SUPPL.4, p. Article number S14, December 2006.
- CHOU, P. Y.; FASMAN, G. D. Conformational parameters for amino acids in helical beta-sheet and random coil regions calculated from proteins. **Biochemistry**, 1974. 211-222.
- CHOU, P. Y.; FASMAN, G. D. Prediction of protein conformation. **Biochemistry**, 1974. 222–245.
- COLE, C.; BARBER, J. D.; BARTON, G. J. The Jpred 3 secondary structure prediction server. **Nucleic Acids Research**, 2008. 197-201.
- COOLEY, R. B.; ARP, D. J.; KARPLUS, P. A. Evolutionary origin of a secondary structure: π -helices as cryptic but widespread insertional variations of α -helices enhancing protein functionality. **J Mol Biol**, 2010. 232–246.

- CUFF, J. A.; BARTON, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. **Proteins**, 1999. 508-19.
- CUFF, J. A.; BARTON, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. **Proteins**, 2000. 502–511.
- DE MORAES, F. R. E. A. Improving predictions of protein-protein interfaces by combining amino acid-specific classifiers based on structural and physicochemical descriptors with their weighted neighbor averages. **PloS One**, 9, 2014. 87-107.
- DONGARDIVE, J.; ABRAHAM, S. Reaching optimized parameter set: protein secondary structure prediction using neural network. **Neural Computing and Applications**, p. 1-28, January 2016.
- DOR, O.; ZHOU, Y. Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training. **Proteins**, 2006. 838–45.
- FRISHMAN, D.; ARGOS, P. Knowledge-based protein secondary structure assignment. **Proteins**, 1995. 566-79.
- FRISHMAN, D.; ARGOS, P. Seventy-five percent accuracy in protein secondary structure prediction. **Proteins**, 1997. 329-35.
- GARBUZYNSKIY, S. O. et al. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? **Proteins**, 2005. 139-47.
- GARNIER, J.; GIBRAT, J. F.; ROBSON, B. GOR method for predicting protein secondary structure from amino acid sequence. **Methods Enzymol**, 1996. 540-5.
- GARNIER, J.; OSGUTHORPE, D. J.; ROBSON, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. **J Mol Biol**, 1978. 97-120.
- GASTON, M. A. et al. The complete biosynthesis of the genetically encoded amino acid pyrrolysine from lysine. **Nature**, 2011. 647–50.
- GURUPRASAD K, P. M. K. G. Analysis of gammabeta, betagamma, gammagamma, betabeta multiple turns in proteins. **J Pept Res.**, 2000. 250-63.
- HAIR, J. F. E. A. **Multivariate data analysis**. Upper Saddle River, NJ: Prentice hall, 1998.
- HERRÁE, A. Biomolecules in the computer: Jmol to the rescue. **Biochemistry and Molecular Biology Education**, 2006. 255-261.
- HOLLEY, H. L.; KARPLUS, M. Protein secondary structure prediction with a neural network. **Proc Natl Acad Sci U S A.**, 1989. 152–156.

- HUMPHREY, W.; DALKE, A.; SCHULTEN, K. VMD: visual molecular dynamics. **J Mol Graph.**, 1996. 33-8, 27-8.
- HUTCHINSON, E. G.; THORNTON, J. M. A revised set of potentials for β -turn formation in proteins. **Protein Science**, 1994. 2207–2216.
- JOHANSSON, L.; GAFVELIN, G.; AMÉR, E. S. J. Selenocysteine in Proteins — Properties and Biotechnological Use. **Biochimica et Biophysica Acta**, 2005. 1–13.
- JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. New Jersey: Prentice-Hall, 1999. 816 p.
- JONES, D. T.; TAYLOR, W. R.; THORNTON, J. M. A new approach to protein fold recognition. **Nature**, 1992. 86–89.
- KABSCH, W.; SANDER, C. How good are predictions of protein secondary structure? **FEBS Lett**, 1983. 179–82.
- KARYPIS, G. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. **Proteins**, 2006. 575-86.
- KEARSE, M. et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. **Bioinformatics**, 2012. 1647–1649.
- KING, R. D.; STERNBERG, M. J. E. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. **Protein Sci.**, 1996. 2298–2310.
- KYTE, J.; DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. **Journal of Molecular Biology**, 1982. 105–32.
- LEE B, R. F. The interpretation of protein structures: estimation of static accessibility. **Journal of Molecular Biology**, 1971. 379-400.
- LEE, B.; RICHARDS, F. M. The interpretation of protein structures: estimation of static accessibility. **Journal of Molecular Biology**, 1971. 379-400.
- LEE, J.; WU, S.; ZHANG, Y. **From Protein Structure to Function with Bioinformatics**. [S.l.]: Springer, 2009.
- LEINONEN, R. et al. UniProt archive. **Bioinformatics**, 2004. 3236–3237.
- LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, v. 22, 2006. p. 1658-1659.
- LIN, H. N. et al. Improving protein secondary structure prediction based on short subsequences with local structure similarity. **BMC Genomics.**, 2010. Suppl 4:S4.
- MACDONALD, J. R.; JOHNSON JR, W. C. Environmental features are important in determining protein secondary structure. **Protein Sci.**, Jun 2001. 1172–1177.

- MANAVALAN, P.; PONNUSWAMY, P. K. A study of the preferred environment of amino acid residues in globular proteins. **Archives of Biochemistry and Biophysics**, 1977. 476–487.
- MANCINI, A. L. E. A. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. **Bioinformatics**, 2004. 2145-7.
- MARTI-RENO, M. A. et al. Comparative protein structure modeling of genes and genomes. **Annu Rev Biophys Biomol Struct**, 2000. 291–325.
- MAZONI, I. E. A. Study of specific nanoenvironments containing α -helices in all- α and (α + β)+(α / β) proteins. **PloS One**, 13, 2018.
- NANNI, L.; BRAHNAM, S.; LUMINI, A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. **Journal of Theoretical Biology**, v. Volume 360, p. 109-116, November 2014.
- NESHICH, G. et al. Calculo de Área acessível por Solvente Utilizando SURFV - Definição de Interface Intramolecular pelo SMS. **Comunicado Técnico 36**, 2002. 1-4.
- NESHICH, G. et al. **BlueStar STING - A multiplatform environment for protein structure analysis**. [S.l.]. 2006.
- NESHICH, G.; ROCCHIA, W. Electrostatic Potential Calculation for biomolecules - creating a database of pre-calculated values reported on a per residue basis for all PDB protein structures. **Genetic Molecular Research**, 2007. 923-936.
- OLIVEIRA, S. R. M. et al. **Sting_RDB: A Relational Database of Structural Parameters for Protein Analysis**. [S.l.]. 2006.
- OLIVEIRA, S. R. M. et al. STING_RDB: A relational database of structural parameters for protein analysis with support for Data Warehousing and Data Mining. **Genetic Molecular Research**, 2007. 911-922.
- ORENGO, C.; JONES, D.; THORNTON, J. **Bioinformatics: Genes, Proteins and Computers**. 1. ed. [S.l.]: Taylor & Francis, 2003.
- PAULING, L.; COREY, R. B.; BRANSON, H. R. The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. **Proceedings of the National Academy of Sciences**, 1951. 205–211.
- PENG, J.; XU, J. RaptorX: exploiting structure information for protein alignment by statistical inference. **PROTEINS**, 2011. 161-71.
- PETERSEN, B. et al. A generic method for assignment of reliability scores applied to solvent accessibility predictions. **BMC Struct Biol.**, 2009. 51.
- PETTERSEN, E. F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. **J Comput Chem.**, 2004. 1605-12.

POLLASTRI, G. et al. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. **Protein Struct. Funct. Genet.**, 2002.

POROLLO, J.; MELLER, A. Prediction-based fingerprints of protein-protein interactions. **Proteins**, 2007. 630-645.

PROTEOPEDIA. **Proteopedia**. Disponível em:
<http://proteopedia.org/wiki/index.php/Secondary_structure>. Acesso em: 25 jun. 2018.

PROXYCHEM. ProXyChem. **ProXyChem**. Disponível em:
<http://www.proxychem.com/macromolecular_crystallography.html>. Acesso em: 18 junho 2013.

RADZICKA, A.; WOLFENDEN, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. **Biochemistry**, 1988. 1664–1670.

RAMACHANDRAN, G. N.; RAMAKRISHNAN, C.; SASISEKHARAN, V. Stereochemistry of polypeptide chain configurations. **Journal of Molecular Biology**, 1963. 95–9.

ROSE, G. D.; GIERASCH, L. M.; SMITH, J. A. Turns in peptides and proteins. **Adv Protein Chem**, 1985. 1-109.

ROST, B. PHD: predicting one-dimensional protein structure by profile based neural networks. **Methods Enzymol.**, 1996.

ROST, B. Protein secondary structure prediction continues to rise.. **J. Struct. Biol.**, 2001.

ROST, B.; SANDER, C. Combining evolutionary information and neural networks to predict protein secondary structure. **Proteins**, 1994. 55-72.

ROST, B.; YACHDAV, G.; LIU, J. The PredictProtein server. **Nucleic Acids Res.**, 2004. (Web Server issue): W321–W326.

RUSKEY, F.; WESTON, M. A survey of Venn diagrams. **Electronic Journal of Combinatorics**, v. 4, 1997. p. 3.

SALIM, J. A. E. A. Multiple structure single parameter: analysis of a single protein nano environment descriptor characterizing a shared loci on structurally aligned proteins. **Bioinformatics**, 32, 2016. 1885-1887.

SAYLE, R. A.; MILNER-WHITE, E. J. RASMOL: biomolecular graphics for all. **Trends Biochem Sci.**, 1995. 374-376.

SCHNEIDER, R.; SANDER, C. The HSSP database of protein structure-sequence alignments. **Nucleic Acids Research**, 1996. 201-5.

SCHRAUBE, H.; EISENHABER, F.; ARGOS, P. Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. **Journal of Molecular Biology**, 1993. 592-612.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, 52, 1965. 591-611.

SIBANDA, B. L.; BLUNDELL, T. L.; THORNTON, J. M. Conformation of β -hairpins in protein structures: A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. **Journal of Molecular Biology**, 1989. 759–777.

SIGRIST, C. J. A. et al. New and continuing developments at PROSITE. **Nucleic Acids Research**, 2012. 1-4.

SILBERSCHATZ, A.; SUNDARSHAN, S.; KORTH, H. F. **Sistema de banco de dados**. [S.l.]: Elsevier Brasil, 2016.

SILVEIRA, C. H. et al. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. **Proteins**, 2009. 727-43.

SMITH, L. J. et al. The concept of a random coil : Residual structure in peptides and denatured proteins. **Folding and Design**, p. R95-R106, 1996.

SRIDHARAN, S.; NICHOLLS, A.; HONIG, B. A new vertex algorithm to calculate solvent accessible surface areas. **Biophys. J**, 1992. A174.

THE PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

TONIOLO, C. Intramolecularly hydrogen-bonded peptide conformations. **CRC Crit Rev Biochem**, 1980. 1-44.

TSILO, L. C. **Protein secondary structure prediction using neural networks and support vector machines**. (Masters thesis). Rhodes University. [S.l.]. 2009.

VENKATACHALAM, C. M. Stereochemical criteria for polypeptides and proteins V Conformation of a system of three linked peptide units. **Biopolymers**, 1968. 1425–36.

VILLANUEVA, J. C. Atoms in the Universe. **Universe today**. Disponível em: <<http://www.universetoday.com/36302/atoms-in-the-universe/>>. Acesso em: 18 junho 2013.

WASSERSTEIN, R. L.; LAZAR, N. A. The ASA's statement on p-values: context, process, and purpose. **The American Statistician**, 70, 2016.

WWPDB. **World Wide PDB**. Disponível em: <<http://www.wwpdb.org/documentation/file-format-content/format23/sect5.html>>. Acesso em: 08 ago. 2018.

XIA, F. et al. Fpga accelerator for protein secondary structure prediction based on the gor algorithm. **BMC Bioinformatics**, v. Volume 12, n. Issue SUPPL. 1, p. Article number S5, February 2011.

ZHONG, L.; JOHNSON, W. C. Environment affects amino acid preference for secondary structure. **Proceedings of the National Academy of Sciences of the United States of America**, 1992. 4462-4465.

APÊNDICE A – NÚMERO DE CADEIAS COM CONSENSO ENTRE PDB_DSSP

As Tabelas 31-35 mostram o número de cadeias das estruturas do tipo all- α , α em $(\alpha+\beta)+(\alpha/\beta)$, all- β , β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada”, respectivamente, com consenso entre PDB e DSSP.

Tamanho da α -hélice	Redundância			
	100%	95%	70%	50%
5	197	63	49	43
6	250	80	61	46
7	212	67	54	41
8	221	71	57	44
9	155	55	46	37
10	223	77	56	38
11	172	68	47	36
12	164	60	43	37
13	223	82	65	54
14	190	64	51	36
15	197	58	39	28
16	111	37	26	20
17	96	30	22	19
18	138	34	21	15
19	150	39	31	22
20	46	17	12	9
21	42	21	16	14
22	42	13	12	11
23	32	11	8	5
24	60	21	14	12
25	55	12	10	10
26	20	6	5	5
27	36	9	9	9
28	35	13	7	4
29	29	13	6	3
30	18	6	4	3
31	15	4	3	2
32	38	7	5	3
33	16	4	3	3
34	1			
35	5	2	2	2
36	6	3	3	3
37	11	4	3	3
38	7	1	1	1
39	3	2	2	2
40	3	1	1	

41	6			
42	1	1	1	1
43	1	1	1	1
44	1			
45	3			
48	1	1	1	1
50	2	1	1	1
51	7			
52	2			
55	1	1	1	1
61	1			
62	3	1	1	1
64	1			
65	1			
67	1	1	1	1
71	1			
74	1	1		
108	1	1	1	1

Tabela 32. Número de cadeias das estruturas do tipo all- α com consenso entre PDB e DSSP.

Tamanho da α -hélice	Redundância			
	100%	95%	70%	50%
5	3038	570	438	343
6	3040	616	479	371
7	2750	566	463	342
8	2318	513	398	300
9	2481	528	407	299
10	2918	580	431	314
11	2705	549	403	290
12	2672	451	313	232
13	2374	402	319	235
14	1981	414	319	224
15	1454	302	215	155
16	1263	241	172	121
17	1094	206	167	114
18	1005	190	142	104
19	825	161	121	86
20	633	114	89	60
21	664	97	72	53
22	355	70	53	42
23	239	48	36	21
24	276	56	35	29
25	237	42	34	29
26	157	35	30	19
27	219	32	24	16
28	153	34	22	12
29	169	30	19	14
30	71	14	10	7
31	100	16	13	10
32	106	19	15	12
33	61	12	7	5
34	31	7	4	2
35	14	4	4	4
36	13	4	3	3
37	23	5	4	3
38	12	3	2	2
39	4	2	2	2
40	6	2	2	1
41	7	1	1	1
42	6	2	2	2
43	8	2	2	2
44	3	1	1	1
45	7	2	2	2
46	1	1		
47	4	1	1	1

48	1	1	1	1
49	2			
50	4	1	1	1
51	7			
52	2			
54	5	2	1	1
55	2	2	2	2
60	1			
61	3	1	1	1
62	3	1	1	1
64	1			
65	1			
66	2			
67	1	1	1	1
71	2			
74	1	1		
107	1	1	1	1
108	1	1	1	1
109	1	1	1	1

Tabela 33. Número de cadeias das estruturas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre PDB e DSSP.

Tamanho da folha- β	Redundância			
	100%	95%	70%	50%
5	837	376	283	211
6	867	340	255	197
7	641	245	173	140
8	520	202	164	139
9	486	176	138	108
10	328	111	86	74
11	209	54	39	28
12	67	26	21	15
13	28	14	12	10
14	23	8	7	6
15	13	5	5	4
16	10	4	3	3
17	13	3	2	2
18	12	4	3	3
19	1	1		
20	18			
22	9	1	1	

Tabela 34. Número de cadeias das estruturas do tipo all- β com consenso entre PDB e DSSP.

Tamanho da folha- β	Redundância			
	100%	95%	70%	50%
5	58294	12422	10168	8370
6	57477	11991	9658	7905
7	46504	9514	7634	6268
8	35799	7168	5746	4729
9	27231	5156	4172	3437
10	21051	3902	3043	2501
11	13531	2572	1919	1611
12	6953	1488	1252	1061
13	4409	802	657	571
14	2676	560	468	403
15	2022	372	311	271
16	1129	224	193	167
17	762	153	135	118
18	485	102	87	75
19	370	73	62	55
20	232	42	39	35
21	124	32	27	27
22	138	29	25	21
23	72	24	22	19
24	73	13	9	9
25	14	6	6	6
26	27	5	5	5
27	7	3	2	2
28	2	1	1	1
30	1			
31	1			
32	1			
33	21			
36	2	1	1	1

Tabela 35. Número de cadeias das estruturas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre PDB e DSSP.

Tamanho	Redundância			
	100%	95%	70%	50%
50	10	1	1	1
51	6	1		
52	3			
53	1			
54	15	2	2	1
55	7	1	1	1
56	25	2	2	2
57	4			
58	8	1	1	1
59	7	1		
60	7	2	1	2
61	19	4	4	2
62	4			
63	3			
64	12	2		
65	7	3	2	1
66	7	1	1	1
67	5			
68	10			
69	26	2	2	1
70	3			
71	23	2	1	1
72	9	3	3	2
73	24	1	1	
74	13			
75	10			
76	21	2	2	1
77	4			
78	21	2	2	2
79	8	1	1	1
80	16	1	1	1
81	31	3	1	1
82	7			
83	3			
84	25	2	2	1
85	5	1		
86	22	2	2	1
87	11			
88	13			
90	3	2	2	2
91	13	2	2	2
92	12	1	1	1
93	3	1	1	1

94	6	1	1	1
95	11	3	3	1
96	12	1	1	1
97	2			
98	4	1	1	1
99	9	1	1	1

Tabela 36. Número de cadeias das estruturas do desordenado com consenso entre PDB e DSSP.

APÊNDICE B – NÚMERO DE CADEIAS COM CONSENSO ENTRE PDB_Stride

As Tabelas 36-39 mostram o número de cadeias das estruturas do tipo all- α , α em $(\alpha+\beta)+(\alpha/\beta)$, all+ β , β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada”, respectivamente, com consenso entre PDB e Stride.

Tamanho da α -hélice	Redundância			
	100%	95%	70%	50%
5	295	76	65	57
6	912	173	141	113
7	629	116	98	79
8	899	189	143	119
9	598	149	122	107
10	699	154	123	97
11	640	159	130	115
12	578	155	126	113
13	576	146	117	101
14	603	133	109	91
15	816	132	108	86
16	447	113	99	84
17	444	97	80	73
18	519	92	69	61
19	485	79	63	52
20	378	74	60	46
21	258	65	53	45
22	228	54	51	42
23	195	38	28	24
24	212	50	39	37
25	230	27	23	20
26	159	30	27	25
27	151	26	21	20
28	171	35	26	18
29	215	34	28	24
30	294	28	20	17
31	154	15	14	11
32	134	15	14	11
33	114	18	13	12
34	35	8	7	5
35	35	6	6	6
36	52	10	10	10
37	21	6	4	4
38	38	8	6	6
39	40	5	4	4

40	11	2	1	1
41	11	2	2	2
42	14	5	3	4
43	22	4	4	3
44	15	5	2	2
45	11	2	2	2
46	10	3	3	2
47	5	3	3	3
48	22	4	3	3
49	15	2	2	2
50	10	2	2	2
51	26	2	2	1
52	8	3	3	3
53	7			
54	4	1	1	1
55	5	3	3	3
56	9	1	1	1
57	6	2	2	1
58	2			
59	7	3	3	2
60	2			
61	7	3	3	3
62	7	3	2	3
63	3			
64	2			
65	3	2	2	2
66	3	1	1	1
67	3	2	2	2
68	1	1	1	1
69	2	1	1	1
70	2			
71	6	3	3	2
72	3			
73	3	1	1	1
74	4	1	1	1
77	1			
78	2	1	1	1
79	1	1	1	1
81	1			
84	1	1	1	1
86	3	2	2	1
87	1	1	1	1
88	1			
89	1	1	1	1
90	1			

93	1	1	1	1
96	2			
99	1			
103	2			
104	1			
106	1			
108	2	1	1	1
109	1			
111	1			
116	1			
120	1	1	1	1
122	2			
143	1			
151	1			
281	1	1	1	1
282	3			

Tabela 37. Número de cadeias das estruturas do tipo all- α com consenso entre PDB e Stride.

Tamanho da α -hélice	Redundância			
	100%	95%	70%	50%
5	7253	1174	910	779
6	17333	2449	2004	1682
7	13610	1962	1703	1404
8	11529	1724	1430	1195
9	11286	1577	1340	1134
10	12590	1722	1448	1179
11	12419	1788	1494	1242
12	11294	1616	1351	1146
13	9786	1253	1096	918
14	9527	1329	1119	933
15	8064	1093	934	764
16	6281	902	779	656
17	5321	718	617	508
18	4589	645	528	442
19	4256	534	435	362
20	3345	385	322	269
21	3276	343	282	233
22	2122	281	248	212
23	1249	180	152	128
24	1278	182	147	127
25	1156	139	124	98
26	971	127	105	91
27	865	116	92	84
28	693	95	75	55
29	758	101	80	67
30	700	80	65	54
31	578	63	60	50
32	419	59	52	45
33	393	51	39	31
34	165	30	25	17
35	214	26	25	20
36	186	22	22	20
37	111	24	17	14
38	144	24	20	19
39	119	19	16	16
40	54	7	6	5
41	58	9	9	8
42	54	8	6	7
43	83	10	10	8
44	43	9	5	4
45	27	7	6	6
46	19	2	2	2
47	7	3	3	3

48	38	8	7	7
49	19	2	2	2
50	22	5	5	4
51	38	6	3	2
52	21	5	5	5
53	16	2	2	2
54	28	7	6	6
55	27	10	10	10
56	22	8	5	5
57	13	3	3	2
58	6	1	1	1
59	8	3	3	2
60	7	1	1	1
61	14	3	3	3
62	21	4	4	4
63	9	1	1	1
64	7			
65	6	2	2	2
66	11	2	2	2
67	4	2	2	2
68	4	3	3	2
69	5	1	1	1
70	4			
71	9	3	3	2
72	5			
73	5	1	1	1
74	7	1	1	1
75	1	1	1	1
77	2			
78	2	1	1	1
79	3	2	2	2
80	1			
81	2			
82	2			
83	2	1	1	1
84	1	1	1	1
86	6	2	2	2
87	2	1	1	1
88	1			
89	4	1	1	1
90	1			
91	2			
92	2			
93	1	1	1	1
96	2			

97	1			
99	1			
100	1			
101	2			
102	1			
103	3			
104	2			
105	1			
106	1			
107	3	3	3	3
108	3	1	1	1
109	3	1	1	1
110	2			
111	3			
115	1	1	1	1
116	1			
120	1	1	1	1
122	3	1	1	1
126	1	1	1	1
143	1			
149	2			
150	2			
151	3			
152	1			
158	4			
166	4			
196	1			
220	1			
281	1	1	1	1
282	3			

Tabela 38. Número de cadeias das estruturas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre PDB e Stride.

Tamanho da folha- β	Redundância			
	100%	95%	70%	50%
5	1213	364	278	210
6	1511	366	274	218
7	1159	254	176	143
8	929	208	160	133
9	888	174	134	112
10	581	110	82	68
11	448	65	47	36
12	127	38	27	22
13	35	11	8	8
14	25	7	5	5
15	22	4	4	4
16	15	1	1	1
17	23	5	5	4
18	24	5	5	5
19	4	3	2	2
20	44	1	1	1
21	2			
22	29	1	1	
24	1	1	1	1

Tabela 39. Número de cadeias das estruturas do tipo all- β com consenso entre PDB e Stride.

Tamanho da folha- β	Redundância			
	100%	95%	70%	50%
5	128524	18319	14733	12124
6	126931	17664	13937	11467
7	102148	13896	10951	9031
8	77260	10577	8261	6797
9	58635	7262	5784	4770
10	44111	5635	4170	3404
11	27813	3862	2700	2231
12	14500	2075	1704	1440
13	9390	1074	899	764
14	5834	710	591	516
15	4814	484	398	343
16	2694	273	228	196
17	1510	197	168	142
18	864	145	127	110
19	870	106	85	69
20	385	70	63	57
21	293	38	29	29
22	215	28	23	17
23	254	27	24	20
24	174	20	15	13
25	40	10	9	9
26	31	9	9	8
27	13	4	3	2
28	5	1	1	1
29	6			
30	2			
31	1			
32	8			
33	22			
36	2	1	1	1

Tabela 40. Número de cadeias das estruturas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre PDB e Stride.

Tamanho	Redundância			
	100%	95%	70%	50%
50	10	1	1	1
51	6	1	2	1
52	2	2		
53	1			
54	13			
55	6	1	1	1
56	23	1	1	1
57	1			
58	8	1	1	1
59	2	1		
60	6	1		
61	15	2	2	1
62	4			
63	3			
64	11	2		
65	7	3	2	1
66	7	1	1	1
67	5			
68	10			
69	26	2	2	1
70	2			
71	23	2	1	1
72	9	3	3	2
73	24	1	1	
74	13			
75	10			
76	21	2	2	1
77	4			
78	21	2	2	2
79	8	1	1	
80	16	1	1	1
81	31	3	1	1
82	7			
83	2			
84	24	2	2	1
85	5			
86	22	1		
87	11			
88	12	1	1	
90	3	2	2	2
91	13	2	2	2
92	11	1	1	1
93	2			

94	6	1	1	1
95	10	2	2	1
96	11	1	1	1
97	1			
98	3			
99	9	1	1	1

Tabela 41. Número de cadeias das estruturas do tipo desordenado com consenso entre PDB e Stride.

APÊNDICE C – NÚMERO DE CADEIAS COM CONSENSO ENTRE DSSP_Stride

As Tabelas 41-45 mostram o número de cadeias das estruturas do tipo all- α , α em $(\alpha+\beta)+(\alpha/\beta)$, all+ β , β em $(\alpha+\beta)+(\alpha/\beta)$ e “desordenada”, respectivamente, com consenso entre DSSP e Stride.

Tamanho da α -hélice	Redundância			
	100%	95%	70%	50%
5	649	206	173	149
6	1597	349	277	231
7	1038	330	287	242
8	1204	393	334	297
9	756	248	215	187
10	991	347	282	237
11	1050	370	306	252
12	731	264	219	191
13	942	306	249	205
14	1033	337	276	241
15	1485	329	261	216
16	717	217	185	154
17	725	214	174	155
18	833	236	175	149
19	688	160	113	94
20	416	99	84	70
21	396	122	94	85
22	426	111	91	80
23	251	89	78	68
24	290	89	57	49
25	571	80	70	64
26	181	46	40	38
27	174	58	54	46
28	315	81	59	53
29	179	44	37	31
30	132	30	23	19
31	142	37	30	23
32	232	24	16	15
33	83	10	9	7
34	30	14	12	9
35	35	14	12	10
36	22	10	7	6
37	33	13	12	11
38	26	3	3	2
39	17	4	3	2

40	12	4	4	3
41	5	4	1	1
42	12	5	3	3
43	11	8	7	6
44	11	3	3	2
45	5			
46	3	1		
47	4	3	3	3
48	5	1	1	1
49	12	2	2	1
50	5	2	2	2
51	11	1	1	1
52	2		1	
53	3	1	1	
54	1	1	1	1
55	1	1	1	1
56	1	1	1	
58	3	1	3	
59	5	3	1	3
60	1			
61	2	1	2	1
62	5	2	1	1
63	3	1	1	1
64	2			
65	1	1	2	1
66	1			
67	7	2	1	1
69	1	1	1	
72	1	1		
76	1	1	1	1
77	1			
78	1			
79	1			
85	3	1	1	1
99	1	1	1	1
100	1	1	1	1
103	3	1	1	1
108	2	1	1	1
109	3	1	1	1
123	1			
134	1	1	1	1
325	1	1	1	1

Tabela 42. Número de cadeias das estruturas do tipo all- α com consenso entre DSSP e Stride.

Tamanho da α -hélice	Redundância			
	100%	95%	70%	50%
5	14018	2974	2518	2090
6	18886	3859	3267	2710
7	18218	3895	3384	2828
8	15184	3386	2912	2481
9	14108	3100	2679	2233
10	19043	4221	3587	2951
11	18603	4022	3413	2778
12	15814	2989	2487	2021
13	13466	2832	2431	2002
14	15511	3378	2901	2396
15	11428	2298	1930	1592
16	7586	1564	1331	1108
17	7994	1550	1332	1104
18	7038	1475	1258	1036
19	5014	921	751	607
20	4169	734	613	504
21	3884	673	559	469
22	2841	577	494	399
23	1584	344	307	258
24	2086	373	300	253
25	1772	296	259	220
26	1191	213	173	148
27	994	175	153	117
28	1002	164	128	108
29	926	139	114	96
30	663	106	87	73
31	579	105	88	70
32	654	94	81	71
33	270	45	35	30
34	200	44	34	23
35	124	33	28	26
36	99	23	18	16
37	68	19	15	13
38	61	14	13	10
39	66	14	12	9
40	37	13	12	10
41	18	10	7	7
42	32	12	9	9
43	37	15	12	11
44	33	10	7	4
45	14	4	3	3
46	9	4	3	3
47	14	8	8	8

48	10	2	2	2
49	21	5	4	3
50	8	4	4	4
51	38	4	3	3
52	40	3	1	1
53	8	2	2	1
54	15	5	4	4
55	15	1	1	1
56	3	2	2	1
57	1			
58	4	2	2	1
59	6	3	3	3
60	7			
61	5	2	2	2
62	5	2	2	1
63	6	2	2	2
64	3			
65	1	1	1	1
66	3			
67	10	3	3	2
68	3	2	1	1
69	3	1		
70	2	1	1	1
71	4	1	1	1
72	1	1	1	
74	1			
75	1			
76	2	2	2	2
77	1			
78	1			
79	1			
82	5	1		
85	3	1	1	1
90	1	1		
92	1	1	1	
99	1	1	1	1
100	1	1	1	1
101	1			
102	2	1		
103	3	1	1	1
107	2	1	1	1
108	2	1	1	1
109	5	2	2	2
111	1			
123	1			

134	1	1	1	1
154	1			
167	3	1		
325	1	1	1	1

Tabela 43. Número de cadeias das estruturas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre DSSP e Stride.

Tamanho da folha- β	Redundância			
	100%	95%	70%	50%
5	595	260	198	148
6	732	258	197	150
7	565	189	134	105
8	422	140	112	89
9	384	118	89	69
10	251	78	55	44
11	199	43	30	22
12	52	20	15	11
13	16	9	6	6
14	12	6	5	4
15	7	3	3	3
16	5	1	1	1
17	9	2	2	2
18	7	1	1	1
19	1	1		
20	23	3	3	1
22	22	4	4	1

Tabela 44. Número de cadeias das estruturas do tipo all- β com consenso entre DSSP e Stride.

Tamanho da folha- β	Redundância			
	100%	95%	70%	50%
5	54059	11198	9141	7510
6	53610	10892	8725	7129
7	42200	8482	6755	5516
8	32655	6419	5090	4153
9	24026	4443	3580	2918
10	19187	3386	2563	2088
11	11999	2265	1640	1369
12	6178	1304	1087	918
13	3865	681	548	475
14	2364	474	396	333
15	1961	324	271	235
16	1008	179	155	131
17	645	133	112	95
18	410	93	73	59
19	373	71	54	44
20	172	37	36	33
21	114	27	23	23
22	129	23	20	15
23	88	23	21	19
24	61	9	8	7
25	15	8	8	8
26	16	6	6	5
27	10	3	2	2
28	8	2	2	2
30	1			
31	1			
32	1			
33	21	1	1	1
36	2	1	1	1

Tabela 45. Número de cadeias das estruturas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$ com consenso entre DSSP e Stride.

Tamanho	Redundância			
	100%	95%	70%	50%
50	10	1	1	1
51	6	1		
52	2			
53	2	1	1	1
54	13	2	2	1
55	6	1	1	1
56	23	1	1	1
57	2			
58	8	1	1	1
59	2	1		
60	6	1		
61	15	2	2	1
62	4			
63	3			
64	11	2		
65	7	3	2	1
66	7	1	1	1
67	5			
68	10			
69	26	2	2	1
70	2			
71	23	2	1	1
72	9	3	3	2
73	24	1	1	
74	13			
75	10			
76	21	2	2	1
77	4			
78	21	2	2	2
79	8	1	1	1
80	16	1	1	1
81	32	4	2	2
82	7			
83	2			
84	24	2	2	1
85	5			
86	22	1		
87	11			
88	13	1	1	
90	3	2	2	2
91	13	2	2	2
92	11	1	1	1
93	2			

94	6	1	1	1
95	10	2	2	1
96	12	1	1	1
97	1			
98	3			
99	9	1	1	1

Tabela 46. Número de cadeias das estruturas do tipo desordenado com consenso entre DSSP e Stride.

APÊNDICE D – DADOS NORMALIZANDOS PELO ICV

Ao prepararmos os dados para o teste MANOVA, nós normalizamos os dados pelo seu ICV (Inverse of Coefficient of Variation). A normalização dos dados foi necessária dada a sua natureza diversa. Por exemplo, os valores do descritor relativo ao potencial eletrostático no carbono- α varia dentro de uma faixa entre -1000 e +1500, enquanto o número de contatos do tipo ligações de hidrogênio entre cadeias principais varia entre 0 e 32. Para exemplificar a dificuldade de trabalhar com os dados brutos, a Fig. 70 introduz um gráfico do tipo radar mostrando a diferença entre as médias de cada um dos 69 descritores presentes do STING_RDB usados neste trabalho, dentro e fora da α -hélice de tamanho 17. É praticamente impossível ver as diferenças, dada a grande amplitude dos valores.

Dados brutos de todos os descritores

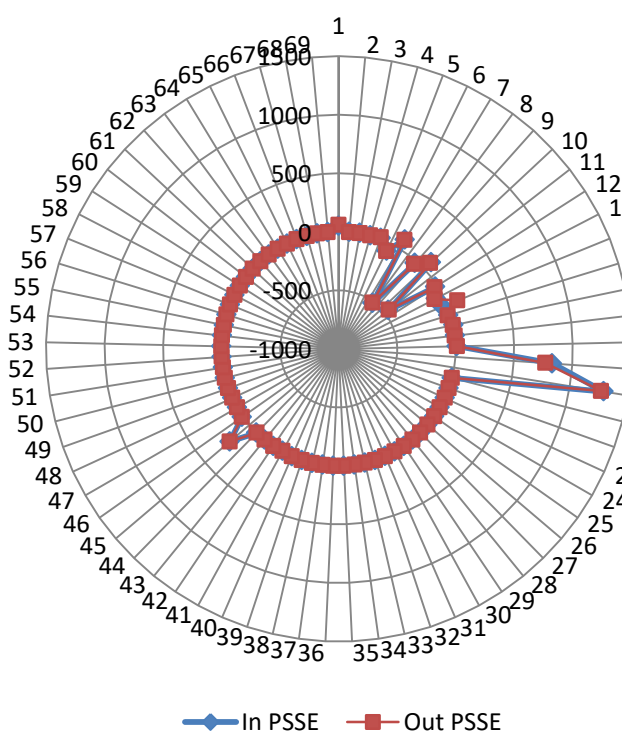


Figura 70. Gráfico do tipo radar mostrando a diferença entre os valores médios de 69 descritores do STING_RDB dentro e fora da α -hélice de tamanho 17. Quando trabalhamos com os dados brutos, é praticamente impossível ver as diferenças, dada a grande amplitude dos valores.

Na Fig. 71 mostramos o mesmo gráfico, mas apenas para aqueles descritores cuja diferença entre os valores médios e fora da α -hélice seja maior que 1, e na Fig. 72 mostramos os descritores com diferença entre 0,1 e 1. Ainda assim é praticamente impossível ver as diferenças.

Dados brutos com diferença > 1

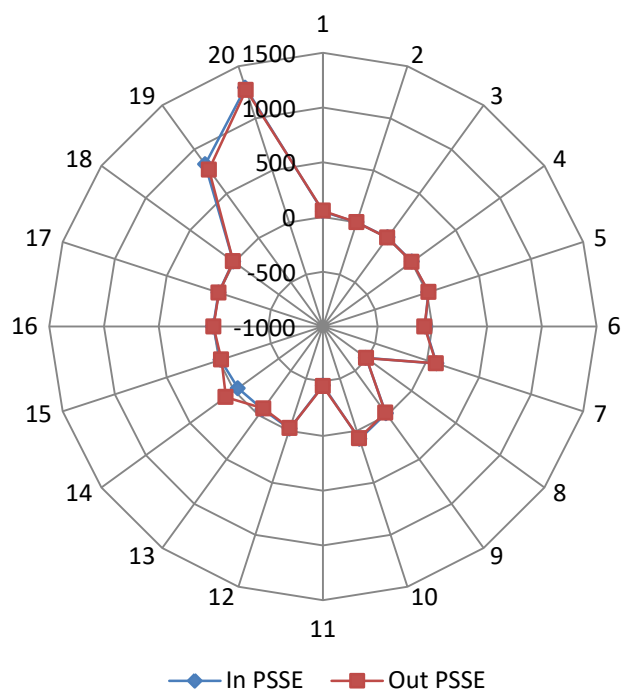


Figura 71. Gráfico do tipo radar mostrando os descritores com diferença maior que 1 entre os valores brutos médios dentro e fora da α -hélice de tamanho 17.

Dados brutos com diferença entre 0,1 e 1

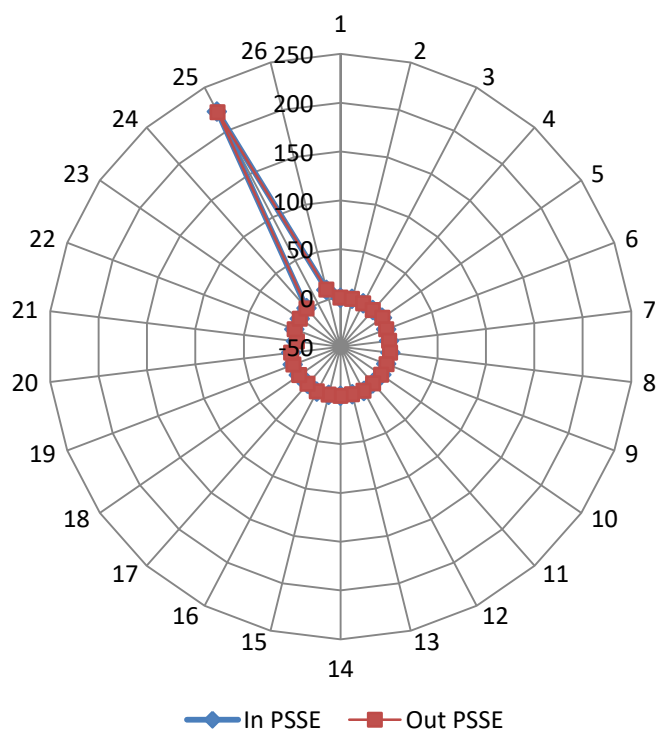


Figura 72. Gráfico do tipo radar mostrando os descritores com diferença entre 0,1 e 1 entre os valores brutos médios dentro e fora da α -hélice de tamanho 17.

Quando trabalhamos com dados tão dispersos, a solução é normalizá-los. Um dos métodos usados para normalizar os dados é o coeficiente de variância (CV), dado pela Eq. 26.

$$CV = \frac{\sigma}{\mu} \quad (26)$$

onde σ é o desvio padrão e μ é a média.

A Fig. 73 mostra o gráfico do tipo radar indicando a diferença entre os valores normalizados pelo CV de cada um dos 69 descritores presentes do STING_RDB usados neste trabalho, dentro e fora da α -hélice. Exceto para três descritores, é praticamente impossível ver alguma diferença entre eles.

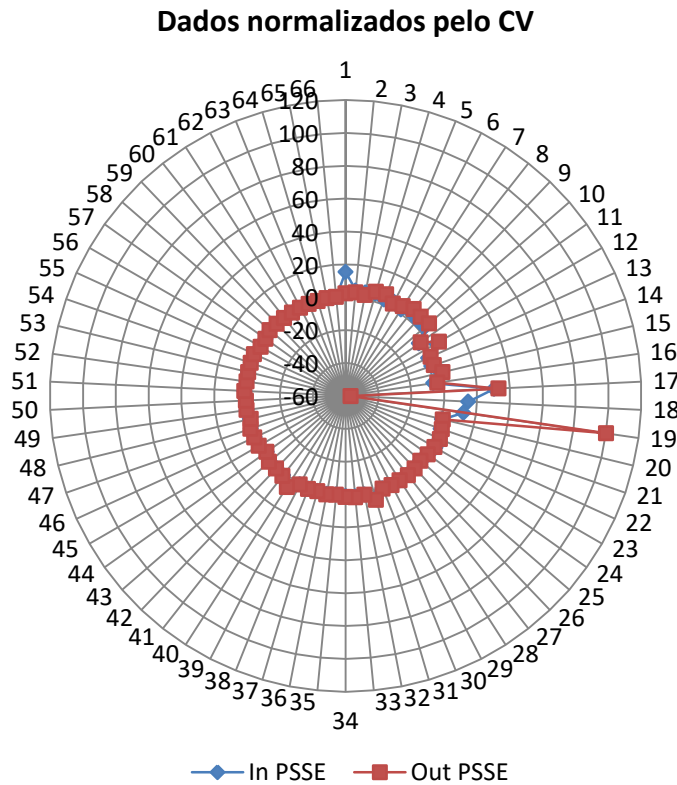


Figura 73. Gráfico do tipo radar mostrando a diferença entre os valores normalizados pelo CV de 69 descritores do STING_RDB dentro e fora da α -hélice de tamanho 17.

Na Fig. 74 mostramos o mesmo gráfico, mas apenas para aqueles descritores cuja diferença entre os valores normalizados pelo CV dentro e fora da α -hélice seja maior que 1, e na Fig. 75 mostramos os descritores com diferença entre 0,1 e 1.

Dados normalizados pelo CV com diferença > 1

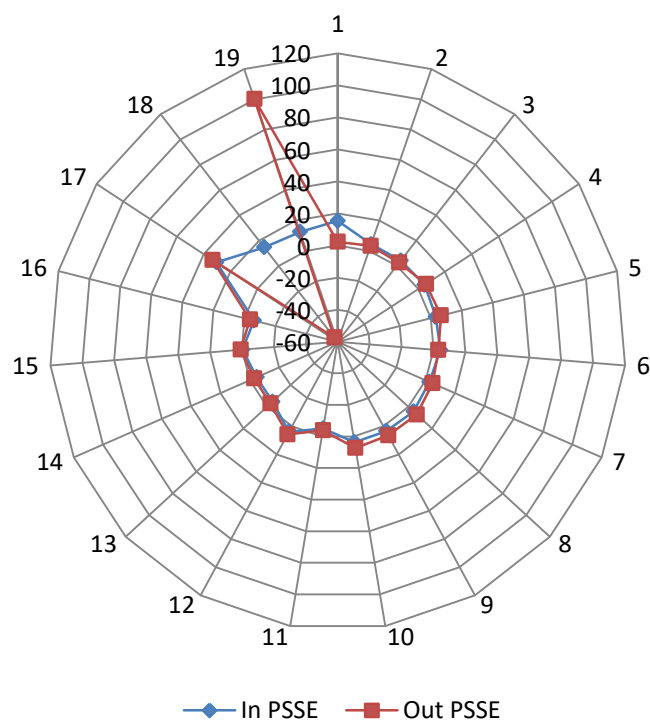


Figura 74. Gráfico do tipo radar mostrando os descritores com diferença maior que 1 entre os valores normalizados pelo CV dentro e fora da α -hélice de tamanho 17.

Dados normalizados pelo CV com diferença entre 0,1 e 1

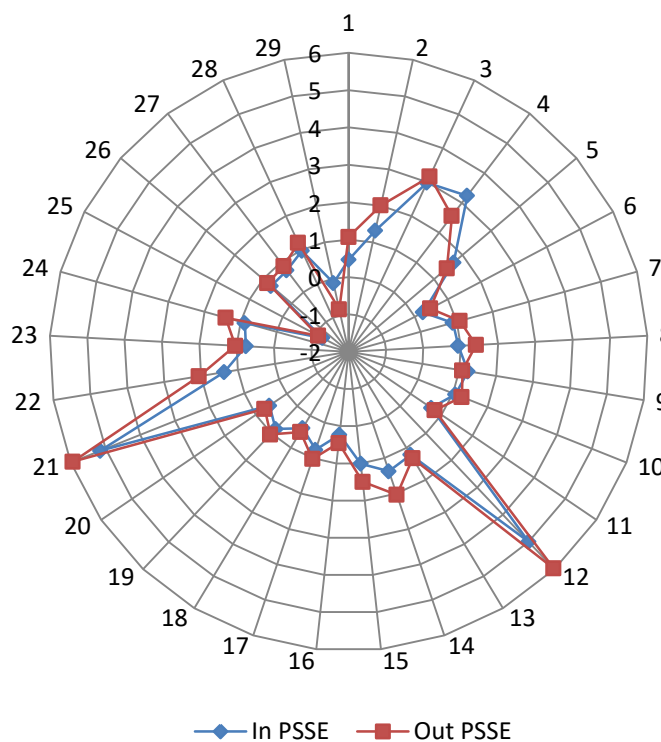


Figura 75. Gráfico do tipo radar mostrando os descritores com diferença entre 0,1 e 1 entre os valores normalizados pelo CV dentro e fora da α -hélice de tamanho 17.

Normalizando os dados pelo CV conseguimos uma melhor visualização do que acontece dentro e fora da α -hélice. Para aprimorar esse resultado, trabalhamos com o inverso do coeficiente de variância (ICV) dado pela Eq. 27.

$$ICV = \frac{\mu}{\sigma} \quad (27)$$

onde σ é o desvio padrão e μ é a média.

A Fig. 76 mostra o gráfico do tipo radar indicando a diferença entre os valores normalizados pelo ICV de cada um dos 69 descritores presentes do STING_RDB usados neste trabalho, dentro e fora da α -hélice. Normalizando os dados pelo ICV observamos melhor as diferenças que nos casos anteriores. Por esse motivo usamos a normalização pelo ICV neste trabalho.

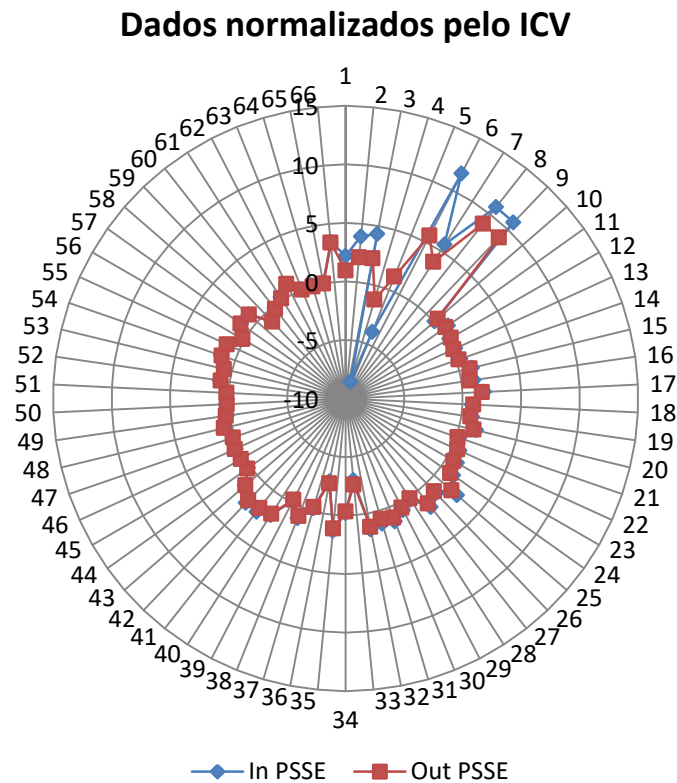


Figura 76. Gráfico do tipo radar mostrando a diferença entre os valores normalizados pelo ICV de 69 descritores do STING_RDB dentro e fora da α -hélice de tamanho 17.

Na Fig. 77 mostramos o mesmo gráfico, mas apenas para aqueles descritores cuja diferença entre os valores normalizados pelo ICV dentro e fora da α -hélice seja maior que 1, e na Fig. 89 mostramos os descritores com diferença entre 0,1 e 1.

Dados normalizados pelo ICV com diferença > 1

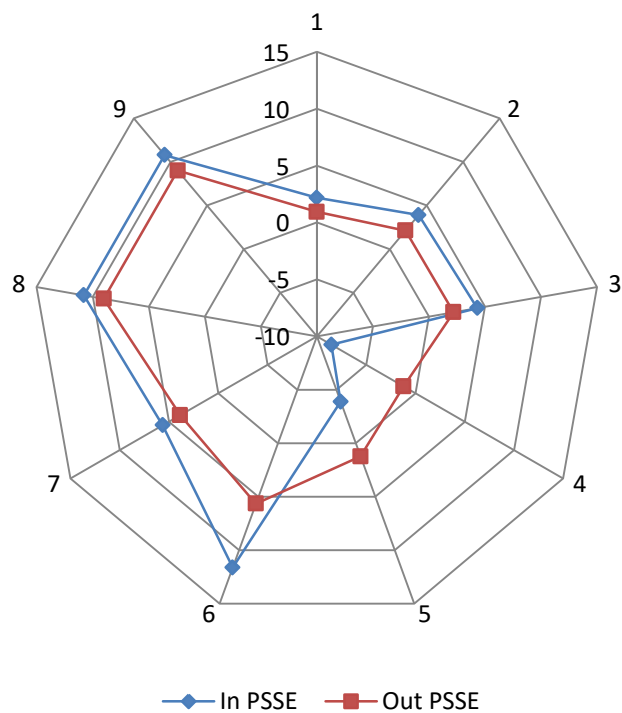


Figura 77. Gráfico do tipo radar mostrando os descritores com diferença maior que 1 entre os valores normalizados pelo ICV dentro e fora da α -hélice de tamanho 17.

Dados normalizados pelo ICV com diferença entre 0,1 e 1

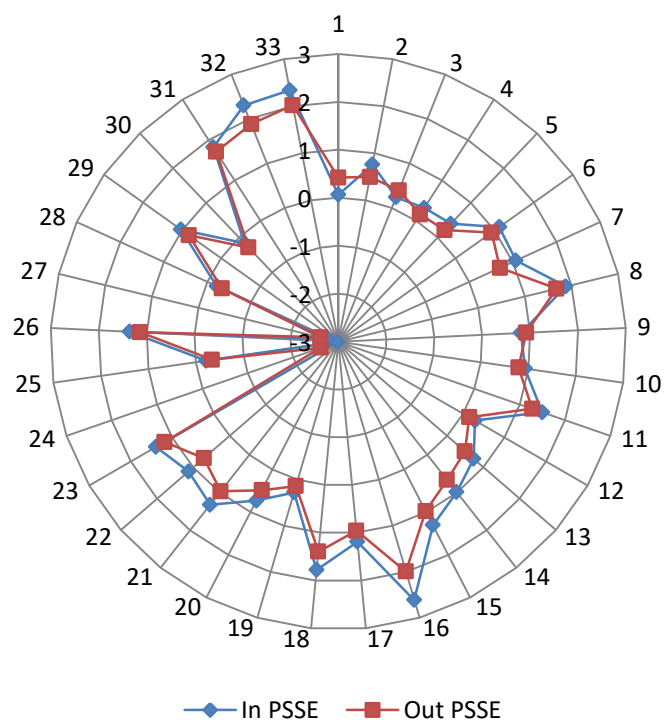


Figura 78. Gráfico do tipo radar mostrando os descritores com diferença entre 0,1 e 1 entre os valores normalizados pelo ICV dentro e fora da α -hélice de tamanho 17.

APÊNDICE E – DESCRITORES USADOS NO TESTE MANOVA PARA AS ESTRUTURAS ALINHADAS POR TAMANHO

Tabela 47. Descritores usados no teste MANOVA para as proteínas do tipo all- α alinhadas por tamanho.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
2	hbmm	2,0955	1,090416	1,6933	1,172711	1,921744	1,443919	VERDADEIRO	FALSO
16	hbmm_WNADist	7,851924	2,475322	6,704244	2,694063	3,172082	2,488525	VERDADEIRO	FALSO
30	hbmm_WNASurf	11,30999	3,851	9,871572	3,724929	2,936897	2,650137	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	23,60101	17,88983	22,57976	21,40013	1,319242	1,055123	VERDADEIRO	FALSO
63	Number_Unused_Contact	214,9774	71,3479	215,4237	68,99295	3,013087	3,122402	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	798,4927	131,8234	773,3958	138,5938	6,057291	5,580306	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1305,773	294,5074	1254,379	297,7674	4,433753	4,212614	VERDADEIRO	FALSO
66	IFR_CA_3	1,0734	0,170898	1,063233	0,196456	6,280928	5,412059	VERDADEIRO	FALSO
67	Internal_CA_3	1,09168	0,097717	1,087713	0,112947	11,17185	9,630296	VERDADEIRO	VERDADEIRO
1	Accessible_Surface_in_Isolation	56,55733	47,62713	59,85746	49,80647	1,187502	1,201801	FALSO	FALSO
3	hbmwm	0,0223	0,16429	0,0404	0,221615	0,135736	0,182298	FALSO	FALSO
4	hbmwwm	0,0096	0,111254	0,02	0,16254	0,086289	0,123047	FALSO	FALSO
5	hbms	0,2261	0,509443	0,2768	0,570807	0,443818	0,484927	FALSO	FALSO
6	hbmws	0,0477	0,245336	0,0633	0,287928	0,194427	0,219847	FALSO	FALSO
7	hbmwws	0,0306	0,21241	0,04	0,243055	0,144061	0,164572	FALSO	FALSO
8	hbss	0,0651	0,281965	0,0734	0,295175	0,23088	0,248666	FALSO	FALSO
9	hbsws	0,0269	0,190701	0,0271	0,190638	0,141058	0,142154	FALSO	FALSO
10	hbswws	0,014	0,14736	0,0124	0,131605	0,095006	0,094222	FALSO	FALSO
11	hydrophobic	1,4897	2,425879	1,465	2,320116	0,614087	0,631434	FALSO	FALSO
12	aromatic	0,0958	0,356412	0,1065	0,389537	0,26879	0,273402	FALSO	FALSO
13	disulfide	0,0049	0,070112	0,004	0,063115	0,069889	0,063376	FALSO	FALSO
14	ch_attractive	0,4696	1,308878	0,4384	1,28876	0,358781	0,340172	FALSO	FALSO
15	ch_repulsive	0,1566	0,736097	0,1771	0,81295	0,212744	0,217849	FALSO	FALSO
17	hbmwm_WNADist	0,120151	0,277506	0,162925	0,344983	0,432966	0,47227	FALSO	FALSO
18	hbmwwm_WNADist	0,060925	0,195563	0,084748	0,264248	0,311538	0,320715	FALSO	FALSO
19	hbms_WNADist	1,030321	0,801844	1,110962	0,854042	1,28494	1,300828	FALSO	FALSO

20	hbmws_WNADist	0,235126	0,411841	0,263565	0,453652	0,570915	0,580985	FALSO	FALSO
21	hbmws_WNADist	0,154047	0,374034	0,184374	0,428385	0,411852	0,430394	FALSO	FALSO
22	hbss_WNADist	0,433345	0,508889	0,438383	0,509411	0,851551	0,860567	FALSO	FALSO
23	hbsws_WNADist	0,026038	0,187877	0,026374	0,188314	0,138591	0,140054	FALSO	FALSO
24	hbswws_WNADist	0,013943	0,146992	0,012264	0,129469	0,094859	0,094729	FALSO	FALSO
25	hydrophobic_WNADist	0,199513	0,911664	0,134467	0,756009	0,218844	0,177865	FALSO	FALSO
26	aromatic_WNADist	0,474327	0,568129	0,473549	0,585473	0,834893	0,808831	FALSO	FALSO
27	disulfide_WNADist	0,01428	0,098647	0,013296	0,089811	0,144756	0,148045	FALSO	FALSO
28	ch_attractive_WNADist	2,190002	2,068598	2,052457	1,989171	1,058689	1,031815	FALSO	FALSO
29	ch_repulsive_WNADist	0,808357	1,160636	0,803818	1,234636	0,696478	0,651057	FALSO	FALSO
31	hbmwm_WNASurf	0,209125	0,405414	0,252686	0,445771	0,515831	0,56685	FALSO	FALSO
32	hbmwwm_WNASurf	0,121412	0,331822	0,144316	0,362887	0,365896	0,397689	FALSO	FALSO
33	hbms_WNASurf	1,535469	1,002782	1,566999	0,992566	1,531209	1,578735	FALSO	FALSO
34	hbmws_WNASurf	0,372372	0,561342	0,404497	0,595983	0,663361	0,678706	FALSO	FALSO
35	hbmwws_WNASurf	0,25526	0,524245	0,294008	0,574232	0,486911	0,512001	FALSO	FALSO
36	hbss_WNASurf	0,645489	0,661255	0,645003	0,646206	0,976157	0,998139	FALSO	FALSO
37	hbsws_WNASurf	0,00399	0,034167	0,003934	0,033846	0,116768	0,116234	FALSO	FALSO
38	hbswws_WNASurf	0,00227	0,026929	0,002052	0,025917	0,084299	0,079179	FALSO	FALSO
39	hydrophobic_WNASurf	0,098588	0,459949	0,068844	0,368585	0,214345	0,18678	FALSO	FALSO
40	aromatic_WNASurf	0,714601	0,665186	0,679367	0,642178	1,074288	1,057911	FALSO	FALSO
41	disulfide_WNASurf	0,013795	0,096524	0,012966	0,088492	0,142914	0,146524	FALSO	FALSO
42	ch_attractive_WNASurf	3,447251	2,523386	3,330695	2,490209	1,366121	1,337516	FALSO	FALSO
43	ch_repulsive_WNASurf	1,381761	1,496134	1,326822	1,487663	0,923554	0,891883	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-40,0995	85,25718	-40,8913	83,69997	-0,47034	-0,48855	FALSO	FALSO
46	Electrostatic_Potential_Average	8,120212	25,43093	6,525534	24,59127	0,319305	0,26536	FALSO	FALSO
47	Hydrophobicity_KDI	0,803541	1,93816	0,64669	1,890155	0,41459	0,342136	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	178,6357	83,89572	167,146	80,9448	2,129259	2,064938	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-359,122	236,9835	-356,138	226,4408	-1,51539	-1,57277	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	63,89277	109,2546	52,81607	102,1078	0,584806	0,517258	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	157,2217	76,56196	147,169	72,70412	2,053522	2,024218	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-323,128	216,4877	-320,387	206,4151	-1,49259	-1,55215	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	56,65359	97,26713	47,17335	90,81268	0,582454	0,519458	FALSO	FALSO

54	Cross_Link_Order_CA	0,3512	0,54279	0,3496	0,541958	0,647027	0,645068	FALSO	FALSO
55	Cross_Pres_Order_CA	0,5672	0,767956	0,59	0,731197	0,738584	0,806896	FALSO	FALSO
58	Dihedral_Chi1	-32,8362	102,0925	-30,7888	97,92079	-0,32163	-0,31443	FALSO	FALSO
59	Dihedral_Chi2	13,48328	101,5674	11,11088	96,32393	0,132752	0,115349	FALSO	FALSO
60	Dihedral_Chi3	-0,76988	57,8007	-0,72793	55,71696	-0,01332	-0,01306	FALSO	FALSO
61	Dihedral_Chi4	0,964474	49,37612	0,480226	48,20007	0,019533	0,009963	FALSO	FALSO
62	Temperature_Factor_CA	23,71194	21,03282	26,88271	24,88351	1,127378	1,080342	FALSO	FALSO
68	Clash	0,301117	0,141301	0,309258	0,15116	2,131026	2,045904	FALSO	FALSO
69	Percent	9,640462	4,456254	9,892538	4,747194	2,163356	2,083871	FALSO	FALSO

Tabela 48. Descritores usados no teste MANOVA para as proteínas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$ alinhadas por tamanho.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
2	hbmm	2,0425	1,072306	1,3957	1,107224	1,904773	1,26054	VERDADEIRO	FALSO
16	hbmm_WNADist	6,782647	2,217584	5,359259	2,283529	3,058575	2,34692	VERDADEIRO	FALSO
30	hbmm_WNASurf	8,599748	3,228016	7,415587	3,133372	2,664097	2,366647	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	22,7936	18,34243	21,10244	23,75457	1,242671	0,888353	VERDADEIRO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	154,8529	72,45506	148,5785	73,30509	2,137227	2,026851	VERDADEIRO	FALSO
63	Number_Unused_Contact	214,1908	71,34466	211,2271	67,68161	3,002198	3,120894	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	771,267	120,8162	747,8056	126,3221	6,383804	5,919832	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1178,753	273,8127	1144,239	279,9179	4,304961	4,087766	VERDADEIRO	FALSO
66	IFR_CA_3	1,012674	0,288753	1,011083	0,297501	3,50706	3,398587	VERDADEIRO	FALSO
67	Internal_CA_3	1,086651	0,09070542	1,087142	0,109988	11,98	9,884223	VERDADEIRO	VERDADEIRO
1	Accessible_Surface_in_Isolation	47,60289	46,54097	48,4396	47,37958	1,022817	1,022373	FALSO	FALSO
3	hbmwm	0,0352	0,2151173	0,0832	0,331257	0,163632	0,251165	FALSO	FALSO
4	hbmwwm	0,0161	0,1468381	0,038	0,233867	0,109645	0,162486	FALSO	FALSO
5	hbms	0,3048	0,5980801	0,3507	0,654471	0,509631	0,535853	FALSO	FALSO
6	hbmws	0,0929	0,3647464	0,121	0,415501	0,254698	0,291215	FALSO	FALSO
7	hbmwws	0,0605	0,3088014	0,0718	0,342443	0,195919	0,20967	FALSO	FALSO
8	hbss	0,0936	0,3377713	0,1026	0,350368	0,277111	0,292835	FALSO	FALSO
9	hbsws	0,0364	0,2201145	0,0392	0,229484	0,165368	0,170818	FALSO	FALSO
10	hbswws	0,0243	0,198732	0,0231	0,190866	0,122275	0,121028	FALSO	FALSO
11	hydrophobic	1,5808	2,224305	1,6458	2,28514	0,710694	0,720218	FALSO	FALSO
12	aromatic	0,1295	0,441858	0,121	0,429197	0,293081	0,281922	FALSO	FALSO
13	disulfide	0,0041	0,06477905	0,0044	0,066393	0,063292	0,066272	FALSO	FALSO
14	ch_attractive	0,5078	1,392176	0,4207	1,294649	0,364753	0,324953	FALSO	FALSO
15	ch_repulsive	0,1665	0,756407	0,1676	0,823524	0,22012	0,203516	FALSO	FALSO
17	hbmwm_WNADist	0,212849	0,3583631	0,310948	0,503393	0,593948	0,617705	FALSO	FALSO
18	hbmwwm_WNADist	0,1128336	0,261915	0,158899	0,37024	0,430802	0,429179	FALSO	FALSO
19	hbms_WNADist	1,287963	0,8790433	1,359902	0,939842	1,465187	1,446948	FALSO	FALSO
20	hbmws_WNADist	0,4142159	0,5814544	0,470766	0,648567	0,712379	0,725855	FALSO	FALSO
21	hbmwws_WNADist	0,2901288	0,5433304	0,307805	0,582286	0,533982	0,528615	FALSO	FALSO
22	hbss_WNADist	0,5769034	0,5858587	0,593046	0,602836	0,984714	0,98376	FALSO	FALSO

23	hbsws_WNADist	0,0357391	0,2182394	0,038432	0,227176	0,163761	0,169172	FALSO	FALSO
24	hbswws_WNADist	0,02320058	0,1931043	0,022279	0,186906	0,120145	0,119199	FALSO	FALSO
25	hydrophobic_WNADist	0,1350287	0,7187238	0,114722	0,675699	0,187873	0,169782	FALSO	FALSO
26	aromatic_WNADist	0,539248	0,6147269	0,514405	0,592156	0,877216	0,868698	FALSO	FALSO
27	disulfide_WNADist	0,01238375	0,08687734	0,012858	0,087403	0,142543	0,147112	FALSO	FALSO
28	ch_attractive_WNADist	2,359794	2,122175	2,10437	1,992604	1,11197	1,05609	FALSO	FALSO
29	ch_repulsive_WNADist	0,7965036	1,190893	0,766736	1,222329	0,668829	0,627275	FALSO	FALSO
31	hbmwm_WNASurf	0,3385108	0,4866085	0,405257	0,54981	0,695653	0,737085	FALSO	FALSO
32	hbmwwm_WNASurf	0,1872131	0,3762873	0,222	0,428793	0,497527	0,517732	FALSO	FALSO
33	hbms_WNASurf	1,71274	0,9602777	1,750564	0,959974	1,783588	1,823553	FALSO	FALSO
34	hbmws_WNASurf	0,5814027	0,7009154	0,620055	0,725828	0,829491	0,854272	FALSO	FALSO
35	hbmwws_WNASurf	0,411197	0,6839469	0,42072	0,692139	0,601212	0,607854	FALSO	FALSO
36	hbss_WNASurf	0,7889587	0,7242586	0,80006	0,729483	1,089333	1,09675	FALSO	FALSO
37	hbsws_WNASurf	0,004535897	0,03537568	0,004765	0,03708	0,128221	0,12852	FALSO	FALSO
38	hbswws_WNASurf	0,003032479	0,03115784	0,002856	0,030956	0,097326	0,092275	FALSO	FALSO
39	hydrophobic_WNASurf	0,06318883	0,3209432	0,056669	0,333078	0,196885	0,170137	FALSO	FALSO
40	aromatic_WNASurf	0,685097	0,5946861	0,661202	0,57859	1,152031	1,142783	FALSO	FALSO
41	disulfide_WNASurf	0,01214096	0,08583128	0,012545	0,08631	0,141451	0,145343	FALSO	FALSO
42	ch_attractive_WNASurf	3,491642	2,493432	3,25598	2,446694	1,400336	1,330767	FALSO	FALSO
43	ch_repulsive_WNASurf	1,220591	1,476887	1,160654	1,42251	0,826462	0,81592	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-44,68282	82,92191	-40,4824	81,36903	-0,53885	-0,49752	FALSO	FALSO
46	Electrostatic_Potential_Average	6,277278	24,26471	4,4428	23,9046	0,2587	0,185855	FALSO	FALSO
47	Hydrophobicity_KDI	-40,6413	643,1156	-36,591	609,8763	-0,06319	-0,06	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-365,6663	194,6614	-346,266	191,2862	-1,87847	-1,8102	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	39,75937	88,17166	34,2243	85,43303	0,450931	0,400598	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	132,3629	65,62262	127,7737	65,09475	2,017032	1,962888	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-322,5129	177,7014	-306,358	173,9178	-1,81491	-1,76151	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	33,57809	75,67011	29,76642	73,03404	0,443743	0,407569	FALSO	FALSO
54	Cross_Link_Order_CA	0,4621	0,6287716	0,5211	0,681625	0,734925	0,764496	FALSO	FALSO
55	Cross_Pres_Order_CA	0,8377	0,8790702	0,9491	0,927464	0,952939	1,023329	FALSO	FALSO
58	Dihedral_Chi1	-30,36802	105,664	-25,0972	100,0503	-0,2874	-0,25085	FALSO	FALSO
59	Dihedral_Chi2	13,65043	106,275	9,250374	98,32611	0,128444	0,094079	FALSO	FALSO

60	Dihedral_Chi3	0,2095477	64,84969	0,787666	61,3295	0,003231	0,012843	FALSO	FALSO
61	Dihedral_Chi4	1,808171	56,48557	1,23701	54,64261	0,032011	0,022638	FALSO	FALSO
62	Temperature_Factor_CA	24,21341	17,79483	25,93252	19,83228	1,360699	1,307591	FALSO	FALSO
68	Clash	0,2912168	0,1326537	0,298464	0,134679	2,195316	2,216114	FALSO	FALSO
69	Percent	9,37301	4,214345	9,602438	4,280925	2,224073	2,243076	FALSO	FALSO

Tabela 49. Descritores usados no teste MANOVA para as proteínas do tipo all- β alinhadas por tamanho.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
1	Accessible_Surface_in_Isolation	40,67865	40,94155	58,60037	48,96966	0,993579	1,196667	VERDADEIRO	FALSO
2	hbmm	1,3758	0,913769	0,9745	0,916944	1,505632	1,06277	VERDADEIRO	FALSO
5	hbms	0,1661	0,468566	0,2745	0,588903	0,354486	0,466121	VERDADEIRO	FALSO
11	hydrophobic	2,3384	2,95344	1,7591	2,588398	0,791755	0,67961	VERDADEIRO	FALSO
16	hbmm_WNADist	4,845132	1,618116	3,82545	1,724322	2,994304	2,218524	VERDADEIRO	FALSO
29	ch_repulsive_WNADist	0,597694	0,970049	0,582066	1,183458	0,616148	0,491835	VERDADEIRO	FALSO
30	hbmm_WNASurf	5,8368	2,075982	4,894067	2,026811	2,811585	2,414664	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	15,54668	18,31868	16,87411	23,40835	0,848679	0,720859	VERDADEIRO	FALSO
54	Cross_Link_Order_CA	0,5477	0,629672	0,4319	0,591498	0,869818	0,73018	VERDADEIRO	FALSO
55	Cross_Pres_Order_CA	0,9362	0,81459	0,7934	0,798054	1,149289	0,994169	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	820,5959	131,7607	761,7621	141,0027	6,227926	5,402465	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1336,855	303,8565	1205,885	301,1754	4,399626	4,003929	VERDADEIRO	FALSO
66	IFR_CA_3	1,109882	0,114232	-1,22E+33	6,45E+35	9,716043	-0,00189	VERDADEIRO	VERDADEIRO
67	Internal_CA_3	1,110373	0,111766	-1,22E+33	6,45E+35	9,934846	-0,00189	VERDADEIRO	VERDADEIRO
3	hbmwm	0,0482	0,258879	0,0717	0,3077	0,186188	0,233019	FALSO	FALSO
4	hbmwwm	0,0381	0,261437	0,0437	0,263539	0,145733	0,16582	FALSO	FALSO
6	hbmws	0,0948	0,369283	0,0959	0,36307	0,256714	0,264137	FALSO	FALSO
7	hbmwws	0,0621	0,315289	0,0665	0,33485	0,196962	0,198596	FALSO	FALSO
8	hbss	0,0912	0,329273	0,0932	0,334043	0,276974	0,279006	FALSO	FALSO
9	hbsws	0,0421	0,235691	0,0353	0,216326	0,178623	0,16318	FALSO	FALSO
10	hbswws	0,029	0,22481	0,0242	0,197187	0,128998	0,122726	FALSO	FALSO
12	aromatic	0,1052	0,370167	0,0823	0,331991	0,284196	0,247898	FALSO	FALSO
13	disulfide	0,0197	0,138988	0,0118	0,108558	0,141739	0,108697	FALSO	FALSO
14	ch_attractive	0,299	1,079781	0,2969	1,04071	0,276908	0,285286	FALSO	FALSO
15	ch_repulsive	0,1004	0,590188	0,1284	0,748642	0,170115	0,171511	FALSO	FALSO
17	hbmwm_WNADist	0,261102	0,437394	0,275687	0,48784	0,59695	0,565116	FALSO	FALSO
18	hbmwwm_WNADist	0,204933	0,465565	0,191446	0,451396	0,440181	0,424119	FALSO	FALSO
19	hbms_WNADist	0,949847	0,791237	1,077711	0,925512	1,200458	1,164449	FALSO	FALSO
20	hbmws_WNADist	0,488776	0,715458	0,434358	0,673485	0,683166	0,644942	FALSO	FALSO
21	hbmwws_WNADist	0,345844	0,678931	0,309621	0,644805	0,509395	0,480178	FALSO	FALSO

22	hbss_WNADist	0,701142	0,736521	0,628952	0,691852	0,951964	0,909085	FALSO	FALSO
23	hbsws_WNADist	0,042744	0,23854	0,03576	0,216505	0,179191	0,165168	FALSO	FALSO
24	hbswws_WNADist	0,029035	0,22517	0,024426	0,197098	0,128947	0,123927	FALSO	FALSO
25	hydrophobic_WNADist	0,196655	0,802803	0,145406	0,707014	0,24496	0,205663	FALSO	FALSO
26	aromatic_WNADist	0,454689	0,549992	0,380004	0,500059	0,826719	0,759918	FALSO	FALSO
27	disulfide_WNADist	0,05356	0,213253	0,036621	0,167616	0,251156	0,218484	FALSO	FALSO
28	ch_attractive_WNADist	1,665835	1,744181	1,471509	1,661458	0,955081	0,885673	FALSO	FALSO
31	hbmwm_WNASurf	0,457093	0,64211	0,415734	0,60621	0,71186	0,685793	FALSO	FALSO
32	hbmwwm_WNASurf	0,361572	0,711574	0,310307	0,627909	0,50813	0,494192	FALSO	FALSO
33	hbms_WNASurf	1,512896	1,025232	1,491896	1,031414	1,475662	1,446457	FALSO	FALSO
34	hbmws_WNASurf	0,749419	0,995532	0,640909	0,888295	0,752783	0,721505	FALSO	FALSO
35	hbmwws_WNASurf	0,577453	1,065894	0,494396	0,949034	0,541755	0,520946	FALSO	FALSO
36	hbss_WNASurf	1,029109	1,02771	0,915758	0,952338	1,001361	0,961589	FALSO	FALSO
37	hbsws_WNASurf	0,00511	0,03663	0,005603	0,041245	0,139515	0,135839	FALSO	FALSO
38	hbswws_WNASurf	0,003844	0,035937	0,004553	0,043854	0,106956	0,103824	FALSO	FALSO
39	hydrophobic_WNASurf	0,08715	0,355688	0,071107	0,355931	0,245019	0,199776	FALSO	FALSO
40	aromatic_WNASurf	0,605171	0,585853	0,524758	0,549271	1,032974	0,955372	FALSO	FALSO
41	disulfide_WNASurf	0,052284	0,209694	0,035833	0,165113	0,249333	0,217019	FALSO	FALSO
42	ch_attractive_WNASurf	2,647608	2,187473	2,295539	2,042426	1,21035	1,123928	FALSO	FALSO
43	ch_repulsive_WNASurf	0,999518	1,353811	0,8881	1,308682	0,7383	0,678622	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-37,424	72,11327	-41,4597	80,35753	-0,51896	-0,51594	FALSO	FALSO
46	Electrostatic_Potential_Average	2,744763	22,68089	1,723383	23,4673	0,121017	0,073438	FALSO	FALSO
47	Hydrophobicity_KDI	-426,366	2023,177	-372,67	1895,557	-0,21074	-0,1966	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	141,008	87,2624	130,3073	81,01153	1,615908	1,608503	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-360,802	220,5149	-330,4	210,9728	-1,63618	-1,56608	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	26,74929	103,882	23,01929	93,87179	0,257497	0,245221	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	127,6166	77,84525	116,8063	71,24717	1,639363	1,639452	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-328,954	200,6587	-298,105	188,542	-1,63937	-1,58111	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	24,57617	92,78291	21,72095	83,39706	0,264878	0,260452	FALSO	FALSO
58	Dihedral_Chi1	-22,1363	104,7608	-23,4341	93,79279	-0,2113	-0,24985	FALSO	FALSO
59	Dihedral_Chi2	15,21616	95,3692	8,980143	90,88869	0,15955	0,098804	FALSO	FALSO
60	Dihedral_Chi3	-1,09626	48,80444	-0,25653	50,30314	-0,02246	-0,0051	FALSO	FALSO

61	Dihedral_Chi4	-1,0467	42,32345	-0,30379	43,16386	-0,02473	-0,00704	FALSO	FALSO
62	Temperature_Factor_CA	21,09098	17,3552	23,86839	19,96993	1,215254	1,195217	FALSO	FALSO
63	Number_Unused_Contact	205,9449	66,76257	212,4156	66,71395	3,084736	3,183976	FALSO	FALSO
68	Clash	0,292666	0,137283	0,304032	0,14101	2,131849	2,156104	FALSO	FALSO
69	Percent	9,389321	4,251208	9,761346	4,4236	2,208624	2,206652	FALSO	FALSO

Tabela 50. Descritores usados no teste MANOVA para as proteínas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$ alinhadas por tamanho.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
1	Accessible_Surface_in_Isolation	27,81939	36,21611	49,88189	47,87905	0,76815	1,041831	VERDADEIRO	FALSO
2	hbmm	1,4542	0,84157	1,2335	1,029867	1,72796	1,197727	VERDADEIRO	FALSO
5	hbms	0,2022	0,518769	0,3374	0,645799	0,389769	0,522454	VERDADEIRO	FALSO
11	hydrophobic	2,3178	2,638848	1,7117	2,35127	0,878338	0,72799	VERDADEIRO	FALSO
16	hbmm_WNADist	5,110569	1,492174	4,554705	1,929715	3,424915	2,360299	VERDADEIRO	VERDADEIRO
30	hbmm_WNASurf	6,703313	2,451395	6,045882	2,480116	2,734489	2,437742	VERDADEIRO	FALSO
40	aromatic_WNASurf	0,746841	0,611001	0,649957	0,582444	1,222323	1,115914	VERDADEIRO	FALSO
42	ch_attractive_WNASurf	3,584317	2,414112	3,140658	2,296884	1,484735	1,367356	VERDADEIRO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-408,753	176,4778	-377,763	171,6608	-2,31617	-2,20064	VERDADEIRO	FALSO
54	Cross_Link_Order_CA	0,7551	0,742335	0,5472	0,680748	1,017195	0,803821	VERDADEIRO	FALSO
55	Cross_Pres_Order_CA	1,2934	0,940517	1,0149	0,91498	1,375202	1,109204	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	789,9434	133,4024	748,2125	129,2142	5,921508	5,790482	VERDADEIRO	FALSO
66	IFR_CA_3	1,091015	0,150621	-3,27E+32	3,34E+35	7,243445	-0,00098	VERDADEIRO	VERDADEIRO
67	Internal_CA_3	1,102145	0,102619	-3,27E+32	3,34E+35	10,74019	-0,00098	VERDADEIRO	VERDADEIRO
68	Clash	0,288586	0,133035	0,302444	0,153223	2,16925	1,973883	VERDADEIRO	FALSO
69	Percent	9,306347	4,140321	9,739571	4,783548	2,247736	2,036056	VERDADEIRO	FALSO
3	hbmwm	0,07	0,308347	0,096	0,355358	0,227017	0,27015	FALSO	FALSO
4	hbmwwm	0,0373	0,236798	0,0507	0,271363	0,157518	0,186835	FALSO	FALSO
6	hbmws	0,1283	0,432838	0,1365	0,440711	0,296416	0,309727	FALSO	FALSO
7	hbmwws	0,0694	0,342384	0,0864	0,379647	0,202696	0,22758	FALSO	FALSO
8	hbss	0,1192	0,381258	0,1034	0,351686	0,312649	0,294013	FALSO	FALSO
9	hbsws	0,0516	0,265379	0,044	0,244013	0,194439	0,180318	FALSO	FALSO
10	hbswws	0,0339	0,237788	0,0313	0,225877	0,142564	0,138571	FALSO	FALSO
12	aromatic	0,1432	0,463648	0,1162	0,420224	0,308855	0,276519	FALSO	FALSO
13	disulfide	0,0075	0,08764	0,0051	0,072093	0,085578	0,070742	FALSO	FALSO
14	ch_attractive	0,411	1,322761	0,4307	1,301735	0,310714	0,330866	FALSO	FALSO
15	ch_repulsive	0,1554	0,789072	0,1637	0,773116	0,19694	0,211741	FALSO	FALSO
17	hbmwm_WNADist	0,348962	0,483441	0,362823	0,543093	0,721829	0,668068	FALSO	FALSO
18	hbmwwm_WNADist	0,209592	0,396513	0,211257	0,436723	0,528588	0,483731	FALSO	FALSO
19	hbms_WNADist	1,138916	0,791113	1,283784	0,932202	1,439638	1,377153	FALSO	FALSO

20	hbmws_WNADist	0,598771	0,732224	0,554282	0,723101	0,817742	0,766534	FALSO	FALSO
21	hbmws_WNADist	0,374804	0,628935	0,371674	0,659739	0,595934	0,563365	FALSO	FALSO
22	hbss_WNADist	0,75438	0,700058	0,654321	0,661773	1,077597	0,988739	FALSO	FALSO
23	hbsws_WNADist	0,050562	0,261891	0,042718	0,239827	0,193064	0,178121	FALSO	FALSO
24	hbswws_WNADist	0,033082	0,234666	0,030238	0,221482	0,140973	0,136526	FALSO	FALSO
25	hydrophobic_WNADist	0,1364	0,662741	0,158015	0,739099	0,205812	0,213795	FALSO	FALSO
26	aromatic_WNADist	0,584957	0,634535	0,503711	0,602278	0,921867	0,836344	FALSO	FALSO
27	disulfide_WNADist	0,016889	0,109158	0,013253	0,092035	0,154722	0,144	FALSO	FALSO
28	ch_attractive_WNADist	2,261866	1,957846	2,0605	1,946311	1,155283	1,058669	FALSO	FALSO
29	ch_repulsive_WNADist	0,786716	1,114423	0,740224	1,104832	0,70594	0,669988	FALSO	FALSO
31	hbmwm_WNASurf	0,555685	0,655066	0,51451	0,635859	0,848289	0,809158	FALSO	FALSO
32	hbmwwm_WNASurf	0,343486	0,576095	0,314494	0,549335	0,596232	0,5725	FALSO	FALSO
33	hbms_WNASurf	1,722338	0,975416	1,655085	0,947292	1,765747	1,747175	FALSO	FALSO
34	hbmws_WNASurf	0,850623	0,919303	0,751109	0,850613	0,925291	0,883021	FALSO	FALSO
35	hbmws_WNASurf	0,580539	0,88248	0,534186	0,839467	0,657849	0,636339	FALSO	FALSO
36	hbss_WNASurf	1,045169	0,922695	0,904352	0,860606	1,132735	1,050831	FALSO	FALSO
37	hbsws_WNASurf	0,004345	0,032078	0,00551	0,040393	0,135447	0,136407	FALSO	FALSO
38	hbswws_WNASurf	0,003278	0,031411	0,004307	0,039829	0,104358	0,108144	FALSO	FALSO
39	hydrophobic_WNASurf	0,058132	0,283252	0,076591	0,357198	0,20523	0,214423	FALSO	FALSO
41	disulfide_WNASurf	0,016596	0,108146	0,012945	0,091016	0,15346	0,142225	FALSO	FALSO
43	ch_repulsive_WNASurf	1,205061	1,260656	1,096867	1,218939	0,9559	0,899854	FALSO	FALSO
44	Electrostatic_Potential_at_CA	13,33038	19,30034	17,85716	23,63549	0,690681	0,755523	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-40,1234	71,43723	-45,5613	79,98771	-0,56166	-0,5696	FALSO	FALSO
46	Electrostatic_Potential_Average	-0,44165	21,05112	0,600701	21,55142	-0,02098	0,027873	FALSO	FALSO
47	Hydrophobicity_KDI	-231,009	1506,97	-224,529	1483,747	-0,15329	-0,15133	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	130,3593	68,69905	127,1079	64,15326	1,897542	1,981316	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	6,024714	67,64426	7,092919	62,32602	0,089065	0,113803	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	113,5698	62,81815	107,7195	57,93455	1,807914	1,859331	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-358,1	173,6462	-327,096	162,8992	-2,06224	-2,00796	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	6,373591	56,29735	6,42425	51,68093	0,113213	0,124306	FALSO	FALSO
58	Dihedral_Chi1	-22,0829	108,0285	-24,115	96,01022	-0,20442	-0,25117	FALSO	FALSO
59	Dihedral_Chi2	15,38284	98,17565	9,897514	95,57007	0,156687	0,103563	FALSO	FALSO

60	Dihedral_Chi3	-0,31297	51,19099	-0,0589	56,36926	-0,00611	-0,00104	FALSO	FALSO
61	Dihedral_Chi4	0,182491	43,69003	0,454073	47,80853	0,004177	0,009498	FALSO	FALSO
62	Temperature_Factor_CA	31,23069	27,76589	35,07972	30,13796	1,124786	1,163971	FALSO	FALSO
63	Number_Unused_Contact	203,3434	66,9847	212,3724	68,0724	3,035669	3,119802	FALSO	FALSO
65	Number_Unused_Contact_WNASurf	1256,388	310,7496	1145,093	278,0957	4,043088	4,117622	FALSO	FALSO

APÊNDICE F – DESCRITORES USADOS NO TESTE MANOVA PARA AS ESTRUTURAS ALINHADAS PELO C-TERMINAL

Tabela 51. Descritores usados no teste MANOVA para as proteínas do tipo all- α alinhadas pelo C-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
2	hbmm	2,0817	1,086718	1,6837	1,179488	1,915584	1,427484	VERDADEIRO	FALSO
16	hbmm_WNADist	7,849432	2,476509	6,695954	2,717837	3,169555	2,463707	VERDADEIRO	FALSO
30	hbmm_WNASurf	11,30977	3,854332	9,817754	3,740423	2,934301	2,624771	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	23,60779	17,9068	22,71091	22,23224	1,31837	1,02153	VERDADEIRO	FALSO
63	Number_Unused_Contact	215,0143	71,34895	214,8538	68,54851	3,013559	3,134332	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	798,4014	131,8307	771,0919	137,3853	6,056263	5,612623	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1305,735	294,5402	1248,913	293,6662	4,43313	4,252832	VERDADEIRO	FALSO
66	IFR_CA_3	1,073363	0,171104	1,063424	0,197578	6,273154	5,382302	VERDADEIRO	FALSO
67	Internal_CA_3	1,091707	0,097732	1,088719	0,11131	11,17046	9,780989	VERDADEIRO	VERDADEIRO
68	Clash	0,303171	0,143516	0,312656	0,155406	2,11246	2,01186	VERDADEIRO	FALSO
1	Accessible_Surface_in_Isolation	56,58149	47,64067	60,70818	49,92135	1,187672	1,216076	FALSO	FALSO
3	hbmwm	0,0236	0,169225	0,0409	0,223919	0,139459	0,182655	FALSO	FALSO
4	hbmwwm	0,0103	0,115837	0,0198	0,162661	0,088918	0,121726	FALSO	FALSO
5	hbms	0,2258	0,509417	0,2693	0,566601	0,443252	0,475291	FALSO	FALSO
6	hbmws	0,048	0,245362	0,0614	0,283838	0,195629	0,21632	FALSO	FALSO
7	hbmwws	0,0319	0,215944	0,0417	0,247853	0,147723	0,168245	FALSO	FALSO
8	hbss	0,0645	0,280177	0,0728	0,295809	0,230211	0,246105	FALSO	FALSO
9	hbsws	0,0264	0,188126	0,0259	0,190528	0,140331	0,135938	FALSO	FALSO
10	hbswws	0,0141	0,146675	0,0124	0,131956	0,096131	0,093971	FALSO	FALSO
11	hydrophobic	1,4868	2,435966	1,4404	2,307577	0,610353	0,624205	FALSO	FALSO
12	aromatic	0,0997	0,369026	0,1078	0,396362	0,270171	0,271974	FALSO	FALSO
13	disulfide	0,0052	0,071835	0,0043	0,065682	0,072389	0,065467	FALSO	FALSO
14	ch_attractive	0,4707	1,314161	0,4027	1,236028	0,358175	0,325802	FALSO	FALSO
15	ch_repulsive	0,158	0,735715	0,1695	0,814691	0,214757	0,208054	FALSO	FALSO
17	hbmwm_WNADist	0,120583	0,277911	0,16581	0,350564	0,43389	0,47298	FALSO	FALSO
18	hbmwwm_WNADist	0,061144	0,19588	0,083696	0,2637	0,312152	0,317391	FALSO	FALSO
19	hbms_WNADist	1,031051	0,802606	1,105834	0,850665	1,28463	1,299964	FALSO	FALSO

20	hbmws_WNADist	0,235971	0,412338	0,258779	0,451	0,572276	0,57379	FALSO	FALSO
21	hbmwws_WNADist	0,1546	0,374591	0,18489	0,426984	0,412717	0,433015	FALSO	FALSO
22	hbss_WNADist	0,432474	0,508715	0,436019	0,509274	0,850131	0,856158	FALSO	FALSO
23	hbsws_WNADist	0,026132	0,188208	0,025938	0,190708	0,138845	0,136011	FALSO	FALSO
24	hbswws_WNADist	0,013994	0,147254	0,012435	0,132021	0,09503	0,09419	FALSO	FALSO
25	hydrophobic_WNADist	0,199234	0,912301	0,124782	0,742717	0,218386	0,168008	FALSO	FALSO
26	aromatic_WNADist	0,473756	0,567867	0,468511	0,582981	0,834273	0,803648	FALSO	FALSO
27	disulfide_WNADist	0,014331	0,09882	0,013276	0,09109	0,145021	0,145749	FALSO	FALSO
28	ch_attractive_WNADist	2,191425	2,070406	1,979813	1,939995	1,058452	1,020525	FALSO	FALSO
29	ch_repulsive_WNADist	0,809093	1,160489	0,798537	1,281647	0,6972	0,623055	FALSO	FALSO
31	hbmwm_WNASurf	0,209877	0,405948	0,257876	0,452968	0,517005	0,569303	FALSO	FALSO
32	hbmwwm_WNASurf	0,121849	0,332338	0,139103	0,355817	0,366642	0,39094	FALSO	FALSO
33	hbms_WNASurf	1,535803	1,003452	1,566793	0,999564	1,53052	1,567477	FALSO	FALSO
34	hbmws_WNASurf	0,373711	0,561905	0,393638	0,57949	0,665079	0,679283	FALSO	FALSO
35	hbmwws_WNASurf	0,256178	0,524962	0,292213	0,57403	0,487993	0,509056	FALSO	FALSO
36	hbss_WNASurf	0,644632	0,661332	0,642006	0,649317	0,974747	0,988741	FALSO	FALSO
37	hbsws_WNASurf	0,004004	0,034228	0,004013	0,034798	0,116981	0,115334	FALSO	FALSO
38	hbswws_WNASurf	0,002278	0,026977	0,002106	0,026948	0,084452	0,078149	FALSO	FALSO
39	hydrophobic_WNASurf	0,098574	0,460556	0,06386	0,360836	0,214033	0,176978	FALSO	FALSO
40	aromatic_WNASurf	0,714023	0,665642	0,672157	0,631856	1,072683	1,063782	FALSO	FALSO
41	disulfide_WNASurf	0,013844	0,096694	0,012971	0,089959	0,143175	0,144187	FALSO	FALSO
42	ch_attractive_WNASurf	3,448534	2,524805	3,258845	2,430415	1,365862	1,340859	FALSO	FALSO
43	ch_repulsive_WNASurf	1,383128	1,496311	1,321717	1,529242	0,924359	0,864296	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-40,0807	85,30796	-39,4283	84,37881	-0,46984	-0,46728	FALSO	FALSO
46	Electrostatic_Potential_Average	8,128178	25,45987	6,621376	24,80411	0,319254	0,266947	FALSO	FALSO
47	Hydrophobicity_KDI	0,80265	1,938156	0,620618	1,886926	0,414131	0,328904	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	178,6901	83,99995	166,1971	81,87131	2,127264	2,02998	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-358,729	237,1681	-350,35	230,2806	-1,51255	-1,52141	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	64,00583	109,4277	52,99029	103,8351	0,584914	0,510331	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	157,2637	76,65705	146,1242	73,47708	2,051523	1,988705	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-322,741	216,6331	-315,882	209,6559	-1,48981	-1,50667	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	56,75664	97,41951	47,19351	92,39982	0,5826	0,510753	FALSO	FALSO

54	Cross_Link_Order_CA	0,3512	0,542716	0,3373	0,538995	0,647116	0,625795	FALSO	FALSO
55	Cross_Pres_Order_CA	0,5663	0,767359	0,5744	0,720319	0,737986	0,797425	FALSO	FALSO
58	Dihedral_Chi1	-32,7762	102,1104	-30,5818	98,71895	-0,32099	-0,30979	FALSO	FALSO
59	Dihedral_Chi2	13,4907	101,6147	11,41776	97,04204	0,132763	0,117658	FALSO	FALSO
60	Dihedral_Chi3	-0,80747	57,85162	-0,44865	57,27718	-0,01396	-0,00783	FALSO	FALSO
61	Dihedral_Chi4	0,941906	49,39605	0,553872	49,97833	0,019068	0,011082	FALSO	FALSO
62	Temperature_Factor_CA	23,76137	21,05137	27,0592	24,97406	1,128733	1,083492	FALSO	FALSO
69	Percent	9,703429	4,523631	10,00001	4,87848	2,145053	2,049821	FALSO	FALSO

Tabela 52. Descritores usados no teste MANOVA para as proteínas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$ alinhadas pelo C-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
2	hbmm	2,0425	1,072306	1,3921	1,107404	1,904773	1,257084	VERDADEIRO	FALSO
16	hbmm_WNADist	6,771627	2,20258	5,326684	2,265906	3,074407	2,350797	VERDADEIRO	FALSO
30	hbmm_WNASurf	8,579823	3,21083	7,341915	3,073118	2,672151	2,389077	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	22,53927	18,28647	21,00106	24,0852	1,232565	0,871949	VERDADEIRO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	153,6058	71,9611	146,373	72,92121	2,134567	2,007276	VERDADEIRO	FALSO
63	Number_Unused_Contact	213,9414	71,41042	210,978	67,6746	2,995941	3,117536	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	770,3679	120,8618	745,55	126,2002	6,373957	5,907677	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1176,873	273,6518	1137,858	278,5236	4,300622	4,08532	VERDADEIRO	FALSO
66	IFR_CA_3	1,009236	0,293645	1,007666	0,303793	3,436922	3,316955	VERDADEIRO	FALSO
67	Internal_CA_3	1,086205	0,090931	1,087601	0,109879	11,94536	9,898206	VERDADEIRO	VERDADEIRO
1	Accessible_Surface_in_Isolation	47,14606	46,34009	48,33775	47,54012	1,017392	1,016778	FALSO	FALSO
3	hbmwm	0,0352	0,215117	0,0834	0,332634	0,163632	0,250726	FALSO	FALSO
4	hbmwwm	0,0161	0,146838	0,0374	0,230342	0,109645	0,162367	FALSO	FALSO
5	hbms	0,3048	0,59808	0,3517	0,656026	0,509631	0,536107	FALSO	FALSO
6	hbmws	0,0929	0,364746	0,1206	0,414457	0,254698	0,290983	FALSO	FALSO
7	hbmwws	0,0605	0,308801	0,0711	0,338167	0,195919	0,210251	FALSO	FALSO
8	hbss	0,0936	0,337771	0,1034	0,353597	0,277111	0,292423	FALSO	FALSO
9	hbsws	0,0364	0,220115	0,0395	0,232304	0,165368	0,170036	FALSO	FALSO
10	hbswws	0,0243	0,198732	0,0233	0,192105	0,122275	0,121288	FALSO	FALSO
11	hydrophobic	1,5808	2,224305	1,636	2,26985	0,710694	0,720752	FALSO	FALSO
12	aromatic	0,1295	0,441858	0,1203	0,428017	0,293081	0,281064	FALSO	FALSO
13	disulfide	0,0041	0,064779	0,0042	0,064879	0,063292	0,064736	FALSO	FALSO
14	ch_attractive	0,5078	1,392176	0,4168	1,282985	0,364753	0,324867	FALSO	FALSO
15	ch_repulsive	0,1665	0,756407	0,1628	0,803778	0,22012	0,202544	FALSO	FALSO
17	hbmwm_WNADist	0,214101	0,358295	0,309562	0,501362	0,597556	0,617441	FALSO	FALSO
18	hbmwwm_WNADist	0,113662	0,261355	0,159645	0,368328	0,434897	0,433433	FALSO	FALSO
19	hbms_WNADist	1,29002	0,878789	1,370436	0,947283	1,467952	1,446702	FALSO	FALSO
20	hbmws_WNADist	0,420744	0,584835	0,480314	0,651446	0,719424	0,737304	FALSO	FALSO
21	hbmwws_WNADist	0,295854	0,549084	0,316341	0,585953	0,538815	0,539875	FALSO	FALSO
22	hbss_WNADist	0,583771	0,589627	0,605453	0,61743	0,990069	0,980602	FALSO	FALSO

23	hbsws_WNADist	0,036465	0,220769	0,039464	0,23253	0,165174	0,169717	FALSO	FALSO
24	hbswws_WNADist	0,023971	0,196998	0,023114	0,190392	0,121682	0,121403	FALSO	FALSO
25	hydrophobic_WNADist	0,135205	0,719653	0,113597	0,715739	0,187876	0,158712	FALSO	FALSO
26	aromatic_WNADist	0,536382	0,612584	0,509818	0,587945	0,875605	0,867118	FALSO	FALSO
27	disulfide_WNADist	0,012156	0,085683	0,011556	0,082547	0,141877	0,139997	FALSO	FALSO
28	ch_attractive_WNADist	2,366683	2,123197	2,115143	1,972813	1,114679	1,072146	FALSO	FALSO
29	ch_repulsive_WNADist	0,791648	1,193334	0,751431	1,202125	0,663391	0,625086	FALSO	FALSO
31	hbmwm_WNASurf	0,340178	0,485156	0,401794	0,540517	0,701174	0,743351	FALSO	FALSO
32	hbmwwm_WNASurf	0,189275	0,378048	0,223691	0,430841	0,500665	0,519196	FALSO	FALSO
33	hbms_WNASurf	1,716874	0,963033	1,762139	0,973283	1,782778	1,810511	FALSO	FALSO
34	hbmws_WNASurf	0,590226	0,70492	0,632872	0,730854	0,837296	0,865934	FALSO	FALSO
35	hbmwws_WNASurf	0,419631	0,692455	0,429603	0,696254	0,606004	0,61702	FALSO	FALSO
36	hbss_WNASurf	0,797898	0,729506	0,81322	0,742884	1,093751	1,09468	FALSO	FALSO
37	hbsws_WNASurf	0,004566	0,035469	0,004769	0,037427	0,128723	0,127418	FALSO	FALSO
38	hbswws_WNASurf	0,003104	0,031611	0,002963	0,03179	0,098179	0,093193	FALSO	FALSO
39	hydrophobic_WNASurf	0,063201	0,320712	0,056657	0,349292	0,197066	0,162205	FALSO	FALSO
40	aromatic_WNASurf	0,680675	0,591108	0,655075	0,571833	1,151524	1,145571	FALSO	FALSO
41	disulfide_WNASurf	0,011917	0,084631	0,011265	0,081518	0,140808	0,138193	FALSO	FALSO
42	ch_attractive_WNASurf	3,501251	2,491179	3,265414	2,415437	1,405459	1,351894	FALSO	FALSO
43	ch_repulsive_WNASurf	1,216009	1,489436	1,146977	1,402023	0,816422	0,818087	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-44,874	82,92479	-40,6512	81,48954	-0,54114	-0,49885	FALSO	FALSO
46	Electrostatic_Potential_Average	5,957643	24,20611	4,143153	23,75197	0,246121	0,174434	FALSO	FALSO
47	Hydrophobicity_KDI	-40,8113	644,5015	-35,411	600,1545	-0,06332	-0,059	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-367,245	194,3078	-348,662	190,9855	-1,89001	-1,8256	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	38,20981	87,47238	31,56476	84,13233	0,436821	0,37518	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	131,2871	65,17759	125,7039	64,60086	2,014298	1,945855	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-323,92	177,3992	-308,635	173,5882	-1,82594	-1,77797	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	32,24446	75,06227	27,37125	71,79397	0,429569	0,381247	FALSO	FALSO
54	Cross_Link_Order_CA	0,4845	0,635037	0,5429	0,687561	0,762948	0,789602	FALSO	FALSO
55	Cross_Pres_Order_CA	0,8486	0,882705	0,9553	0,925442	0,961363	1,032263	FALSO	FALSO
58	Dihedral_Chi1	-31,5092	101,0142	-26,1874	95,37619	-0,31193	-0,27457	FALSO	FALSO
59	Dihedral_Chi2	13,01182	101,8668	8,815882	93,81751	0,127734	0,093968	FALSO	FALSO

60	Dihedral_Chi3	-0,49015	56,99651	-0,30167	53,39736	-0,0086	-0,00565	FALSO	FALSO
61	Dihedral_Chi4	0,85266	47,27547	0,42534	45,89609	0,018036	0,009267	FALSO	FALSO
62	Temperature_Factor_CA	23,88658	17,73843	25,73214	19,61901	1,346601	1,311592	FALSO	FALSO
68	Clash	0,291217	0,132654	0,298101	0,134006	2,195316	2,224542	FALSO	FALSO
69	Percent	9,37301	4,214345	9,59258	4,264106	2,224073	2,249611	FALSO	FALSO

Tabela 53. Descritores usados no teste MANOVA para as proteínas do tipo all- β alinhadas pelo C-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
1	Accessible_Surface_in_Isolation	40,67865	40,94155	58,24028	49,09566	0,993579	1,186261	VERDADEIRO	FALSO
2	hbm	1,3758	0,913769	0,9847	0,915003	1,505632	1,076172	VERDADEIRO	FALSO
5	hbms	0,1661	0,468566	0,2753	0,590667	0,354486	0,466083	VERDADEIRO	FALSO
16	hbm_WNADist	4,845132	1,618116	3,85304	1,704187	2,994304	2,260926	VERDADEIRO	FALSO
29	ch_repulsive_WNADist	0,597694	0,970049	0,609476	1,278254	0,616148	0,476804	VERDADEIRO	FALSO
30	hbm_WNASurf	5,8368	2,075982	4,920387	2,008704	2,811585	2,449533	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	15,54668	18,31868	16,40888	23,39007	0,848679	0,701532	VERDADEIRO	FALSO
54	Cross_Link_Order_CA	0,5477	0,629672	0,4311	0,588855	0,869818	0,732099	VERDADEIRO	FALSO
55	Cross_Pres_Order_CA	0,9362	0,81459	0,7994	0,808423	1,149289	0,988838	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	820,5959	131,7607	764,309	139,7321	6,227926	5,469817	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1336,855	303,8565	1211,027	304,1688	4,399626	3,981431	VERDADEIRO	FALSO
66	IFR_CA_3	1,109882	0,114232	1,101463	0,12463	9,716043	8,837885	VERDADEIRO	FALSO
67	Internal_CA_3	1,110373	0,111766	1,101425	0,123695	9,934846	8,90439	VERDADEIRO	VERDADEIRO
3	hbmwm	0,0482	0,258879	0,0718	0,307231	0,186188	0,2337	FALSO	FALSO
4	hbmwwm	0,0381	0,261437	0,0418	0,254874	0,145733	0,164003	FALSO	FALSO
6	hbmws	0,0948	0,369283	0,0934	0,355298	0,256714	0,262878	FALSO	FALSO
7	hbmwws	0,0621	0,315289	0,066	0,332759	0,196962	0,198342	FALSO	FALSO
8	hbss	0,0912	0,329273	0,096	0,338649	0,276974	0,28348	FALSO	FALSO
9	hbsws	0,0421	0,235691	0,0359	0,216276	0,178623	0,165992	FALSO	FALSO
10	hbswws	0,029	0,22481	0,0246	0,198328	0,128998	0,124037	FALSO	FALSO
11	hydrophobic	2,3384	2,95344	1,7708	2,5548	0,791755	0,693127	FALSO	FALSO
12	aromatic	0,1052	0,370167	0,0834	0,335806	0,284196	0,248358	FALSO	FALSO
13	disulfide	0,0197	0,138988	0,0126	0,111768	0,141739	0,112734	FALSO	FALSO
14	ch_attractive	0,299	1,079781	0,3025	1,056881	0,276908	0,28622	FALSO	FALSO
15	ch_repulsive	0,1004	0,590188	0,1392	0,814547	0,170115	0,170893	FALSO	FALSO
17	hbmwm_WNADist	0,261102	0,437394	0,273371	0,482378	0,59695	0,566715	FALSO	FALSO
18	hbmwwm_WNADist	0,204933	0,465565	0,187869	0,440456	0,440181	0,426533	FALSO	FALSO
19	hbms_WNADist	0,949847	0,791237	1,082083	0,927176	1,200458	1,167074	FALSO	FALSO
20	hbmws_WNADist	0,488776	0,715458	0,428411	0,661901	0,683166	0,647244	FALSO	FALSO
21	hbmwws_WNADist	0,345844	0,678931	0,307991	0,638605	0,509395	0,482286	FALSO	FALSO

22	hbss_WNADist	0,701142	0,736521	0,639189	0,70002	0,951964	0,913101	FALSO	FALSO
23	hbsws_WNADist	0,042744	0,23854	0,036422	0,216826	0,179191	0,167978	FALSO	FALSO
24	hbswws_WNADist	0,029035	0,22517	0,024738	0,198964	0,128947	0,124332	FALSO	FALSO
25	hydrophobic_WNADist	0,196655	0,802803	0,154185	0,726877	0,24496	0,212119	FALSO	FALSO
26	aromatic_WNADist	0,454689	0,549992	0,38708	0,503591	0,826719	0,76864	FALSO	FALSO
27	disulfide_WNADist	0,05356	0,213253	0,035963	0,168216	0,251156	0,213792	FALSO	FALSO
28	ch_attractive_WNADist	1,665835	1,744181	1,500914	1,681753	0,955081	0,89247	FALSO	FALSO
31	hbmwm_WNASurf	0,457093	0,64211	0,412761	0,600471	0,71186	0,687395	FALSO	FALSO
32	hbmwwm_WNASurf	0,361572	0,711574	0,307116	0,624172	0,50813	0,492038	FALSO	FALSO
33	hbms_WNASurf	1,512896	1,025232	1,498841	1,031489	1,475662	1,453085	FALSO	FALSO
34	hbmws_WNASurf	0,749419	0,995532	0,635951	0,882841	0,752783	0,720345	FALSO	FALSO
35	hbmwws_WNASurf	0,577453	1,065894	0,492882	0,948562	0,541755	0,519609	FALSO	FALSO
36	hbss_WNASurf	1,029109	1,02771	0,924768	0,965041	1,001361	0,958268	FALSO	FALSO
37	hbsws_WNASurf	0,00511	0,03663	0,005622	0,040878	0,139515	0,137522	FALSO	FALSO
38	hbswws_WNASurf	0,003844	0,035937	0,004502	0,043838	0,106956	0,102704	FALSO	FALSO
39	hydrophobic_WNASurf	0,08715	0,355688	0,072846	0,355028	0,245019	0,205183	FALSO	FALSO
40	aromatic_WNASurf	0,605171	0,585853	0,534619	0,553537	1,032974	0,965824	FALSO	FALSO
41	disulfide_WNASurf	0,052284	0,209694	0,035016	0,16485	0,249333	0,21241	FALSO	FALSO
42	ch_attractive_WNASurf	2,647608	2,187473	2,340569	2,058674	1,21035	1,13693	FALSO	FALSO
43	ch_repulsive_WNASurf	0,999518	1,353811	0,916143	1,354657	0,7383	0,676292	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-37,424	72,11327	-42,294	80,26889	-0,51896	-0,5269	FALSO	FALSO
46	Electrostatic_Potential_Average	2,744763	22,68089	1,347229	23,52757	0,121017	0,057262	FALSO	FALSO
47	Hydrophobicity_KDI	-426,366	2023,177	-383,165	1921,003	-0,21074	-0,19946	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	141,008	87,2624	129,3115	80,30555	1,615908	1,610244	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-360,802	220,5149	-335,701	210,5813	-1,63618	-1,59416	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	26,74929	103,882	21,29324	93,15005	0,257497	0,228591	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	127,6166	77,84525	116,1763	70,66579	1,639363	1,644025	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-328,954	200,6587	-302,653	188,5603	-1,63937	-1,60507	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	24,57617	92,78291	20,27783	82,70561	0,264878	0,245181	FALSO	FALSO
58	Dihedral_Chi1	-22,1363	104,7608	-23,2688	94,16315	-0,2113	-0,24711	FALSO	FALSO
59	Dihedral_Chi2	15,21616	95,3692	9,451631	91,64758	0,15955	0,10313	FALSO	FALSO
60	Dihedral_Chi3	-1,09626	48,80444	-0,18268	50,70036	-0,02246	-0,0036	FALSO	FALSO

61	Dihedral_Chi4	-1,0467	42,32345	-0,29026	43,67982	-0,02473	-0,00665	FALSO	FALSO
62	Temperature_Factor_CA	21,09098	17,3552	23,93959	19,84077	1,215254	1,206586	FALSO	FALSO
63	Number_Unused_Contact	205,9449	66,76257	212,3904	66,8972	3,084736	3,174877	FALSO	FALSO
68	Clash	0,292531	0,137307	0,304068	0,141507	2,130487	2,148781	FALSO	FALSO
69	Percent	9,385343	4,252824	9,762855	4,434333	2,20685	2,201651	FALSO	FALSO

Tabela 54. Descritores usados no teste MANOVA para as proteínas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$ alinhadas pelo C-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
1	Accessible_Surface_in_Isolation	28,84209	37,02921	49,48567	47,81468	0,778901	1,034947	VERDADEIRO	FALSO
2	hbmm	2,1366	1,285512	1,926	1,392854	1,662061	1,382772	VERDADEIRO	FALSO
5	hbms	0,219	0,54781	0,3622	0,676427	0,399773	0,535461	VERDADEIRO	FALSO
16	hbmm_WNADist	8,432341	4,690055	7,671604	4,569257	1,797919	1,678961	VERDADEIRO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-405,467	179,7145	-375,926	175,8033	-2,25617	-2,13834	VERDADEIRO	FALSO
54	Cross_Link_Order_CA	0,661	0,741219	0,476	0,665113	0,891774	0,715668	VERDADEIRO	FALSO
55	Cross_Pres_Order_CA	1,311	0,96951	1,003	0,933067	1,35223	1,074949	VERDADEIRO	FALSO
67	Internal_CA_3	1,056469	0,234189	1,045546	0,237129	4,511185	4,409188	VERDADEIRO	FALSO
68	Clash	0,288586	0,133035	0,302264	0,153761	2,16925	1,965811	VERDADEIRO	FALSO
69	Percent	9,306347	4,140321	9,733425	4,798177	2,247736	2,028567	VERDADEIRO	FALSO
3	hbmwm	0,0603	0,287043	0,0844	0,33657	0,210073	0,250765	FALSO	FALSO
4	hbmwwm	0,0322	0,220998	0,0432	0,252261	0,145703	0,171251	FALSO	FALSO
6	hbmws	0,1136	0,412369	0,1204	0,419029	0,275482	0,287331	FALSO	FALSO
7	hbmwws	0,0613	0,325516	0,0761	0,358938	0,188316	0,212014	FALSO	FALSO
8	hbss	0,1142	0,372057	0,1024	0,351119	0,306942	0,291639	FALSO	FALSO
9	hbsws	0,0441	0,246561	0,0385	0,229898	0,178861	0,167466	FALSO	FALSO
10	hbswws	0,0282	0,217557	0,0271	0,212755	0,129621	0,127377	FALSO	FALSO
11	hydrophobic	5,0422	4,63858	4,3135	4,357741	1,087014	0,989848	FALSO	FALSO
12	aromatic	0,1392	0,457704	0,1167	0,421755	0,304126	0,276701	FALSO	FALSO
13	disulfide	0,006	0,078315	0,0037	0,06158	0,076614	0,060085	FALSO	FALSO
14	ch_attractive	0,3333	1,278375	0,3329	1,226343	0,260722	0,271457	FALSO	FALSO
15	ch_repulsive	0,1554	0,872257	0,1499	0,800063	0,178158	0,18736	FALSO	FALSO
17	hbmwm_WNADist	0,317076	0,47606	0,334133	0,537249	0,666041	0,621933	FALSO	FALSO
18	hbmwwm_WNADist	0,189246	0,386865	0,191272	0,421971	0,489178	0,453283	FALSO	FALSO
19	hbms_WNADist	1,258189	0,877467	1,420847	1,02542	1,433888	1,385624	FALSO	FALSO
20	hbmws_WNADist	0,547131	0,731606	0,511772	0,724779	0,747849	0,706108	FALSO	FALSO
21	hbmwws_WNADist	0,348225	0,630246	0,347881	0,659885	0,552521	0,527184	FALSO	FALSO
22	hbss_WNADist	0,73053	0,703369	0,643954	0,670777	1,038615	0,960012	FALSO	FALSO
23	hbsws_WNADist	0,045425	0,250278	0,039789	0,233617	0,181498	0,170318	FALSO	FALSO
24	hbswws_WNADist	0,029136	0,221239	0,028108	0,216673	0,131693	0,129725	FALSO	FALSO

25	hydrophobic_WNADist	0,149806	0,717602	0,179144	0,804136	0,208759	0,222778	FALSO	FALSO
26	aromatic_WNADist	0,589624	0,639653	0,510049	0,603439	0,921787	0,845236	FALSO	FALSO
27	disulfide_WNADist	0,012475	0,094739	0,009246	0,076992	0,131674	0,120096	FALSO	FALSO
28	ch_attractive_WNADist	1,838109	2,131501	1,62232	2,045588	0,862354	0,793082	FALSO	FALSO
29	ch_repulsive_WNADist	0,837895	1,426348	0,727917	1,310181	0,587441	0,555585	FALSO	FALSO
30	hbmm_WNASurf	11,75288	7,344665	10,69991	6,80756	1,600193	1,571769	FALSO	FALSO
31	hbmwm_WNASurf	0,50179	0,6536	0,469083	0,634079	0,767732	0,739787	FALSO	FALSO
32	hbmwwm_WNASurf	0,308068	0,560367	0,284864	0,537205	0,549762	0,530271	FALSO	FALSO
33	hbms_WNASurf	1,904227	1,102148	1,848484	1,078215	1,727742	1,714393	FALSO	FALSO
34	hbmws_WNASurf	0,772496	0,920785	0,690063	0,85671	0,838954	0,805481	FALSO	FALSO
35	hbmwws_WNASurf	0,538304	0,887575	0,499602	0,846747	0,606489	0,590025	FALSO	FALSO
36	hbss_WNASurf	1,004605	0,923056	0,880363	0,86609	1,088347	1,01648	FALSO	FALSO
37	hbsws_WNASurf	0,003973	0,030942	0,005046	0,038901	0,128406	0,129724	FALSO	FALSO
38	hbswws_WNASurf	0,002926	0,029676	0,003969	0,038682	0,098611	0,102615	FALSO	FALSO
39	hydrophobic_WNASurf	0,064662	0,312235	0,08753	0,390888	0,207094	0,223927	FALSO	FALSO
40	aromatic_WNASurf	0,763946	0,634183	0,66971	0,605673	1,204614	1,105729	FALSO	FALSO
41	disulfide_WNASurf	0,012236	0,093778	0,009019	0,07603	0,130482	0,118617	FALSO	FALSO
42	ch_attractive_WNASurf	3,010259	2,910217	2,572617	2,695702	1,034376	0,95434	FALSO	FALSO
43	ch_repulsive_WNASurf	1,381684	2,002269	1,165947	1,792732	0,690059	0,650374	FALSO	FALSO
44	Electrostatic_Potential_at_CA	13,34183	19,276	17,86398	23,68246	0,692147	0,754313	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-39,3973	71,72738	-44,9531	80,2168	-0,54926	-0,5604	FALSO	FALSO
46	Electrostatic_Potential_Average	-0,28982	21,25878	0,691107	21,8746	-0,01363	0,031594	FALSO	FALSO
47	Hydrophobicity_KDI	-494,132	2170,394	-487,605	2155,14	-0,22767	-0,22625	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	130,2486	69,85625	128,4833	65,93972	1,864523	1,948496	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	6,565645	70,45474	7,970138	65,372	0,09319	0,12192	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	108,0869	66,74145	103,7618	62,53484	1,619487	1,659264	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-338,714	187,8414	-310,09	176,5952	-1,80319	-1,75594	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	6,446615	57,42589	6,832488	53,04021	0,11226	0,128817	FALSO	FALSO
58	Dihedral_Chi1	-2,18579	176,4758	-5,10104	167,8883	-0,01239	-0,03038	FALSO	FALSO
59	Dihedral_Chi2	33,83182	166,8839	28,13945	163,5933	0,202727	0,172009	FALSO	FALSO
60	Dihedral_Chi3	19,14459	146,8966	18,63835	145,9237	0,130327	0,127727	FALSO	FALSO
61	Dihedral_Chi4	19,5676	144,4916	19,08526	143,0376	0,135424	0,133428	FALSO	FALSO

62	Temperature_Factor_CA	37,94361	38,55705	42,11276	41,38389	0,98409	1,017612	FALSO	FALSO
63	Number_Unused_Contact	163,4581	88,94734	171,8737	90,45839	1,837695	1,900031	FALSO	FALSO
64	Number_Unused_Contact_WNADist	625,397	273,9529	596,2637	260,3808	2,282863	2,289968	FALSO	FALSO
65	Number_Unused_Contact_WNASurf	1001,487	460,9452	919,668	421,2984	2,172681	2,182937	FALSO	FALSO
66	IFR_CA_3	0,729893	0,527165	0,731962	0,520148	1,384562	1,40722	FALSO	FALSO

APÊNDICE G – DESCRITORES USADOS NO TESTE MANOVA PARA AS ESTRUTURAS ALINHADAS PELO N-TERMINAL

Tabela 55. Descritores usados no teste MANOVA para as proteínas do tipo all- α alinhadas pelo N-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
2	hbmm	2,0411	1,091091	1,6843	1,165862	1,870696	1,444682	VERDADEIRO	FALSO
16	hbmm_WNADist	7,720433	2,485192	6,711898	2,669193	3,106574	2,51458	VERDADEIRO	FALSO
30	hbmm_WNASurf	11,15295	3,836378	9,926398	3,712427	2,907156	2,67383	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	23,03813	17,82004	22,44666	20,5412	1,292821	1,092763	VERDADEIRO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-354,561	235,4559	-361,08	222,9568	-1,50585	-1,61951	VERDADEIRO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-318,871	215,0728	-324,02	203,5052	-1,48262	-1,59219	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	794,2036	132,5386	775,7475	139,8482	5,992244	5,547068	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1297,193	294,3776	1259,865	301,9953	4,406562	4,171803	VERDADEIRO	FALSO
66	IFR_CA_3	1,071648	0,174565	1,062852	0,195982	6,138959	5,42321	VERDADEIRO	FALSO
67	Internal_CA_3	1,090524	0,101226	1,08674	0,114662	10,7732	9,477811	VERDADEIRO	VERDADEIRO
68	Clash	0,304018	0,143249	0,314553	0,158197	2,122298	1,988361	VERDADEIRO	FALSO
69	Percent	9,732133	4,515688	10,05765	4,955693	2,155183	2,029514	VERDADEIRO	FALSO
1	Accessible_Surface_in_Isolation	56,39868	47,45141	58,99776	49,64719	1,188556	1,18834	FALSO	FALSO
3	hbmwm	0,0243	0,173005	0,0383	0,216547	0,140459	0,176867	FALSO	FALSO
4	hbmwwm	0,0106	0,121159	0,0209	0,166192	0,087489	0,125758	FALSO	FALSO
5	hbms	0,2348	0,522712	0,2794	0,570664	0,449196	0,489605	FALSO	FALSO
6	hbmws	0,0485	0,246292	0,0627	0,283689	0,196921	0,221017	FALSO	FALSO
7	hbmwws	0,0329	0,220128	0,0405	0,24757	0,149459	0,16359	FALSO	FALSO
8	hbss	0,0636	0,278021	0,0742	0,295892	0,228759	0,250767	FALSO	FALSO
9	hbsws	0,0264	0,187511	0,0271	0,186086	0,140792	0,145631	FALSO	FALSO
10	hbswws	0,0138	0,144351	0,0121	0,12661	0,095601	0,095569	FALSO	FALSO
11	hydrophobic	1,4848	2,418375	1,4729	2,346925	0,613966	0,627587	FALSO	FALSO
12	aromatic	0,1033	0,377755	0,1073	0,38742	0,273457	0,27696	FALSO	FALSO
13	disulfide	0,0052	0,072137	0,0041	0,063595	0,072085	0,06447	FALSO	FALSO
14	ch_attractive	0,4597	1,305295	0,4563	1,310883	0,352181	0,348086	FALSO	FALSO
15	ch_repulsive	0,1605	0,752236	0,176	0,774428	0,213364	0,227265	FALSO	FALSO
17	hbmwm_WNADist	0,122714	0,283176	0,16159	0,340598	0,43335	0,474431	FALSO	FALSO

18	hbmwwm_WNADist	0,062965	0,204581	0,086601	0,265903	0,307773	0,325685	FALSO	FALSO
19	hbms_WNADist	1,040609	0,812822	1,117726	0,85961	1,280243	1,300272	FALSO	FALSO
20	hbmws_WNADist	0,237156	0,415485	0,270835	0,457612	0,570794	0,591843	FALSO	FALSO
21	hbmwws_WNADist	0,155622	0,38081	0,185605	0,431421	0,408661	0,430219	FALSO	FALSO
22	hbss_WNADist	0,427108	0,50501	0,440035	0,510095	0,845742	0,862652	FALSO	FALSO
23	hbsws_WNADist	0,026176	0,187757	0,027059	0,186778	0,139414	0,144871	FALSO	FALSO
24	hbswws_WNADist	0,013739	0,144906	0,01221	0,12749	0,094811	0,095776	FALSO	FALSO
25	hydrophobic_WNADist	0,190814	0,889886	0,144145	0,770578	0,214425	0,18706	FALSO	FALSO
26	aromatic_WNADist	0,476071	0,576344	0,479177	0,589123	0,826019	0,813373	FALSO	FALSO
27	disulfide_WNADist	0,014736	0,099834	0,013442	0,088938	0,147604	0,151136	FALSO	FALSO
28	ch_attractive_WNADist	2,156353	2,057314	2,12838	2,037892	1,04814	1,044403	FALSO	FALSO
29	ch_repulsive_WNADist	0,806479	1,182385	0,814418	1,189215	0,682078	0,684837	FALSO	FALSO
31	hbmwm_WNASurf	0,211246	0,408523	0,249902	0,439891	0,517097	0,568099	FALSO	FALSO
32	hbmwwm_WNASurf	0,123307	0,334897	0,150882	0,371126	0,368193	0,406553	FALSO	FALSO
33	hbms_WNASurf	1,535568	1,003119	1,568097	0,987206	1,530793	1,588419	FALSO	FALSO
34	hbmws_WNASurf	0,374501	0,566475	0,419157	0,613198	0,661108	0,683559	FALSO	FALSO
35	hbmwws_WNASurf	0,254634	0,523901	0,298581	0,576414	0,486035	0,517996	FALSO	FALSO
36	hbss_WNASurf	0,637728	0,656283	0,646473	0,643754	0,971728	1,004223	FALSO	FALSO
37	hbsws_WNASurf	0,003986	0,033898	0,003892	0,033031	0,117583	0,117833	FALSO	FALSO
38	hbswws_WNASurf	0,002246	0,026791	0,002018	0,024972	0,083814	0,080798	FALSO	FALSO
39	hydrophobic_WNASurf	0,095147	0,450379	0,074006	0,377409	0,211259	0,196089	FALSO	FALSO
40	aromatic_WNASurf	0,709949	0,664015	0,686335	0,653604	1,069176	1,050076	FALSO	FALSO
41	disulfide_WNASurf	0,014247	0,097733	0,013084	0,087421	0,14578	0,14967	FALSO	FALSO
42	ch_attractive_WNASurf	3,40474	2,510316	3,407074	2,549298	1,356299	1,336475	FALSO	FALSO
43	ch_repulsive_WNASurf	1,375563	1,495996	1,338789	1,446611	0,919496	0,925466	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-39,7114	84,49963	-42,3462	83,09256	-0,46996	-0,50963	FALSO	FALSO
46	Electrostatic_Potential_Average	8,030319	25,27523	6,438613	24,44592	0,317715	0,263382	FALSO	FALSO
47	Hydrophobicity_KDI	0,782047	1,921335	0,671281	1,892319	0,407033	0,35474	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	177,5736	83,62067	168,2112	80,24652	2,123561	2,096181	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	64,01433	108,627	52,88473	100,7902	0,589304	0,524701	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	156,6415	76,30665	148,3075	72,13508	2,05279	2,055969	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	56,87472	96,78721	47,38046	89,58247	0,587626	0,528903	FALSO	FALSO

54	Cross_Link_Order_CA	0,3451	0,53892	0,3627	0,545126	0,640355	0,66535	FALSO	FALSO
55	Cross_Pres_Order_CA	0,5586	0,761376	0,6026	0,740989	0,733671	0,813238	FALSO	FALSO
58	Dihedral_Chi1	-32,8377	101,2302	-30,9486	97,13819	-0,32439	-0,3186	FALSO	FALSO
59	Dihedral_Chi2	12,82665	100,4523	10,86901	95,68814	0,127689	0,113588	FALSO	FALSO
60	Dihedral_Chi3	-0,76034	57,40597	-1,06904	54,21041	-0,01324	-0,01972	FALSO	FALSO
61	Dihedral_Chi4	0,80662	49,12567	0,334824	46,40924	0,01642	0,007215	FALSO	FALSO
62	Temperature_Factor_CA	23,87976	21,26215	26,84163	24,85874	1,123111	1,079766	FALSO	FALSO
63	Number_Unused_Contact	214,7015	70,83524	216,1124	69,38521	3,030998	3,114675	FALSO	FALSO

Tabela 56. Descritores usados no teste MANOVA para as proteínas do tipo α em $(\alpha+\beta)+(\alpha/\beta)$ alinhadas pelo N-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
2	hbmm	2,0521	1,079606	1,3967	1,114039	1,900786	1,253726	VERDADEIRO	FALSO
16	hbmm_WNADist	6,798634	2,217873	5,348396	2,283131	3,065385	2,342571	VERDADEIRO	FALSO
30	hbmm_WNASurf	8,625826	3,236919	7,385295	3,107532	2,664826	2,376579	VERDADEIRO	FALSO
44	Electrostatic_Potential_at_CA	22,71286	18,20872	21,08036	24,03095	1,247362	0,877217	VERDADEIRO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	154,5552	72,22369	147,0892	73,08877	2,139952	2,012473	VERDADEIRO	FALSO
63	Number_Unused_Contact	214,1752	71,41002	211,2357	67,77749	2,999232	3,116606	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	771,7337	121,0458	746,816	126,644	6,375551	5,896971	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1180,108	274,3123	1140,864	279,0796	4,30206	4,087952	VERDADEIRO	FALSO
66	IFR_CA_3	1,012845	0,288133	1,011127	0,298444	3,515206	3,387995	VERDADEIRO	FALSO
67	Internal_CA_3	1,086432	0,090604	1,087656	0,109595	11,99101	9,924339	VERDADEIRO	VERDADEIRO
1	Accessible_Surface_in_Isolation	47,59036	46,53359	48,6495	47,60322	1,02271	1,021979	FALSO	FALSO
3	hbmwm	0,0344	0,21272	0,0816	0,329264	0,161715	0,247826	FALSO	FALSO
4	hbmwwm	0,0161	0,148768	0,0365	0,227384	0,108222	0,160522	FALSO	FALSO
5	hbms	0,3013	0,594851	0,349	0,653012	0,506513	0,534447	FALSO	FALSO
6	hbmws	0,0905	0,359779	0,1176	0,409312	0,251543	0,287312	FALSO	FALSO
7	hbmwws	0,059	0,304475	0,0695	0,334176	0,193776	0,207975	FALSO	FALSO
8	hbss	0,092	0,334915	0,1024	0,351629	0,274697	0,291216	FALSO	FALSO
9	hbsws	0,0356	0,217751	0,0387	0,229805	0,16349	0,168404	FALSO	FALSO
10	hbswws	0,0234	0,194887	0,0227	0,189155	0,12007	0,120008	FALSO	FALSO
11	hydrophobic	1,5992	2,246656	1,6429	2,277111	0,711813	0,721484	FALSO	FALSO
12	aromatic	0,1316	0,445201	0,1209	0,429231	0,295597	0,281667	FALSO	FALSO
13	disulfide	0,0039	0,063184	0,004	0,063555	0,061724	0,062937	FALSO	FALSO
14	ch_attractive	0,5073	1,389892	0,4175	1,281892	0,364992	0,32569	FALSO	FALSO
15	ch_repulsive	0,1689	0,756595	0,1654	0,805911	0,223237	0,205234	FALSO	FALSO
17	hbmwm_WNADist	0,210906	0,355907	0,305356	0,498491	0,592586	0,612561	FALSO	FALSO
18	hbmwwm_WNADist	0,111257	0,258668	0,156661	0,364676	0,430116	0,42959	FALSO	FALSO
19	hbms_WNADist	1,283052	0,876446	1,364198	0,94663	1,463926	1,441111	FALSO	FALSO
20	hbmws_WNADist	0,414674	0,58143	0,473105	0,647457	0,713197	0,730713	FALSO	FALSO
21	hbmwws_WNADist	0,291049	0,544563	0,310952	0,580979	0,534464	0,53522	FALSO	FALSO
22	hbss_WNADist	0,578873	0,587198	0,601015	0,616534	0,985823	0,974827	FALSO	FALSO

23	hbsws_WNADist	0,03599	0,219284	0,038944	0,230919	0,164124	0,168646	FALSO	FALSO
24	hbswws_WNADist	0,023452	0,194461	0,02266	0,188379	0,120601	0,120291	FALSO	FALSO
25	hydrophobic_WNADist	0,134337	0,716693	0,112761	0,710368	0,18744	0,158736	FALSO	FALSO
26	aromatic_WNADist	0,541282	0,617072	0,514121	0,591523	0,877179	0,869148	FALSO	FALSO
27	disulfide_WNADist	0,011728	0,084147	0,011203	0,081341	0,139378	0,137733	FALSO	FALSO
28	ch_attractive_WNADist	2,37012	2,126172	2,118613	1,976345	1,114736	1,071985	FALSO	FALSO
29	ch_repulsive_WNADist	0,800986	1,193105	0,761049	1,205382	0,671346	0,631376	FALSO	FALSO
31	hbmwm_WNASurf	0,335403	0,483031	0,396579	0,537789	0,694372	0,737424	FALSO	FALSO
32	hbmwwm_WNASurf	0,185445	0,374393	0,21978	0,426877	0,495321	0,514854	FALSO	FALSO
33	hbms_WNASurf	1,708697	0,960741	1,754999	0,972687	1,77852	1,80428	FALSO	FALSO
34	hbmws_WNASurf	0,582027	0,701461	0,623962	0,728002	0,829735	0,857088	FALSO	FALSO
35	hbmwws_WNASurf	0,412963	0,686647	0,422569	0,690643	0,601419	0,611849	FALSO	FALSO
36	hbss_WNASurf	0,791882	0,726833	0,806976	0,741078	1,089497	1,088922	FALSO	FALSO
37	hbsws_WNASurf	0,004534	0,035356	0,004725	0,037194	0,128243	0,127037	FALSO	FALSO
38	hbswws_WNASurf	0,003057	0,031347	0,002917	0,031502	0,097527	0,092608	FALSO	FALSO
39	hydrophobic_WNASurf	0,062725	0,318894	0,056212	0,346399	0,196697	0,162275	FALSO	FALSO
40	aromatic_WNASurf	0,687964	0,597017	0,662447	0,577806	1,152335	1,146488	FALSO	FALSO
41	disulfide_WNASurf	0,011498	0,083115	0,010921	0,08033	0,138334	0,135954	FALSO	FALSO
42	ch_attractive_WNASurf	3,50578	2,4934	3,27071	2,41691	1,406024	1,353261	FALSO	FALSO
43	ch_repulsive_WNASurf	1,228722	1,484292	1,159785	1,402129	0,827817	0,82716	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-44,7341	82,9411	-40,5419	81,45561	-0,53935	-0,49772	FALSO	FALSO
46	Electrostatic_Potential_Average	6,197923	24,20137	4,277658	23,74935	0,256098	0,180117	FALSO	FALSO
47	Hydrophobicity_KDI	-41,7062	651,2852	-36,4752	608,8518	-0,06404	-0,05991	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-366,822	194,3728	-348,457	191,5471	-1,88721	-1,81917	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	39,31033	87,83584	32,47743	84,48671	0,447543	0,384409	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	132,079	65,527	126,2748	64,85171	2,015642	1,947131	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-323,435	177,5538	-308,331	174,2091	-1,82162	-1,76989	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	33,19555	75,40952	28,15957	72,13937	0,440204	0,39035	FALSO	FALSO
54	Cross_Link_Order_CA	0,4621	0,62854	0,5192	0,681585	0,735195	0,761754	FALSO	FALSO
55	Cross_Pres_Order_CA	0,8357	0,88025	0,942	0,923373	0,949389	1,020173	FALSO	FALSO
58	Dihedral_Chi1	-30,2712	105,8432	-25,2159	100,6254	-0,286	-0,25059	FALSO	FALSO
59	Dihedral_Chi2	13,76092	106,5459	9,574311	98,75325	0,129155	0,096952	FALSO	FALSO

60	Dihedral_Chi3	0,219888	65,11008	0,657365	61,94097	0,003377	0,010613	FALSO	FALSO
61	Dihedral_Chi4	1,806887	56,79955	1,353581	55,61842	0,031812	0,024337	FALSO	FALSO
62	Temperature_Factor_CA	24,32237	17,88737	26,16649	19,86168	1,359751	1,317436	FALSO	FALSO
68	Clash	0,292914	0,132707	0,299594	0,134406	2,207228	2,229022	FALSO	FALSO
69	Percent	9,421668	4,220846	9,635371	4,281299	2,232175	2,250572	FALSO	FALSO

Tabela 57. Descritores usados no teste MANOVA para as proteínas do tipo all- β alinhadas pelo N-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
1	Accessible_Surface_in_Isolation	43,74417	42,78046	58,95833	48,84145	1,022527	1,207137	VERDADEIRO	FALSO
2	hbmm	1,3177	0,926972	0,9643	0,918761	1,421511	1,049565	VERDADEIRO	FALSO
11	hydrophobic	2,2227	2,887144	1,7474	2,621391	0,769861	0,666593	VERDADEIRO	FALSO
16	hbmm_WNADist	4,681072	1,654875	3,798035	1,743667	2,828656	2,178188	VERDADEIRO	FALSO
30	hbmm_WNASurf	5,674839	2,076418	4,867915	2,044307	2,732995	2,381205	VERDADEIRO	FALSO
54	Cross_Link_Order_CA	0,5302	0,623587	0,4327	0,594118	0,850242	0,728307	VERDADEIRO	FALSO
55	Cross_Pres_Order_CA	0,9193	0,81609	0,7874	0,787541	1,12647	0,999821	VERDADEIRO	FALSO
64	Number_Unused_Contact_WNADist	809,797	134,9141	759,2314	142,2088	6,002316	5,33885	VERDADEIRO	FALSO
65	Number_Unused_Contact_WNASurf	1312,878	306,0554	1200,776	298,0835	4,289674	4,028321	VERDADEIRO	FALSO
66	IFR_CA_3	1,108314	0,116701	-2,44E+33	9,11E+35	9,497072	-0,00268	VERDADEIRO	VERDADEIRO
67	Internal_CA_3	1,108817	0,114216	-2,44E+33	9,11E+35	9,708079	-0,00268	VERDADEIRO	VERDADEIRO
3	hbmwm	0,0533	0,268706	0,0715	0,308166	0,198358	0,232018	FALSO	FALSO
4	hbmwwm	0,0384	0,257984	0,0455	0,271884	0,148847	0,167351	FALSO	FALSO
5	hbms	0,194	0,514272	0,2736	0,587139	0,377233	0,465988	FALSO	FALSO
6	hbmws	0,0937	0,364933	0,0985	0,370633	0,25676	0,265762	FALSO	FALSO
7	hbmwws	0,0634	0,32051	0,067	0,336921	0,197809	0,19886	FALSO	FALSO
8	hbss	0,0935	0,333541	0,0904	0,329367	0,280326	0,274466	FALSO	FALSO
9	hbsws	0,0411	0,233563	0,0348	0,216373	0,17597	0,160833	FALSO	FALSO
10	hbswws	0,0289	0,223442	0,0239	0,196043	0,12934	0,121912	FALSO	FALSO
12	aromatic	0,0998	0,360929	0,0812	0,328143	0,276508	0,247453	FALSO	FALSO
13	disulfide	0,0185	0,134713	0,011	0,105258	0,137329	0,104506	FALSO	FALSO
14	ch_attractive	0,3025	1,07779	0,2912	1,024313	0,280667	0,284288	FALSO	FALSO
15	ch_repulsive	0,1155	0,681172	0,1176	0,676453	0,169561	0,173848	FALSO	FALSO
17	hbmwm_WNADist	0,264688	0,447916	0,277987	0,493197	0,590933	0,563643	FALSO	FALSO
18	hbmwwm_WNADist	0,200799	0,457966	0,194999	0,461982	0,438459	0,422092	FALSO	FALSO
19	hbms_WNADist	0,978675	0,837525	1,073367	0,923834	1,168533	1,161861	FALSO	FALSO
20	hbmws_WNADist	0,477603	0,707837	0,440267	0,684749	0,674736	0,642962	FALSO	FALSO
21	hbmwws_WNADist	0,339423	0,67519	0,311242	0,650903	0,502708	0,478169	FALSO	FALSO
22	hbss_WNADist	0,685533	0,731217	0,618781	0,683488	0,937524	0,905329	FALSO	FALSO
23	hbsws_WNADist	0,041719	0,23634	0,035102	0,216184	0,176522	0,162369	FALSO	FALSO

24	hbswws_WNADist	0,029022	0,224192	0,024116	0,195225	0,129452	0,123529	FALSO	FALSO
25	hydrophobic_WNADist	0,188143	0,783536	0,136684	0,686597	0,24012	0,199075	FALSO	FALSO
26	aromatic_WNADist	0,438362	0,539324	0,372973	0,496425	0,812798	0,751318	FALSO	FALSO
27	disulfide_WNADist	0,051703	0,207908	0,037276	0,167015	0,248683	0,223187	FALSO	FALSO
28	ch_attractive_WNADist	1,632589	1,737266	1,442292	1,640522	0,939746	0,879167	FALSO	FALSO
29	ch_repulsive_WNADist	0,605644	1,101074	0,554831	1,080377	0,550048	0,513553	FALSO	FALSO
31	hbmwm_WNASurf	0,449712	0,637472	0,418688	0,611843	0,705461	0,684305	FALSO	FALSO
32	hbmwwm_WNASurf	0,351096	0,695745	0,313478	0,631585	0,504633	0,496336	FALSO	FALSO
33	hbms_WNASurf	1,502912	1,024151	1,484994	1,031294	1,467471	1,439933	FALSO	FALSO
34	hbmws_WNASurf	0,728527	0,977987	0,645835	0,893653	0,744925	0,722692	FALSO	FALSO
35	hbmwws_WNASurf	0,560423	1,043199	0,4959	0,9495	0,537216	0,522275	FALSO	FALSO
36	hbss_WNASurf	1,002229	1,013048	0,906805	0,93946	0,98932	0,965241	FALSO	FALSO
37	hbsws_WNASurf	0,005318	0,037872	0,005584	0,041608	0,140411	0,134207	FALSO	FALSO
38	hbswws_WNASurf	0,004164	0,038018	0,004604	0,043871	0,109533	0,104936	FALSO	FALSO
39	hydrophobic_WNASurf	0,085526	0,356488	0,069378	0,356818	0,239913	0,194437	FALSO	FALSO
40	aromatic_WNASurf	0,588396	0,578222	0,514961	0,544823	1,017595	0,945189	FALSO	FALSO
41	disulfide_WNASurf	0,05041	0,204123	0,036644	0,16537	0,246957	0,221589	FALSO	FALSO
42	ch_attractive_WNASurf	2,58094	2,163301	2,250797	2,025162	1,193056	1,111416	FALSO	FALSO
43	ch_repulsive_WNASurf	0,9831	1,360648	0,860236	1,260724	0,722523	0,682335	FALSO	FALSO
44	Electrostatic_Potential_at_CA	15,5737	19,19273	17,33761	23,41735	0,811437	0,740375	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-38,712	73,82677	-40,6284	80,43714	-0,52436	-0,5051	FALSO	FALSO
46	Electrostatic_Potential_Average	2,572657	23,02186	2,098141	23,40109	0,111748	0,08966	FALSO	FALSO
47	Hydrophobicity_KDI	-418,557	2005,128	-362,214	1869,802	-0,20874	-0,19372	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	138,712	86,49474	131,2968	81,69495	1,603704	1,607159	FALSO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-355,375	219,3495	-325,132	211,2293	-1,62013	-1,53924	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	26,21984	102,5745	24,73434	94,55227	0,255618	0,261594	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	125,4325	77,01161	117,4322	71,81472	1,628748	1,635211	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-322,901	198,7921	-293,587	188,4152	-1,62432	-1,55819	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	24,1712	91,43454	23,15486	84,05394	0,264355	0,275476	FALSO	FALSO
58	Dihedral_Chi1	-22,6971	102,6174	-23,5988	93,42205	-0,22118	-0,2526	FALSO	FALSO
59	Dihedral_Chi2	14,01214	94,89027	8,510403	90,12381	0,147667	0,09443	FALSO	FALSO
60	Dihedral_Chi3	-0,79333	48,64942	-0,33011	49,90414	-0,01631	-0,00661	FALSO	FALSO

61	Dihedral_Chi4	-0,86258	41,8262	-0,31727	42,6436	-0,02062	-0,00744	FALSO	FALSO
62	Temperature_Factor_CA	21,5489	17,80473	23,79745	20,09753	1,210291	1,184098	FALSO	FALSO
63	Number_Unused_Contact	206,9409	66,36156	212,4407	66,53085	3,118385	3,193116	FALSO	FALSO
68	Clash	0,294091	0,137667	0,303711	0,140466	2,136247	2,16217	FALSO	FALSO
69	Percent	9,437137	4,272178	9,751684	4,412914	2,208976	2,209806	FALSO	FALSO

Tabela 58. Descritores usados no teste MANOVA para as proteínas do tipo β em $(\alpha+\beta)+(\alpha/\beta)$ alinhadas pelo N-Terminal.

Descritores		\bar{x}_{dentro}	σ_{dentro}	\bar{x}_{fora}	σ_{fora}	D_{norm}	F_{norm}	$ D_{norm} - F_{norm} < 0,1$	$ D_{norm} - F_{norm} < 1,0$
1	Accessible_Surface_in_Isolation	31,03323	38,67496	50,29988	48,05858	0,802411	1,046637	VERDADEIRO	FALSO
2	hbm	2,082	1,299027	1,9084	1,388511	1,602738	1,374422	VERDADEIRO	FALSO
49	Electrostatic_Potential_at_LHA_WNADist	-401,825	179,4696	-375,557	175,7192	-2,23896	-2,13726	VERDADEIRO	FALSO
54	Cross_Link_Order_CA	0,6526	0,739435	0,4755	0,66279	0,882566	0,717421	VERDADEIRO	FALSO
55	Cross_Pres_Order_CA	1,3047	0,969279	1,0066	0,92683	1,346052	1,086068	VERDADEIRO	FALSO
66	IFR_CA_3	0,729884	0,52675	-3,12E+32	3,26E+35	1,385636	-0,00096	VERDADEIRO	VERDADEIRO
67	Internal_CA_3	1,055362	0,235258	-3,12E+32	3,26E+35	4,485987	-0,00096	VERDADEIRO	VERDADEIRO
68	Clash	0,290498	0,134617	0,302635	0,152654	2,157959	1,982493	VERDADEIRO	FALSO
69	Percent	9,368293	4,198839	9,746057	4,768054	2,231163	2,044032	VERDADEIRO	FALSO
3	hbmwm	0,0697	0,308907	0,0822	0,332357	0,225634	0,247325	FALSO	FALSO
4	hbmwwm	0,0348	0,228998	0,0433	0,252534	0,151966	0,171462	FALSO	FALSO
5	hbms	0,2574	0,600401	0,3543	0,671019	0,428714	0,528003	FALSO	FALSO
6	hbmws	0,1179	0,420318	0,119	0,415899	0,280502	0,286127	FALSO	FALSO
7	hbmwws	0,0652	0,335272	0,0756	0,356917	0,194469	0,211814	FALSO	FALSO
8	hbss	0,1153	0,373365	0,0983	0,343076	0,308813	0,286526	FALSO	FALSO
9	hbsws	0,0443	0,247434	0,0371	0,224645	0,179038	0,16515	FALSO	FALSO
10	hbswws	0,0285	0,219201	0,0261	0,206539	0,130018	0,126369	FALSO	FALSO
11	hydrophobic	4,9471	4,609854	4,2875	4,353568	1,073158	0,984824	FALSO	FALSO
12	aromatic	0,1351	0,451607	0,115	0,418048	0,299154	0,275088	FALSO	FALSO
13	disulfide	0,0056	0,075558	0,0038	0,062324	0,074115	0,060972	FALSO	FALSO
14	ch_attractive	0,3407	1,291443	0,3559	1,296059	0,263813	0,274602	FALSO	FALSO
15	ch_repulsive	0,1643	0,903869	0,1698	0,905331	0,181774	0,187556	FALSO	FALSO
16	hbm WNADist	8,27475	4,670504	7,621372	4,557277	1,771704	1,672352	FALSO	FALSO
17	hbmwm WNADist	0,328132	0,49957	0,328123	0,53172	0,656829	0,617096	FALSO	FALSO
18	hbmwwm WNADist	0,191669	0,395473	0,189471	0,422375	0,484658	0,448584	FALSO	FALSO
19	hbms WNADist	1,302989	0,930099	1,401985	1,017013	1,400914	1,378532	FALSO	FALSO
20	hbmws WNADist	0,547386	0,737267	0,508925	0,723162	0,742452	0,703749	FALSO	FALSO
21	hbmwws WNADist	0,35068	0,639443	0,346632	0,658985	0,548414	0,52601	FALSO	FALSO
22	hbss WNADist	0,723947	0,703146	0,632049	0,660377	1,029583	0,957103	FALSO	FALSO
23	hbsws WNADist	0,045722	0,251266	0,03834	0,228266	0,181966	0,167963	FALSO	FALSO

24	hbswws_WNADist	0,029519	0,222995	0,027024	0,210411	0,132376	0,128432	FALSO	FALSO
25	hydrophobic_WNADist	0,149484	0,718073	0,17175	0,787122	0,208174	0,2182	FALSO	FALSO
26	aromatic_WNADist	0,57689	0,633346	0,502617	0,598967	0,910862	0,839141	FALSO	FALSO
27	disulfide_WNADist	0,011957	0,091967	0,009389	0,07779	0,130011	0,120694	FALSO	FALSO
28	ch_attractive_WNADist	1,819646	2,136605	1,700732	2,142709	0,851653	0,79373	FALSO	FALSO
29	ch_repulsive_WNADist	0,840221	1,452837	0,791254	1,483441	0,578331	0,533391	FALSO	FALSO
30	hbmms_WNASurf	11,55512	7,273231	10,65657	6,795551	1,588719	1,568169	FALSO	FALSO
31	hbmwm_WNASurf	0,499182	0,652203	0,464839	0,630054	0,765379	0,737777	FALSO	FALSO
32	hbmwwm_WNASurf	0,30602	0,558244	0,28121	0,535737	0,548183	0,524903	FALSO	FALSO
33	hbms_WNASurf	1,899247	1,101079	1,83459	1,075577	1,724896	1,70568	FALSO	FALSO
34	hbmws_WNASurf	0,76205	0,912478	0,689581	0,858753	0,835144	0,803003	FALSO	FALSO
35	hbmwws_WNASurf	0,533107	0,882307	0,500017	0,849144	0,604219	0,588849	FALSO	FALSO
36	hbss_WNASurf	0,989895	0,916702	0,873309	0,860844	1,079843	1,01448	FALSO	FALSO
37	hbsws_WNASurf	0,004122	0,031844	0,005046	0,039064	0,12943	0,129172	FALSO	FALSO
38	hbswws_WNASurf	0,003048	0,030675	0,003956	0,03877	0,099371	0,102028	FALSO	FALSO
39	hydrophobic_WNASurf	0,06554	0,316781	0,084157	0,381853	0,206895	0,220391	FALSO	FALSO
40	aromatic_WNASurf	0,749745	0,63024	0,66073	0,602516	1,189618	1,096617	FALSO	FALSO
41	disulfide_WNASurf	0,011721	0,091004	0,009172	0,076908	0,128797	0,119262	FALSO	FALSO
42	ch_attractive_WNASurf	2,953894	2,883011	2,687881	2,794883	1,024586	0,961715	FALSO	FALSO
43	ch_repulsive_WNASurf	1,36371	1,988461	1,258973	1,977153	0,685812	0,636761	FALSO	FALSO
44	Electrostatic_Potential_at_CA	13,55891	20,10952	17,91958	23,2618	0,674253	0,770344	FALSO	FALSO
45	Electrostatic_Potential_at_LHA	-40,8112	73,47934	-44,5405	80,12876	-0,55541	-0,55586	FALSO	FALSO
46	Electrostatic_Potential_Average	-0,43955	21,67905	0,888775	21,66539	-0,02028	0,041023	FALSO	FALSO
47	Hydrophobicity_KDI	-493,579	2168,966	-488,176	2156,32	-0,22756	-0,22639	FALSO	FALSO
48	Electrostatic_Potential_at_CA_WNADist	129,363	69,47139	127,482	65,17028	1,862105	1,956137	FALSO	FALSO
50	Electrostatic_Potential_Average_WNADist	6,269488	69,90916	8,09986	65,03656	0,08968	0,124543	FALSO	FALSO
51	Electrostatic_Potential_at_CA_WNASurf	107,2215	66,20778	102,7958	62,05538	1,61947	1,656517	FALSO	FALSO
52	Electrostatic_Potential_at_LHA_WNASurf	-334,56	186,3277	-309,995	176,802	-1,79555	-1,75335	FALSO	FALSO
53	Electrostatic_Potential_Average_WNASurf	6,291972	56,81541	6,828023	52,89332	0,110744	0,12909	FALSO	FALSO
58	Dihedral_Chi1	-2,46903	175,1893	-5,59765	167,5797	-0,01409	-0,0334	FALSO	FALSO
59	Dihedral_Chi2	32,46951	166,2462	28,18229	163,7362	0,19531	0,17212	FALSO	FALSO
60	Dihedral_Chi3	19,10784	146,6827	18,49791	146,0509	0,130266	0,126654	FALSO	FALSO

61	Dihedral_Chi4	19,53082	144,315	19,03042	143,0075	0,135335	0,133073	FALSO	FALSO
62	Temperature_Factor_CA	38,45127	38,89366	42,25672	41,37712	0,988626	1,021258	FALSO	FALSO
63	Number_Unused_Contact	164,8254	89,05972	171,8055	90,4743	1,850729	1,898943	FALSO	FALSO
64	Number_Unused_Contact_WNADist	621,609	272,3635	594,6501	259,317	2,282277	2,29314	FALSO	FALSO
65	Number_Unused_Contact_WNASurf	989,102	456,2969	917,6727	418,081	2,167672	2,194964	FALSO	FALSO

ANEXO 1 – ARTIGO PUBLICADO

RESEARCH ARTICLE

Study of specific nanoenvironments containing α -helices in all- α and $(\alpha+\beta)+(\alpha/\beta)$ proteins

Ivan Mazoni¹, Luiz César Borro², José Gilberto Jardine³, Inácio Henrique Yano¹, José Augusto Salim⁴, Goran Neshich^{1*}

1 Embrapa Agricultural Informatics, Campinas, São Paulo, Brazil, 2 Institute of Biology, University of Campinas, Campinas, São Paulo, Brazil, 3 Embrapa Territorial Management, Campinas, São Paulo, Brazil, 4 Research Center on Biodiversity and Computing, University of São Paulo, São Paulo, São Paulo, Brazil

* goran.neshich@embrapa.br



Abstract

Protein secondary structure elements (PSSEs) such as α -helices, β -strands, and turns are the primary building blocks of the tertiary protein structure. Our primary interest here is to reveal the characteristics of the nanoenvironment formed by both PSSEs and their surrounding amino acid residues (AARs), which might contribute to the general understanding of how proteins fold. The characteristics of such nanoenvironments must be specific to each secondary structure element, and we have set our goal here to gather the fullest possible description of the α -helical nanoenvironment. In general, this postulate (the existence of specific nanoenvironments for specific protein substructures/neighbourhoods/regions with distinct functionality) was already successfully explored and confirmed for some protein regions, such as protein-protein interfaces and enzyme catalytic sites. Consequently, PSSEs were the obvious next choice for additional work for further evidence showing that specific nanoenvironments (having characteristics fully describable by means of structural and physical chemical descriptors) do exist for the corresponding and determined intraprotein regions. The nanoenvironment of α -helices (nEo α H) is defined as any region of the protein where this secondary structure element type is detected. The nEo α H, therefore, includes not only the α -helix amino acid residues but also the residues immediately around the α -helix. The hypothesis that motivated this work is that it might in fact be possible to detect a postulated “signal” or “signature” that distinguishes the specific location of α -helices. This “signal” must be discernible by tracking differences in the values of physical, chemical, physicochemical, structural and geometric descriptors immediately before (or after) the PSSE from those in the region along the α -helices. The search for this specific nanoenvironment “signal” was made possible by aligning previously selected α -helices of equal length. Afterward, we calculated the average value, standard deviation and mean square error at each aligned residue position for each selected descriptor. We applied Student’s t-test, the Kolmogorov-Smirnov test and MANOVA statistical tests to the dataset constructed as described above, and the results confirmed that the hypothesized “signal”/“signature” is both existing/identifiable and capable of distinguishing the presence of an α -helix inside the specific nanoenvironment, contextualized as a specific region within the

OPEN ACCESS

Citation: Mazoni I, Borro LC, Jardine JG, Yano IH, Salim JA, Neshich G (2018) Study of specific nanoenvironments containing α -helices in all- α and $(\alpha+\beta)+(\alpha/\beta)$ proteins. PLoS ONE 13(7): e0200018. <https://doi.org/10.1371/journal.pone.0200018>

Editor: Eugene A. Permyakov, Russian Academy of Medical Sciences, RUSSIAN FEDERATION

Received: February 23, 2018

Accepted: June 18, 2018

Published: July 10, 2018

Copyright: © 2018 Mazoni et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All initial data are drawn from STING RDB which could be accessed at http://www.cbi.cnptia.embrapa.br/SMS/index_s.html. Data are additionally available at: https://figshare.com/projects/Structural_and_physical-chemical_characterization_of_alpha-helices/35462.

Funding: Funded by Embrapa Internal Financing.

Competing interests: The authors have declared that no competing interests exist.

whole protein. However, such conclusion might rarely be reached if only one descriptor is considered at a time. A more accurate signal with broader coverage is achieved only if one applies multivariate analysis, which means that several descriptors (usually approximately 10 descriptors) should be considered at the same time. To a limited extent (up to a maximum of 15% of cases), such conclusion is also possible with only a single descriptor, and the conclusion is also possible in general for up to 50–80% of cases when no less than 5 nonlinear descriptors are selected and considered. Using all the descriptors considered in this work, provided all assumptions about data characteristics for this analysis are met, multivariate analysis regularly reached a coverage and accuracy above 90%. Understanding how secondary structure elements are formed and maintained within a protein structure could enable a more detailed understanding of how proteins reach their final 3D structure and consequently, their function. Likewise, this knowledge may also improve the tools used to determine how good a structure is by means of comparing the “signal” around a selected PSSE with the one obtained from the best (resolution and quality wise) protein structures available.

Introduction

Crick [1] explained that proteins are uniquely essential to maintaining life. Although proteins can perform almost any type of role in animals, plants or microorganisms, so far, the most studied protein function is definitely its enzymatic role. A precisely folded 3D structure is necessary for a protein to perform its function, and if a protein has been unfolded, e.g., by heat or some chemical agent, the protein will lose its biological function. Before a protein assumes its final and correct 3D structure, the construction of its PSSEs must have been completed.

Anfinsen [2] experimentally confirmed that the amino acid sequence of a protein provides all the necessary information for the protein to assume its precise and final 3D structure. Although currently it is known that this is only a specific case in such an experiment, Anfinsen stimulated several attempts to predict the 3D structure with the amino acid sequence alone. Several methodologies appeared during the last 4–5 decades, such as homology modelling [3–7], *ab initio* modelling [8–12], and threading [13–17]. Another possible approach to predicting the final 3D structure of a protein is to understand how the protein forms and maintains the different PSSEs during and after folding. Several methods were developed to predict PSSEs. For example, PSSE predictions have been based on circular dichroism data [18–20], on multi-step learning coupled with a prediction of the solvent accessible surface area, and on backbone torsion angles [21–23], as well as on machine learning techniques [24–26]. The results of the aforementioned techniques are much more accurate now than the early results (from the 1980s) for this particular scientific area. For example, the accuracy of these methods increased from 56% in 1983 [27] to more than 80% in 2015 [28].

The pattern of formed hydrogen bonds between the amino and carboxyl groups along the PSSE, together with their ϕ and ψ dihedral angles inside a particular region of the Ramachandran plot [29], fully defines the PSSE. There are three groups of secondary structure elements: helical structures, β -sheets, and turns. Helical structures are subdivided into helix 2₇, α -helix, helix 3₁₀, and π -helix. β -sheet structures may be parallel, anti-parallel and β -bridge. Finally, a turn may be a tight *turn*, multiple *turns*, hairpins and *turns* of type I, II, VIII, I', II', VIa1, VIa2, VIb, and IV.

Several algorithms explore the regularity of hydrogen-bond patterns of PSSEs and are frequently employed to identify and distinguish PSSE types. The DSSP algorithm uses the

hydrogen-bond pattern recognition and the geometric characteristics extracted from the spatial coordinates of the atoms of the amino acid residues to characterize a secondary structure element [30]. The Stride algorithm uses, in addition to the information used by DSSP, the dihedral angle potentials to characterize a secondary structure element [31]. Cuff [32] showed that the DSSP and Stride definitions agree with each other in 95% of the explored cases. An excellent tool to observe the PSSE definition differences is ^{Java}Protein Dossier (¹PD) [33] of the BlueStar STING suite of programs.

Proteins can be grouped into SSE classes using the SCOP structure classification [34]: all- α , all- β , $\alpha+\beta$ and α/β . “All- α ” proteins are those that have only helical secondary structure elements present in the protein (along with some turns and irregular portions) but crucially, no β -sheets. The “all- β ” structures are those that have only β -sheets present in the protein (along with some turns and irregular portions) but no helical structures. In $\alpha+\beta$ proteins, both α -helices and β -strands are present but are largely segregated. The most frequently detected β -strands in the $\alpha+\beta$ type of proteins are antiparallel ones [35]. In the α/β type of proteins (when the α -helices and β -strands are alternately following each other in the protein structure), the β -strands are mostly organized in a parallel fashion [35]. Finally, there are intrinsically disordered proteins, where neither α -helices nor β -strands are present in the protein’s 3D structure [36]. Using the Structural Classification of Proteins (SCOP) database [37] and the Protein Data Bank (PDB) [38], we constructed a *Datamart*, called General SSEs, which performed the following classifications: 1606 protein chains as all- α – 12 containing a single or exclusive helix and 1594 nonexclusive helical chains (meaning they have more than one helix); 19407 protein chains as $(\alpha+\beta)+(\alpha/\beta)$ – 99 “exclusive helix” chains and 19308 nonexclusive helix chains (based on PDB data from August 8, 2016).

This work focuses on α -helical elements and their nanoenvironment in the all- α proteins and the α -helices in $(\alpha+\beta)+(\alpha/\beta)$ classes of proteins.

Neshich and coworkers [39] introduced the concept of an intraprotein nanoenvironment, and Moraes and coworkers [40] tested this idea for the first time in a protein interface study.

In this work, we considered the set of amino acid residues located within and around α -helices—the α -helical nanoenvironment. The amino acid residues that form the secondary structure element plus the amino acid residues in their vicinity yield a complete neighbourhood with its own, very specific nanoenvironment characteristics (Fig 1).

We used a somewhat arbitrary number, which is still empirically considered the most suitable number, of thirty-two amino acid residues, before and after the secondary structure element in the primary sequence to define the total length of the protein sequence to be studied. The region defined this way is used for analysis so that the corresponding structure fragments can be aligned during a procedure and so that descriptor values can then be inspected for the appearance of the hypothesized specific nanoenvironment “signal”. The spatial inclusion of AARs other than those belonging to the PSSE itself was achieved by defining the radius of a sphere of an “AAR enclosure” drawn from the α -carbon atom of any residue of the selected PSSE. As previously hinted, this procedure was performed for the stretch of thirty-two AARs before a PSSE’s N-terminal and after its C-terminal.

Regarding the importance of this work, we certainly argue for possible positive implications of the act of acquiring detailed knowledge about PSSEs and their nanoenvironments and therefore making it possible to perform the following:

- a. better understand the protein folding process;
- b. improve existing and design new computational tools for quality validation of the secondary structure element location, extension, and internal geometry in protein structure models obtained either by employing protein structure modelling software, such as Modeller

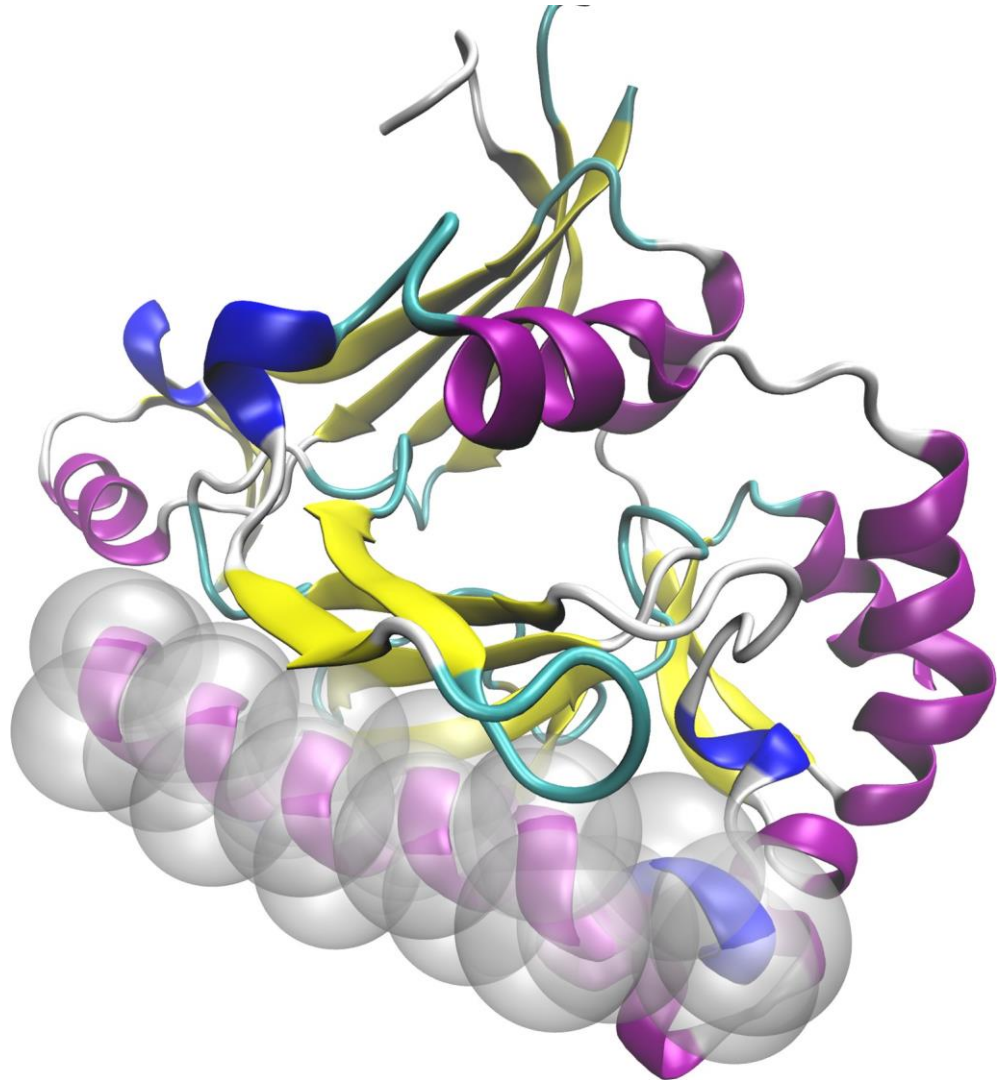


Fig 1. An example of an α -helix (in a specific $\alpha+\beta$ protein) and its nanoenvironment: The synthetic gene encoded DepS bound to the inhibitor DG157493 (3bl9.pdb) has fourteen α -helices, and each helix has its own nanoenvironment. Highlighted inside the transparent spheres is an α -helix (ribbon, purple). The nanoenvironment includes the amino acid residues of the α -helix and the amino acid residues around the helix that are within reach of the probing sphere, whose radius was previously selected. The pre- and postregions (extension by 32 AARs each) are not shown here for the sake of clarity of the basic definition.

<https://doi.org/10.1371/journal.pone.0200018.g001>

[41] and Swiss-Model [42], or by X-ray crystallography, NMR, and electronic microscopy—and in all cases by inspecting the PSSE environment’s characteristics and their “signal-like” behaviour compared to that of a “signal” obtained from the best quality reference structures.

The knowledge about the relationship between protein amino acid sequences, their 3D structure, and their function will enable improvements in many applications, eventually allowing the proposal of substances such as new vaccines, drugs, veterinary drugs, insecticides, all with more efficacy [43]. For example, if one starts from a known genome (human, animal, plant or microorganism) for which it is already possible to obtain an annotated protein sequence, then with the acquired knowledge we are compiling and reporting in this work, the

improved secondary structure element prediction will also improve the prediction of the whole protein's 3D structure. Accordingly, this information will lead to finding out how to better develop new protein function inhibitors (e.g., bactericides, pesticides, insecticides, and vaccines). It could therefore be feasible to more precisely simulate enzymatic reactions, protein-protein interactions, and protein-substrate interactions, with all of the simulations being faster, more accurate and quite possibly, less expensive [44].

Materials and methods

We extracted the data for analysis of the PSSE nanoenvironment from STING_RDB [45]. The STING RDB had 9,320,604,319 records in 98 tables (based on PDB data from August 8, 2016) and included physical, chemical, physicochemical, structural and geometric descriptors (reported in “per amino acid residue” fashion, for each protein chain) of each protein structure in the PDB.

All of the raw data, ready for type of processing that we did in this paper, are available at: https://figshare.com/projects/Structural_and_physical-chemical_characterization_of_alpha-helices/35462

Two new STING modules for evaluating the PSSE nanoenvironment: PS³A and PS³DV

This work expanded the Blue Star STING platform [46] by adding two new modules. The first one is called Protein Secondary Structure STING Analyzer (PS³A). The PS³A, written in the JS and PHP programming languages, allows a user to set some options for the fine tuning of PSSE analysis: PSSE length, consensus type (relative to the definition of the PSSE), redundancy level, and a selector for depicting one of the 69 different descriptors available for the PSSE nanoenvironment. Additionally, the user may obtain the data reliability plot, the sequence “logo” (indicating the local conservation of amino acid residues) and the empirical cumulative distribution function (ECDF) curve. The ECDF curve shows how the descriptor value levels inside the PSSE environment are different from the values outside this environment. Users may access the PS³A at <https://www.ps3a.cbi.cnptia.embrapa.br/>.

PS³DataVizualizer is the second new STING module that was added as a visualization tool for PS³A, and it offers users the possibility to visualize any of the 69 different STING descriptors available in separate plots that are produced for the nanoenvironments of α -helices (later, we will also have PS³DV for β -strands and turns as well). The capabilities of PS³DataVizualizer make it possible to observe the PSSE in two ways: exclusive α -helices (only a single helical structure segment is present in the protein's whole structure) and nonexclusive α -helices (more than one helix is present per protein analysed). Users may select any image and keep viewing a carousel of more than two thousand images representing combinations of the selected size of the PSSE and a multitude of protein descriptors drawn from STING_DB. These plots were produced using the R package with a high-quality image generator enabled. Images may be saved by the user for later analysis. We developed PS³DV in HTML5 integrated with JS and also used the jQuery and Bootstrap technologies. PS³DV is accessible at <https://www.ps3dv.cbi.cnptia.embrapa.br/>

Initial hypothesis verification

The hypothesis explored and evaluated here (and contextualized in the biological sense) was as follows: the nanoenvironment descriptor values inside the PSSE are, or are not, statistically equal to the values outside the PSSE. The same hypothesis contextualized statistically was as

follows: H_0 —the descriptor value distributions are the same, and H_1 —the descriptor value distributions are not the same (again: inside vs outside the PSSE nanoenvironment).

Three statistical tests were applied to verify the hypothesis. The first one, the Kolmogorov-Smirnov test (KSt) [47], was used to compare a sample of interest to one with a reference probability distribution. As a nonparametric test of the equality of continuous, the KSt may also be used to compare two samples. The Kolmogorov-Smirnov statistics quantify the distance between the empirical distribution function of the cumulative distribution for the reference distribution and that of the analysed sample or the distance between the two empirical distribution functions of two samples. The empirical distribution function F_n for n observations X_i is defined as shown in the equation:

$$F_0(x) = \frac{1}{n} \sum_{i=1}^n I_{[-\infty, x]}(X_i)$$

where $I_{[-\infty, x]}(X_i)$ is the indicator function, is equal to 1 if $X_i < x$ and is otherwise equal to 0.

The Kolmogorov-Smirnov statistic for F_n is calculated by finding the supremum (or the greatest lower bound) absolute value of the differences between the two samples:

$$D_n = \sup |F_n(x) - F(x)|$$

where \sup_x is the supremum of the set of distances.

The distance D_n is used to calculate the p-value that indicates whether the numbers differ significantly. The null hypothesis H_0 is rejected if the p-value is close to zero.

The second statistical test applied was the Q-Q plot [48], which we used here to support Student's t-test [49]. The Q-Q plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the data have a normal distribution, then Student's t-test may be used. Student's t-test for different sample sizes may be calculated using the equation below:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{x_1 x_2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where \bar{x}_i is the sample average, n_i is the sample size, $S_{x_1 x_2} = \sqrt{\frac{(n_1-1)S_{x_1}^2 + (n_2-1)S_{x_2}^2}{n_1 + n_2 - 2}}$, and the degrees of freedom used in this test are $n_1 + n_2 - 2$.

Descriptor value sliding window

Another approach used in this work to confirm the existence of the hypothesized “signal” (where the word “signal” is used here to designate “something that shows that something else exists or is likely to happen”) was to run a “sliding window” along the positional/structural alignment for average descriptor values of amino acid residues belonging to the pre-, post- and PSSE regions. For each descriptor, a limited (PSSE size wise) sliding window test was performed by varying the PSSE length between 1 and the maximum number of amino acid residues found (this number can be some of the PSSE sizes found in the STING RDB and actually found in some specific derived *Datamarts*, which are described below). For such selected lengths, we collected all amino acid residue descriptor values, and the average value was calculated. We then grouped and stored these data using an R script. Student's t-test was applied while dividing the data into two sets: inside and outside the “sliding window”. If the p-value within the region where the window matches (exactly covers) the PSSE length, approached zero, then the environment/sample inside the window region was considered statistically different from the region located outside that window. The results obtained

confirm that the descriptor values inside the PSSE are statistically distinct from the values outside (Fig 2).

Elimination of redundancies

The presence of redundant sequences inevitably introduces bias in the statistical tests, masking the real variance in the descriptor values along the PSSE positional alignment. We eliminated sequence redundancy (along the PSSE length) at the 95%, 70%, and 50% levels of similarity.

Building PS³A Datamarts

The first step undertaken in this work was to obtain all the protein structures from STING_RDB containing at least one α -helix. Based on the PDB, DSSP and Stride definitions, the all- α and $(\alpha+\beta)+(\alpha/\beta)$ proteins were filtered and stored for further analysis. The second step was to eliminate the protein sequence redundancy at the 95%, 70%, and 50% levels of similarity using CD-HIT software [50]. The third step was to group the α -helices according to the consensus detected among the different algorithms used here to identify/characterize the PSSEs. The most rigid consensus requires a consensus for all three selected algorithms: PDB, DSSP and

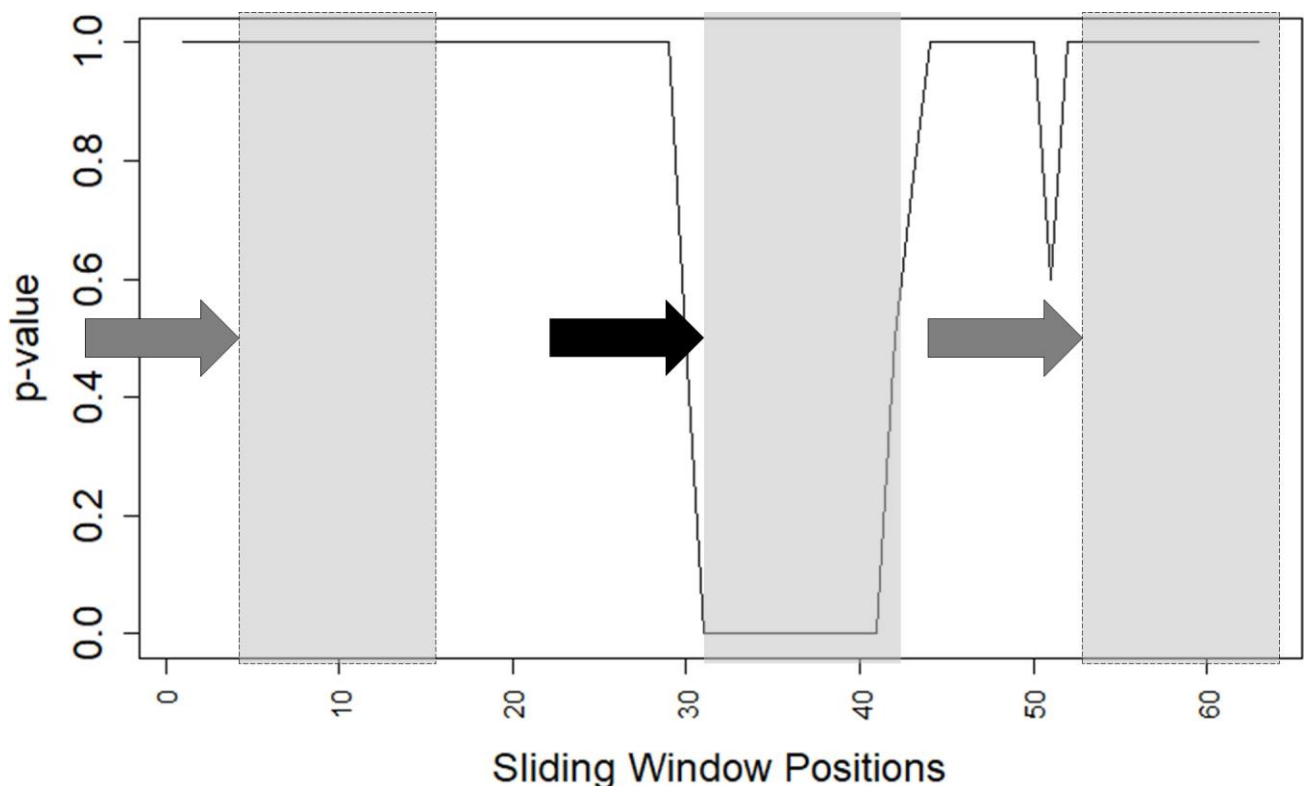


Fig 2. The p-value of Student's t-test evaluation for a selected descriptor value along the "sliding window" for positionally aligned PSSE sequences. The coverage of the sequence containing a PSSE is from the N- to the C-terminal ends (± 32 AAR). The sequences includes the PSSE plus 32 residues before its N-terminal and 32 residues after its C-terminal. The "sliding window" size in this particular case is the same size as the selected PSSE length (12 AAR). Student's t-test is used for each position of the sliding window. This test measures how much the data inside the "sliding window" differ from the data outside the windows. The p-values are shown along the y-axes. A p-value that approaches zero in any particular region means that within this region, the descriptor values differ from the values outside the region in a statistically significant manner. The arrows indicate the direction of movement for the "sliding window box" (shown here before, at and after the PSSE), and the solid arrow indicates the exact position of the N-terminal of the PSSE. Shaded boxes indicate the size of the sliding window placed at three specific positions. The region with a p-value approximating zero coincides with the positional alignment of the α -helix that has the exact same size. The sharp invagination around AAR position 52 is not as representative (too short compared to the PSSE under investigation) as the one directly on top and over the whole analysed PSSE.

<https://doi.org/10.1371/journal.pone.0200018.g002>

Stride, i.e., the α -helices should have the same length according to all three definitions. Another possible consensus would be based on the definitions of only two algorithms: PDB- DSSP, PDB-Stride, and DSSP-Stride (Fig 3). The fourth step in this work was to positionally/ structurally align the α -helices of equal length. The statistics for such PSSE compilations (tabulated below) and some examples of plots using the PDB-DSSP, PDB-Stride and DSSP-Stride consensus are available in the Supporting Information (Figure A to Figure C in [S1 File](#)).

Before proceeding to the next step, the names of the basic *Datamarts* used frequently in this work are listed: DM1, DM1_e, DM1_ne, DM2, DM2_e, and DM2_ne. There is a complete description of these *Datamarts* in [Table 1](#), “Number of Helices of Different Sizes in All α and α + β and α/β Proteins”. “DM1_e” shows the number of helices for each length in all- α proteins for the so-called “exclusive” helices (or just one helix in the whole protein structure). “DM1_ne” shows the same information but for the cases involving more than one helix (*nonexclusive* helices). DM1 is the sum of DM1_e and DM1_ne. DM2 follows the same nomenclature but considers α + β and α/β proteins.

Positional/structural alignment of α -helices

The positional/structural alignment was made considering fixed lengths of α -helices for each of the sizes found in our *Datamart* starting with PSSEs that were a minimum of five amino acid residues long and ending with the maximum encountered value. As we mentioned above, to analyse the nanoenvironment of the α -helices, the observation field was extended by 32 amino acid residues before and after the N- and C-terminal ends (marked with “*”, meaning that any residue might be found at that position). However, those residues might be in any PSSE class, except for the H right before the N-terminal end and after the C-terminal end of the analysed PSSE. However, in cases of missing residues (to complete the desired 32 ones before or after the selected PSSE), the residues were noted by using gaps in these positions: “-” (Fig 4).

Descriptor selection

We selected 69 descriptors from the extensive list of STING_RDB descriptors and grouped them in seven categories ([Table 2](#)). These descriptors describe the characteristics of the nanoenvironment where the α -helices form. In the Supporting Information, we offer an explanation and detailed description of these descriptors.

It is not within the scope of this manuscript to discuss in detail the descriptors used here. The readers are welcome to see the description in [40]; in [51]; and at the Sting web site: http://www.cbi.cnpia.embrapa.br/SMS/STINGm/help/MegaHelp_JPD.html

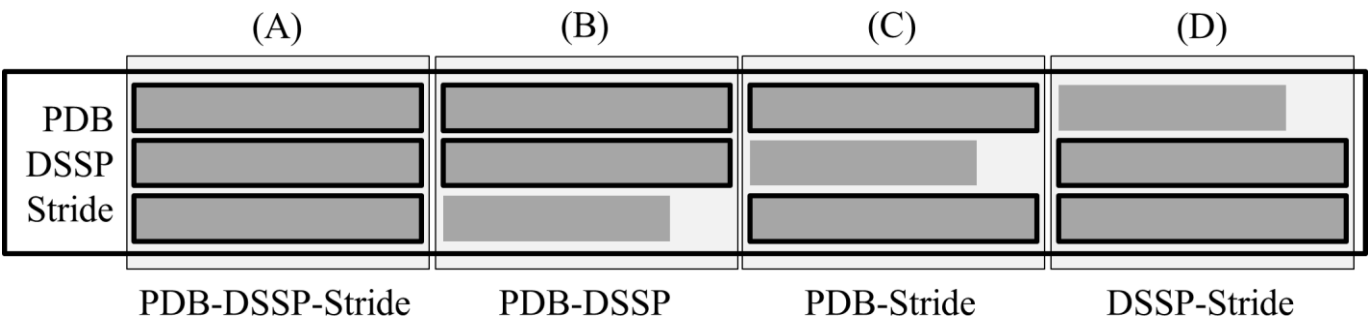


Fig 3. Grouping of same-length α -helices using consensus definitions based on the PDB, DSSP and Stride classifications. There are four possible consensus groups. (A) PDB-DSSP-Stride: when the secondary structure element starts and finishes at the same corresponding amino acid residue location and hence, has the same length according to the PDB, DSSP and Stride definitions. (B), (C) and (D) when the secondary structure elements start but do NOT finish at the same amino acid residue, as defined by one of the three criteria used: PDB-DSSP, PDB-Stride, and DSSP-Stride definitions, respectively.

<https://doi.org/10.1371/journal.pone.0200018.g003>

Table 1. Number of helices of different lengths in all α and $\alpha+\beta$ and α/β proteins.

DM1		DM1_e		DM1_ne		DM2		DM2_e		DM2_ne	
all- α (exc. + nexc.)		all- α (exc.)		all- α (nexc.)		$\alpha+\beta$ (exc. + nexc.)		$\alpha+\beta$ (exc.)		$\alpha+\beta$ (nexc.)	
Helix size	# of helices	Helix size	# of helices	Helix size	# of helices	Helix size	# of helices	Helix size	# of helices	Helix size	# of helices
5	81			5	81	5	1463	5	1	5	1462
6	177	6	1	6	176	6	1768	6	14	6	1754
7	97	7	2	7	95	7	1549	7	10	7	1539
8	133	8	1	8	132	8	1173	8	3	8	1170
9	62			9	62	9	1212	9	2	9	1210
10	100			10	100	10	1619	10	6	10	1613
11	107	11	1	11	106	11	1692	11	12	11	1680
12	87			12	87	12	1657	12	3	12	1654
13	111			13	111	13	1222	13	1	13	1221
14	92	14	2	14	90	14	1215	14	5	14	1210
15	131	15	1	15	130	15	882	15	6	15	876
16	44			16	44	16	643	16	1	16	642
17	27			17	27	17	572	17	28	17	544
18	50			18	50	18	490	18	1	18	489
19	52			19	52	19	392			19	392
20	20	20	2	20	18	20	323	20	3	20	320
21	21			21	21	21	428	21	1	21	427
22	23			22	23	22	194			22	194
23	12			23	12	23	99			23	99
24	18			24	18	24	123			24	123
25	41			25	41	25	126			25	126
26	4			26	4	26	61			26	61
27	14			27	14	27	106			27	106
28	19			28	19	28	56			28	56
29	10			29	10	29	84			29	84
30	3	30	1	30	2	30	41	30	1	30	40
31	9			31	9	31	46			31	46
32	18			32	18	32	60			32	60
33	12			33	12	33	37			33	37
						34	17			34	17
35	4			35	4	35	6			35	6
36	2			36	2	36	4			36	4
37	1			37	1	37	5			37	5
38	1			38	1	38	3			38	3
39	1			39	1	39	1			39	1
40	2	40	1	40	1	40	4	40	1	40	3
						41	1			41	1
42	1			42	1	42	1			42	1
43	1			43	1	43	7			43	7
44	1			44	1	44	1			44	1
45	2			45	2	45	3			45	3
48	1			48	1	48	1			48	1
50	2			50	2	50	2			50	2
51	7			51	7	51	7			51	7
						54	1			54	1

(Continued)

Table 2. List of STING_RDB descriptors used in this work. Although the Supporting Information contains a detailed description, here, we explain some of the acronyms used.

Structural (I)	18. HBMWS * (56)	35. HBMS * (55)
1. Temperature_Factor_CA *	19. HBMWWS *	36. HBMWS * (56)
2. Dihedral_Angle_PHI *	20. HBSS	37. HBMWWS * (57)
3. Dihedral_Angle_PSI *	21. HBSWS	38. HBSS * (58)
4. Dihedral_Chi1 *	22. HBSWWS	39. HBSWS (59)
5. Dihedral_Chi2	23. Hydrophobic	40. HBSWWS * (60)
6. Dihedral_Chi3	24. Aromatic	41. Hydrophobic (61)
7. Dihedral_Chi4	25. Ch_attractive *	42. Aromatic (62)
8. Density IFR	26. Ch_repulsive	43. Ch_attractive * (63)
9. Density Internal *	27. Disulfide *	44. Ch_repulsive (64)
10. Space Clash number of clashes *	Unused Contacts (IV)	45. Disulfide * (65)
11. Space Clash percent *	28. Number_Unused_Contact	46. Number_Unused_Contact * (66)
Geometric (II)	Physical Chemical (V)	47. Electrostatic_Potential_at_CA (67)
12. Cross_Link_Order_CA *	29. Electrostatic_Potential_at_CA *	48. Electrostatic_Potential_Average (68)
13. Cross_Pres_Order_CA	30. Electrostatic_Potential_Average	49. Electrostatic_Potential_at_LHA (69)
Contacts (III)	31. Electrostatic_Potential_at_LHA	Others (VII)
14. HBMM *	WNA ^(*) by Distance and at Surface (VI)	50. Accessible_in_Isolation
15. HBMWM *	32. HBMM * (52)	51. Hydrophobicity_KDI
16. HBMWWM	33. HBMWM * (53)	
17. HBMS *	34. HBMWWM * (54)	

HBMM: main chain to main chain hydrogen bond. HBMWM: main chain to (one H₂O) to main chain hydrogen bond. HBMWWM: main chain to (2 H₂O) to main chain hydrogen bond. HBMS: main chain to side chain hydrogen bond. HBMWS: main chain to (one H₂O) to side chain hydrogen bond. HBMWWS: main chain to (two H₂O) to side chain hydrogen bond. HBSS: means side chain to side chain hydrogen bond. HBSWS: side chain to (one H₂O) to side chain hydrogen bond. HBSWWS: side chain to (two H₂O) to side chain hydrogen bond. Ch_attractive means an attractive charge interaction, and Ch_repulsive means a repulsive charge interaction.

(*) The weighted neighbour average (WNA) is calculated by a weighting according to distance among interacting atoms and the accessibility at the surface (Equations 9 and 10, respectively, in the Supporting Information). Hence, the number of WNA descriptors should be counted twice (numbers in ascending order within brackets on the right side of that column). Descriptors whose order numbers are marked with a * are those that passed the radar plot test.

<https://doi.org/10.1371/journal.pone.0200018.t002>

Results

A constitutive set of parameters defining the PSSE nanoenvironment

It is intuitively clear that a set of parameters rather than just one descriptor (or a couple descriptors) is definitely more suitable for a full description of the nanoenvironment in the context defined above. Figs 6 and 7 demonstrate how 42 parameters (of the 69 selected for this work) differ (in normalized values) inside versus outside the PSSE. In this particular case, we analysed a set of 69 parameters for 28 α + β protein structures containing a single helix of 17 AAR per PSSE and calculated the average and standard deviation values for each parameter. Normalized values were calculated by dividing the average by the standard deviation. This process in fact is normalization by the inverse coefficient of variation, which permits easier reading of the difference between two data sets, which otherwise have very similar values [52]. To demonstrate the extent of the difference in the parameters inside and outside the PSSE, we calculated the difference between the normalized values. From a total of 69 parameters, 34 show a difference (between 0.1 and 1.0) between the average values inside versus outside the helix, and for 8 parameters, this difference is greater than 1, totalling 42 parameters (approximately 60.9% of the set of 69 parameters). Those parameters appear to constitute a statistically significant “signal” (two of examples are depicted at Figure D and Figure E in [S1 File](#) in Supporting

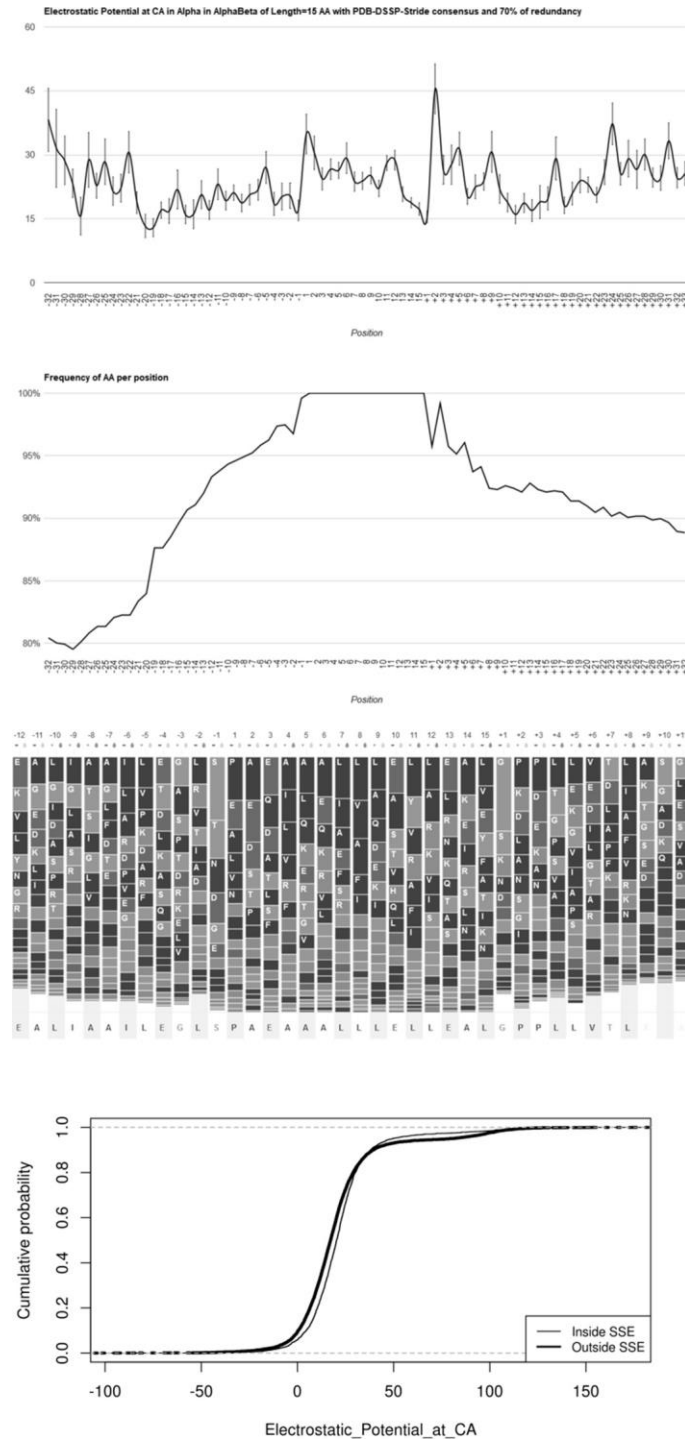


Fig 5. The Protein Secondary Structure Sting Analyzer (PS3A) panels contain four types of plots. In the case shown, 987 α -helices that are 15 amino acid residues long were examined from the *Datamart* in which we removed 70% of the redundancy at the whole protein sequence level, and all instances of α -helices were taken from both all- α and (α + β)+(α / β) proteins. The consensus definition used to determine the presence of an α -helical structure within proteins was the PDB-DSSP-Stride—the most rigorous one. The total number of such proteins is indicated in the Supporting Information in Figure B in [S1 File](#). Plots produced by the PS3A software: A) XY plot for average values (\pm SEM) for the selected descriptor: electrostatic potential at the α -carbon atom (CA). Negative numbers along the x-axes indicate locations to the left of the N-terminal of the examined/central PSSE, and positive ones follow its C-terminal end. B) The degree of occupancy per AAR position or “reliability”, which is the estimate of how accurately the signal

may be observed in A) above. This estimate is only based on how many amino acid residues are present at any location of the positional alignment of the PSSE. The maximum value (100% reliability) is assumed for the ensemble of studied samples along the PSSE. Outside the PSSE, the reliability is usually lower than 100%. C) The sequence logo presents which amino acid type is more frequently found at each positional alignment location—basically indicating the consensus sequence of the PSSE for a selected length (also shown at the bottom part of the logo). The amino acid position numbers (shown on the upper part of the plot) follow the same convention described for A) above. D) The ECDF curve shows how the descriptor average values inside the PSSE region are different from the corresponding values outside the selected PSSE. All of these plots (for each selected PSSE length, type of protein and redundancy level) may be accessed at <https://www.ps3a.cbi.cnptia.embrapa.br>.

<https://doi.org/10.1371/journal.pone.0200018.g005>

Information) for defining the α -helical nanoenvironment. The coverage for exclusive helices is rather broad, the same number of SSE signals containing 42 descriptors remains nearly the same for PSSE sizes of 11 to 40 (decreasing to 38 for helix size 6). On the other hand, for

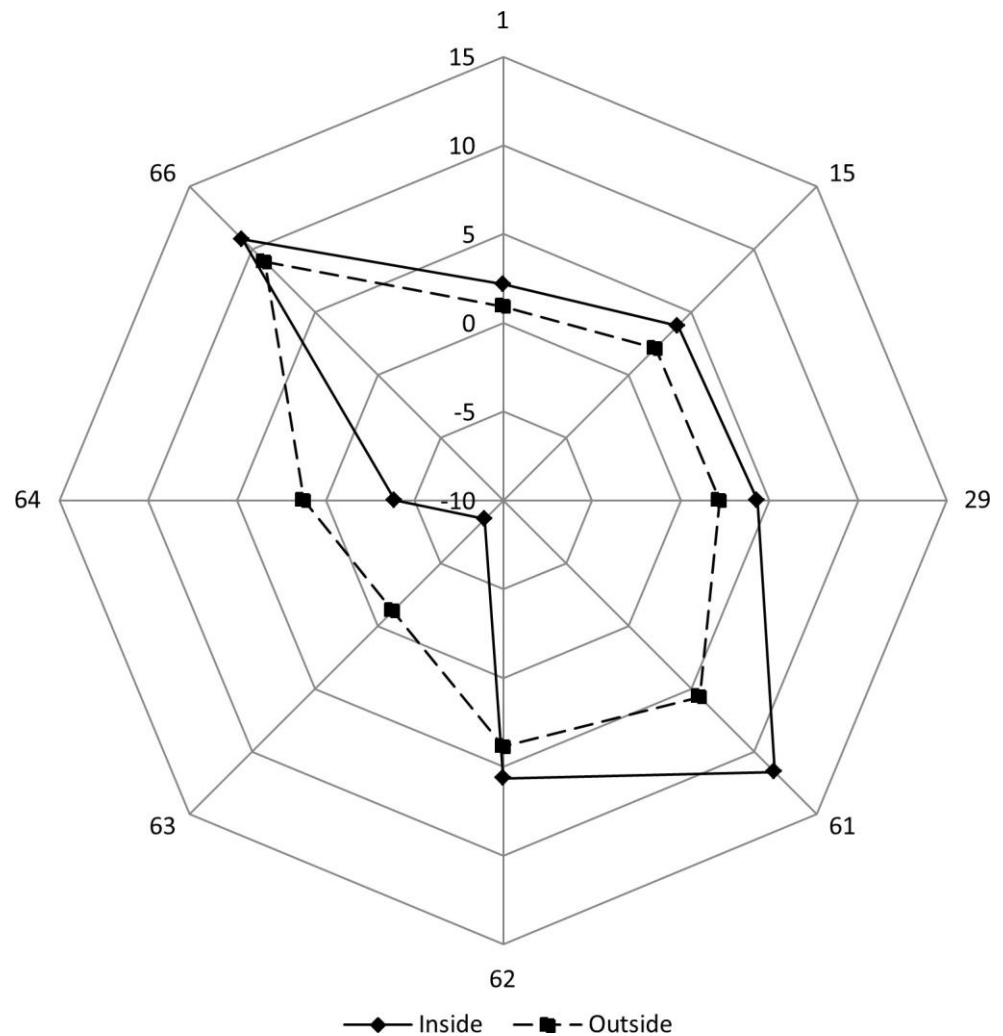


Fig 6. Comparison of the average values of 8 descriptors, normalized (by inverse coefficient of variation) done by dividing the parameter values with the corresponding standard deviation, and calculated for regions inside (17 AAARs) and outside the PSSE. The following descriptors are likely to show the postulated “signal” (the differences between the inside and outside descriptor values per position are higher than 1): 1. Hbmm, 15. Hbmm_WNADist, 29. Hbmm_WNASurf, 61. Number_Used_Contact_WNADist, 62. Number_Used_Contact_WNASurf, 63. Dihedral_Angle_PHI, 64. Dihedral_Angle_PSI, 66. Density. The two shadowed descriptors are expected to show differences, as these descriptors are basically part of the definition of the investigated PSSE.

<https://doi.org/10.1371/journal.pone.0200018.g006>

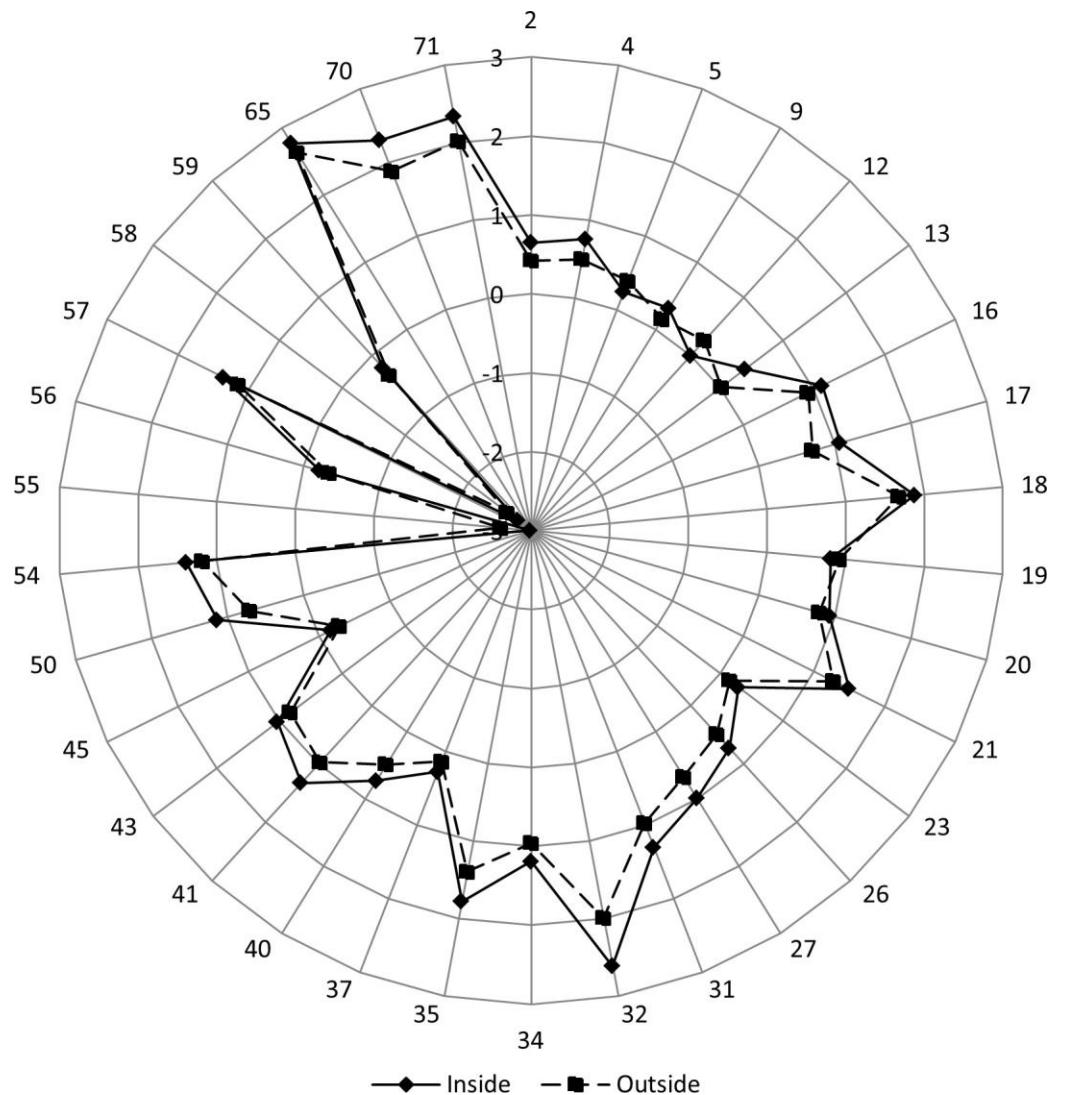


Fig 7. Comparison of the average values of 34 descriptors, normalized (by inverse coefficient of variation) done by dividing the parameter values with corresponding standard deviation, and calculated for regions inside (17 AAARs) and outside the PSSE. The following descriptors are likely to show the postulated “signal” (the differences between the inside and outside descriptor values per position are higher than 0.1 and lower than 1): 2. Hbmwm, 4. Hbms, 5. Hbmws, 9. Hbssws, 12. Disulfide, 13. Ch_attractive, 16.hbmwm_WNADist, 17. Hbmwm_WNADist, 18. Hbms_WNADist, 19. Hbmws_WNADist, 20. Hbmws_WNADist, 21. Hbss_WNADist, 23. Hbssws_WNADist, 26. Disulfide_WNADist, 27. ch_attractive_WNADist, 31. Hbmwm_WNASurf, 32. Hbms_WNASurf, 34. Hbmwm_WNASurf, 35. Hbss_WNASurf, 37. Hbssws_WNASurf, 40. Disulfide_WNASurf, 41. Ch_attractive_WNASurf, 43. Cross_Link_Order_CA, 45. Dihedral_Chi1, 50. Electrostatic_Potential_at_CA, 54. Electrostatic_Potential_at_CA_WNADist, 55. Electrostatic_Potential_at_LHA_WNADist, 56. Electrostatic_Potential_Average_WNADist, 57. Electrostatic_Potential_at_CA_WNASurf, 58. Electrostatic_Potential_at_LHA_WNASurf, 59. Electrostatic_Potential_Average_WNASurf, 65. Temperature_Factor_CA, 70. SC_Clash, 71. SC_Percent.

<https://doi.org/10.1371/journal.pone.0200018.g007>

nonexclusive helices, the coverage decreases from 54 descriptors for helix size 6, to 42 descriptors for helix size 40.

Multiple variable analyses

In the multivariate analysis of variance (MANOVA), it was necessary to first remove those descriptors that did not obey a normal distribution. In addition, linearly correlated descriptors

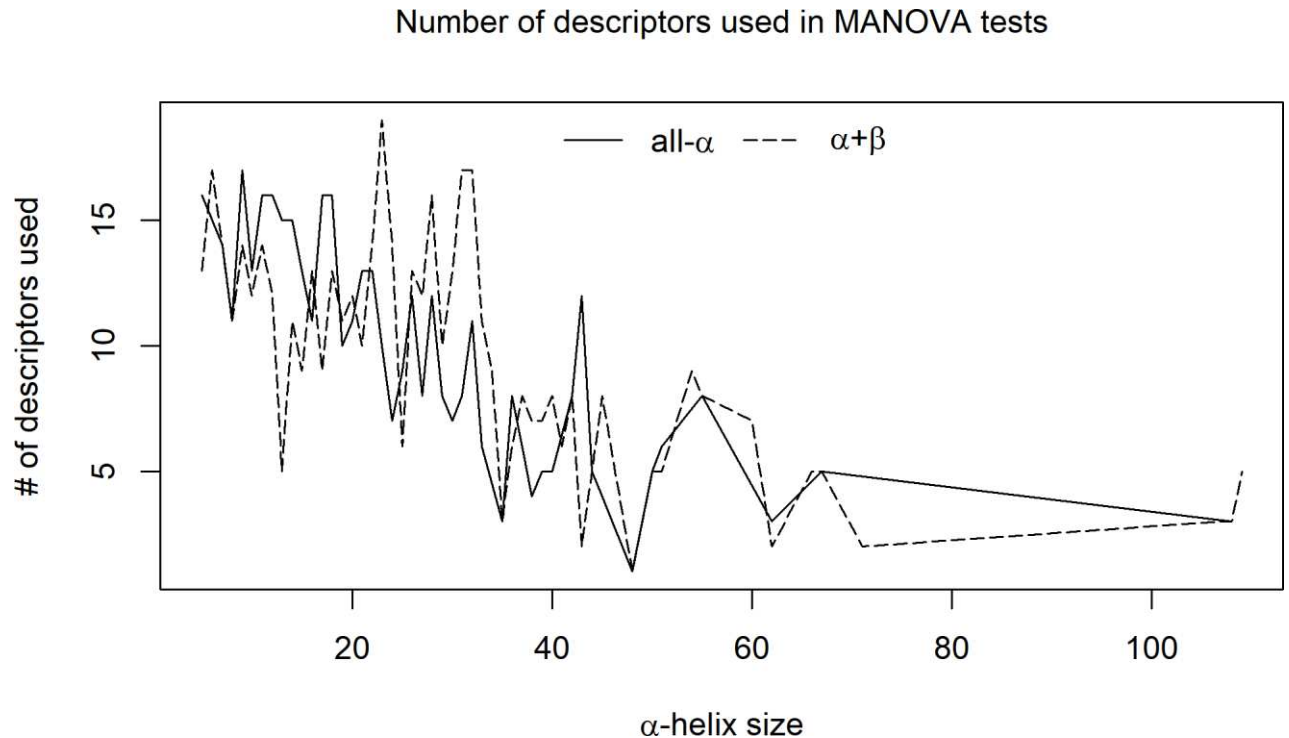


Fig 8. Variation in the number of descriptors that passed both the normal distribution test and the no mutual correlation test for different helix sizes.

<https://doi.org/10.1371/journal.pone.0200018.g008>

(defined after using an R script and default threshold/cutoff of 0.9) were also eliminated. As a result, in almost all of the tests, the number of descriptors actually used in the MANOVA was lower than the number of descriptors provided to the test input. As described in Fig 8, as the size of the SSE increases, the number of descriptors that passed both the normal distribution test and the no mutual correlation test decreases. Consequently, the MANOVA test had less descriptors: approximately 10 to 15 for helix sizes up to 25 AARs and then decreasing to approximately 5 descriptors for sizes of 40 AARs and above. The lowest number of descriptors used in this test was 3 (for a helix size of 108 AARs). The average number of descriptors for all helix sizes was 10. For all helix sizes, the four applied MANOVA tests showed p-values lower than 1×10^{-6} in at least 83% of the cases and lower than 1×10^{-3} in at least 95% of the cases. Such results clearly support our initial hypothesis that a selected number of qualified descriptors may fully identify and describe the internal nanoenvironment of the protein region containing the helical structure.

The ten most frequent descriptors used by MANOVA (after the descriptors passed the double test on the input) were as follows (see also Fig 9): 1. Electrostatic_Potential_Average_WNASurf (30, 65%), 2. Number_Unused_Contact_WNADist (30, 65%), 3. Hbms_WNASurf (26, 57%), 4. Hbmm_WNASurf (25, 54%), 5. Number_Unused_Contact_WNASurf (25, 54%), 6. Hbmm_WNADist (24, 52%), 7. Electrostatic_Potential_at_CA_WNADist (23, 50%), 8. Electrostatic_Potential_Average_WNADist (22, 48%), 9. Dihedral_Chi1 (20, 43%), and 10. Electrostatic_Potential_at_CA_WNASurf (20, 43%). The numbers within parentheses represent the total number of appearances among the descriptors used and the percentage of the total number of helix sizes where that particular descriptor was used by MANOVA, respectively.

The situation in $\alpha+\beta$ protein structures is similar; however, the 10 most frequently found descriptors are as follows: 1. Number_Unused_Contact_WNADist (37, 69%), 2. Electrostatic_Potential_at_LHA_WNADist (30, 56%), 3. Electrostatic_Potential_Average_WNASurf (29,

Descriptor distribution vs SSE size in MANOVA tests for all- α proteins

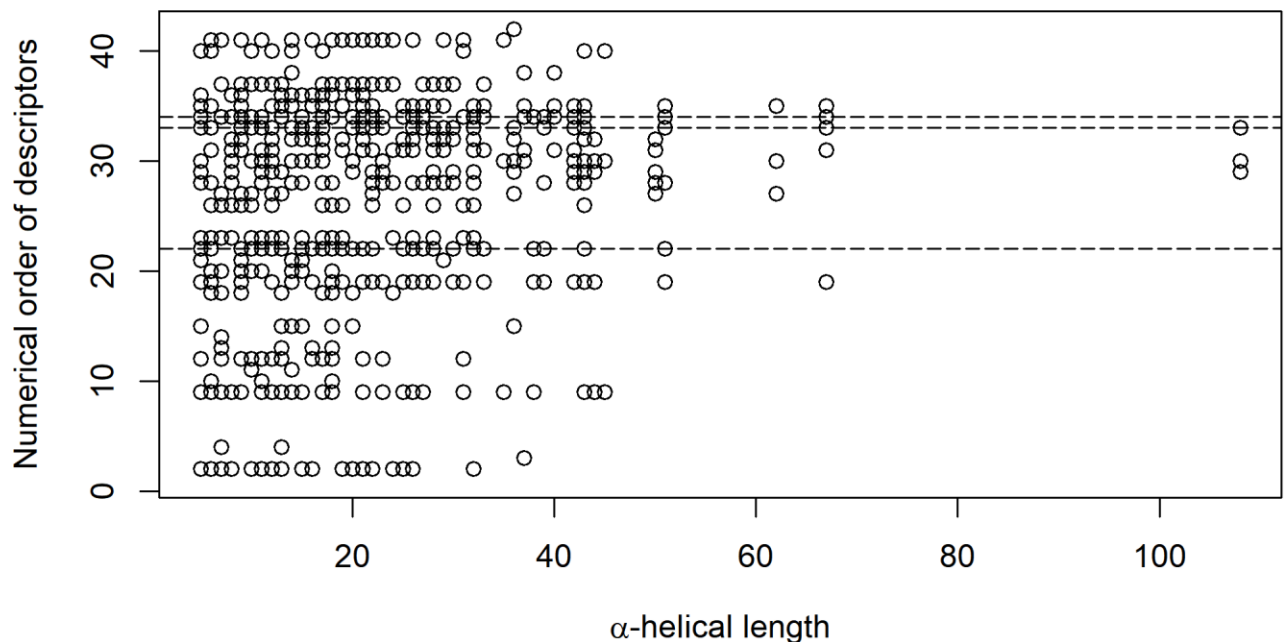


Fig 9. Representation of 42 different descriptors used for the MANOVA input and then filtered by a double test: A normal distribution of data and a lack of mutual correlation. The points, which are plotted for each helical size (x-axes), represent those descriptors used by MANOVA for that particular size. The 42 descriptors found on the y-axes are as follows: 1. Hbmm, 2. Hbmwm, 3. Hbms, 4. Hbmws, 5. Hbswws, 6. Disulfide, 7. Ch_attractive, 8. Hbmm_WNADist, 9. hbmwm_WNADist, 10. Hbmwwm_WNADist, 11. Hbms_WNADist, 12. Hbmws_WNADist, 13. Hbmwws_WNADist, 14. Hbss_WNADist, 15. Hbswws_WNADist, 16. Disulfide_WNADist, 17. ch_attractive_WNADist, 18. Hbmm_WNASurf, 19. Hbmwm_WNASurf, 20. Hbmwwm_WNASurf, 21. Hbms_WNASurf, 22. Hbss_WNASurf, 23. Hbswws_WNASurf, 24. Disulfide_WNASurf, 25. Ch_attractive_WNASurf, 26. Electrostatic_Potential_at_CA, 27. Electrostatic_Potential_at_CA_WNADist, 28. Electrostatic_Potential_at_LHA_WNADist, 29. Electrostatic_Potential_Average_WNADist, 30. Electrostatic_Potential_at_CA_WNASurf, 31. Electrostatic_Potential_at_LHA_WNASurf, 32. Electrostatic_Potential_Average_WNASurf, 33. Number_Used_Contact_WNADist, 34. Number_Used_Contact_WNASurf, 35. Cross_Link_Order_CA, 36. Dihedral_Chi1, 37. Dihedral_Angle_PHI, 38. Dihedral_Angle_PSI, 39. Temperature_Factor_CA, 40. Internal_CA_3, 41. Clash, 42. Percent. Finally, the three most frequently plotted descriptors are as follows (designated by the three horizontal dashed lines, from top to bottom): 1. Electrostatic_Potential_Average_WNASurf (order number: 32) $>$ (30, 65%), 2. Number_Used_Contact_WNADist (order number: 33) $>$ (30, 67%) and 3. Hbms_WNASurf (order number: 21) $>$ (26, 58%).

<https://doi.org/10.1371/journal.pone.0200018.g009>

54%), 4. Number_Used_Contact_WNASurf (29, 54%), 5. Electrostatic_Potential_at_LHA_WNASurf (28, 52%), 6. Electrostatic_Potential_at_CA_WNADist (25, 46%), 7. Electrostatic_Potential_at_CA_WNASurf (24, 44%), 8. Hbms_WNASurf (21, 39%), 9. Cross_Link_Order_CA (20, 37%), 10. Dihedral_Chi1 (20, 37%). The lowest number of descriptors used in this test was 2 (for a helix size of 109 AARs). The average number of descriptors for all helix sizes in $\alpha+\beta$ proteins was 10 again.

For $\alpha+\beta$ proteins and all helix sizes, the four applied MANOVA tests showed p-values lower than 1×10^{-6} in at least 85% of the cases and lower than 1×10^{-3} in at least 92% of the cases. Once again, the null hypothesis was easily ruled out.

We conducted one more test in order to estimate the power of the most frequently used descriptors (listed above) to distinguish the nanoenvironment of an α -helix from the non-helical environments. In the case of all- α proteins, the 10 most frequently used descriptors would give us approximately 70% coverage and p-values lower than 10^{-3} in more than 90% of the cases. The first four most used descriptors gave 63% coverage and p-values less than 10^{-3} in

more than 83% of the cases. For nonexclusive α -helices, using the 10 corresponding most frequently used descriptors yielded an 84% coverage and p-value less than 10^{-3} in more than 96% of the cases. The first 5 of the aforementioned descriptors gave 55% coverage and p-values less than 10^{-3} in more than 73% of the cases.

Moving/sliding window of the average values for a selected descriptor

[Fig 10](#) is a compilation of plots, and each set (A-B) has three different graphs. The first graph shows the variation in the average value of the descriptor: in this case—the number of contacts of HBMM type(53), weighted by the distance measured over the surface to which the AAR belongs. Sequence redundancy was removed at the 70% similarity level, revealing 24 helices in the all- α structures (on the left side of the figure). The same level of redundancy was set for the α -helices in the $(\alpha+\beta) + (\alpha/\beta)$ structures, revealing 203 helices, which also had 12 AAR (on the right side of the figure). The average descriptor value is shown for each location of the positional alignment of α -helix size. Along the horizontal axis, “negative” positions (“-”) correspond to the amino acid residues located before the PSSE N-terminal end, while “positive” positions (“+”) correspond to the amino acid residues after the PSSE C-terminal end. The second graph shows the degree of occupancy per position of the encountered amino acid residues (of any type) that belong to the α -helical structure at any particular position presented in that plot (where that number is 100 along the PSSE location itself and lower is lower to the left and right of the PSSE). Therefore, the reliability is 100% along the examined PSSE. In addition, some residues might be missing (to the left and the right of the analysed PSSE) as the stretch might not reach the desired limits of the N-terminal -32 to C-terminal +32 residues. The temporary decrease immediately before and immediately after the examined PSSE, followed by an increase in reliability, is because the PSSE limits must be delimited by AARs, which definitely may not belong to a helix (that being the definition of PSSE limit). Further from the N- and C-termini, the presence of other α -helices is possible, and the reliability therefore oscillates until positions -32 and +32. The reduction in reliability at the very N-terminal and C-terminal ends of the extended PSSE region results from the surge in gaps, which become ever more frequent and which we introduce in such cases, usually at the positions discussed here where the original sequence does not reach the desired limit position (± 32). The third plot represents the data resulting from the sliding window test. Namely, using an R script, the descriptor values were grouped in the window with the same length of the selected PSSE and that window was run (slide) from the left to the right along the positional alignment of the α -helical element and its extended region. Additionally, other sliding windows of varying length, from one AAR to double the PSSE size, were also made to slide along the same path. This experiment proved that it is possible to identify the PSSE region by a sliding window test, preferably using a window with the minimum possible length, as proven by the results depicted in Figure H in [S1 File](#). Finally, Student’s t-test was applied to data that were divided into two sets: data for AARs from within and from outside the sliding window range. As shown in [Fig 10](#), the p-value from Student’s t-test approaches zero exactly within the region where the sliding window matches the PSSE position. That observation corroborates the hypothesis we postulated at the beginning of this work. In general, if, as in the third plot of [Fig 10A and 10B](#), the p-value demonstrates such a clear variation, then the ensemble inside the sliding window at a selected position is significantly different from the ensemble outside that window region. Although some regions other than the central PSSE regions appear to have a p-value approaching zero, these regions are mostly outside the PSSE where the p-value exhibits a different/specific behaviour compared to the behaviour observed for the PSSE flanking regions. Even though the PDB does not contain any single helix protein structure with a long enough sequence to satisfy the conditions of our

Descriptor = HBMM_WNASurf; PSSE length = 12 AARs; Consensus = PDB-DSSP-Stride; Redundancy = 70%

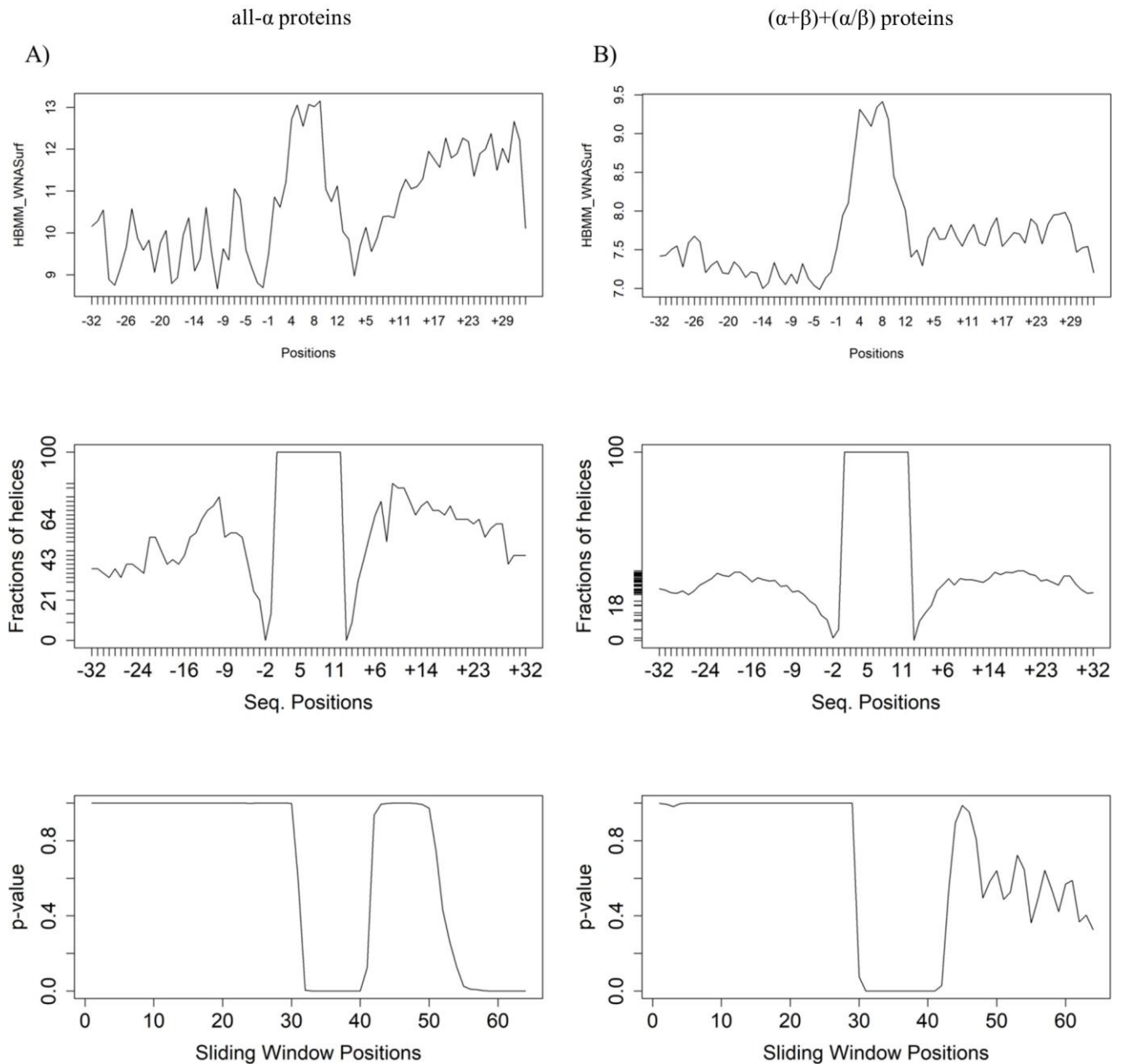


Fig 10. Composite graphs showing the following: Descriptor variation along the regions before, at and after the analysed PSSE; the reliability value (or % of helical structure at each loci) and the p-value for the descriptor: Number of contacts, type “HBMM”. Data are drawn from the *Datamart* containing PSSEs of length = 12 AARs; the consensus definition of a helix element is from “PDB-DSSP-Stride”, and the redundancy is 70% similarity at the sequence level.

<https://doi.org/10.1371/journal.pone.0200018.g010>

experiment, which would create the situation where we would have flanking regions with no perturbation from other helical structures or turn structures, it is quite clear from our plots that such signals / perturbations do exist in the experimental setup. This possibility clearly might explain partially why we have p-values fluctuating in the flanking regions.

After being convinced of the validity of the initial hypothesis of this work (valid for multiple descriptors in multivariate analysis), we needed to closely examine some specific SSE sizes and selected parameters/descriptors, aiming to identify the upper limit for coverage in the case of using only a single descriptor.

The Kolmogorov-Smirnov test was applied in order to statistically confirm the previous visual observations for some of the data analysed here regarding the “existence” of a nanoenvironment-specific single descriptor “signal”. A total of 46 different helical lengths (sizes) (DM1) were subjected to statistical examination using one out of available 69 descriptors (see Table 2) for a total of 3165 available tests (9 tests were excluded because they showed an input data problem). A total of 125 tests have a p-value equal to or less than 1×10^{-6} (3.9%) exactly over the region fully matching the extent/position of the examined α -helix. Additionally, 426 tests show p-values along the encountered α -helices equal to or less than 1×10^{-3} (13.5%). These results indicate that the analysed tests show p-values compatible with the conclusion that the helical region (in terms of nanoenvironment) is significantly different from the regions outside the helix. Such relatively low performance of the success indicator becomes more favourable for the postulated existence of a “signal” for α -helices found in $(\alpha+\beta)+(\alpha/\beta)$ proteins: a total of 54 different helical sizes (DM2) were subjected to statistical examination using the same 69 descriptors, for a total of 3704 available tests (22 tests were excluded because they showed an input data problem). For this situation, a total of 766 tests have a p-value equal to or less than 1×10^{-3} (20.7%) and exactly over the region fully matching the extent/position of the examined α -helix, and 298 tests have a p-value smaller than 1×10^{-6} (8%). A more detailed analysis (Table 3) identified very clear-cut situations in which the statistics indicate the existence of a nanoenvironment “signal” in approximately one fifth of the abovementioned cases. In addition, as explained in the “Hypothesis verification section”, when the p-values are close to 0, as in the two mentioned sample marts, then we may reject the H_0 hypothesis (which claims that the two datasets—the data from inside the PSSE compared to outside the PSSE—are statistically not distinguishable from each other). The remaining approximately 80% of the studied cases, do not have such low p-values within the region of the analysed PSSE. However, this finding was expected as there is no such unique parameter that has the power to singlehandedly fully describe the $nEo\alpha H$.

In conclusion, most of the statistical analysis points towards the possibility of clearly distinguishing a helical nanoenvironment from the environment of the region outside the helix. An apparent upper limit is clearly seen for the usage of a single descriptor. On the other hand, the number of descriptors in multivariate analysis comfortably points to a coverage level greater than 90%.

Single descriptor “signature” comparison: The case of two different types of PSSE

To demonstrate how representative the analysis we presented so far is, it is also necessary to verify the behaviour for a single descriptor value for an entirely different PSSE—the “ β -strand”. If our hypothesis holds, then the same principles should hold as well, while the signal/signature

Table 3. The KS test applied for sliding windows in all size helices using single parameter analysis.

Datamarts	p-value $\leq 1 \times 10^{-6}$	$1 \times 10^{-6} > \text{p-value} \leq 1 \times 10^{-4}$	$1 \times 10^{-4} > \text{p-value} \leq 1 \times 10^{-2}$	p-value $> 1 \times 10^{-2}$
all- α exclusive	1,1%	2,4%	7,6%	88,9%
all- α nonexclusive	3,9%	4,5%	13,1%	78,5%
$\alpha+\beta$ exclusive	2,9%	3,9%	10,5%	82,7%
$\alpha+\beta$ nonexclusive	8,1%	6,8%	14,3%	70,8%

<https://doi.org/10.1371/journal.pone.0200018.t003>

might be different in shape, content (type of descriptor), and intensity. A comparison between the corresponding datasets for α -helical regions against β -strand regions in $(\alpha+\beta)+(\alpha/\beta)$ proteins demonstrated that there is a convincing difference in the PSSE “signal” types describing their respective PSSE regions (the compilation for complete β -strand analysis is still under construction and will be published separately). Fig 11 presents the data pointing to the difference between the nanoenvironments for α -helices and β -strand regions. The descriptors and datasets compared were as follows: A) (on the left side of this figure panel) EP@C α , and B) (on the right side of this figure panel) the number of contacts of HBMM_WNASurf type, in 1657 α -helices in $(\alpha+\beta)+(\alpha/\beta)$ proteins against 20790 β -strands in $(\alpha+\beta)+(\alpha/\beta)$ proteins chains. The comparison between the EP@C α moving average values for α -helices in $(\alpha+\beta)+(\alpha/\beta)$ proteins and β -strands in $(\alpha+\beta)+(\alpha/\beta)$ proteins shows that while the PSSE “signal” resembles the letter “N” in the case of α -helices, at the same time, the “signal” for β -strands resembles the letter “U” more. The HBMM_WNASurf average values for α -helices in $(\alpha+\beta)+(\alpha/\beta)$ proteins and β -strands in $(\alpha+\beta)+(\alpha/\beta)$ proteins are higher for α -helices than for β -strands.

Knowledge of the PSSE nanoenvironment applied to quality assessment of protein structures: The case of predicted vs experimentally obtained 3D structures

To demonstrate one example of a practical use of this knowledge and the potential value of analysing the collected data in everyday situations for computational structural biologists, in

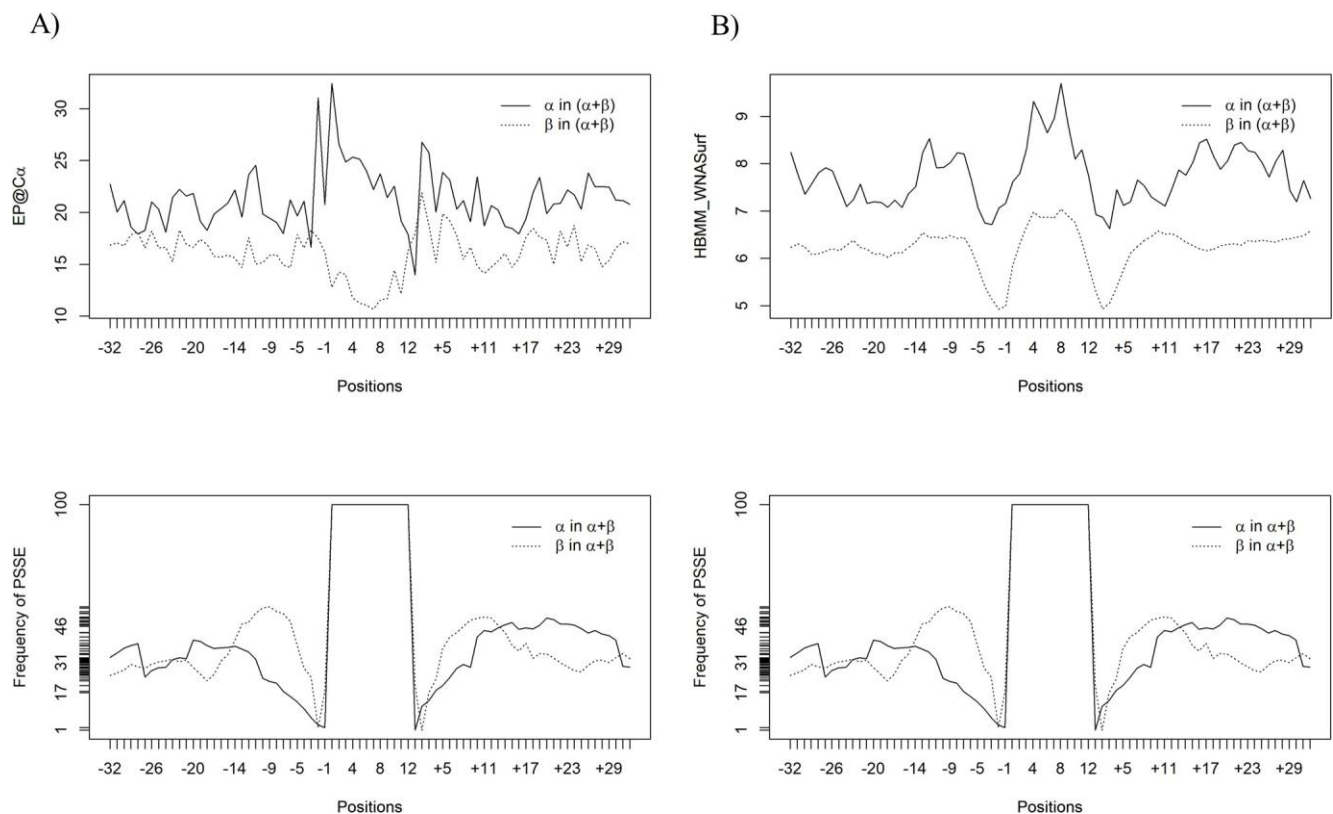


Fig 11. Differences in the variation behaviour of two selected descriptors around α -helices (solid lines) and β -strands (dotted lines). The plots above present the behaviour of the A) EP@C α average values for 1811 α -helices in $(\alpha+\beta)+(\alpha/\beta)$ proteins and 7773 β -strands in $(\alpha+\beta)+(\alpha/\beta)$ proteins. B) HBMM_WNASurf average values for α -helices in $(\alpha+\beta)+(\alpha/\beta)$ proteins and β -strands in $(\alpha+\beta)+(\alpha/\beta)$ proteins. The average number of this contact type is higher in and around α -helices than in and around β -strands. As shown, there are clear differences in signal pattern in the cases presented in A and B.

<https://doi.org/10.1371/journal.pone.0200018.g011>

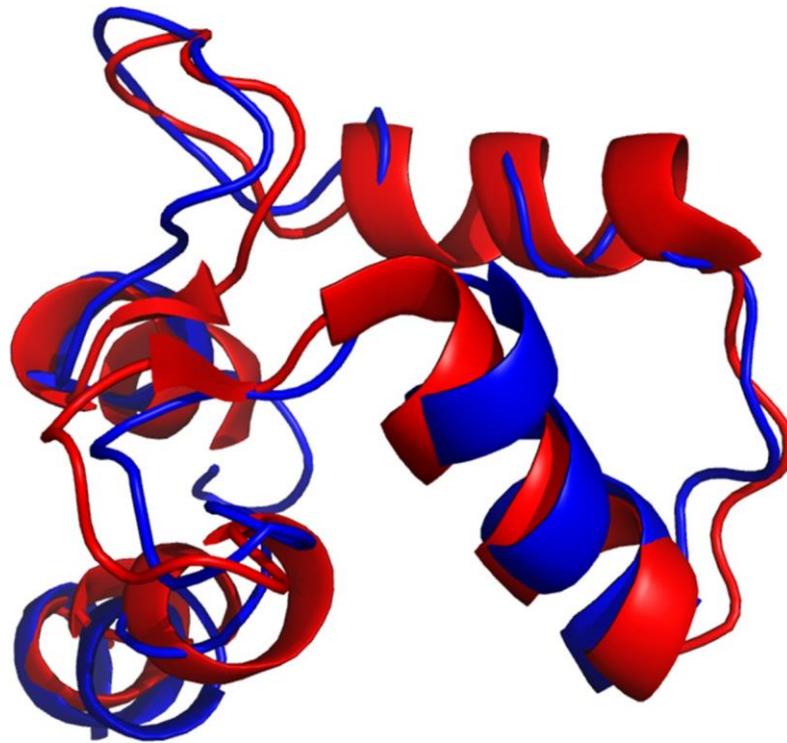


Fig 13. The superposition of two identical proteins whose structures were solved at two very different resolutions. The 1fw4.pdb (red ribbon) structure has a 1.7 Å resolution and 1trc.pdb (blue ribbon), 3.6 Å. Both structures have the very same amino acid sequence, but 1trc.pdb is an older structure, and its low resolution (3.6 Å) causes some errors in the α -helix definition and positioning. The region between 117 and 127 AAR, at the top right, demonstrate that in both cases there is a helical element there but the lower resolution structure does not have a corresponding assignment for it.

<https://doi.org/10.1371/journal.pone.0200018.g013>

particular characteristics of the nanoenvironment of each type of secondary structure element. The concept of a nanoenvironment was already explored earlier with successful results reported for protein-protein interfaces and enzyme catalytic sites. This work employed the same idea and applied the concept to the case of the PSSE nanoenvironment.

The individual plots in Fig 11 demonstrate a different behaviour inside the PSSE for the “signal” of α -helices, which is not the same as the “signal” of β -strands. However, considering that the nanoenvironment is not fully defined by a unique/single descriptor, a different approach was necessary to confirm the hypothesis of the existence of a “signal” inside a specific PSSE. The application of multivariate analysis of variance (MANOVA) to the same dataset confirmed the existence of a “signal” for α -helices. Based on these tests, we conclude that a set of specific parameters, such as contacts, physical-chemical, geometrical and structural descriptors, describes a nanoenvironment, in this case, the nanoenvironment of α -helices. Three descriptor categories were found to be among the most frequently used for nEo α H identification: Number_Unused_Contacts, Electrostatic_Potential and number of contacts of Hbms type, meaning that the potential for forming contacts, the number of hydrogen bonds (of main chain to side chain type) and the electrostatic potential of the involved AARs are crucial descriptors of the nEo α H.

The next step to continue our work is to confirm the hypothesis for β -strands, turns, and coils. Preliminary tests indicated that the hypothesis would be confirmed in the same way as our confirmation for α -helices.

Supporting information

S1 File.
(DOCX)

Author Contributions

Conceptualization: José Augusto Salim, Goran Neshich.

Data curation: Ivan Mazoni, Luiz César Borro, Goran Neshich.

Formal analysis: Ivan Mazoni, Luiz César Borro, José Augusto Salim, Goran Neshich.

Funding acquisition: José Gilberto Jardine, Goran Neshich.

Investigation: Ivan Mazoni, Luiz César Borro.

Methodology: Ivan Mazoni, Luiz César Borro, Goran Neshich.

Project administration: José Gilberto Jardine, Goran Neshich.

Resources: José Gilberto Jardine, Inácio Henrique Yano, Goran Neshich.

Software: Ivan Mazoni.

Supervision: José Gilberto Jardine, Goran Neshich.

Validation: Ivan Mazoni, Luiz César Borro, José Augusto Salim, Goran Neshich.

Visualization: Ivan Mazoni, José Augusto Salim, Goran Neshich.

Writing – original draft: Ivan Mazoni, Goran Neshich.

Writing – review & editing: Goran Neshich.

References

1. Crick FHC. On protein synthesis. 1958; p. 138–63.
2. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973; p. 223–230. PMID: [4124164](https://pubmed.ncbi.nlm.nih.gov/4124164/)
3. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*. 2000; p. 291–325. <https://doi.org/10.1146/annurev.biophys.29.1.291> PMID: [10940251](https://pubmed.ncbi.nlm.nih.gov/10940251/)
4. Benjamin W, Sali A. Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*. 2014; p. 5–6.
5. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 2006; p. 195–201.
6. Kim DE, Dylan C, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic acids research*. 2004; p. W526–W531. <https://doi.org/10.1093/nar/gkh468> PMID: [15215442](https://pubmed.ncbi.nlm.nih.gov/15215442/)
7. Elmar K, Koraimann G, Vriend G. Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. *Proteins: Structure, Function, and Bioinformatics*. 2002; p. 393–402.
8. Lee J, Wu S, Zhang Y. *From Protein Structure to Function with Bioinformatics*: Springer; 2009.
9. Thévenet P, Shen Y, Maupetit J, Guyon F, Derreumaux P, Tufféry P. PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic acids research*. 2012; p. W288–W293. <https://doi.org/10.1093/nar/gks419> PMID: [22581768](https://pubmed.ncbi.nlm.nih.gov/22581768/)
10. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge. *Structure, Function, and Bioinformatics*. 2012; p. 1715–1735.
11. Rohl C, Strauss C, Misura K, Baker D. Protein structure prediction using Rosetta. *Methods in enzymology*. 2004; p. 66–93. [https://doi.org/10.1016/S0076-6879\(04\)83004-0](https://doi.org/10.1016/S0076-6879(04)83004-0) PMID: [15063647](https://pubmed.ncbi.nlm.nih.gov/15063647/)

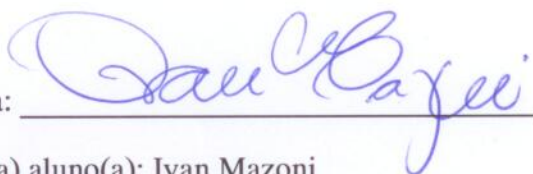
12. Yang Z, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical journal*. 2003; p. 1145–1164. [https://doi.org/10.1016/S0006-3495\(03\)74551-2](https://doi.org/10.1016/S0006-3495(03)74551-2) PMID: [12885659](#)
13. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature*. 1992; p. 86–89.
14. A Kelley L, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*. 2015; p. 845–858. <https://doi.org/10.1038/nprot.2015.053> PMID: [25950237](#)
15. Söding J, Biegert A, Lupas A. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*. 2005; p. W244–W248. <https://doi.org/10.1093/nar/gki408> PMID: [15980461](#)
16. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. *Journal of bioinformatics and computational biology*. 2003; p. 95–117. PMID: [15290783](#)
17. Jian P, Xu J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*. 2011; p. 161–171.
18. LouisiJeune C, Andrade C. Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins: Structure, Function, and Bioinformatics*. 2012; p. 374–381.
19. Christoph W, Bellstedt P, Görlach M. CAPITO-A web server based analysis and plotting tool for circular dichroism data. *Bioinformatics*. 2013.
20. Lin K, Yang H, Gao Z, Li F, Yu S. RETRACTED ARTICLE: Overestimated accuracy of circular dichroism in determining protein secondary structure. *European Biophysics Journal*. 2013; p. 455–461. <https://doi.org/10.1007/s00249-013-0896-y> PMID: [23467783](#)
21. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of computational chemistry*. 2012; p. 259–267.
22. Shen Y, Bax A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of biomolecular NMR* 56.3. 2013; p. 227–241. <https://doi.org/10.1007/s10858-013-9741-y> PMID: [23728592](#)
23. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports* 5. 2015.
24. Seeley M, Clement M, Snell Q. Feature identification and reduction for improved generalization accuracy in secondary-structure prediction. 13th IEEE International Conference on Bioinformatics and Bio-Engineering. 2013.
25. Spencer M, Eickholt J, Cheng J. A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 12.1. 2015; p. 103–112.
26. Magnan CN, Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* 30.18. 2014; p. 2592–2597. <https://doi.org/10.1093/bioinformatics/btu352> PMID: [24860169](#)
27. Kabsch W, Sander C. How good are predictions of protein secondary structure? *FEBS letters*. 1983; p. 179–182.
28. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. *Nucleic acids research*. 2015; p. gkv332.
29. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*. 1963; p. 95–9.
30. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; p. 2577–637. <https://doi.org/10.1002/bip.360221211> PMID: [6667333](#)
31. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins*. 1995; p. 566–79. <https://doi.org/10.1002/prot.340230412> PMID: [8749853](#)
32. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*. 1999; p. 508–19. PMID: [10081963](#)
33. Neshich G, Togawa R, Rocchia W, Mancini AL, Kuser PR, Yamagishi MEB, et al. STING MILLENNIUM SUITE v.3 and JAVA PROTEIN DOSSIER: a novel concept in data visualization and analysis of the protein structure/function relationship. In ; 2003; Brisbane, Australia.
34. Murzin AG, et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* 247.4. 1995; p. 536–540. <https://doi.org/10.1006/jmbi.1995.0159> PMID: [7723011](#)

35. Efimov A. Structural similarity between two-layer α/β and β -proteins. *Journal of molecular biology*. 1995; p. 402–415.
36. Sugase K, Dyson HJ, Wright PE. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*. 2007; p. 1021–1025. <https://doi.org/10.1038/nature05858> PMID: [17522630](#)
37. Alexey G, Steven M, Brenner E, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*. 1995; p. 536–540.
38. PDB. Protein Data Bank. ; 1971.
39. Neshich G, Neshich IAP, Moraes F, Salim JA, Borro L, Yano IH, et al. Using Structural and Physical–Chemical Parameters to Identify, Classify, and Predict Functional Districts in Proteins—The Role of Electrostatic Potential Walter Rocchia MS, editor.: Springer International Publishing; 2014.
40. Moraes FR, Neshich IAP, Mazoni I, Yano IH, Pereira JGC, Salim JA, et al. Improving predictions of protein-protein interfaces by combining amino acid-specific classifiers based on structural and physico-chemical descriptors with their weighted neighbor averages. *Plos One*. 2014 January 28; p. 87–107.
41. Benjamin W, Sali A. Protein structure modeling with MODELLER. *Protein Structure Prediction*. 2014; p. 1–15.
42. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research*. 2014; p. gku340.
43. Wadood A, Ahmed N, Shah L, Ahmad A, Hassan H, Shams S. In-silico drug design: An approach which revolutionarised the drug discovery process. *OA Drug Design & Delivery*. 2013 Apr.
44. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics*. 2016; p. 20–33. <https://doi.org/10.1016/j.jhealeco.2016.01.012> PMID: [26928437](#)
45. Oliveira SRM, Almeida GV, Souza KRR, Rodrigues DN, Falcão PRK, Yamagishi MEB, et al. STING_RDB: A relational database of structural parameters for protein analysis with support for Data Warehousing and Data Mining. *Genetic Molecular Research*. 2007; p. 911–922.
46. Neshich G, Togawa RC, Mancini AL, Kuser PR, Yamagishi MEB, Pappas G, et al. STING Millennium: A web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic acids research* 31.13. 2003; p. 3386–3392. PMID: [12824333](#)
47. Chakravarti, Laha, Roy. *Handbook of Methods of Applied Statistics*: John Wiley and Sons; 1967.
48. Wilk M, Gnanesikan R. Probability plotting methods for the analysis for the analysis of data. *Biometrika* 55.1. 1968; p. 1–17. PMID: [5661047](#)
49. Haynes W. Student's t-Test. In *Encyclopedia of Systems Biology*. New York: Springer; 2013. p. 2023–2025.
50. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006.
51. Neshich G, Mancini AL, Yamagishi MEB, Kuser PR, Fileto R, Pinto IP, et al. STING Report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the STING database. *Nucleic Acids Research*. 2005; p. D269–D274. <https://doi.org/10.1093/nar/gki111> PMID: [15608194](#)
52. Sharma K, Krishna H. Asymptotic sampling distribution of inverse coefficient-of-variation and its applications. *IEEE Transactions on Reliability*. 1994; p. 630–633.

ANEXO 2 - TERMO DE BIOÉTICA/BIOSSEGURANÇA**DECLARAÇÃO**

Em observância ao §5º do Artigo 1º da Informação CCPG-UNICAMP/001/15, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada ***“ANÁLISE DO NANO-AMBIENTE PROPÍCIO PARA NUCLEAÇÃO E MANUTENÇÃO DOS ELEMENTOS DA ESTRUTURA SECUNDÁRIA NO CONTEXTO ESTRUTURAL DAS PROTEÍNAS FUNCIONAIS”***, desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: _____



Nome do(a) aluno(a): Ivan Mazoni

Assinatura: _____



Nome do(a) orientador(a): Goran Neshich

Data: 10 de dezembro de 2018

ANEXO 3 - DECLARAÇÃO DE DIREITOS AUTORAIS

DECLARAÇÃO

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **ANÁLISE DO NANO-AMBIENTE PROPÍCIO PARA NUCLEAÇÃO E MANUTENÇÃO DOS ELEMENTOS DA ESTRUTURA SECUNDÁRIA NO CONTEXTO ESTRUTURAL DAS PROTEÍNAS FUNCIONAIS**, não infringem os dispositivos da Lei n.º 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 10 de dezembro de 2018

Assinatura : _____

Nome do(a) autor(a): **Ivan Mazoni**

RG n.º 282292755

Assinatura : _____

Nome do(a) orientador(a): **Goran Neshich**

RG n.º 1535326 SSP/DF