**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS**

Antonio Marinho da Silva Neto

**Geometria diferencial e teoria da informação aplicada a
análise de ensembles conformacionais de proteínas**

São Carlos

2017

Antonio Marinho da Silva Neto

**Geometria diferencial e teoria da informação aplicada a
análise de ensembles conformacionais de proteínas**

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Física Aplicada
Opção: Física Biomolecular

Orientador: Prof. Dr. Glaucius Oliva
Coorientador: Dr. Rinaldo Wander Montalvão

Versão corrigida
(Versão original disponível na Unidade que aloja o Programa)

São Carlos
2017

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Silva Neto, Antonio Marinho da
Geometria diferencial e teoria da informação aplicada a
análise de ensembles conformacionais de proteínas. /
Antonio Marinho da Silva Neto; orientador Glaucius Oliva;
co-orientador Rinaldo Wander Montalvão - versão corrigida
-- São Carlos, 2017.
141 p.

Tese (Doutorado - Programa de Pós-Graduação em Física
Aplicada Biomolecular) -- Instituto de Física de São
Carlos, Universidade de São Paulo, 2017.

1. Geometria diferencial. 2. Teoria da informação. 3.
Flexibilidade de proteínas. 4. Espaço conformacional de
proteínas. I. Oliva, Glaucius, orient. II. Montalvão,
Rinaldo Wander, co-orient. III. Título.

Dedico esta tese aos meus pais, Sandoval Oliveira da Silva e Edilene Cristina Nogueira dos Santos Bassetti.

AGRADECIMENTOS

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, pelo financiamento (processo número: 2013/18398-0) sem o qual o desenvolvimento deste projeto não seria possível.

Agradeço ao Prof. Rinaldo Wander Montalvão por ter aceitado o desafio e pela paciência de orientar um aluno sem experiência em física, programação e matemática para desenvolver um projeto que envolve conceitos de física, programação e matemática.

Agradeço ao Prof. Glaucius Oliva pela disponibilidade e prontidão em assumir a minha orientação e todo o apoio dado ao desenvolvimento deste projeto, apoio sem o qual a finalização deste projeto não seria possível.

Agradeço ao Prof. Michele Vendruscolo, da Universidade de Cambridge, e ao Prof. Carlo Camillone, da Universidade de Milão, por compartilhar conosco seus dados sobre o complexo c-Myb-KIX utilizados neste trabalho.

Agradeço ao Dr. Carlos Henrique Madeiros Castelletti, Profa. Dra. Danyelly Bruneka Gondim Martins e Prof. Dr. José Luiz de Lima Filho, do Laboratório de Imunopatologia Keizo Azami (LIKA) da Universidade Federal de Pernambuco (UFPE), e ao Prof. Dr. Elton Carvalho, da Universidade Federal do Rio Grande Norte (UFRN), pelo apoio e por compartilharem seus recursos computacionais que permitiram a realização das simulações em tempo hábil desenvolvidas neste projeto.

Agradeço à minha esposa, Katarina Holanda, pelo apoio incondicional, companheirismo e por dividir o peso de todas as etapas associadas ao desenvolvimento do doutorado.

Agradeço aos meus pais, Sandoval Oliveira e Edilene Nogueira, por todos os sacrifícios, trabalho árduo e tempo dedicado para viabilizar a melhor educação possível para mim e para meu irmão, Emerson Tiago.

Agradeço a Amanda Souza Camâra pela amizade, discussões, apoio e críticas que contribuíram para o desenvolvimento deste projeto.

Agradeço a todos os amigos e amigas que me apoiaram e dividiram comigo as dificuldades dessa jornada de quatro anos, em especial Evandro Ares de Araújo, Samuel Silva, Renata Vieira Bueno, Fernando Ruggiero Bachega e Larissa Gonçalves Maciel.

Agradeço aos companheiros do podcast PODentender, Fabio Neves, Carol Lacerda, Katarina Holanda e Fernando Bachega por terem abraçado junto comigo o desafio de fazer divulgação científica (DC) no Brasil. Também agradeço aos amigos(as) que fazem o portal Dragões de Garagem pelo apoio dado ao PODentender, em especial Luciano Queiroz e Vitor Nascimento.

"...It from bits.

*Otherwise put, every **it** - ... - derives its function, its meaning, its very existence entirely,
even if in some contexts indirectly, from the apparatus, elicited answers to yes or no
questions, binary choices, **bits**"*

- John Archibald Wheeler

RESUMO

SILVA NETO, A. M. da **Geometria diferencial e teoria da informação aplicada a análise de ensembles conformacionais de proteínas**. 2017. 141p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2017.

Um dos maiores desafios atuais na biologia estrutural é como lidar com flexibilidade de proteínas. Além do desafio experimental, uma limitação teórica é a falta de uma linguagem matemática conveniente para representação do espaço conformacional de proteínas. As representações mais populares apresentam diversas limitações, que se refletem nas dificuldades associadas à análise de *ensembles* conformacionais. Nesse contexto, a aplicação de geometria diferencial (GD) e teoria da informação (TI) foi pouco explorada. Neste trabalho investigamos o uso de descritores de GD e TI como uma representação matemática do espaço conformacional de proteínas aplicada à análise de *ensembles* conformacionais. O cálculo dos descritores de GD consiste em representar o *backbone* de proteínas como curvas espaciais e caracterizá-las utilizando os seus valores de curvatura, κ , e torção, τ . Baseado nesses valores, definimos medidas de flexibilidade, de distância entre conformações e aplicamos uma estratégia de *clustering* para identificação de estados conformacionais. Para permitir a aplicação de TI, desenvolvemos um sistema de codificação desses descritores para expressar cada conformação por uma sequência de símbolos finitos. A partir dessas sequências, definimos uma medida da informação associada a um resíduo, R_{res} , e a uma conformação, R_{conf} . Para investigar sua eficácia, aplicamos os métodos propostos aos *ensembles* conformacionais de três sistemas testes: 1) Ubiquitina, 2) E1-DBD do HPV18 e 3) as etapas de formação do complexo c-Myb-KIX. A análise da representação por geometria diferencial se mostrou igualmente eficaz ou superior aos métodos comumente utilizados em todos os sistemas analisados. O método é especialmente útil para monitoramento de estabilidade de hélices e para análise de proteínas e regiões muito flexíveis, pois evita a necessidade de sobreposição estrutural. Os valores de R_{conf} se apresentaram úteis para análise de processos de enovelamento e resíduos próximos a regiões funcionais tendem a apresentar maiores valores R_{res} . No entanto, o papel desses resíduos é incerto e mais estudos são necessários para determinar se há e qual é seu real significado. Apesar disso, as medidas de informação se mostraram úteis para comparação de estados conformacionais e permitem levantar hipóteses testáveis em laboratório. Por fim, a representação por GD é computacionalmente conveniente, intuitiva, evita todas as limitações dos métodos popularmente utilizados e se mostrou eficaz para análise de *ensembles* conformacionais.

Palavras-chave: Geometria diferencial. Teoria da informação. Estrutura de proteínas. Análise conformacional. Espaço conformacional.

ABSTRACT

SILVA NETO, A. M. da **Differential geometry and information theory application to protein conformational analyses**. 2017. 141p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2017.

One of the major challenges of modern structural biology is how to deal with protein flexibility. Besides the experimental difficulties, a relatively overlooked theoretical challenge is the lack of a proper mathematical language to represent protein conformational space. The most popular representations have severe limitations, which reflects on the difficulties associated with conformational ensemble analyses. However, differential geometry (GD) and information theory (TI) can help to overcome such difficulties and were not well explored in this context. Here we investigate the usage of DG and TI as a mathematical representation of protein conformational space applied to the analyses of conformational ensembles. The DG descriptors calculation consists of representing protein backbone as a spatial curve and describes it by its curvature, κ , and torsion, τ . Based on those values, the distance between conformation and flexibility measurements were defined and a clustering algorithm was applied to identify conformational states. For the application of TI, a coding system for DG descriptors was developed to express each conformation as a sequence of finite symbols. Based on those sequences, information measurements associated to a residue, R_{res} , and to a conformation, R_{conf} , were defined. To investigate its efficacy, the proposed method was applied to conformation ensembles of three test systems: 1) Ubiquitin, 2) E1-DBD of HPV18 and 3) the steps of c-Myb-KIX binding. The DG analyses show equally good or superior performance when compared with popular methods on all tested system. In addition, the methods are especially useful to monitoring helix stability and analyses of very flexible proteins (or regions), since avoids the necessity of superposing structures. The values of R_{conf} are useful to compare different steps of a folding process and residues near regions involved in binding events tend to present higher values of R_{res} . However, those residues importance is uncertain and further studies are necessary to determinate if and how those can contribute to protein function. Nevertheless, the information measurements were informative on the comparison of compare conformational states and allow to formulate a testable hypothesis. On the other hand, the GD representation is computationally convenient, intuitive and avoid most of the limitations of the popular method applied to conformational ensemble analyses.

Keywords: Differential geometry. Information theory. Flexibility. Protein structure. Conformational analyses.

LISTA DE FIGURAS

Figura 1 – Diagrama do funil de enovelamento de proteínas. Existem muitos estados não enovelados de alta energia que tendem aos estados de menor energia por diferentes trajetórias e, eventualmente, alcançam a região de mais baixa energia. Essas regiões de mais baixa energia determina o que consiste o estado nativo.	25
Figura 2 – Diagrama do modelo de seleção conformacional estendido para duas moléculas (A e B) capazes de se ligar. Essas moléculas são representadas pelos pontos vermelho e amarelo (A e B, respectivamente) e sua posição no fundo do funil de enovelamento representa os estados conformacionais e a energia associada a cada um deles. A seta com duas pontas indica mecanismos de seleção conformacional. As linhas pontilhadas representam trajetórias. Na parte inferior da figura encontram-se representadas quatro possíveis cenários de ligação na perspectiva do ligante A.	26
Figura 3 – Representação de proteínas por A) coordenadas atômicas, B) variáveis coletivas e C) ângulos ϕ e ψ . Na figura é ilustrado um ângulo como variável coletiva, no entanto, qualquer quantidade calculada a partir da estrutura pode ser utilizada como variável coletiva.	29
Figura 4 – Projeção de uma distribuição uniforme de ângulos ϕ - ψ , após transformação para o sistema de coordenadas proposto por, ¹ no plano determinado pelos dois primeiros PCAs.	31
Figura 5 – Representação de estruturas de proteínas por A) ligações virtuais, B) <i>fitting</i> de hélices e C) interpolação, todas baseadas apenas nas coordenadas de carbonos α	32
Figura 6 – Decomposição de eventos para escolha de três possibilidades. Se a mesma coleção de eventos pode ser decomposta como o resultado de escolhas sucessivas, a mensagem original será igual a soma ponderada dos valores das escolhas.	35
Figura 7 – Logo do sítio de ligação reconhecidos pela Fis de Escherichia coli. Na parte superior da figura estão ilustrado 12 sítios e seus respectivos valores de $R_{sequência}$, Na parte inferior, o logo que reflete a quantidade de informação para cada base do sítio, $R_{sequência}(l)$	40
Figura 8 – Conceito de reparametrização de curvas paramétricas. A representação $\vec{r}(s)$ é equivalente a $\vec{r}(h(t))$	49
Figura 9 – <i>Frenet-Serret frame</i> composto por três vetores (vetor tangente, $\vec{T}$, vetor normal, $\vec{N}$, e vetor binormal, $\vec{B}$).	50

Figura 10 – Significado dos valores de curvatura κ e torção τ e sua relação com os vetores $\vec{T}$, $\vec{N}$ e $\vec{B}$	53
Figura 11 – Esquema do processo de codificação dos descritores de geometria diferencial em símbolos. Um alfabeto é definido para representar intervalos de κ e τ a partir da repartição de intervalos que possuam a mesma área no histograma. A partir desse alfabeto é possível expressar os valores de κ e τ de cada resíduo por símbolos e, portanto, cada conformação i será representada por duas sequências de símbolos.	57
Figura 12 – Ensemble conformacional da Ubiquitina determinado por metadinâmica restringida por RMN, disponível no PDB (2LJ5).	61
Figura 13 – Modelos do DBD da E1 com o loop de ligação na conformação ligada ao DNA (DBL "ligado") e em uma conformação "livre"(DBL "livre") . . .	63
Figura 14 – <i>Ensembles</i> do complexo c-Myb-KIX de diferentes etapas do enovelamento da c-Myb (azul) utilizados em um estudo que demonstrou a existência de um mecanismo de " <i>templated folding</i> "para proteínas não estruturadas. Três mutações (I72V,I26V e L43A) da KIX (verde) permitiram travar o processo de enovelamento da c-Myb em diferentes etapas. A etapa 1) demonstra que a interação entre as duas proteínas não começa no sítio de ligação final.	64
Figura 15 – Estados conformacionais da c-Myb de todas as etapas de enovelamento alinhadas, com destaque para a conformação final em α -hélice.	65
Figura 16 – Valores de curvatura (Azul) e torção (Vermelho) das conformações do <i>ensemble</i> da Ubiquitina e da conformação da Ubiquitina quando ligada a E2 observada na estrutura cristalográfica (3JW0).	68
Figura 17 – Gráficos "em violino"da distribuição dos valores de curvatura e torção observados no <i>ensemble</i> da Ubiquitina.	69
Figura 18 – A) Conformações da cauda C-terminal e da B) região do resíduos 10, uma glicina, do <i>ensemble</i> da Ubiquitina.	70
Figura 19 – <i>Root Mean Square Fluctuation</i> dos $C\alpha$ resíduos calculados para o <i>ensemble</i> da Ubiquitina.	71
Figura 20 – <i>Z-score</i> das distância κ/τ máxima observada nos resíduos do <i>ensemble</i> da Ubiquitina.	71
Figura 21 – "Mapa de calor"das distâncias euclidianas observadas entre a primeira e as demais conformações do <i>ensemble</i> da Ubiquitina.	72
Figura 22 – "Mapa de calor"das distâncias euclidianas observadas entre a primeira e as demais conformações do <i>ensemble</i> da Ubiquitina.	73
Figura 23 – Sobreposição das conformações do <i>ensemble</i> da Ubiquitina e a conformação da Ubiquitina ligada a E2 (3JW0 ²).	74

Figura 24 – A) Valores de informação associada a observação dos resíduos, R_{res} , em bits do <i>ensemble</i> da Ubiquitina. B) Superfície da ubiquitina destacando duas regiões hidrofóbicas envolvidas com interação com outras proteínas (RH1 e RH2), as lisinas que determinam o tipo de ubiquitinação (Amarelo) e os resíduos que apresentaram $R_{res} \geq 5.48bits$ (Vermelho).	76
Figura 25 – A) Informação conformacional, R_{conf} , e B) matriz de identidade entre as seqüências das conformações do <i>ensemble</i> da Ubiquitina.	77
Figura 26 – Distribuição de ângulos de abertura do <i>loop</i> de ligação ao DNA observados nas simulações dos modelos do E1DBD do HPV18.	79
Figura 27 – Valores de κ e τ por resíduos observados no <i>ensemble</i> obtido por simulações dos modelos do E1DBD do HPV18.	80
Figura 28 – Gráfico de "violino" das distribuições dos valores de κ e τ por resíduos observados no <i>ensemble</i> obtido por simulações dos modelos do E1DBD do HPV18. Resíduos com $Z - score > 1.0$ estão marcados em vermelho	81
Figura 29 – <i>Root Mean Square Flucuation</i> do <i>ensemble</i> obtido por simulações dos modelos do E1DBD do HPV18.	82
Figura 30 – O $Z - score$ das distâncias κ/τ máximas por resíduo do <i>ensemble</i> obtido por simulações dos modelos do E1DBD do HPV18.	82
Figura 31 – A) Resultados de <i>Clustering</i> pelo HDBScan dos valores de curvatura e torção observados para o resíduo A4 (amarelo) no <i>ensemble</i> conformacional do E1DBD do HPV18. Dois estados foram identificados, no primeiro (azul) a maior parte das conformações mantém uma α -hélice e no segundo (verde) a maioria apresenta a mesma hélice com os quatro primeiros resíduos desenovelados. B) Valores de curvatura e torção das conformações do resíduo A4 segundo os estados "Fechado"(roxo), "Semi-fechado"(azul), "Aberto"(amarelo) e o da conformação ligada ao DNA (vermelho)	83
Figura 32 – A) <i>Clusters</i> encontrados pelo HDBScan dos valores de curvatura e torção observados para o resíduo D27 no <i>ensemble</i> conformacional do E1DBD do HPV18. B) Valores de curvatura e torção das conformações do resíduo D27 segundo os estados "Fechado"(roxo), "Semi-fechado"(azul), "Aberto"(amarelo) e o da conformação ligada ao DNA (vermelho).	84
Figura 33 – Resultados de <i>Clustering</i> por HDBScan dos valores de curvatura e torção observados para o resíduo T30 do <i>ensemble</i> conformacional do E1DBD do HPV18. B) Valores de curvatura e torção das conformações do resíduo T30 segundo os estados "Fechado"(roxo), "Semi-fechado"(azul), "Aberto"(amarelo) e o da conformação ligada ao DNA (vermelho).	85

Figura 34 – Resultados de <i>Clustering</i> por HDBScan dos valores de curvatura e torção observados para o resíduo 4 (amarelo) do <i>ensemble</i> conformacional do E1DBD do HPV18. Dois estados foram identificados, no primeiro (azul) a maior parte das conformações mantém uma α -hélice e no segundo (verde) a maioria apresenta a mesma hélice com somente os quatro primeiros resíduos desenovelados.	86
Figura 35 – A) Distribuição dos valores de R_{res} e B) os valores de R_{res} por resíduo do <i>ensemble</i> do E1DBD do HPV18.	88
Figura 36 – Superfície do E1DBD do HPV18 com o <i>loop</i> de ligação ao DNA, DBL(azul), a hélice de ligação ao DNA, DBH (laranja), a hélice de dimerização (verde) e resíduos que apresentaram $R_{res} \geq 4.06bits$ (vermelho). A maioria dos resíduos com $R_{res} \geq 4.06bits$ encontram-se próximas à regiões diretamente envolvidas com a interação com outras moléculas.	89
Figura 37 – Gráfico dos valores de R_{conf} e ângulo do DBL para cada conformação do <i>ensemble</i> do E1DBD do HPV18 obtido por simulação. Em destaque, as duas conformações com menores valores de R_{conf}	90
Figura 38 – Variância dos valores de κ e τ entre os estados "fechado"(laranja) e "semi-fechado"(azul).	91
Figura 39 – Média e desvio padrão de κ e τ observados para a região do <i>DNA Binding loop</i> entre os estados "fechado"(laranja) e "semi-fechado"(azul claro) e os valores de κ e τ da conformação do DBL observada quando ligada ao DNA.	92
Figura 40 – Posição na superfície do E1DBD do DBL(azul), do DBH(laranja), 59-70 (vermelho) e 121-127(amarelo) e a posição do domínio de oligomerização.	93
Figura 41 – Mecanismo proposto envolvendo resíduos determinantes para abertura e dinâmica do DBL. A R85 forma ligações de hidrogênio entre a D27 e a Q106 (azul) e os resíduos F24, P107 e W34 formam uma região hidrofóbica interna que permite que o DBL mantenha sua forma característica. O HPV18 possui uma alanina na posição 89 (vermelho), enquanto os demais HPVs apresentam outros resíduos capazes de interagir com a arginina central.	94
Figura 42 – Valores de curvatura (Azul) e torção (Vermelho) das conformações da c-Myb pertencentes aos <i>ensembles</i> das etapas de ligação à KIX.	97
Figura 43 – Gráficos em "violino" da distribuição dos valores de curvatura e torção observados nas conformações da c-Myb pertencentes aos <i>ensembles</i> das etapas de ligação à KIX.	98
Figura 44 – <i>Root Mean Square Flucuation</i> dos carbonos α das conformações da c-Myb pertencentes aos <i>ensembles</i> das etapas de ligação à KIX.	99

Figura 45 – O Z – score das distâncias κ/τ máximas por resíduo das conformações da c-Myb pertencentes aos <i>ensembles</i> das etapas de ligação à KIX. . . .	99
Figura 46 – A) Resultados de <i>clustering</i> e B) distribuição das 4 etapas de ligação dos valores de curvatura e torção do resíduos 101 observados nas conformações da c-Myb pertencentes aos <i>ensembles</i> das etapas de ligação à KIX. Há quatro ensembles referidos de acordo a mutação da KIX utilizada para determiná-los. Os resultados de <i>clustering</i> permitem identificar quase todos os estados individualmente.	101
Figura 47 – <i>Ensemble</i> conformacional da c-Myb-KIX(I72V). Os resíduos 88-100 da c-Myb forma uma alça relativamente estável (<i>rainbow</i>) devido ao contato com a T11 da KIX, enquanto os demais resíduos, 101-112, permanecem com posicionamento instável (cinza).	102
Figura 48 – Rede de ligação de hidrogênio envolvendo a S101,T102,N104 e E105 da c-Myb e K21 da KIX. Essa rede só é formada na segunda etapa de enovelamento (cMyb-KIX-I26V).	102
Figura 49 – A) Resultados de <i>clustering</i> e B) distribuição das 4 etapas de ligação dos valores de curvatura e torção do resíduos E96 (azul) observados nas conformações da c-Myb pertencentes aos <i>ensembles</i> das etapas de ligação à KIX. Há quatro ensembles referidos de acordo a mutação da KIX utilizada para determiná-los. Os resultados de <i>clustering</i> permitem identificar quase todos os estados individualmente. A região 88-95 (cinza) e o resíduo 101 (vermelho) encontram-se destacadas.	103
Figura 50 – Interação entre a R61 e a E96 auxilia na estabilização da c-Myb na primeira etapa da ligação com a KIX.	104
Figura 51 – A) Resultados de <i>clustering</i> e B) distribuição das 4 etapas de ligação dos valores de curvatura e torção do resíduos K107 (amarelo) observados nas conformações da c-Myb pertencentes aos <i>ensembles</i> das etapas de ligação à KIX. Há quatro ensembles, todos referidos de acordo a mutação da KIX utilizada para determiná-los. O resultado de <i>clustering</i> permite identificar apenas as duas últimas etapas de ligação e reflete, o processo de enovelamento da última volta de hélice. A região 88-95 (cinza) e os resíduos S101(vermelho) e E96(azul) encontram-se destacados.	105
Figura 52 – Distribuição dos valores de R_{conf} da das quatro etapas de ligação da A)c-Myb com a B)KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4).	106

Figura 53 – Valores de $\langle R_{conf} \rangle$ dos <i>ensembles</i> da KIX correspondente e valores da constante de associação reportada para cada etapa de ligação com a KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4). Valores de k_{on} divulgados no material suplementar em ³ . . .	108
Figura 54 – Distribuição dos valores de R_{res} da das quatro etapas de ligação da A)c-Myb com a B)KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4).	109
Figura 55 – A) R_{res} calculados para as conformações da KIX do <i>ensemble</i> da primeira etapa de ligação da c-Myb com a KIX (c-Myb-KIXI72V). B) Distribuição dos resíduos $R_{res} \geq 5.0$ (rosa), $R_{res} \geq 6.0$ (vermelho), do sítio de ligação a c-Myb (azul) e sítio de ligação ao MLL(amarelo) na superfície da KIX. Uma conformação da c-Myb está representada na forma de <i>ribbon</i>	110
Figura 56 – Tempo em segundos para processar 9006 conformações do <i>ensembles</i> do E1-DBD do HPV18 (145 resíduos) por <i>core</i> utilizado pelo protótipo desenvolvido para esta tese. Performance em um intel core i7-4770 3.40GHz, 16Gb de memória RAM, Ubuntu 16.04.	112
Figura 57 – Matriz de identidade das sequências conformacionais das etapas da ligação da c-Myb com a KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4)	115

SUMÁRIO

1	INTRODUÇÃO	21
1.1	A importância da inclusão da flexibilidade em proteínas, uma breve revisão histórica	22
1.2	Representação do espaço conformacional, suas dificuldades e geometria diferencial como alternativa	27
1.2.1	Representação de proteínas por geometria diferencial	31
1.3	Teoria da Informação	34
1.3.1	Informação e incerteza	34
1.3.2	Teoria da Informação aplicada aos sistemas biomoleculares	38
2	OBJETIVOS	45
2.1	Objetivo geral	45
2.2	Objetivos específicos	45
3	MÉTODOS E METODOLOGIA	47
3.1	Cálculo dos descritores de geometria diferencial	47
3.1.1	Representação do <i>backbone</i> de estruturas proteicas por curvas paramétricas	48
3.1.2	Cálculo de curvatura e torção	50
3.1.3	Representação de geometria diferencial de ensembles conformacionais de proteínas	53
3.2	Métrica de distância e medidas de variabilidade estrutural local	53
3.2.1	Variabilidade estrutural local	54
3.2.2	Identificação de resíduos flexíveis	55
3.3	Identificação de estados conformacionais	55
3.4	Teoria da Informação	56
3.4.1	Codificação dos descritores de geometria diferencial	56
3.4.2	$H(n)$, R_{res} e R_{conf}	57
3.5	Conjunto teste	59
3.5.1	Ubiquitina	59
3.5.2	Domínio de Ligação ao DNA da E1 de Papilomavirus Humano	60
3.5.2.1	Modelagem molecular por homologia	62
3.5.2.2	Simulação por dinâmica molecular	62
3.5.2.3	Definição de variável coletiva e análise de forma do DBL	63
3.5.3	c-Myb-KIX	63
4	RESULTADOS E DISCUSSÃO	67

4.1	Ubiquitina	67
4.1.1	Medidas de flexibilidade	67
4.1.2	Comparação com estados de referência	70
4.1.3	Medidas de informação conformacional	72
4.2	Domínio de Ligação ao DNA da E1	75
4.2.1	Qualidade dos modelos por homologia	76
4.2.2	Distribuição de ângulos do <i>loop</i> de ligação ao DNA	77
4.2.3	Medidas de Flexibilidade	78
4.2.4	Identificação de estados conformacionais	79
4.2.5	Medidas de Informação	81
4.2.6	Análise de acessibilidade da conformação do DBL ligada ao DNA	87
4.2.7	Resíduos chaves para a transição entre os estados “fechado” e “semi-fechado” do DBL	91
4.3	c-Myb-KIX	96
4.3.1	Medidas de flexibilidade	96
4.3.2	Identificação de estados conformacionais	98
4.3.3	Medidas de Informação	104
4.4	Vantagens, limitações e perspectivas para os métodos propostos	107
4.4.1	Representação do espaço conformacional de proteínas por geometria diferencial	107
4.4.2	Teoria da Informação aplicada a análise conformacional	113
5	CONCLUSÃO	117
	REFERÊNCIAS	119
	APÊNDICES	131
	APÊNDICE A – CÁLCULO DA CURVATURA EM FUNÇÃO DE UMA CURVA $\gamma(t)$	133
	APÊNDICE B – CÁLCULO DA TORÇÃO EM FUNÇÃO DE UMA CURVA $\gamma(t)$	135
	APÊNDICE C – PROVA DO TEOREMA FUNDAMENTAL DAS CURVAS ESPACIAIS	137

1 INTRODUÇÃO

Organismos vivos são capazes de realizar desde tarefas relativamente simples, como identificar a presença de um composto químico no ambiente, até tarefas mais complexas, como coordenar o comportamento de células de vários tecidos e sistemas diferentes em resposta à uma determinada condição ambiental. Todos esses processos são baseados na obtenção e processamento de informação providas do ambiente e, em última análise, isso só é possível em sistemas biológicos através de interações entre duas ou mais biomoléculas. Dentre as biomoléculas, as proteínas são de grande interesse acadêmico e tecnológico, pois estão envolvidas em virtualmente todos os processos biológicos. Uma parte importante de seu papel nesses processos é a capacidade de formar complexos com outras moléculas, como outras proteínas, pequenas moléculas, carboidratos, lipídios, DNA e RNA. A formação desses complexos é importante para a construção de macroestruturas capazes de realizar funções estruturais na célula. Tais estruturas compreendem desde microtúbulos, importantes vias para trânsito de organelas e divisão celular,⁴ até a formação de maquinários moleculares capazes de catalisar complexas reações, como os ribossomos responsáveis pela síntese de proteínas.⁵ Mesmo interações relativamente simples entre duas proteínas ou uma proteína e uma molécula pequena estão no centro do funcionamento de sistemas essenciais para manutenção da vida. Dentre os inúmeros exemplos, podemos citar o reconhecimento de diferentes neurotransmissores na fenda sináptica entre neurônios,⁶ o transporte de colesterol por lipoproteínas no sangue,⁷ resposta do sistema imune a organismos invasores pelo reconhecimento da presença de antígenos.⁸ Sendo assim, eventos de ligação são mecanismos explorados universalmente por sistemas biológicos para receber e processar informações e/ou enviá-las ao ambiente externo. Portanto, ser capaz de descrever e prever o comportamento desses eventos é importante não apenas para permitir o desenvolvimento de novas tecnologias e soluções para problemas biomédicos e biotecnológicos, mas também significa ser capaz de descrever como processos essenciais para o surgimento e manutenção da vida nos seus mais diversos graus de complexidade funcionam.

Apesar de vários avanços nas últimas décadas, a descrição e predição da formação de complexos envolvendo proteínas apresenta muitos desafios, especialmente quando envolve flexibilidade estrutural e estimativas da entropia.⁹ Avanços recentes no uso de simulações de dinâmica molecular para obtenção de *ensembles* conformacionais que reproduzem dados de Ressonância Magnética Nuclear (RMN) abrem a possibilidade de acessar experimentalmente conjuntos de conformações que representam os vários estados conformacionais acessíveis em solução e movimentos que ocorrem em escalas de tempo biologicamente relevantes.¹⁰⁻¹² Apesar de ainda em desenvolvimento, essas técnicas já vêm sendo utilizadas e se provaram úteis em vários cenários, por exemplo, para obtenção conjuntos de confor-

mações representativos do equilíbrio em solução de proteína¹³ e de RNA,¹⁴ determinação de mudanças conformacionais associada à formação de complexos proteína-proteína,¹⁵ caracterização de movimentos entre domínios¹² e estudo de mecanismos de enovelamento relacionados a reconhecimentos de sítio de ligação por proteínas não estruturadas.³ No entanto, mesmo nos casos em que é possível obter esses conjuntos, a análise dos dados ainda envolve vários desafios. Parte desses desafios se deve a limitações dos sistemas métricos comumente utilizados para esse tipo de análise. De modo que a determinação de características importantes como a identificação de estados conformacionais relevantes para função da proteína ou estabelecer medidas de flexibilidade global e local podem se tornar limitadas ou até mesmo impraticáveis. Uma alternativa relativamente pouco explorada neste contexto é o uso de descritores de geometria diferencial (GD) e teoria da informação (TI).

Dentro desse contexto, o nosso objetivo geral é explorar a aplicação de GD e TI para obtenção de características estruturais relevantes a partir de *ensembles* conformacionais de proteínas. Portanto, se faz necessária a apresentação e discussão de alguns conceitos e dificuldades relacionados a flexibilidade em proteínas (seção 1.1), as atuais alternativas para representar e analisar *ensembles* conformacionais e suas limitações (seção 1.2), o histórico da aplicação de GD para estudo de estruturas proteicas (subseção 1.2.1) e, por fim, uma breve revisão e discussão sobre paralelos entre conceitos de teoria da informação (seção 1.2) e sua aplicação para análise de sistemas biomoleculares (subseção 1.3.2).

1.1 A importância da inclusão da flexibilidade em proteínas, uma breve revisão histórica

Antes de discutir a importância da flexibilidade de proteínas é preciso definir do modo mais preciso o significado deste conceito. Na literatura é possível encontrar termos como “desordem” e “flexibilidade” muitas vezes se referindo à mesma ideia, isto é, a variação da posição de átomos de uma determinada estrutura em determinadas condições. Jöel Janin e Michael J.E. Sternberg argumentam que o termo “desordem” costuma implicar na falta de restrições em alguns ou todos os graus de liberdade de uma estrutura, enquanto o termo “flexibilidade” conota características associadas à variação dentro de restrições de um ou mais graus de liberdade dessa estrutura.¹⁶ Baseado nisto, daremos preferência ao uso do termo “flexibilidade” devido à conotação de normalidade atribuída ao mesmo.

Entender o funcionamento de proteínas é um grande desafio e podemos citar como primeiro grande avanço na compreensão do comportamento dessas biomoléculas a construção do paradigma estrutura-função. Com suas raízes nos trabalhos de Louis Pasteur sobre isômeros ópticos, que demonstrou a importância e necessidade de interpretar moléculas como objetos tridimensionais, o paradigma estrutura-função postula que a estrutura molecular tridimensional de uma molécula determina o seu comportamento em

um determinado contexto, portanto, estaria diretamente relacionada ao que atribuímos como sua função em um determinado sistema. Este conceito é um dos pilares da biologia estrutural. Após a publicação da estrutura da Mioglobina, a primeira estrutura de proteína determinada por cristalografia por difração de raios-X, em 1958 em um trabalho pioneiro de Kendrew et al.,¹⁷ iniciou-se uma grande revolução na forma de entender o funcionamento de proteínas e atualmente existem 132.000 estruturas depositadas no Protein DataBank (PDB), das quais 119.000 foram determinadas por cristalografia por difração de raios-X. Décadas de análise dessas estruturas e comparação com resultados experimentais resultaram no acúmulo de evidências que fundamentam a ideia de que é possível entender e prever o comportamento de biomoléculas, especialmente proteínas, através da análise de suas estruturas moleculares. Um bom exemplo é a elucidação de mecanismos associados a mutação que levam ao desenvolvimento de câncer de mama desvendado pela análise da estrutura do complexo RAD51-BRCA2.¹⁸ As informações obtidas a partir dessa estrutura são usadas como base para desenvolvimento de novos fármacos¹⁹ e auxiliaram a compreensão de mecanismos gerais associados a doenças humanas.²⁰

Como as estruturas cristalográficas representam a principal fonte de informação experimental para a construção do conhecimento da biologia estrutural é natural que as limitações dessa técnica acabem refletindo nas atuais limitações em descrever e entender o comportamento de biomoléculas. Uma das dificuldades de se obter informação de flexibilidade por cristalografia é que a medida que aumentam as variações de posição de um conjunto de átomos entre as réplicas que compõem o cristal utilizado no experimento de difração de raios-X, mais difícil será obter um mapa de densidade eletrônica definido para a região da variação.²¹ Soma-se a maior dificuldade de obtenção de cristais para proteínas que apresentam significativas mudanças conformacionais. Por esses motivos, a utilização da principal ferramenta experimental para biologia estrutural disponível atualmente é limitada para caracterizar flexibilidade de proteínas. A importância desse desafio é reconhecida por cristalógrafos pelo menos desde a década de 1970 e 1980.^{22–24} A existência de conformações diferentes das observadas em estruturas obtidas por difração de raios-X de cristais de proteínas e que não são capturados em cristais é aceita na comunidade científica.^{25,26} Apesar das limitações, estratégias para utilizar os dados de difração de raios-X para obter informações sobre o *ensemble* conformacional associado ao cristal vêm sendo desenvolvidas.²¹

Apesar de hoje ser consenso a importância da flexibilidade em proteínas, a noção de que proteínas em seu estado nativo devem ser representadas por um único estado conformacional, proposta inicialmente por Mirsky e Paulin em 1936,²⁷ foi a visão dominante por muito tempo. O primeiro trabalho a sugerir a inclusão da flexibilidade de proteínas como fator necessário para explicar características funcionais foi publicado por Fred Karush em 1950,²⁸ contrariando a visão proposta por Mirsky e Paulin. Karush sugere que para melhor explicar dados termodinâmicos da ligação de albumina seria preciso incluir o que

ele chamou de “adaptabilidade configuracional”, isto é, assumir que um número de sítios de ligação poderia assumir várias “configurações” em equilíbrio com aproximadamente a mesma energia associada.²⁸ Em seguida, outros dois trabalhos propuseram de formas distintas a inclusão da flexibilidade para explicar o comportamento de proteínas. Em 1958, Koshland, Ray e Erwin propuseram o modelo de encaixe induzido, que apesar de considerar o estado nativo como uma única conformação, assumia que a ligação com outra molécula poderia induzir mudanças conformacionais que favorecessem a interação.²⁹

Em 1965, Monod, Wyman e Changeux propuseram que dois ou mais estados conformacionais reversíveis são acessíveis em solução para explicar alosterismo e fenômenos de cooperação entre oligômeros.³⁰ Ambos os modelos consistiam em explicações fenomenológicas e ainda havia a necessidade de propor um modelo que explicasse atomisticamente esses fenômenos. Estudos de RMN nas décadas de 70 e 80 já apontavam para a natureza complexa da dinâmica de estruturas de proteínas.^{31,32} Após o acúmulo de evidências experimentais a favor do modelo de Monod, quase 30 anos depois foi formulado o conceito de Folding Funnel 1991.³³ A ideia de Folding Funnel consiste em descrever o perfil de energia potencial de uma proteína em função de um sistema de coordenadas conformacionais assumindo o princípio de frustração mínima. Este conceito resolveu o famoso paradoxo de Levinthal³⁴ e permitiu a comunidade científica avançar significativamente no estudo experimental e teórico dos mecanismos de enovelamento de proteínas.³⁵ Uma das implicações diretas para a descrição de flexibilidade em proteínas é que a forma e rugosidade da região de menor energia do funil, i. e., o fundo do funil, determina o que deve ser considerado como estado nativo da proteína [Figura 1](#). Consequentemente, a distribuição de estados conformacionais acessíveis em solução e as barreiras energéticas que determinam a taxa de transição entre esses estados em uma dada condição ambiental estão diretamente relacionadas à topologia do fundo do funil. De modo geral, podemos assumir que funis de proteínas mais rígidas tendem a apresentar um fundo com um único ou alguns poucos mínimos de energia separados por barreiras energéticas relativamente grandes. Proteínas flexíveis apresentariam uma superfície com muitos mínimos de energia e pequenas barreiras energéticas entre eles, permitindo rápidas transições entre os estados acessíveis. Naturalmente, as implicações desse modelo para eventos de ligação foram logo exploradas e um paralelo para ligação, o *binding funnel*, foi proposto.³⁶

Mais recentemente uma extensão desse modelo foi proposta para descrever o cenário mais geral ao também incluir nessa descrição os modelos clássicos de chave-fechadura e encaixe induzido³⁷ [Figura 2](#). Apesar dos avanços, casos como os das proteínas intrinsecamente não estruturadas que podem ter enovelamentos induzidos por eventos de ligação^{3,38} exigem aprimoramento desse modelo. Existem discussões sobre a possibilidade da flexibilidade observada *in vitro* em proteínas não estruturadas seja na verdade causada pelo próprio método de produção em laboratório e não necessariamente representativas do seu comportamento no ambiente celular,¹⁶ mas a importância da flexibilidade para

essas proteínas é evidente especialmente pela sua capacidade de interagir com múltiplas moléculas por meio de conformações diferentes.

Figura 1 – Diagrama do funil de enovelamento de proteínas. Existem muitos estados não enovelados de alta energia que tendem aos estados de menor energia por diferentes trajetórias e, eventualmente, alcançam a região de mais baixa energia. Essas regiões de mais baixa energia determina o que consiste o estado nativo.

Fonte: DILL; MACCALLUM..³⁹

Apesar de interessante do ponto de vista teórico, pode-se questionar o quão relevante realmente é esse aspecto para o entendimento do comportamento de sistemas biológicos. A relevância desse aspecto se deve em parte à sua predominância em sistemas biológicos e processos de interesses biomédico e biotecnológico. Há estimativas que sugerem que aproximadamente 35% de todas as proteínas humanas possuem regiões flexíveis (ou “desordenadas”) e 15% são completamente não estruturadas.⁴⁰ Há evidências da flexibilidade ser um fator relevante para a evolução de proteínas⁴¹⁻⁴³ e, consequentemente, ela também é um aspecto comprovadamente relevante para o estudo de várias doenças humanas.⁴⁴⁻⁴⁶ Por exemplo, a participação de proteínas intrinsecamente não estruturadas em doenças neurodegenerativas como Alzheimer, Parkinson e Huntington gerou a proposição do conceito “D3”,⁴⁷ do inglês, *protein intrinsic Disorder in neuroDegenerative Diseases*, que coloca em destaque a participação de proteínas intrinsecamente não estruturadas como fatores importantes para essa classe de doenças. Outro exemplo é a interação da BRCA2 e RAD51, cuja flexibilidade da região da BRCA2 capaz de interagir com uma interface de

Figura 2 – Diagrama do modelo de seleção conformacional estendido para duas moléculas (A e B) capazes de se ligar. Essas moléculas são representadas pelos pontos vermelho e amarelo (A e B, respectivamente) e sua posição no fundo do funil de enovelamento representa os estados conformacionais e a energia associada a cada um deles. A seta com duas pontas indica mecanismos de seleção conformacional. As linhas pontilhadas representam trajetórias. Na parte inferior da figura encontram-se representadas quatro possíveis cenários de ligação na perspectiva do ligante A.

Fonte: Adaptada de CSERMELY; PALOTAI; NUSSINOV..³⁷

dimerização da RAD51 é determinante para que esse complexo se forme. Essa interação é um dos pontos chaves que servem como mecanismo para desenvolvimento de câncer de mama e tem sido alvo para busca de fármacos para essa doença.^{19,48}

Um exemplo emblemático, que ilustra a importância de se considerar a flexibilidade de proteínas para desenho racional de fármacos, é o caso da Gleevec®, uma droga utilizada contra vários tipos de câncer que apresenta alto grau de especificidade para a onco-proteína BCR-abl. Recentemente, aspectos dessa alta especificidade foram desvendados utilizando técnicas de reconstrução de sequências ancestrais e caracterização cinética.⁴⁹ A análise de estruturas cristalográficas de diferentes quinases não revelaram mudanças significativas na interface de ligação o suficiente para explicar a especificidade dessa droga e nesse estudo foi demonstrado experimentalmente que é possível explicar esse fenômeno por características de mudanças conformacionais e mecanismos de ‘encaixe induzido’.⁴⁹ Portanto, a princípio, o desenho de drogas específicas para outras proteínas da mesma família poderiam explorar diferenças entre esses aspectos dinâmicos das quinases. Em um artigo de opinião, um dos autores do trabalho citado aponta a maior atenção dada a etapa de ligação física entre a biomolécula e os ligantes nos estudos de Docking como um dos fatores que atrasaram o entendimento desses mecanismos, enquanto mudanças conformacionais antes e depois da ligação foram relativamente negligenciadas.⁵⁰ Os autores também sugerem que esforços experimentais e computacionais devem ser mais centrados na dinâmica do alvo e do complexo droga-alvo para que melhores resultados sejam obtidos em desenho racional de drogas.⁵⁰

Existem vários exemplos que poderiam ser citados que exemplificam a importância de considerar a flexibilidade para explicar características funcionais de proteínas, alguns explorados em mais detalhes em uma revisão publicada na Chemical Reviews em 2016.²⁶ Apesar de comprovadamente relevante, ainda existem várias dificuldades experimentais e teóricas associadas a obter informações sobre este aspecto das estruturas proteicas.

1.2 Representação do espaço conformacional, suas dificuldades e geometria diferencial como alternativa

A necessidade de lidar com a flexibilidade de proteínas é evidente, mas envolve vários desafios e, de modo geral, podemos alocá-los em duas grandes categorias: amostragem e análise conformacional. O principal objetivo deste projeto é explorar soluções alternativas para os problemas de análise, mas precisamos fazer algumas breves considerações sobre como a integração de métodos computacionais e dados experimentais podem auxiliar a resolver o problema da amostragem.

O problema da amostragem é um reflexo do espaço conformacional de uma proteína típica ser, literalmente, astronômico,³⁴ o que torna a realização de amostragem exaustiva desse espaço impraticável. Por exemplo, considere uma proteína com N resíduos e cada resíduo pode acessar somente dois estados igualmente prováveis, isto é, $p_{na} = p_{nb} = 0.5$. O número de estados conformacionais possíveis é 2^N . Para uma proteína típica, $N = 300$, haveria 10^{90} conformações acessíveis, algo da mesma ordem de grandeza de

variáveis utilizadas em astronomia e astrofísica. Para uma proteína real, muitos estados são acessíveis para cada resíduo e não conhecemos as probabilidades de cada estado. Realizar amostragem puramente teórica nesse espaço conformacional é desafiador. Apesar das limitações associadas, alternativas puramente computacionais como simulações de dinâmica molecular, especialmente técnicas avançadas de amostragem⁵¹ têm se mostrado úteis na elucidação de características funcionais de proteínas, principalmente quando é difícil obter informação experimental, como por exemplo, estudo de proteínas de membrana,⁵² alosteria⁵³ e interação de proteínas com glicolipídios.⁵⁴

Do ponto de vista experimental, existem várias alternativas para obter informações sobre os estados acessíveis em solução, como por exemplo, SAXS, espectrometria de massa, crio-microscopia eletrônica e cristalografia.²¹ No entanto, a obtenção de dados de Ressonância Magnética Nuclear (RMN) ainda é uma das principais alternativas para capturar informação experimental de dinâmica de proteínas com resolução atômica em escala de nano- a milissegundos usando diferentes tipos de medidas, como por exemplo *Residual Dipolar Couplings* (RDC), *Nuclear Overhauser Effect* (NOEs) e *Chemical Shifts* (CS).⁵⁵ Obter esses dados nem sempre é possível ou viável e as informações obtidas normalmente precisam ser complementadas com outros experimentos para que seja possível chegar a conclusões robustas.⁵⁵ A integração de dados experimentais em protocolos de simulação tem se mostrado uma alternativa promissora para a obtenção direta de *ensembles* conformacionais. Por exemplo, avanços recentes na inclusão do uso de RDCs e CS como viés no campo de força utilizado em simulações de dinâmica molecular permitiram a obtenção de *ensembles* conformacionais que capturam movimentos em escalas de tempo biologicamente relevantes e são compatíveis com dados experimentais.¹⁰⁻¹² Apesar de ainda em desenvolvimento, a inclusão de dados experimentais em simulações de dinâmica molecular se mostra uma alternativa viável e promissora para resolver o problema de amostragem conformacional e obtenção de *ensembles* representativos do equilíbrio em solução de proteínas.

Mesmo que seja possível obter esses conjuntos de conformações, o próximo desafio é a extração e análise de informações úteis para entender o funcionamento dessas proteínas. Uma das dificuldades é a falta de uma linguagem adequada e computacionalmente conveniente para descrever o espaço conformacional. Para ilustrar melhor essa deficiência, vamos analisar brevemente os métodos para identificar quais os estados mais populados em um *ensemble* conformacional. Identificar esses estados é útil para inferir mecanismos estruturais que influenciam vários aspectos da função de uma proteína. Esse é essencialmente um problema de agrupamento (ou *clustering*) de dados. Portanto, uma medida adequada de similaridade ou dissimilaridade entre cada ponto no conjunto de dados precisa ser estabelecida. Em termos mais precisos, se faz necessário determinar um vetor cujas dimensões representem graus de liberdade relevantes para o sistema e/ou para o problema que se deseja resolver e, ao comparar esses vetores, diferenças estruturais

relevantes sejam refletidas nessas medidas. As três alternativas mais comuns para definir medidas de similaridade ou dissimilaridade entre duas conformações de proteínas são: a utilização direta das coordenadas dos átomos, variáveis coletivas (CVs) que representem o espaço conformacional de forma intuitiva e/ou os ângulos ϕ e ψ definidos de acordo com o *backbone* da estrutura (Figura 3).

Figura 3 – Representação de proteínas por A) coordenadas atômicas, B) variáveis coletivas e C) ângulos ϕ e ψ . Na figura é ilustrado um ângulo como variável coletiva, no entanto, qualquer quantidade calculada a partir da estrutura pode ser utilizada como variável coletiva.

Fonte: Elaborada pelo autor.

A primeira alternativa é utilizar diretamente as coordenadas dos átomos e apresenta algumas dificuldades. Neste caso uma proteína seria representada por $3N - 6$ graus de liberdade, onde N é o número de átomos, normalmente N é da ordem de alguns milhares a dezenas de milhares. Esse grande número de graus de liberdade torna difícil a interpretação dos resultados de *clustering*. Técnicas de redução de dimensionalidade, como Análise de Componentes Principais (PCA), costumam ser empregadas pois permitem separar movimentos de acordo com sua amplitude. Há dificuldades quanto à interpretação devido à necessidade de uma janela de amostragem adequada e à possibilidade de movimentos de difusão poderem ser erroneamente interpretados como estados de transição.⁵⁶ Para os casos de PCA baseado em coordenadas cartesianas (cPCA), há a necessidade de eliminar movimentos de translação e rotação da trajetória, o que limita sua utilidade para sistemas muito flexíveis.⁵⁷ Apesar disso, o PCA já se mostrou útil para análise de dinâmica estrutural em vários tipos de estudos, como, por exemplo, variabilidade estrutural em subdomínios associados a efeitos de mutações⁵⁸ e na análise de subunidades catalíticas.⁵⁹

Uma métrica de escolha popular é o desvio médio quadrático (*Root Mean Square*

Deviation, RMSD) das coordenadas dos átomos que compõem a proteína e é definido como:

$$RMSD_{A,B} = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\vec{A}_i - \vec{B}_i\|^2} \quad (1.1)$$

onde N é o número de átomos alinhados, $\vec{A}_i$ é o vetor da posição do átomo i da estrutura A , $\vec{B}_i$ é o vetor da posição do átomo i da estrutura B . Uma das maiores limitações do RMSD se deve a necessidade de solucionar o problema da sobreposição estrutural,^{60,61} o que pode ser problemático especialmente para proteínas com alto grau de flexibilidade ou com movimentos entre domínios. Outro problema é a alta sensibilidade para grandes desvios em pequenas regiões. Desse modo, valores relativamente pequenos de RMSD são fáceis de interpretar, enquanto valores relativamente grandes não necessariamente significam grandes diferenças estruturais. Isso se deve ao fato do RMSD não obedecer a desigualdade triangular, isto é:

$$d(x, y) + d(y, z) \geq d(x, z) \quad (1.2)$$

Em termos práticos, se a métrica de distância utilizada viola essa regra, significa que se sua métrica indica similaridade entre a estrutura x e y , e também entre a estrutura y e z , você não tem informação para inferir similaridade entre x e z . Portanto, para que o RMSD se comporte como uma métrica adequada, ele precisa obedecer a desigualdade triangular. Isso só é possível quando se obtém um RMSD calculado a partir de uma superposição ótima (ou, RMSDopt).⁶² É possível obter bons resultados para sobreposição adequada de estruturas rígidas,⁶³ mas há dificuldades para proteínas com alto grau de flexibilidade e alternativas vem sendo desenvolvidas.⁶⁴ Apesar disso, existem métodos já disponíveis que buscam minimizar essas limitações, como o MaxSub⁶⁵ e o TM-score.^{66,67}

Uma outra alternativa é o uso de CVs, que consiste em descrever o espaço conformacional através de variáveis que refletem aspectos essenciais da dinâmica da estrutura, como ângulos, distâncias ou similaridade com uma estrutura de referência. No entanto, exige que algumas características do espaço conformacional sejam conhecidas *a priori* e possam ser representadas adequadamente por um conjunto de CVs, o que na prática é difícil de determinar.⁶⁸ Apesar dessa restrição, nos casos em que é possível determinar um conjunto de CVs adequado, a aplicação de CVs permite obter informações estruturais úteis sobre o funcionamento do sistema de interesse, por exemplo, por simulações de metadinâmica.⁶⁹

O uso de ângulos ϕ - ψ é uma das alternativas de coordenadas internas comumente utilizada, que consiste em descrever o *backbone* de uma proteína como um conjunto de n ângulos ϕ e ψ , n é igual ao número de resíduos. Esses ângulos assumem valores entre -180° e $+180^\circ$, mas choques estéricos restringem esses valores, como demonstrado no gráfico de Ramachandran.⁷⁰ Essa estratégia evita a necessidade de sobrepor as estruturas e já foi sugerido que se dois fragmentos de oito resíduos não diferirem em mais de 120° para todos os resíduos equivalentes, esses fragmentos terão $RMSD < 1.4\text{\AA}$.⁷¹ Há algumas

difficultades para calcular médias e definir medidas de distância, em virtude desses ângulos de diedros serem quantidades periódicas. Uma das propostas para evitar esse problema é realizar uma operação de transformação dessas coordenadas e uma análise de componentes principais (dPCA) para identificar mínimos de energia livre no espaço conformacional.¹ No entanto, esse método também não evita as dificuldades de se trabalhar com quantidades periódicas. Usando dPCA, uma amostragem de várias combinações possíveis entre 0 e 2π representando uma única dupla de ângulos $\phi - \psi$ foi realizada e revelou que a projeção dessas coordenadas no plano definido pelos dois componentes principais gera padrões circulares⁷²(Figura 4). De modo que o espaço definido por essas coordenadas não é uniforme e obedece uma topologia imposta pela natureza periódica dos ângulos, o que dificulta o *clustering* de estruturas.⁷² Apesar dessa limitação, ainda é possível extrair informação úteis para análise de dinâmica utilizando estratégias como essa.⁵⁷

Figura 4 – Projeção de uma distribuição uniforme de ângulos $\phi-\psi$, após transformação para o sistema de coordenadas proposto por,¹ no plano determinado pelos dois primeiros PCAs.

Fonte: HINSEN..⁷²

Desse modo, podemos afirmar que um dos grandes desafios atuais em análise conformacional de proteínas é a falta uma linguagem matemática adequada e computacionalmente conveniente para descrever o espaço conformacional. Uma das alternativas relativamente pouco explorada para este fim é o uso de descritores de GD.

1.2.1 Representação de proteínas por geometria diferencial

A representação de conformações de proteínas baseada em descritores de GD foi inicialmente proposta por S. Rackovsky e H. A. Scheraga em 1978.⁷³ Baseado em vetores de ligações virtuais entre carbonos alfa Figura 5A, eles estabeleceram medidas de curvatura, κ , e torsão, τ , para cada carbono α .⁷³ Detalhes de como estabelecer a representação

matemática por GD serão discutidos na seção XX, no momento, vamos nos ater apenas a apresentar e discutir brevemente alguns aspectos e a evolução história desse tema. Neste trabalho, os autores apontam o RMSD como uma solução insatisfatória para comparar estruturas mesmo para valores baixos de RMSD, pois já havia sido demonstrado que duas proteínas poderiam apresentar baixo RMSD mesmo que uma apresente enovelamento incorreto.⁷⁴ Os autores também argumentam que utilizar um único número para comparar estruturas de proteínas é uma solução inerentemente insatisfatória.⁷³ Na época que o trabalho foi escrito, o termo topologia era inapropriadamente utilizado para se referir ao *backbone* de estruturas proteicas, pois, desconsiderando a ocorrência de pontes dissulfeto, todas as proteínas são curvas sem nós e, portanto, topologicamente equivalentes. Sendo assim, GD seria uma ferramenta mais adequada para descrever e comparar essas curvas.⁷³ Neste mesmo trabalho, os autores também argumentam que medidas locais de diferença entre resíduos baseadas em valores de κ e τ são mais adequadas para comparar estruturas, especialmente por não haver necessidade de sobreposição.⁷³

Figura 5 – Representação de estruturas de proteínas por A) ligações virtuais, B) *fitting* de hélices e C) interpolação, todas baseadas apenas nas coordenadas de carbonos α .

Fonte: A) RACKOVSKY; SCHERAGA⁷³; B) LOUIE; SOMORJAI; KLUG⁷⁵; C) Elaborada pelo autor.

Paralelamente aos trabalhos de Rackovsky e Scheraga, também foi proposta uma abordagem de geometria diferencial do *backbone* de proteínas ao aproximá-lo por uma curva regular, parametrizada e contínua que passa pelos carbonos α .⁷⁶ Neste trabalho, os autores

descrevem α -hélices e folhas β como estruturas que repousam em uma superfície regular parametrizada, no caso, helicoides e catenoides, respectivamente. Essa análise permitiu a demonstração de que estruturas regulares de proteínas estão relacionadas a superfícies mínimas e que as conformações do *backbone* de proteínas correspondem à geodésicas dessas superfícies.⁷⁶ No ano seguinte, os mesmos autores descreveram um algoritmo para obter aproximações das conformações de *backbone* de proteínas por curvas contínuas através de ajuste de hélices por fragmentos (em inglês, *piece-wise fitted helices*)⁷⁵ (Figura 5B).

Os anos seguintes dessa representação foram essencialmente voltados para identificação de padrões gerais em estruturas de proteínas⁷⁷ e alinhamento de estruturas de proteínas baseado na abordagem propostas em 1978.^{73,78} Somente em 2005, o primeiro programa de modelagem molecular, o CHORAL, baseado na representação do espaço conformacional utilizando descritores de GD foi publicado,⁷⁹ com uma implementação mais intuitiva que os métodos propostos até então. Na abordagem utilizada no CHORAL, o *backbone* é representado por uma curva construída por interpolação *spline* de fragmentos utilizando as coordenadas dos carbonos α como nós Figura 5C. Assim é possível obter uma curva contínua e suave que passa por todos os $C\alpha$, de forma análoga à representação popular do *backbone* em *ribbon*. O CHORAL apresenta resultados semelhantes ou superiores ao popular MODELLER.⁷⁹ Uma representação semelhante à utilizada pelo CHORAL foi publicada na forma de *pre-print* no Arxiv em 2008.⁸⁰ Paralelamente, modelos do espaço conformacional utilizando perfis de curvatura para uma arquitetura de repetições de hélices permitiu definir um conjunto de parâmetros utilizando geometria diferencial para explorar o espaço de enovelamentos possíveis para essa arquitetura.⁸¹ Neste trabalho foram identificados enovelamentos já determinados experimentalmente e potenciais novos enovelamentos ainda não identificados ou jamais explorados pela natureza.⁸¹ A representação do *backbone* da proteína usado naquele trabalho é baseado em poli-hélices,⁸¹ semelhante à representação proposta por Louie e colegas em 1983.⁷⁵ Esse trabalho de Hausrath e Goriely é importante para demonstrar como a aplicação de técnicas analíticas baseadas em uma representação contínua de estruturas de proteínas (como a baseada em curvas regulares) pode ser vantajosa para estudar estruturas de proteínas. No entanto, a representação natural de moléculas é a baseada em coordenadas de átomos e várias características de proteínas só serão compreendidas ao analisar essas coordenadas. Diante disso, Hausrath e Goriely propuseram um método para conversão entre essas duas representações por um processo iterativo de *fitting*.⁸² Há uma proposta de aplicação de GD para a obtenção de uma função de energia livre baseada em modelos de física de polímeros, dependente unicamente dos valores de κ para o enovelamento de hélices,⁸³ posteriormente expandido para considerar os valores de τ .⁸⁴ No entanto, como os próprios autores comentam, há dificuldades para validar experimentalmente as conclusões dessas funções.⁸⁴

Além dos exemplos comentados até então, conceitos de GD associada a teoria de *coiled-coils* já foram utilizados para análise conformacional envolvendo formação de hélices

por complexos cucCFBA bacterianos e oligomerização da Adiponectina.⁸⁵ A eficácia de métodos como o *Automated Protein Structure Analyses*, APSA, para análises de motivos de estrutura secundária,⁸⁰ avaliar alinhamento estrutural⁸⁶ e análise de coleções de estruturas de diferentes proteínas⁸⁷ atestam a efetividade dessa representação.

1.3 Teoria da Informação

Um outro ramo explorado neste trabalho é a aplicação de teoria da informação (TI) à representação baseada em GD de conformações. Por esta razão, nesta seção vamos apresentar e discutir brevemente alguns conceitos de TI e, em seguida, sua aplicação a sistemas biológicos.

1.3.1 Informação e incerteza

Em 1948, Claude E. Shannon publica “*A Mathematical Theory of Communication*” abordando o problema fundamental de comunicação, que é a reprodução exata ou aproximada por um receptor de um valor escolhido por um emissor, com a proposta de tratar matematicamente a informação por uma perspectiva estatística.⁸⁸ Ao considerar que aspectos semânticos da mensagem irrelevantes do ponto de vista de engenharia, Shannon propõe uma medida quantitativa para o conceito de informação associada a uma mensagem, o que tornou possível tratar e viabilizar a transmissão de uma determinada quantidade de informação por um dado canal de comunicação com um determinado nível de ruído associado. Este trabalho lançou as bases do que viria a se tornar TI. Por esta razão nesta seção discutiremos com mais detalhes as ideias lançadas por Shannon. No momento vamos nos ater apenas à formulação do conceito de informação, alguns dos conceitos associados a processos de comunicação serão discutidos posteriormente.

Shannon afirma que, do ponto de vista de engenharia, um sistema de comunicação depende de ser capaz de selecionar uma mensagem de um conjunto total de mensagens possíveis.⁸⁸ Portanto, esse sistema precisa ser capaz de selecionar cada uma do conjunto de mensagens possíveis. Para um número finito de mensagens possíveis, este número, ou uma função baseada nesse número, pode ser utilizado para estabelecer uma medida da informação obtida ao selecionar uma mensagem entre todas as possíveis. Dentre as várias funções possíveis para esta medida, Shannon aponta a função logarítmica como conveniente, pois vários parâmetros usados em problemas de engenharia tendem a variar linearmente com o logaritmo do número de possibilidades, um valor que varia linearmente é mais intuitivo, e as operações matemáticas com logaritmos são relativamente simples.⁸⁸ Além disso, considerando a probabilidade de n mensagens como $p_1, p_2, \dots, p_n$, para uma medida de informação $H(p_1, p_2, \dots, p_n)$ também seria conveniente que:

- a) H fosse contínua em p_i ;

- b) se todos os p_i forem iguais ($p_i = \frac{1}{n}$), H cresce monotonicamente em função de n ;
- c) se a escolha de uma mensagem puder ser também expressa como o resultado de escolhas sucessivas, H da mensagem original deve ser uma soma ponderada dos valores individuais de H para as etapas de escolhas (Figura 6). Por exemplo,

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3}, \frac{1}{3}\right) \quad (1.3)$$

O coeficiente $\frac{1}{2}$ é adicionado pois a segunda escolha só ocorre metade das vezes.

Figura 6 – Decomposição de eventos para escolha de três possibilidades. Se a mesma coleção de eventos pode ser decomposta como o resultado de escolhas sucessivas, a mensagem original será igual a soma ponderada dos valores das escolhas.

Fonte: SHANNON.⁸⁸

Shannon propõe que a única função que satisfaz essas conveniências é

$$H = -K \sum_{i=1}^n p_i \log_b p_i \quad (1.4)$$

, onde K é uma constante positiva e b é a base do log.⁸⁸ A unidade de H é arbitrária e depende apenas do valor de K , normalmente $\frac{1}{\ln(b)}$. Para $b = 2$, se convencionou usar o termo bit. Claramente, existe relação entre H e o conceito de entropia utilizado em mecânica estatística. Por esta razão, Shannon decide chamar essa quantidade de entropia, mas nessa mesma publicação ele se refere à H como "incerteza", várias vezes. No momento, utilizaremos o termo "entropia", posteriormente será discutido porque o termo é inadequado.

Shanno argumenta que a entropia H pode ser utilizada para definir medidas de informação, escolha e incerteza, pois possui algumas propriedades convenientes,⁸⁸ são elas:

- a) $H = 0$, se e somente se $\forall i : p_i = 0$ com exceção de apenas um estado i . Isto é, $H = 0$ somente quando é possível observar apenas um estado, caso contrário, $H > 0$.
- b) Para n observações possíveis, H é máximo e igual a $\log(n)$ somente quando todos os p_i forem iguais ($p_i = 1/n, \forall i$). Isso implica que H reflete uma noção muito intuitiva, a incerteza sobre o possível resultado de uma observação é máxima somente se todos os estados possíveis forem igualmente prováveis.

- c) H de eventos conjuntos é sempre menor ou igual a soma dos H eventos individuais e H só é igual a soma se os eventos forem independentes, isto é,

$$H(x, y) = - \sum_{i,j} p(i, j) \log(p(i, j)) \quad (1.5)$$

enquanto,

$$H(x) = - \sum_{i,j} p(i, j) \log\left(\sum_j p(i, j)\right) \quad (1.6)$$

$$H(y) = - \sum_{i,j} p(i, j) \log\left(\sum_i p(i, j)\right) \quad (1.7)$$

é possível demonstrar que,

$$H(x, y) \leq H(x) + H(y) \quad (1.8)$$

- d) Qualquer mudança nos valores de p_i na direção de torná-los mais próximos da igualdade para todos os i resultará em aumento de H .
- e) No caso de eventos condicionais, por exemplo, observar y dado x , é possível definir um H condicional baseado nas probabilidades condicionais. Para cada valor i que x pode assumir existe uma probabilidade condicional para $p_i(j)$ para quando y assume o valor j ,

$$p_i(j) = \frac{p(i, j)}{\sum_j p(i, j)} \quad (1.9)$$

Desse modo, podemos definir entropia condicional de y , $H_x(y)$, como a média da entropia da observação de j de y para cada valor i de x , ponderado de acordo com a probabilidade de obter cada valor de i , isto é,

$$H_x(y) = - \sum_{i,j} p(i, j) \log(p_i(j)) \quad (1.10)$$

Esta quantidade reflete a média de incerteza de y quando se conhece o valor de x . Ao substituir o valor de $p_i(j)$ (Equação 1.9), obtemos

$$H_x(y) = - \sum_{i,j} p(i, j) \log(p(i, j)) + \sum_{i,j} p(i, j) \log\left(\sum_j p(i, j)\right) \quad (1.11)$$

$$H_x(y) = H(x, y) - H(x) \quad (1.12)$$

logo,

$$H(x, y) = H(x) + H_x(y) \quad (1.13)$$

A Equação 1.13 reflete uma noção intuitiva sobre incertezas, isto é, a incerteza de um evento conjunto x e y é a incerteza de x mais a incerteza de y quando x é conhecido.

f) Segue de c) e e) que

$$H(x) + H(y) \geq H(x, y) = Hx + Hx(y) \quad (1.14)$$

logo,

$$H(y) \geq Hx(y) \quad (1.15)$$

Isto é, a incerteza de y nunca aumenta ao saber o valor de x .

Todas as características mencionadas são convenientes para estabelecer uma medida quantitativa de incerteza. Em seu clássico trabalho, Shannon chama H de entropia, mas em algumas passagens se refere também como “incerteza”.⁸⁸ O termo entropia pode causar confusão e ambiguidades, especialmente ao discutir temas na interface entre grandezas termodinâmica e grandezas de TI. Inclusive, ao ser questionado sobre o porquê de ter escolhido o termo entropia, Shannon respondeu: *“Minha maior preocupação era como chamá-la. Eu pensei em chamá-la ‘informação’, mas a palavra é muito utilizada, então decidi chamá-la de ‘incerteza’. Ao discutir esse tópico com John von Newman, ele teve uma ideia melhor. Von Newmann me disse, ‘você deveria chamá-la de entropia por duas razões. Em primeiro lugar, sua função de incerteza tem sido usada em mecânica estatística sob esse termo, então já possui um nome. Em segundo lugar, e mais importante, ninguém sabe o que entropia realmente é, então em um debate você sempre terá a vantagem.’”*⁸⁹ (Tradução livre)

Apesar disso, Shannon entendia a diferença entre sua medida e o conceito de entropia aplicada aos sistemas físicos. Evidência disso é que dentro do contexto de canais de comunicação ele renomeia $H_y(x)$ como “equívoco”, na página 20 de sua publicação. Nesse contexto, $H_y(x)$ é a ambiguidade média do sinal recebido em relação ao sinal enviado. Para evitar o alongamento desse tópico, apesar de informal, sugerimos ao leitor uma discussão na página do grupo de Teoria da Informação Molecular do *National Institutes of Health* (NIH), dirigido por Tom D. Schneider sobre o tema (<https://schneider.ncifcrf.gov/information.is.not.uncertainty.html>). Seguiremos as indicações de Tom Schneider sobre a utilização do termo, comentadas brevemente em,⁹⁰ para evitar possíveis ambiguidades, portanto, chamaremos H de “incerteza”.

Com base na medida de incerteza estabelecida, Shannon discute que parte da informação, R , enviada a um receptor pode ser perdida devido à presença de ruído no canal de comunicação e define R como sendo a diferença entre a incerteza da mensagem original e a incerteza da mensagem recebida, isto é,

$$R = H(x) - H_y(x) \quad (1.16)$$

Portanto, a medida de informação definida por Shannon reflete a diminuição da incerteza do receptor sobre a mensagem original ao considerar a incerteza associada à observação da mensagem recebida por um canal com ruído.⁸⁸ É importante ressaltar que R é sempre uma

diferença entre duas incertezas, normalmente, o “antes” e “depois” de uma observação ou operação, fato que é ressaltado por Shannon na página 38 do “*A Mathematical Theory of Communication*”.⁸⁸

É importante estabelecer que essa medida de informação definida por Shannon não é entropia física e não é incerteza. É possível encontrar na literatura autores considerando informação como equivalente de “incerteza”⁹¹ ou como uma medida de entropia física,⁹² um bom exemplo são os comentários de Sahotra Sarkar para a BioScience em 1996.⁹³ Sarka afirma que informação conota incerteza e que sequências de DNA conservadas teriam menos informação, pois a incerteza para essas sequências é menor.⁹³ Como Thomas D. Schneider argumenta em seu site, considerar informação semelhante a incerteza implica que quanto mais aleatório é o sistema, mais informação estaria associada a ele. No entanto, a própria noção de informação implica o oposto. Parte dessa confusão possivelmente se deve ao fato de, para o caso especial de comunicação sem ruído, $H_y(x) = 0$, portanto,

$$R = H(x) - 0 \quad (1.17)$$

Como a definição da incerteza H é formalmente igual à de entropia, é possível concluir erroneamente a partir da Equação 1.17 que “informação = incerteza = entropia”. Para evitar essa confusão, basta seguir a definição formal dada por Shannon, i. e., *informação é a diminuição da incerteza de um receptor sobre uma mensagem transmitida por um dado canal de comunicação com determinado nível de ruído*.

O impacto dos conceitos e ideias desenvolvidos por Shannon não se limitaram apenas ao campo de tecnologia da comunicação, mas também à física, química, criptografia e muitos outros campos. Naturalmente, suas implicações para problemas biológicos foram logo percebidas.

1.3.2 Teoria da Informação aplicada aos sistemas biomoleculares

Desde a determinação da estrutura do DNA, a noção de “informação” biológica como algo organizado em uma molécula foi incorporado pela comunidade científica. A discussão filosófica sobre a noção de informação em um sistema biológico ser apenas um conceito teleológico⁹⁴ ou um conceito objetivamente real ainda estar em debate,⁹⁵ tratar sistemas vivos como máquinas que processam informação já se mostrou útil em diferentes contextos. Abordar sistemas biológicos da perspectiva de teoria da informação e comunicação é algo que vem sendo discutido na literatura científica desde a década de 1950.^{96–100} Apesar de simplista, a primeira proposta de associação de informação a topologia de moléculas orgânicas em sistemas biológicos foi proposta em 1955.⁹⁶ Em 1956, Yockey propõe tratar o dano ao DNA causado por radiação em tecidos como a perda de informação, i. e., como equivalente ao “equivoco”, $H_y(x)$, proposto por Shannon.^{98,101} Somente em 1974, a primeira aplicação de TI ao estudo de sequências biológicas, no caso, de DNA, foi proposta.⁹⁹ Em 1981, conceitos de TI são utilizados para demonstrar que nenhum

dos cenários propostos para o surgimento da vida até então era plausível.¹⁰⁰ Podemos citar o método proposto por Thomas D. Schneider em 1986¹⁰² como a primeira aplicação desses conceitos em estratégias de análises a identificação de sítios de ligação proteína-DNA baseados unicamente em medidas de informação de sequências. Nessa mesma publicação é proposta uma medida da informação contida em uma sequência, $R_{sequência}$, e quanto de informação é necessário para encontrar aquela sequência em um dado genoma, $R_{frequência}$. O raciocínio por trás dessas medidas é relativamente simples. Uma proteína ao se ligar a um trecho de uma molécula de DNA é ignorante quanto ao seu conteúdo, portanto, possui “incerteza” máxima antes da ligação. Ao observar uma determinada sequência, essa incerteza diminui. Logo,

$$R_{sequência}(l) = H_{antes} - H_{depois}(bits/base) \quad (1.18)$$

Considerando que o DNA é composto por quatro bases (A, C, T e G), logo, $H(x)$ máxima é igual a 2 bits por posição; a informação ao identificar a base na posição l , $R_{sequência}(l)$, será:

$$R_{sequência}(l) = 2 - H_{depois}(bits/base) \quad (1.19)$$

Uma proteína que reconhece uma sequência específica de DNA, normalmente, é capaz de reconhecer variações dessas sequências. Portanto, ao reconhecer um determinado sítio de ligação, existe uma certa incerteza em relação a qual dos diferentes sítios possíveis foi observado. Portanto, H_{depois} será baseado nas probabilidades de bases possíveis para cada posição. A variação em uma posição desses sítios geralmente é independente de variações nas demais posições,¹⁰³ o que permite definir a informação associada a um sítio como a soma da informação individual associada a cada base,

$$R_{sequência} = \sum_l R_{sequência}(l)(bits/sítio) \quad (1.20)$$

Baseado nessas propostas foi definido o “logos” de uma sequência,^{102,104} popularmente utilizado em análises de sequências de DNA [Figura 7](#). Usando essa estratégia foi possível identificar 16 genes regulados pela p53, dos quais 15 foram confirmados experimentalmente.¹⁰⁵

Figura 7 – Logo do sítio de ligação reconhecidos pela Fis de *Escherichia coli*. Na parte superior da figura estão ilustrado 12 sítios e seus respectivos valores de $R_{sequência}$. Na parte inferior, o logo que reflete a quantidade de informação para cada base do sítio, $R_{sequência}(l)$.

Fonte: Schneider.¹⁰⁶

Uma outra grandeza definida por Schneider é a $R_{frequência}$, uma medida da quantidade de informação necessária para encontrar Γ sítios de ligação em um genoma de tamanho G , definido segundo:

$$R_{frequência} = \log_2 G - \log_2 \Gamma = -\log_2 \Gamma / G (\text{bits/sítio}) \quad (1.21)$$

E sítios de ligação naturais tendem a apresentar valores $R_{sequência}$ próximos dos valores de $R_{frequência}$.^{102,107} Isto é, sítios tendem a apresentar valores de informação próximos da quantidade de informação necessária para encontrar esses sítios no genoma.

Implicações mais profundas da aplicação de teoria da informação e comunicação para comportamento físico de moléculas também foram exploradas por Schneider em dois trabalhos publicados em 1991 intitulados “*The Theory of Molecular Machines*”.^{108,109} Dentre elas podemos citar as relações entre informação e energia de ligação, definição de “capacidade de máquina” baseado no conceito de “capacidade de canal” e várias

considerações sobre mudança de estados de uma molécula e os paralelos com o processo de codificação e decodificação de mensagens representadas em um espaço hiperdimensional, de forma equivalente à discutida por Shannon. Essas relações foram exploradas no contexto ligação proteína-DNA e, baseado no conceito de “capacidade de máquina”, permitiram identificar e explicar limites evolucionários relacionados à eficiência de ligação desses sistemas.^{90,106} Entre os vários tópicos discutidos nessas duas publicações, no momento, vamos apresentar brevemente apenas a relação entre energia, incerteza e informação proposta por Schneider.¹⁰⁸

Considere, a entropia de Boltzmann-Gibbs que é definida como:

$$S = k_B \sum_{i=1}^{\omega} p_i \ln(p_i) (\text{joules}/K * \text{microestado}) \quad (1.22)$$

onde, k_B é a constante de Boltzmann em J/K, ω é o número total de microestados e p_i é a probabilidade do microestado i . Se considerarmos os microestados da [Equação 1.22](#) como os mesmos estados discutidos para a [Equação 1.4](#) e $\log_2(x) = \frac{\ln(x)}{\ln(2)}$ teremos

$$S = k_B \ln(2) \left[\sum_{i=1}^{\omega} p_i \ln(p_i) \right] \quad (1.23)$$

$$S = k_B \ln(2) H \quad (1.24)$$

Após uma máquina molecular realizar uma operação a mudança de entropia será

$$\Delta S = S_{\text{depois}} - S_{\text{antes}} (J/K * \text{operação}) \quad (1.25)$$

e a redução da incerteza após essa operação será

$$\Delta H = H_{\text{depois}} - H_{\text{antes}} (\text{bits}/\text{operação}) \quad (1.26)$$

Substituindo a [Equação 1.24](#) na [Equação 1.25](#), obteremos

$$\Delta S = k_B \ln(2) H_{\text{depois}} - k_B \ln(2) H_{\text{antes}} \quad (1.27)$$

$$\Delta S = k_B \ln(2) \Delta H \quad (1.28)$$

Por definição, ao reduzir a incerteza de uma máquina após uma operação ela ganhará informação R

$$R \equiv -\Delta H (\text{bits}/\text{operação}) \quad (1.29)$$

Substituindo [Equação 1.29](#) na [Equação 1.28](#),

$$\Delta S = -k_B \ln(2) R \quad (1.30)$$

A [Equação 1.30](#) estabelece uma relação quantitativa entre a informação e a redução de entropia associada a operação de uma máquina molecular. Utilizando essa relação é possível estimar a energia mínima necessária para ganhar uma determinada quantidade de informação em bits. Considere a desigualdade de Clausius:

$$dS \geq dq/T \quad (1.31)$$

onde dq é o calor que entra no pequeno volume no qual a máquina molecular opera e T é a temperatura absoluta em K . Normalmente, máquinas moleculares operam a temperatura constante, então podemos integrar [Equação 1.31](#) para obter:

$$\Delta S \geq q/T \quad (1.32)$$

onde q é calor total que entra no volume. Substituindo [Equação 1.30](#) na [Equação 1.32](#),

$$-k_B \ln(2)R \geq q/T \quad (1.33)$$

$$k_B \ln(2)R \leq -q/T \quad (1.34)$$

$$k_B T \ln(2) \leq -q/R (J/bit) \quad (1.35)$$

A [Equação 1.35](#) estabelece a quantidade mínima de energia, ϵ_{min} , que precisa ser dissipada ($-q$) para obter 1 bit de informação, isto é:

$$\epsilon_{min} = k_B T \ln(2) (J/bit) \quad (1.36)$$

ϵ_{min} é um limite mínimo e mais energia poderia ser gasta para obter 1 bit de informação, neste caso a energia extra poderia ser considerada como desperdício. Usando a [Equação 1.36](#) é possível definir um fator de conversão entre energia e informação. Naturalmente, é possível utilizar ϵ_{min} como uma medida de eficiência ao comparar com a energia de fato utilizada pelo sistema. Schneider identificou que um limite evolucionário de 70% de eficiência para reações isotérmicas é relativamente comum na natureza e propõe uma possível explicação baseada no conceito de “capacidade de máquina”.¹⁰⁶

Apesar dos avanços e recentes aplicações nos mais variados campos da biologia,¹¹⁰⁻¹¹⁴ utilização de TI para análise de sistemas biomoleculares ainda está, de modo geral, limitada a análises de sequências, especialmente DNA e RNA. Considerando que é possível aplicar diretamente os conceitos de TI em sistemas que podem ser representados por um conjunto finito de estados bem definidos e mutualmente excludentes, não é uma surpresa que a aplicação de TI às sequências de DNA e RNA seja mais explorada. Parte disso se deve ao fato que essas sequências apresentam algumas conveniências do ponto de vista de análises baseada em TI. As principais conveniências são: 1) é possível representar toda a

complexidade física das bases dessa molécula utilizando quatro estados e 2) é relativamente fácil e acessível identificar experimentalmente cada base de DNA através de técnicas de sequenciamento e 3) essas bases estão dispostas linearmente. Consequentemente, a aplicação de TI já é incorporada nos protocolos de análise de genomas.^{104,115–117} Sequências de aminoácidos apresentam vantagens semelhantes às citadas anteriormente e aplicações desses mesmos conceitos também foram logo desenvolvidas e incorporadas em protocolos de análise de famílias de proteína para explorar informações associadas à evolução de proteínas.^{104,113,118–125} Mais recentemente, abordagens baseadas em TI para funções de *scoring* para *docking* vêm sendo exploradas.¹²⁶

Apesar de pouco exploradas, existem aplicações baseadas em TI no contexto de análise conformacional. Por exemplo, em 2008, tratar interação entre domínios como uma instância de comunicação em um canal com ruído se mostrou útil para explicar aspectos alostéricos envolvendo os domínios SH2 e SH3 da Fyn.¹²⁷ O conceito de informação mútua foi utilizado para medir a relação entre cadeias laterais através do estabelecimento de um alfabeto para representar um número de estados finitos de cadeias laterais, determinados por uma biblioteca de rotâmeros.¹²⁷ Em 2014, um método baseado em teoria da informação para análise de mecanismos de alosteria usando coordenadas atômicas foi proposto para identificar resíduos envolvidos na comunicação entre sítios e atualmente é o único método disponível para esse tipo de análise que considera movimentos correlacionados em nível de *N*-corpos e não apenas “*pair-wise*”.¹²⁸ Essas aplicações demonstram que o uso de conceitos e ideias de teoria da informação e comunicação para análise conformacional de proteínas já se mostrou útil.

2 OBJETIVOS

2.1 Objetivo geral

Desenvolver um método baseado descritores de geometria diferencial e teoria da informação para análise conformacional de proteínas.

2.2 Objetivos específicos

- a) Implementar um programa para realizar medidas de curvatura e torção em ensembles conformacionais;
- b) Explorar o uso de descritores de GD para identificação e quantificação de aspectos estruturais relacionados a flexibilidade de proteínas;
- c) Explorar o uso de descritores de GD para identificação de estados conformacionais;
- d) Desenvolver medidas de quantidade de informação em *ensembles* conformacionais baseadas na representação por geometria diferencial;
- e) Explorar o uso de medidas de incerteza e informação para obtenção de características estruturais relacionadas a *ensembles* conformacionais de proteínas.
- f) Avaliar a eficácia de métodos desenvolvidos e comparar com outros métodos já estabelecidos na análise de *ensembles* conformacionais do domínio de ligação ao DNA da E1, das etapas de ligação do complexo cMib-KIX e a Ubiquitina.

3 MÉTODOS E METODOLOGIA

O método desenvolvido neste trabalho consiste em 1) estabelecer uma representação de conformações de proteínas baseada em GD e utilizá-la para explorar características estruturais observadas em *ensembles* conformacionais (seção 3.1), 2) estabelecer medidas de informação ligadas a *ensembles* (seção 3.4), 3) aplicar o método a diferentes sistemas e comparar os resultados obtidos com técnicas popularmente utilizadas (seção 3.5). Nesta seção todos os aspectos metodológicos e seus fundamentos serão apresentados e discutidos.

3.1 Cálculo dos descritores de geometria diferencial

Uma representação de estruturas de proteínas que pudesse ser caracterizada com a menor quantidade de informação possível e permitisse ignorar movimentos de corpos rígidos seria muito conveniente para lidar com o problema da flexibilidade em proteínas. Felizmente, descrever objetos geométricos dessa forma é uma prática comum em geometria. Por exemplo, é possível comparar circunferências independente de suas posições no espaço utilizando apenas o seu raio. De forma equivalente, de acordo com o *teorema fundamental de curvas espaciais* (TFCE), curvas regulares podem ser precisamente descritas como uma função dos valores de curvatura, κ , e torção, τ , em função do comprimento de arco, s . Segundo o TFCE, se duas curvas apresentam os mesmos valores de curvatura e torção elas serão iguais e só diferem por movimentos de corpos rígidos. Apesar de sua prova matemática não ser trivial (Apêndice C), o TFCE expressa uma noção intuitiva para nossa experiência diária. Imagine dois fios (A e B) semirrígidos e retilíneos, é possível produzir apenas dois tipos de distorção na forma desses fios, curvá-los ou torcê-los em diferentes pontos. Se as mesmas distorções forem realizadas nos fios A e B, ambos terão exatamente a mesma forma final independente de sua localização no espaço. Logo, se for possível estabelecer uma curva r que represente adequadamente as conformações de um *ensemble* de proteínas, os valores de κ e τ do conjunto poderão ser utilizados para analisar os movimentos observados no *ensemble* sem a necessidade de sobreposição das conformações.

Para explorar a representação por geometria diferencial, o primeiro desafio é estabelecer uma curva que represente a estrutura da proteína e calcular os valores de comprimento de arco, curvatura e torção que caracterizem essa curva. Neste trabalho utilizamos essencialmente a mesma estratégia implementada no ARABESQUE,⁸⁶ que é uma evolução da utilizada no CHORAL,⁷⁹ para gerar essa curva. O método consiste em 1) gerar curvas paramétricas em um espaço euclidiano tridimensional, E^3 , por interpolação *spline* usando coordenadas de $C\alpha$ como nós, e 2) calcular os valores de curvatura e torção em função do comprimento de arco para a posição de cada $C\alpha$ na curva paramétrica. Devido

ao aspecto central dessa representação para este trabalho, se faz necessário apresentar ao leitor mais detalhadamente os diferentes aspectos da metodologia implementada.

3.1.1 Representação do *backbone* de estruturas proteicas por curvas paramétricas

A representação de estruturas de proteínas por uma curva que possui a mesma trajetória do *backbone* da proteína é informativa e captura visualmente a posição de hélices, folhas β e *loops* e precisamos definir requerimentos para que uma curva seja considerada apropriada. Portanto, precisamos de uma definição quantitativa dessa curva antes de determinar como ela será calculada. Como um conceito matemático, uma curva r pode ser definida como uma função contínua que mapeia um intervalo I de números reais em pontos no espaço E^3 ($r : I \rightarrow I$). Essa definição permite que muitas curvas sejam definidas, mas somente as que podem ser utilizadas para representar estruturas de proteínas nos interessam. Precisamos definir uma função paramétrica, $\vec{r}(t)$, que define uma curva r baseada em um parâmetro t de uma estrutura proteica. Para a nossa aplicação, a solução mais intuitiva para t é o número dos resíduos da proteína, e não a posição em um alinhamento como ocorre no CHORAL.⁷⁹ É conveniente que a curva utilizada para representar o *backbone* possua duas propriedades, ser diferenciável até a terceira ordem e nunca parar ou reverter sua trajetória em um dado intervalo. Podemos garantir essas duas propriedades se:

- a) $\vec{r}(t)$ é da classe C^3 em I
- b) $\frac{d\vec{r}}{dt} \neq 0, \forall t \in I$

Agora podemos definir uma curva regular por meio de uma base vetorial utilizando uma equação paramétrica:

$$\vec{r}(t) = x(t)\hat{x} + y(t)\hat{y} + z(t)\hat{z} \quad (3.1)$$

Uma alternativa para obter essa curva e atender as propriedades desejadas é realizar uma interpolação *spline* cúbica utilizando as coordenadas dos carbonos α como “nós”, como implementada no ARABESQUE.⁸⁶ *Splines* são funções polinomiais que podem ser utilizada para *fitting* de funções arbitrárias. De modo que podemos expressar cada coordenada individual de $\vec{r}(t)$ (Equação 3.1) como uma função *spline*. Assim é possível obter pontos que formam uma curva tridimensional suave que passa por todos os $C\alpha$. A partir dessa curva representativa do *backbone*, podemos explorar suas características geométricas.

Para tirar proveito do TFCE, precisamos estabelecer uma equivalência entre a nossa curva em função do comprimento de arco s e a curva paramétrica $\vec{r}(t)$ (Equação 3.1). Assim será possível caracterizar o *backbone* de uma estrutura de proteína utilizando os valores de κ e τ calculados a partir de $\vec{r}(t)$. Para estabelecer essa equivalência podemos explorar a propriedade de permissibilidade de troca de parâmetros (ou reparametrização) que curvas

paramétricas possuem. Considere a curva $\vec{r}(t) : I \rightarrow R^3$ e uma função diferenciável $h(s) = t$. A curva gerada por uma função composta $\beta = \vec{r}(s) \circ h(t)$ seria uma reparametrização de $\vec{r}$ por h (Figura 8).

Figura 8 – Conceito de reparametrização de curvas paramétricas. A representação $\vec{r}(s)$ é equivalente a $\vec{r}(h(t))$

Fonte: Elaborada pelo autor.

Podemos definir s em função de t da seguinte maneira, uma medida do comprimento de arco entre os pontos $\vec{r}(t)$ e $\vec{r}(t + \delta t)$ será aproximadamente.

$$s(t, t + \delta t) \approx \|\vec{r}(t + \delta t) - \vec{r}(t)\| \approx \left\| \frac{d\vec{r}}{dt} \right\| \delta t \quad (3.2)$$

Dividindo a curva r em vários segmentos correspondentes a um pequeno aumento δt , quando $\delta \rightarrow 0$, obteremos o comprimento exato dessa curva. Portanto, o comprimento de arco s entre t_0 e t pode ser definido como

$$s = s(t) = \int_{t_0}^t \left\| \frac{d\vec{r}}{dt} \right\| dt \quad (3.3)$$

A Equação 3.3 permite a troca de parâmetros da seguinte maneira:

$$\frac{ds}{dt} = \frac{d \int_{t_0}^t \left\| \frac{d\vec{r}}{dt} \right\| dt}{dt} \quad (3.4)$$

$$\frac{ds}{dt} = \left\| \frac{d\vec{r}}{dt} \right\| \quad (3.5)$$

Como $\left\| \frac{d\vec{r}}{dt} \right\| \neq 0$, $\vec{r}(s(t))$ é uma representação válida para nossa curva regular $\vec{r}(t)$. Observe que a curva $\vec{r}$ só refletirá características do *backbone* da proteína. Portanto, uma das limitações do método implementado neste trabalho é que informações associadas ao movimento de cadeias laterais que não afetem a posição dos $C\alpha$ não serão refletidas nos descritores calculados a partir dessas curvas.

3.1.2 Cálculo de curvatura e torção

Baseado na curva paramétrica discutida na [subseção 3.1.1](#), podemos definir as medidas de curvatura, κ , e torção, τ . Essas são medidas intrínsecas da forma que a curva possui, portanto precisamos de um *frame* de coordenadas que seja invariante quanto a movimentos euclidianos e que se adapte à forma da curva. Apesar de existirem outras soluções possíveis,¹²⁹ a solução clássica para esse problema é a utilização do *Frenet-Serret frame* e consiste em definir três vetores ortonormais (vetor unitário tangente $\vec{T}$, vetor unitário principal $\vec{N}$ e vetor unitário binormal $\vec{B}$) para cada ponto da curva ([Figura 9](#)).

Figura 9 – *Frenet-Serret frame* composto por três vetores (vetor tangente, $\vec{T}$, vetor normal, $\vec{N}$, e vetor binormal, $\vec{B}$).

Fonte: Elaborada pelo autor.

Podemos começar definindo o vetor unitário tangente $\vec{T}$ como

$$\vec{T} = \frac{d\vec{r}}{ds} \quad (3.6)$$

Ao aplicar a regra da cadeia e considerando a [Equação 3.5](#) é possível definir $\vec{T}$ de acordo com um parâmetro arbitrário t :

$$\vec{T} = \frac{d\vec{r}}{dt} \frac{dt}{ds} \quad (3.7)$$

$$\vec{T} = \frac{d\vec{r}}{dt} / \frac{ds}{dt} \quad (3.8)$$

$$\vec{T} = \frac{d\vec{r}}{dt} / \left\| \frac{d\vec{r}}{dt} \right\| \quad (3.9)$$

A [Equação 3.9](#) é equivalente a normalizar o vetor tangente usando o parâmetro t . Se a curva analisada fosse a trajetória de uma partícula, $\vec{T}$ seria o vetor velocidade dessa partícula. Antes de prosseguir com a definição dos demais vetores, baseado em $\vec{T}$ já podemos definir uma medida da curvatura κ para cada ponto da curva. No caso da

trajetória, essa medida seria o equivalente a magnitude do vetor aceleração dessa partícula. Podemos definir o vetor curvatura, $\vec{\kappa}$, como

$$\vec{\kappa} = \vec{\kappa}(s) = \frac{d\vec{T}}{ds} \quad (3.10)$$

Consequentemente, a quantidade escalar κ em função de s será:

$$\kappa = \|\vec{\kappa}(s)\| = \left\| \frac{d\vec{T}}{ds} \right\| \quad (3.11)$$

É conveniente expressar esse valor em função de t e é possível demonstrar ([Apêndice A](#)) que:

$$\kappa = \frac{\|\dot{\vec{r}} \times \ddot{\vec{r}}\|}{\|\dot{\vec{r}}\|^3} \quad (3.12)$$

Observe que o valor de κ é equivalente à magnitude da mudança na direção de $\vec{T}$ de acordo com o comprimento de arco s e reflete o quanto a curva difere de uma linha reta em um determinado ponto ([Figura 10](#)). De modo similar, precisamos definir uma medida τ que reflita o quanto a curva desvia de um plano em um determinado ponto. Para isso, precisamos definir mais dois vetores ortogonais a $\vec{T}$, de modo que seja possível definir um plano e monitorar a variação de um vetor normal a esse plano ao longo da curva. Vamos começar definindo o vetor principal unitário $\vec{N}$, considerando que se dois vetores são ortogonais entre si, eles possuem produto escalar igual a zero ($\vec{a} \cdot \vec{a}^\perp = 0$). Podemos demonstrar que a derivada de $\vec{T}$ resulta em um vetor perpendicular a $\vec{T}$. O produto escalar de um vetor unitário com ele mesmo será

$$\vec{T} \cdot \vec{T} = \|\vec{T}\|^2 = 1^2 = 1 \quad (3.13)$$

Derivando ambos os lados,

$$\frac{d(\vec{T} \cdot \vec{T})}{ds} = \frac{d(1)}{ds} \quad (3.14)$$

$$\vec{T} \cdot \frac{d\vec{T}}{ds} + \frac{d\vec{T}}{ds} \cdot \vec{T} = 0 \quad (3.15)$$

$$2\left(\vec{T} \cdot \frac{d\vec{T}}{ds}\right) = 0 \quad (3.16)$$

Logo,

$$\vec{T} \cdot \frac{d\vec{T}}{ds} = 0 \quad (3.17)$$

Como desejamos que $\vec{N}$ seja um vetor unitário, então podemos defini-lo como:

$$\vec{N} = \frac{\frac{d\vec{T}}{ds}}{\left\| \frac{d\vec{T}}{ds} \right\|} \quad (3.18)$$

Considerando a [Equação 3.10](#),

$$\vec{N} = \frac{\vec{\kappa}(s)}{\|\vec{\kappa}(s)\|} = \frac{1}{\kappa} \frac{d\vec{T}}{ds} \quad (3.19)$$

Podemos obter o nosso terceiro vetor $\vec{B}$ ortogonal à $\vec{T}$ e $\vec{N}$ a partir do produto vetorial entre $\vec{T}$ e $\vec{N}$. Portanto, o vetor unitário binormal $\vec{B}$ pode ser definido como

$$\vec{B} = \vec{T} \times \vec{N} \quad (3.20)$$

De forma similar a definição de κ , agora podemos definir nossa medida de torção τ como a magnitude do vetor $\vec{\tau}$ como

$$\tau = \|\vec{\tau}\| = \left\| \frac{d\vec{B}}{ds} \right\| \quad (3.21)$$

Um aspecto interessante de $\vec{\tau}$ é ser sempre paralelo ao $\vec{N}$. Pela [Equação 3.20](#), observe que

$$\frac{d\vec{B}}{ds} = \frac{d\vec{T}}{ds} \times \vec{N} + \vec{T} \times \frac{d\vec{N}}{ds} \quad (3.22)$$

Rearranjando a [Equação 3.19](#), temos que

$$\frac{d\vec{B}}{ds} = \kappa \vec{T} \times \vec{N} + \vec{T} \times \frac{d\vec{N}}{ds} \quad (3.23)$$

$$\frac{d\vec{B}}{ds} = \kappa(\vec{T} \times \vec{N}) + \vec{T} \times \frac{d\vec{N}}{ds} \quad (3.24)$$

Como o $\vec{T} \times \vec{T} = 0$,

$$\frac{d\vec{B}}{ds} = \vec{T} \times \frac{d\vec{N}}{ds} \quad (3.25)$$

A [Equação 3.25](#) permite concluir que $\frac{d\vec{B}}{ds}$ é necessariamente ortogonal a $\vec{T}$. Consequentemente, $\vec{\tau}$ (definido em [Equação 3.21](#)) precisa ser paralelo a $\vec{N}$. A partir desse fato é possível demonstrar ([Apêndice B](#)) que:

$$\tau = \frac{(\dot{\vec{r}} \times \ddot{\vec{r}}) \cdot \ddot{\vec{r}}}{\|\dot{\vec{r}} \times \ddot{\vec{r}}\|^2} \quad (3.26)$$

τ reflete o quanto a curva em um determinado ponto desvia de um plano ([Figura 10](#)). A partir dos valores de κ e τ das curvas que representam o *backbone* das conformações podemos expressar quantitativamente distorções locais. Como κ e τ são absolutos no espaço, não há a necessidade de sobreposição das estruturas, o que facilita sua aplicação para casos extremos de flexibilidade.

Figura 10 – Significado dos valores de curvatura κ e torção τ e sua relação com os vetores $\vec{T}$, $\vec{N}$ e $\vec{B}$.

Fonte: Adaptada de PITT; MONTALVÃO; BLUNDELL.⁸⁷

3.1.3 Representação de geometria diferencial de ensembles conformacionais de proteínas

Na nossa representação discutida na [subseção 3.1.2](#) podemos atribuir valores de κ e τ para cada ponto da curva que seja equivalente à posição dos α na estrutura. Assim podemos representar as conformações em um espaço multidimensional definido pelo número do resíduo n , κ , τ e s . Portanto, cada resíduo n de uma conformação será representado por um vetor $\vec{x}_n$ de quatro dimensões (n , κ , τ e s). Consequentemente, cada conformação c será representado como um conjunto X_c de pontos x_n , isto é:

$$\vec{x}_n = \{n, \kappa, \tau, s\} \quad (3.27)$$

$$X_c = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\} \quad (3.28)$$

Logo, um ensemble conformacional E será definido como

$$E = \{X_1, X_2, \dots, X_c\} \quad (3.29)$$

3.2 Métrica de distância e medidas de variabilidade estrutural local

Para poder comparar diferentes conformações e avaliar a variabilidade estrutural em um *ensemble* precisamos estabelecer uma métrica de dissimilaridade $d(\vec{x}_i, \vec{x}_j)$ entre os vetores contidos em E de modo que

$$d : X \times X \rightarrow R \quad (3.30)$$

$$\exists d_0 \in R : -\infty < d_0 \leq d(\vec{x}_i, \vec{x}_j) < +\infty, \forall \vec{x}_i, \vec{x}_j \in E \quad (3.31)$$

$$d(\vec{x}_i, \vec{x}_i) = d_0, \forall \vec{x}_i \in E \quad (3.32)$$

$$d(\vec{x}_i, \vec{x}_j) = d_0 \iff \vec{x}_i = \vec{x}_j \quad (3.33)$$

$$d(\vec{x}_i, \vec{x}_m) \leq d(\vec{x}_i, \vec{x}_j) + d(\vec{x}_j, \vec{x}_m), \forall \vec{x}_i, \vec{x}_j, \vec{x}_m \in E \quad (3.34)$$

Obedecendo todos esses requisitos, d será uma métrica apropriada.¹³⁰ Observe que a Equação 3.34 é a desigualdade triangular discutida na seção 1.2. Uma métrica que obedece a todos esses requisitos é a distância Euclideana, d^2 ,¹³⁰ então podemos utilizá-la como métrica de dissimilaridade entre os $\vec{x}_i$ e $\vec{x}_j$, de modo que

$$d^2(\vec{x}_i, \vec{x}_j) = \sqrt{(\kappa - \kappa)^2 + (\tau - \tau)^2 + (s - s)^2} \quad (3.35)$$

Para d^2 , $d_0 = 0$, logo $d^2(\vec{x}_i, \vec{x}_j) \rightarrow [0, \infty)$. Os valores da numeração dos resíduos serão ignorados para o cálculo de $d^2(\vec{x}_i, \vec{x}_j)$. Quando uma das dimensões do vetor assume valores em uma escala significativamente maior que as demais, a distância nesta dimensão dominará a métrica de dissimilaridade. Para evitar que isso ocorra, os valores de κ , τ e s foram normalizados de modo que todos os valores fossem expressos entre 0 e 1. O reescalonamento dos descritores foi realizado seguindo

$$u'_i = \frac{u_i - u_{min}}{u_{max} - u_{min}} \quad (3.36)$$

onde u_i é o valor do dado i , u'_i é o valor de i reescalado, u_{min} é o menor valor observado no conjunto de dados e u_{max} é o maior valor observado no conjunto de dados. Após o reescalonamento, podemos utilizar d^2 para comparar os vetores $\vec{x}_n$.

3.2.1 Variabilidade estrutural local

Uma informação importante que poderia ser obtida a partir desses *ensembles* é o quão flexível um determinado resíduo é. Observar os valores de κ , τ e s ou a distribuição de medidas de distância entre as conformações observadas é útil para julgar este aspecto. Por esta razão, gráficos para visualizar os valores de cada um dos descritores e suas distribuições foram gerados. Além disso, outra informação útil para julgar a flexibilidade de um resíduo é a distância (Equação 3.35) entre uma conformação de referência, ref , arbitrária e todas as demais conformações $c \in E$. Essas distâncias foram calculadas e gráficos da distribuição desses valores foram produzidos para os conjuntos testes (seção 3.5) utilizados para ilustrar a utilidade dessa abordagem. A princípio qualquer conformação ref conveniente para o problema abordado pode ser utilizada, como por exemplo, a conformação inicial de uma trajetória obtida por simulação ou uma estrutura cristalográfica representando um determinado estado. A distribuição das distâncias observada refletirá a

variabilidade estrutural de um resíduo i no *ensemble* E . Várias medidas que permitem caracterizar uma distribuição (por exemplo, média e variância) poderiam ser calculadas baseadas nos descritores, mas todas assumem um determinado tipo de distribuição que não necessariamente será a observada, por isso evitamos utilizar tais descritores.

Para fins de comparação, *Root Mean Square Fluctuation*, $RMSF$, também foi calculado. O $RMSF$ é uma métrica de variabilidade local popular definida como

$$RMSF_{C\alpha,n} = \sqrt{\frac{1}{T} \sum (\vec{p}_{C\alpha,n,c} - \langle \vec{p}_{C\alpha,n} \rangle)^2} \quad (3.37)$$

onde T é o número total de conformações, $\vec{p}_{C\alpha,n,c}$ é o vetor posição do $C\alpha$ do resíduo n da conformação c e $\langle \vec{p}_{C\alpha,n} \rangle$ é a média de $\vec{p}_{C\alpha}$ em T .

3.2.2 Identificação de resíduos flexíveis

Utilizando a estratégia discutida na [subseção 3.2.1](#) é possível identificar resíduos manualmente, mas para proteínas com muitos resíduos é conveniente ter um método para a identificação automática desses resíduos. O método implementado consiste em 1) identificar os resíduo com maior variação de posição e 2) estabelecer um limite a partir do qual um resíduo flexível será selecionado. Devido à possibilidade da ocorrência de *outliers*, um histograma inicialmente é gerado, cujo o tamanho de banda é definido após um processo iterativo de otimização.¹³¹ Pontos pertencentes às bandas da extrema direita e extrema esquerda que possuem menos de 1% das conformações foram considerados *outliers*. Após desconsiderar os *outliers*, a distância máxima para cada resíduo foi calculada usando apenas os valores máximos de κ e τ , a partir da [Equação 3.35](#) e desconsiderando s . Baseado na distribuição de distância máxima para todos os resíduos, uma função de Z-score foi definida e um valor z foi atribuído à distância de cada resíduo.

$$z = \frac{d_{max,i} - \langle d_{max} \rangle}{\sigma} \quad (3.38)$$

onde z é o valor de z-score, $d_{max,i}$ é a distância máxima do resíduo i , $\langle d_{max} \rangle$ e σ é o desvio padrão. Um limite arbitrário de z-score pode ser definido para considerar ou não um resíduo como flexível. Para dar mais flexibilidade a escolha dos resíduos, uma representação com múltiplos níveis de resoluções é construída por uma transformação Haar *wavelet*.

3.3 Identificação de estados conformacionais

Uma das informações úteis de se obter a partir de um *ensemble* conformacional é quais estados estáveis estão presentes e o quão populados eles são. No contexto do funil de enovelamento, um estado conformacional estável seria equivalente a um poço no fundo do funil. As proteínas que visitarem esses mínimos do fundo do funil tenderão a

permanecerem neles por mais tempo em comparação aos estados de transição entre mínimos. Portanto, esse comportamento se refletirá na densidade das coordenadas observadas no *ensemble*. Desse modo, identificar estados conformacionais é essencialmente um problema de *clustering*. Vários algoritmos de *clustering* podem ser utilizados, dentre eles optamos por utilizar algoritmos que necessitassem o mínimo possível de parâmetros de entrada e retornassem bons resultados. Os valores de κ e τ foram utilizados para realizar o *clustering* local utilizando o algoritmo de agrupamento espacial hierárquico baseado em densidade com ruído (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*, HBDSCAN).¹³² O HBDSCAN consiste em quatro etapas, 1) transformação do espaço de acordo com a relação densidade e dispersibilidade, 2) obtenção de uma árvore de extensão mínima das distâncias ponderadas do grafo, 3) construção de solução agrupamento hierárquico dos componentes conectados, 4) Condensar a árvore de extensão mínima baseado no valor de tamanho mínimo de *cluster* e 5) extrair *clusters* estáveis da árvore condensada. .

3.4 Teoria da Informação

Um dos objetivos deste trabalho é definir medidas de quantidade de informação e incerteza para *ensembles* conformacionais e explorar o uso dos conceitos discutido na [seção 1.3](#), já aplicados com sucesso em sequências de DNA e proteínas. A estratégia implementada neste trabalho consiste em 1) codificar a informação espacial dos descritores de κ e τ utilizando um sistema finito de símbolos, 2) representar a informação espacial na forma de sequências e 3) definir medidas de informação baseadas nessas sequências.

3.4.1 Codificação dos descritores de geometria diferencial

Não existe método padrão para codificar um conjunto de descritores quantitativos e contínuos usando um determinado número de símbolos. Portanto, nos baseamos no método implementado por Peter L. Chang para codificar valores de *writhing number*.¹³³ O método se baseia no princípio de que o sistema de codificação mais informativo é o que permite que todos os estados codificados possuam a mesma probabilidade de aparição em uma dada sequência de estados. Além de compatível com os critérios de informação definidos por Shannon,⁸⁸ esse princípio é intuitivo, pois, se um receptor de uma mensagem tem certeza da aparição da letra ‘a’ em um dado ponto da mensagem, não há ganho de informação para o receptor. Se todas as letras são igualmente prováveis será mais difícil prever o conteúdo da mensagem, conseqüentemente, toda mensagem observada resultará em ganho de informação para o receptor.

Para definir um sistema de codificação para um determinado *ensemble* conformacional respeitando o princípio de informação máxima, primeiro um histograma para os valores de κ e τ observados é obtido. O histograma é particionado em L intervalos com

áreas iguais e uma letra do alfabeto é atribuída para cada intervalo. A princípio o valor de L é arbitrário e utilizamos $L = 20$ nesta aplicação. Outros valores de L foram testados nos conjuntos utilizados neste trabalho (dados não mostrados) e as conclusões obtidas foram essencialmente as mesmas para $L \geq 14$. Apesar disso, uma melhor justificativa e embasamento teórico para a determinação deste valor precisa ser elaborada. Como as áreas para cada intervalo representado por uma letra são iguais, todas as letras possuem a mesma probabilidade de serem designadas a um determinado resíduo.

Figura 11 – Esquema do processo de codificação dos descritores de geometria diferencial em símbolos. Um alfabeto é definido para representar intervalos de κ e τ a partir da repartição de intervalos que possuam a mesma área no histograma. A partir desse alfabeto é possível expressar os valores de κ e τ de cada resíduo por símbolos e, portanto, cada conformação i será representada por duas sequências de símbolos.

Fonte: Elaborada pelo autor.

A partir dos alfabetos gerados para κ e τ podemos expressar cada conformação como uma combinação de sequências formadas por essas letras. Assim, cada resíduo é representado por dois símbolos e cada conformação pela combinação dessas duas sequências (Figura 11).

3.4.2 $H(n)$, R_{res} e R_{conf}

A partir das sequências podemos estabelecer medidas de informação e incerteza, análogas à discutidas na subseção 1.3.2, para conformações e resíduos de proteínas. A medida de incerteza para o estado de um resíduo n , $H(n)$, será essencialmente a mesma definida na Equação 1.4, isto é,

$$H(n) = -K \sum_{i=1}^M p_{i,n} \log_2 p_{i,n} \quad (3.39)$$

onde, $p_{i,n}$ é a probabilidade de observar a letra i na posição n e M é o número total de estados possíveis. De forma análoga a $R_{sequência}(l)$ (Equação 1.18), a partir da Equação 3.39 podemos definir uma medida que reflete o quanto de informação, $R_{res}(n)$, está associada à observação de um resíduo n no estado i , pois

$$R_{res}(n) = H_{antes}(n) - H_{depois}(n) \text{ (bits/resíduo)} \quad (3.40)$$

Como cada alfabeto utilizado possui 20 letras e assumindo ignorância máxima do receptor antes de realizar a observação de uma sequência, isto é, todos os estados possuem a mesma probabilidade, teremos

$$H_{antes}(n) = H_{max}(n) \quad (3.41)$$

$$H_{max}(n) = - \sum_{i=1}^M p_{i,n} \log_2 p_{i,n} \quad (3.42)$$

$$H_{max}(n) = -20 \left(\frac{1}{20} \log_2 \left(\frac{1}{20} \right) \right) \quad (3.43)$$

$$H_{max}(n) = -20 * \frac{1}{20} (\log_2 \left(\frac{1}{20} \right)) \quad (3.44)$$

$$H_{max}(n) = -1 * (\log_2(1) - \log_2(20)) \quad (3.45)$$

$$H_{max}(n) = -1 * (0 - \log_2(20)) \quad (3.46)$$

$$H_{antes}(n) = H_{max} = 4.3219280949 \quad (3.47)$$

Os estados conformacionais de cada resíduo são representados por dois símbolos, um para κ e outro para τ , de modo que a combinação das duas sequências resulta em 400 símbolos possíveis. Sendo assim,

$$H_{antes}(n) = -1 * (0 - \log_2(20 * 20)) \quad (3.48)$$

$$H_{antes}(n) = -1 * (0 - (\log_2(20) + \log_2(20))) \quad (3.49)$$

$$H_{antes}(n) = 8.6438561898 \quad (3.50)$$

Portanto,

$$R_{res}(n) = 8.6438561898 - H_{depois}(n)(bits/resíduo) \quad (3.51)$$

Pela [Equação 3.51](#), a incerteza máxima associada a observação de um resíduo em um determinado estado é de ≈ 8.6 bits. H_{depois} será definido também de acordo com [Equação 3.39](#). Agora que estabelecemos uma medida de informação para um resíduo n , podemos usar o mesmo raciocínio para estabelecer uma medida de informação para uma conformação c , $R_{conf}(c)$. Pois,

$$R_{conf}(c) = H_{antes}(c) - H_{depois}(c) \quad (3.52)$$

Assumindo ignorância máxima antes da observação de uma determinada conformação, todas os estados i são igualmente prováveis para todos os resíduos n , portanto,

$$H_{antes}(c) = N * 8.6438561898 \quad (3.53)$$

E $H_{depois}(c)$ será definido como,

$$H_{depois}(c) = \sum_{n=1}^N \sum_{i=1}^M p_{n,i} \log_2(p_{n,i}) \quad (3.54)$$

Então,

$$R_{conf}(c) = (N * 8.6438561898) - \sum_{n=1}^N H_{depois}(n)(c)(bits/sítio) \quad (3.55)$$

Considerando a [Equação 3.51](#),

$$R_{conf}(c) = \sum_{n=1}^N R_{res}(n) \quad (3.56)$$

O valor de $R_{conf}(c)$ reflete a quantidade de informação de associada a uma determinada sequência.

3.5 Conjunto teste

Para avaliar a eficácia da representação por geometria diferencial e das medidas de informações propostas nessa tese, decidimos aplicar o método em representantes de diferentes casos do espectro de flexibilidade. Os sistemas selecionados foram a Ubiquitina, o Domínio de Ligação ao DNA da E1 de HPVs e o complexo cMyb-KIX.

3.5.1 Ubiquitina

A ubiquitina(Ubq) é uma proteína globular relativamente rígida de 76 resíduos e é um bom representante dessa região do espectro de flexibilidade, pois possui importância fundamental para funcionamento celular de eucariotos. A Ubq está envolvida na marcação

de proteínas e essa marcação possui consequências em variados processos celulares,¹³⁴ dentre eles a marcação de proteínas para degradação via proteossomo é a mais bem caracterizada.¹³⁵ A marcação de proteínas por Ubiquitinas é um processo que envolve várias enzimas. Inicialmente, a Ubq é sintetizada na forma inativa e somente após ter a cauda C-terminal processada pela enzima E1, o que expõe duas glicinas que formam o substrato de conjugação, a Ubq estará na forma ativa. A forma ativa passa a se ligar à proteína E2, também conhecida como *Ubq-conjugating enzyme*, pela interação da cauda C-terminal da Ubq com o sítio ativo de cisteínas da E2. O complexo E2-Ubq se liga a proteína ligase E3 e um substrato ligado a E3 será covalentemente ligado a Ubq.¹³⁶ A ligação final normalmente ocorre entre lisinas ou a cauda N-terminal da proteína substrato e lisinas ou a metionina inicial da Ubq. Dentre as possíveis lisinas da Ubq (K6, K11, K27, K29, K33, K48 e K63), a K48 é a mais comum e está associada a marcação para degradação via proteossomo. A segunda mais abundante é a K63 e esta envolvida na formação reversível de vários complexos de sinalização e esta envolvida em reparo de DNA, respostas do sistema imune inato, *splicing* de RNA e muitas outras.¹³⁴ No entanto, marcação envolvendo todas as lisinas já foram identificadas e podem se tornar tão abundantes quando a K48 e K63 dependendo do contexto celular.¹³⁴ Além disso, uma proteína pode ser marcada por cadeias de poliubiquitina e a ligação entre as cadeias de ubiquitina normalmente se dá por lisinas da mesma posição, mas diferentes combinações entre as lisinas também podem ser observadas.¹³⁴ Devido a sua importância e relativa rigidez, adicionamos o *ensemble* da Ubq obtido por metadinâmica restringida por dados de RMN¹¹ (Figura 12), publicamente disponível no PDB (2LJ5), ao nosso conjunto teste. Por apresentar somente uma região muito flexível, o *ensemble* da Ubq permite avaliar a eficácia dos métodos propostos nessa tese podem ser utilizado para caracterizar a dinâmica estrutural em casos relativamente simples.

3.5.2 Domínio de Ligação ao DNA da E1 de Papilomavirus Humano

O estudo da interação proteína-DNA mediada pelo domínio de ligação ao DNA da E1 (E1-DBD) é um caso interessante para nosso conjunto de testes, pois a flexibilidade do loop de ligação ao DNA (*DNA Binding Loop*, DBL) está diretamente relacionado ao reconhecimento do sítio de ligação no genoma viral. A E1 é uma proteína conservada entre HPVs e é um importante fator de replicação viral e também atua como helicase.¹³⁷ A primeira etapa do processo de replicação viral é o reconhecimento da origem de replicação (*ori*) no genoma, mediado pelo Domínio de Ligação ao DNA da E1 (E1-DBD).¹³⁷ A formação do complexo E1-DBD-DNA é crucial para o ciclo de replicação viral de HPVs. Uma vez que a interação entre o E1-DBD e o DNA é mediada pelo DBL e a hélice de ligação ao DNA (*DNA Binding Helix*, DBH),¹³⁸ o comportamento desses elementos podem afetar significativamente a dinâmica de replicação viral. Antes da publicação das primeiras estruturas do E1-DBD, estudos já haviam demonstrado que a especificidade

Figura 12 – Ensemble conformacional da Ubiquitina determinado por metadinâmica restringida por RMN, disponível no PDB (2LJ5).

Fonte: Elaborada pelo autor.

de ligação ao sítio da E1 é resultado de poucas interações base-específicas e, portanto, dependeria majoritariamente de um mecanismo de reconhecimento da forma que o DNA apresenta.^{137,139} A estrutura do E1-DBD do Papillomavirus Bovino (BPV) ligada ao DNA revelou que o principal responsável pelo reconhecimento da sequência de DNA pelo E1-DBD é o DBL e a maior parte dos contatos proteína-DNA envolvia o *backbone* do DNA.¹³⁸ No entanto, a estrutura do E1-DBD do BPV livre em solução apresentou um DBL rígido com a mesma conformação que apresenta ligada ao DNA.^{138,140} Em contra partida, a única estrutura do E1-DBD de HPV depositada não apresenta densidade eletrônica definida na região do DBL, indicando que o DBL não deve ser representado por uma única conformação quando em solução, pelo menos para o HPV18.¹⁴¹ Como o E1-DBD do HPV18 também apresenta a menor afinidade pela *ori* entre os HPVs e pertence ao grupo de HPVs de alto risco (i. e., fortemente associado a casos de câncer do colo do útero)¹³⁷ é possível que esta seja uma característica particular apenas do HPV18. Diante do exposto, decidimos estudar a possível flexibilidade do DBL de diferentes HPVs. Entender o comportamento do DBL nessa interação é importante pode auxiliar a entender características estruturais que determinam o papel da E1 na etapa de promoção da replicação genômica de HPVs. Da perspectiva de avaliação do método proposto nesta tese, o E1-DBD representa uma oportunidade de comparar estados em um *ensemble* com uma conformação de referência. Como o estado ligado do DBL é conhecido, será possível identificar a acessibilidade desse estado em um *ensemble* ao comparar os valores de κ e τ observados com a assinatura da conformação ligada do DBL.

Infelizmente, obter informações experimentais com resolução atômica é geralmente muito difícil ou mesmo impossível, para esses casos, estudos teóricos são uma boa alternativa para investigar aspectos de estrutura-dinâmica-função de proteínas e propor hipóteses

testáveis. Por esta razão decidimos explorar o espaço conformacional do E1DBD do HPV16, -18 e BPV1 por métodos computacionais. Dois modelos da E1-DBD foram construídos, um com o DBL “livre” e o DBL “ligado” e utilizados como pontos de partida para amostragem do espaço conformacional por simulações de dinâmica molecular. Para a análise de dados, uma variável coletiva para monitorar o movimento de abertura do DBL foi definida. Os detalhes de cada etapa estão descritos abaixo.

3.5.2.1 Modelagem molecular por homologia

O aspecto que pretendemos investigar é o quão flexível o DBL pode se apresentar em solução. Para poder explorar diferentes regiões do espaço conformacional dois modelos do E1DBD do HPV18 foram construídos, um com o DBL “ligado”, construído de acordo com a mesma conformação observada para o E1-DBD de BPV quando ligado ao DNA (1F08 e 1KSY), e outro com o DBL “livre”, construído com uma conformação que minimizasse a interação entre os resíduos do loop e o restante da proteína (Figura 13). Utilizamos as sequências para construção dos modelos iniciais utilizando o I-TASSER¹⁴² utilizando as estruturas 1F08,¹⁴⁰ 1KSY¹³⁸ e 1R9W¹⁴¹ do PDB como molde. O ciclo de refinamento dos modelos consistiu em ajustes manuais utilizando o COOT¹⁴³ e minimização de energia. Utilizamos o MolProbity¹⁴⁴ o TM-score, calculado pelo TM-align⁶⁶ para validação dos modelos. Os modelos iniciais construídos pelo I-Tasser foram gerados com a conformação do DBL “ligado”, para obter o modelo “livre” realizamos um primeiro ajuste do *backbone* do DBL utilizando o Pymol¹⁴⁵ e as cadeias laterais foram modeladas com auxílio do COOT.¹⁴³

3.5.2.2 Simulação por dinâmica molecular

Realizamos três simulações independentes de dinâmica molecular clássicas de 150 ns para cada modelo da E1-DBD para cada um dos modelos, total de 450 ns para cada modelo. As simulações foram realizadas utilizando o GROMACS 5.0.¹⁴⁶ O sistema de simulação consistiu em um modelo do E1-DBD em solvente explícito (SPC) dentro de uma caixa dodecaédrica de $\approx 240nm^3$. Como campo de força utilizamos o GROMOS 54a7.¹⁴⁷ Para equilibrar as cargas adicionamos íons Na⁺ e Cl⁻ ao solvente. Em seguida, realizamos uma minimização de energia utilizando o algoritmo *steepest decent*. A fase de equilíbrio consistiu em: 1) equilíbrio de temperatura, o sistema foi inicialmente aquecido até 300K utilizando uma versão modificada do termostato de Berendsen¹⁴⁸; 2) equilíbrio de pressão, uma simulação de 1ns utilizando o método de Parribello-Rahman para controle da pressão¹⁴⁹ até atingir o equilíbrio em 1bar. As coordenadas do *backbone* da proteína foram restringidas durante toda a fase de equilíbrio. Somente após o equilíbrio de temperatura e pressão ser obtido que realizamos a simulação de dinâmica molecular clássica. Ao término da simulação, extraímos as conformações da trajetória em intervalos de 100ps, um total de 1500 conformações por simulação.

Figura 13 – Modelos do DBD da E1 com o loop de ligação na conformação ligada ao DNA (DBL "ligado") e em uma conformação "livre"(DBL "livre") .

Fonte: Elaborada pelo autor.

3.5.2.3 Definição de variável coletiva e análise de forma do DBL

Definimos um ângulo (Ângulo DBL) como variável coletiva para cada modelo utilizando os resíduos D185, P265 e R243 do BPV como referência ([Figura 13](#)) com o objetivo de monitorar o movimento de abertura e fechamento do DBL. A partir da distribuição dos ângulos foi possível definir diferentes macroestados presente nos *ensembles* obtidos. A conformação que o DBL assume ao se ligar ao DNA é conhecida, então podemos monitorar se as conformações visitadas nas simulações visitaram estados semelhantes. Para este fim aplicamos a representação de geometria diferencial ([seção 3.1](#)), calculamos os valores de curvatura e torção do DBL no estado “ligado” e comparamos os valores do DBL observados nas simulações.

3.5.3 c-Myb-KIX

Para estudar aplicabilidade do método em cenários mais extremos do espectro de flexibilidade aplicamos o método desenvolvido aos *ensembles* conformacionais obtidos por Metadinâmica restringida por dados de RMN do complexo c-Myb-KIX. O domínio de transativação da c-Myb, uma proteína intrinsecamente não estruturada de 25 resíduos, interage com a KIX, um domínio globular de 88 resíduos da proteína de ligação ao CREB

(*CREB-binding protein*, CBP).¹⁵⁰ A c-Myb ao se ligar a KIX assume a estrutura de α hélice e o processo de enovelamento da c-Myb envolvido nesta ligação foi estudado a partir de *ensembles* conformacionais obtidos por metadinâmica restringida por dados de RMN de diferentes mutantes da KIX.³ Esses *ensembles* capturaram diferentes etapas do processo de enovelamento (Figura 14) e o processo se inicia em uma região diferente do sítio de ligação final.³ Esses *ensembles* foram gentilmente cedidos pelo Prof. Michele Vendruscolo (Cambridge University, Inglaterra) e Prof. Carlo Camilloni (University of Millan, Itália). Para os objetivos do presente texto, aplicaremos os métodos apenas as conformações da c-Myb nas diferentes etapas. Todas as conformações da c-Myb obtidas refletem o processo de enovelamento de uma proteína não estrutura em uma α -hélice (Figura 15), então podemos utilizar esse *ensemble* para investigar a aplicabilidade dos métodos propostos ao estudo de processos de enovelamento.

Figura 14 – *Ensembles* do complexo c-Myb-KIX de diferentes etapas do enovelamento da c-Myb (azul) utilizados em um estudo que demonstrou a existência de um mecanismo de "templated folding" para proteínas não estruturadas. Três mutações (I72V, I26V e L43A) da KIX (verde) permitiram travar o processo de enovelamento da c-Myb em diferentes etapas. A etapa 1) demonstra que a interação entre as duas proteínas não começa no sítio de ligação final.

Fonte: Elaborada pelo autor.

Figura 15 – Estados conformacionais da c-Myb de todas as etapas de enovelamento alinhadas, com destaque para a conformação final em α -hélice.

Fonte: Elaborada pelo autor.

4 RESULTADOS E DISCUSSÃO

Para melhor apresentar e discutir a eficácia e limitações dos métodos propostos, os resultados serão apresentados por sistema analisado. Cada sistema foi selecionado para representar a aplicação dos métodos propostos em casos representativos de diferentes possibilidades dentro do espectro de flexibilidade em proteínas. Exploraremos a aplicação em três sistemas: 1) A Ubiquitina, uma proteína relativamente rígida, 2) o E1-DBD, uma proteína globular relativamente rígida que possui parte da região diretamente envolvida na sua função biológica com um alto grau de flexibilidade e 3) a interação da c-Myb com a KIX, representando o processo de ligação via *templated folding* de uma proteína não estruturada.

4.1 Ubiquitina

O *ensemble* da ubiquitina utilizado representa um caso de dinâmica relativamente simples, uma proteína rígida com apenas uma região aparentemente “desordenada”. Portanto, uma boa maneira de identificar a efetividade dos descritores de GD e TI e compará-los com outros métodos já estabelecidos diante de um sistema onde a flexibilidade é relativamente baixa. Nesta seção exploraremos a aplicação dos métodos propostos com o objetivo de avaliar seu uso para 1) quantificar a flexibilidade observada, 2) identificar resíduos flexíveis, 3) comparar as conformações do *ensemble* com um determinado estado de referência.

4.1.1 Medidas de flexibilidade

Uma inspeção visual no *ensemble* (Figura 12) permite concluir que é uma proteína que não apresenta grandes mudanças na maior parte de sua estrutura, com exceção da região C-terminal que é “desordenada”. Uma medida que, de modo geral, reflita essas características pode ser considerada adequada. A mesma noção geral é refletida nos valores de κ e τ (Figura 16). Ao observar a distribuição dos valores é possível identificar que a região C-terminal é a que apresenta a distribuição mais ampla de valores em relação aos demais resíduos, no entanto, há outras regiões aparentemente rígidas que apresentam variação de curvatura comparável (Figura 17). Os descritores são absolutos no espaço e refletem a forma da curva, portanto, movimentos de corpos rígidos não afetam esses valores. Ao alinhar somente a cauda C-terminal da Ubiquitina, é possível observar que há uma certa forma preferencial, apesar de alta variação dos resíduos 71, 72 e 73 (Figura 18A). Esses resíduos estão na região inicial da cauda C-terminal da Ubq e servem como “dobradiças”, pois determinam o direcionamento dos demais resíduos posteriores em relação ao restante da proteína. Observe que os resíduos 71, 72 e 73 apresentam variação nos valores de

curvatura comparável à observada para o resíduo 10 (Figura 18B), mas somente esses três resíduos apresentam uma distribuição mais ampla nos valores de torção. O que torna a cauda C-terminal aparentemente “desordenada” em relação aos demais resíduos é a combinação mais variada dos valores de κ e τ . Os valores de τ determinam o desvio em relação a um plano e desvios mais intensos são necessário para que um resíduo seja capaz de rotacionar os resíduos posteriores em mais de uma direção no espaço tridimensional.

Figura 16 – Valores de curvatura (Azul) e torção (Vermelho) das conformações do *ensemble* da Ubiquitina e da conformação da Ubiquitina quando ligada a E2 observada na estrutura cristalográfica (3JW0).

Fonte: Elaborada pelo autor.

A inspeção dos valores de κ e τ é adequada para julgar a flexibilidade, pois são medidas intrínsecas do *backbone* da proteína e, portanto, evitam os problemas relacionados à sobreposição estrutural. É possível que conclusões semelhantes fossem obtidas a partir da inspeção das coordenadas cartesianas individuais dos $C\alpha$ de cada resíduo, no entanto, além de ter que lidar com um espaço de 3 dimensões, ainda seria preciso ter que resolver o problema da sobreposição das estruturas. Apesar de a observação da distribuição dos valores de κ e τ permitir julgar o grau de flexibilidade de uma proteína, é conveniente ter uma medida que forneça um único número que reflita essa característica. Uma solução popularmente utilizada para este fim é a flutuação média quadrática (*Root Mean Square*

Figura 17 – Gráficos "em violino" da distribuição dos valores de curvatura e torção observados no *ensemble* da Ubiquitina.

Fonte: Elaborada pelo autor.

Fluctuation, RMSF). O RMSF do ensemble da Ubq reflete adequadamente a mesma noção geral que a inspeção visual do *ensemble* fornece, isto é, a maior parte dos resíduos apresenta menos de $\approx 1\text{\AA}$ de flutuação média, enquanto a região C-terminal apresenta aumento exponencial chegando a $\approx 8\text{\AA}$ (Figura 19). Apesar de refletir bem a variação da maioria dos resíduos, o RMSF superestima a flexibilidade dos resíduos da cauda C-terminal. Isso ocorre pois os resíduos 71, 72 e 73 determinam o direcionamento da cauda C-terminal. Nesse caso, ainda que os resíduos subsequentes fossem perfeitamente rígidos, o movimento de corpo rígido seria computado e os valores de RMSF seriam diferentes de zero. No mesmo espírito do RMSF, podemos utilizar o $Z - score$ da distância euclidiana máxima observada nos descritores, discutido na subseção 3.2.2, com o mesmo objetivo. O $Z - score$ da distância κ/τ do *ensemble* da Ubq (Figura 20) permite chegar a conclusões semelhantes as obtidas a partir da observação da distribuição dos valores de κ e τ discutidas anteriormente. Isto é, a região da cauda C-terminal é de fato a região com maior flexibilidade, no entanto, apresenta variação comparável a outras regiões da proteína, como a região em torno do resíduo 10. Ao estabelecer um limite de $Z - score = 1.0$ para resíduos flexíveis é possível identificar os resíduos 10, 42, 60, 72, 74 e 75 como os mais flexíveis.

Figura 18 – A) Conformações da cauda C-terminal e da B) região do resíduos 10, uma glicina, do *ensemble* da Ubiquitina.

Fonte: Elaborada pelo autor.

Podemos estimar também a estabilidade estrutural do *ensemble* ao calcular as distâncias das demais conformações em relação a uma conformação arbitrária pertencente ao *ensemble*. Por exemplo, ao calcular as distâncias entre todas as conformações e a primeira conformação observada é possível quantificar o quanto os resíduos se distanciam da posição inicial (Figura 21). Os valores de distância também refletem a rigidez da maior parte dos resíduos, pois se mantêm próximos a posição inicial. Naturalmente, os resíduos identificados anteriormente como mais flexíveis são os também os que mais se distanciam da posição inicial, apesar das distâncias ainda serem relativamente baixas.

4.1.2 Comparação com estados de referência

Se uma conformação da proteína é conhecida previamente, pode ser útil poder compará-la com as conformações observadas em um *ensemble*, por exemplo, para investigar possíveis mecanismos de transição ou acessibilidade da conformação de referência em solução. A região C-terminal identificada como a mais flexível é capaz de interagir com cisteínas do sítio ativo da enzima E2, portanto é a região que viabiliza diretamente a formação do

Figura 19 – *Root Mean Square Fluctuation* dos $C\alpha$ resíduos calculados para o *ensemble* da Ubiquitina.

Fonte: Elaborada pelo autor.

Figura 20 – *Z - score* das distância κ/τ máxima observada nos resíduos do *ensemble* da Ubiquitina.

Fonte: Elaborada pelo autor.

complexo E2-Ubq.² A estrutura da Ubq ligada a E2 é conhecida e está disponível no PDB (3JW0²) e podemos utilizá-la para identificar se conformações semelhantes são observadas no nosso ensemble. A comparação dos valores de κ e τ da Ubq ligada a E2 e o *ensemble* da Ubq (Figura 16) revela que a maior parte dos da conformação ligada possui resíduos com valores dentro do intervalo observado no *ensemble*. No entanto, alguns resíduos apresentam assinaturas $\kappa - \tau$ em regiões pouco ocupadas, em especial para os valores de κ dos resíduos

Figura 21 – "Mapa de calor" das distâncias euclidianas observadas entre a primeira e as demais conformações do *ensemble* da Ubiquitina.

Fonte: Elaborada pelo autor.

8, 10, 69, 70 e 71. Assumindo que o ensemble da Ubq utilizado é um bom representante do equilíbrio em solução, esses resíduos apresentam posições que são favorecidas somente após a ligação com a E2. Ao calcular as distâncias entre todas as conformações observadas e a conformação da Ubq ligada a E2 (Figura 22) é possível concluir que esses resíduos de fato se distanciam mais da estrutura de referência, no entanto, os valores são relativamente baixos de modo que as estruturas observadas são semelhantes (Figura 23).

4.1.3 Medidas de informação conformacional

Aplicamos as medidas de informação conformacional, discutidas na seção 3.4, com o intuito de explorar se e como essas medidas poderiam ser utilizadas para obter informações a partir de *ensembles* conformacional em um caso de baixa variabilidade no *backbone* da proteína. Uma das hipóteses investigadas é se resíduos envolvidos diretamente em interações importantes para a função da proteína apresentariam mais (ou menos) bits de informação quando comparados com os demais. A Ubq possui sete lisinas que determinam diferentes tipos de ubiquitinação e cada tipo é envolvida com diferentes sistemas celulares.¹³⁴ Além das lisinas, três regiões hidrofóbicas da superfície viabilizam a interação da Ubq com diferentes

Figura 22 – "Mapa de calor" das distâncias euclidianas observadas entre a primeira e as demais conformações do *ensemble* da Ubiquitina.

Fonte: Elaborada pelo autor.

proteínas. A primeira região hidrofóbica identificada é formada por Leu8/Ile44/Val70,¹⁵¹ posteriormente estudos de mutações permitiram identificar outros dois sítios formados por Ile36/Leu71/Leu73 e Gln2/Phe4/Thr12.¹⁵² Os valores de informação dos resíduos, R_{res} , (Figura 24A) do *ensemble* da Ubq apresentam em média 4.17 bits, desvio padrão de 1.04 bits. Entre as lisinas, somente a K33 ($R_{res,33} = 7.16$ bits) e K63 ($R_{res,6} = 7.07$ bits) acima de 5.48 bits. Ao investigar a disposição dos resíduos que apresentam $R_{res} \geq 5.48$ bits (S20, N25, K33, F45, G53, R54, N60, Q62, K63, E64) de informação na superfície da proteína identificamos que esses resíduos formam uma região contínua conectando K63 e as regiões hidrofóbicas, enquanto os demais resíduos encontram-se próximos das demais lisinas (Figura 24B). Apesar de duas lisinas estarem no grupo com os maiores valores de R_{res} , a maior parte desses resíduos não faz parte de regiões envolvidas diretamente com

Figura 23 – Sobreposição das conformações do *ensemble* da Ubiquitina e a conformação da Ubiquitina ligada a E2 (3JW0²).

Fonte: Elaborada pelo autor.

interação com outras proteínas. No entanto, esses resíduos ou estão próximos dessas regiões funcionais ou formam uma região contínua conectando essas regiões. Essa proximidade torna plausível uma participação nesses processos, mesmo que de forma indireta. Sendo essa hipótese verdadeira, mutações nesses resíduos deveriam ter algum efeito na afinidade ou na cinética de ligação dessas interações.

Infelizmente, não encontramos nenhum estudo do efeito de mutações desses resíduos na afinidade de ligação da Ubq por alguma proteína. No entanto, em 2013 um estudo do efeito de diversas mutações pontuais da Ubq na taxa de crescimento de leveduras foi publicado.¹⁵³ Apesar de não medir o efeito diretamente na ligação com outras proteínas, podemos utiliza-lo como base para identificar resíduos sensíveis para o papel da Ubq no crescimento de leveduras. Entre os resíduos com $R_{res} \geq 5.48bits$, apenas três mutações no resíduo F45 (F45K, F45D, F45E), uma no N25 (N25P) e uma no K33 (K33P) impediam completa ou severamente o crescimento da levedura,¹⁵³ indicando uma alta tolerância a mutações sem que afete severamente o papel da Ubq. Apesar de ser possível que esse grupo de resíduos atuem facilitando a acomodação de outra proteína na região final da interação, os dados publicados apontam que se houver algum papel, mutações nesses resíduos não tendem a inviabilizar o papel da Ubq no crescimento de leveduras. Infelizmente, sem dados

experimentais sobre o efeito de mutações nas interações com outras proteínas não é possível comprovar a hipótese levantada, de modo que o real significado desses resíduos é incerto. Sendo a Ubq uma proteína relativamente rígida, a variação do *backbone* da proteína é relativamente baixa. Uma limitação dessa abordagem é ser baseada unicamente em dados do *backbone* da proteína. É possível que uma medida de informação que inclua os estados das cadeias laterais pode ser mais sensível para casos como este.

Apesar dessa limitação, ao calcular os valores de informação conformacional, R_{conf} , verificamos um aumento de ≈ 2 bits a partir da conformação 150 (Figura 25A). O que sugere que os estados visitados a partir da conformação 150 são de modo geral mais improváveis que os anteriores. Isto significa que os resíduos dessas conformações assumem estados que são menos prováveis no contexto do *ensemble*. Ao calcular a identidade entre todos os pares de conformação, observamos que as sequências das conformações a partir da 150 são também ligeiramente mais parecidas entre si (Figura 25B). Coincidentemente, é possível observar uma discreta redução nas distâncias entre as conformações e o estado da Ubq quando ligada à E2 a partir da conformação 150 na região dos resíduos 11-14 e 63-67 (Figura 22). Como a conformação ligada a E2 possui valores de $\kappa - \tau$ em regiões pouco visitadas (Figura 17) e o valor de R_{conf} é diretamente proporcional ao valor da incerteza, H_n , dos estados observados no resíduo, conformações com valores semelhantes tenderão a apresentar maiores valores de R_{conf} . O que não significa que todas as conformações com altos valores de R_{conf} são semelhantes ao estado ligado a E2, mas todas as conformações em regiões pouco visitadas no *ensemble* tenderão a apresentar maiores valores de R_{conf} . Por definição, conformações em região de transição entre "poços" do espaço conformacional são menos prováveis que as demais, o que sugere que os valores de R_{conf} poderiam ser utilizados para identificar esses estados conformacionais.

4.2 Domínio de Ligação ao DNA da E1

O E1DBD representa um caso em que uma proteína globular relativamente rígida que apresenta regiões flexíveis envolvidas diretamente com função que a proteína exerce no contexto biológico. O E1DBD é capaz de se ligar ao DNA e a forma do *backbone* do *loop* de ligação ao DNA (DBL) está diretamente relacionada ao reconhecimento do sítio específico de ligação. Esse aspecto do DBL pode ser caracterizado diretamente pelos valores de κ e τ , portanto utilizamos este caso para explorar como esses descritores e os métodos propostos podem ser utilizados em um caso onde a forma do *backbone* afeta diretamente sua função. Obtivemos um *ensemble* do E1DBD do HPV18 por simulações de dinâmica clássica e, para auxiliar as análises, também definimos uma variável coletiva para caracterizar os movimentos de abertura e fechamento do *loop* de ligação ao DNA.

Figura 24 – A) Valores de informação associada a observação dos resíduos, R_{res} , em bits do *ensemble* da Ubiquitina. B) Superfície da ubiquitina destacando duas regiões hidrofóbicas envolvidas com interação com outras proteínas (RH1 e RH2), as lisinas que determinam o tipo de ubiquitinação (Amarelo) e os resíduos que apresentaram $R_{res} \geq 5.48bits$ (Vermelho).

Fonte: Elaborada pelo autor.

4.2.1 Qualidade dos modelos por homologia

A avaliação do MolProbity e o TM-score (Tabela 1) demonstra que os modelos possuem qualidade stereoquímica aceitável e são similares ao molde utilizado (1R9W).

Figura 25 – A) Informação conformacional, R_{conf} , e B) matriz de identidade entre as sequências das conformações do *ensemble* da Ubiquitina.

Fonte: Elaborada pelo autor.

4.2.2 Distribuição de ângulos do *loop* de ligação ao DNA

A distribuição de ângulos de abertura do DBL (Figura 26) observadas nos *ensembles* obtidos por simulação permitem identificar três estados: DBL “fechado” ($< \approx 34^\circ$), “semi-fechado” ($\approx 35^\circ - \approx 53^\circ$) e “aberto” ($> \approx 53^\circ$) com uma região metaestável entre 35° e 40° . Transições entre os estados “fechados” e “semi-fechados” foram amostradas em todas as simulações individuais, independente se a conformação inicial utilizada foi o modelo

Tabela 1 – Análise de qualidade estereoquímica dos modelos do E1DBD realizada pelo MolProbity.

	HPV18 "ligado"	HPV18 "livre"
Clashscore	3.44	4.3
Rotâmeros não favoráveis	0	1
Ramachandran outliers	2	3
Ramachandran permitido	138	134
Desvio $C\beta > 0.25\text{\AA}$	0	0
Ligações ruins do <i>backbone</i>	4/1173	5/1173
Ângulos ruins do <i>backbone</i>	2/1594	6/1594

Fonte – Elaborada pelo autor.

com o DBL “ligado” ou “livre”. Essa transição observada em todas as simulações sugere uma barreira energética de transição relativamente pequena entre esses dois estados. Em contrapartida, o estado “aberto” não foi visitado em todas as simulações, de modo que ou a barreira energética de transição para o estado “aberto” não é computacionalmente acessível para uma simulação de dinâmica clássica com o tempo de simulação utilizado ou ela não é populada em solução. Apesar do método de simulação utilizado não permitir uma caracterização completa do espaço conformacional devido a dificuldades de garantir convergência na amostragem conformacional,¹⁵⁴ os estados e transições observadas refletem características intrínsecas ao espaço conformacional do E1DBD. Por esta razão, focaremos a análise nos estados “fechado” e “semi-fechado” e as transições entre eles. As conformações do estado “fechado” são similares às estruturas cristalográficas do E1DBD ligado ao DNA (PDB id: 1F08, 1KSX), sua acessibilidade mesmo para simulações partindo dos modelos com DBL “aberto” indicam um baixo custo de transição para esse estado. Esse baixo custo entre os estados “fechado” e “semi-fechado” implicam em um DBL flexível, o que é compatível com a estrutura cristalográfica do HPV18 (PDB id: 1R9W) não apresentar um DBL estável.

4.2.3 Medidas de Flexibilidade

Os valores de κ e τ (Figura 27) apresentam regiões com variações significativas de curvatura, especialmente entre os resíduos 1-10, 13-15, 22-33 (DBL), 64-69, 121-130 e 140-143. Os valores de τ apresentam variação relativamente baixa, com exceção do resíduo 70 e 132. No entanto, a análise da distribuição dos valores de κ e τ excluindo os *outliers* desses resíduos da visualização (Figura 28), revela que a maior parte dos valores de τ estão entre 0.2 e -0.4. A distribuição desses valores também permite concluir que boa parte da região do DBL (22-30) e a região entre os resíduos 120 e 130, apresentam mudanças mais significativas quando comparadas com o restante da proteína. As regiões de hélice apresentam baixa variação geral dos valores de κ e τ , com exceção da região inicial da primeira hélice, indicando instabilidade somente nesta hélice. O RMSF (Figura 29)

Figura 26 – Distribuição de ângulos de abertura do *loop* de ligação ao DNA observados nas simulações dos modelos do E1DBD do HPV18.

Fonte: Elaborada pelo autor.

apresenta média de $2.65\text{\AA}$ e, considerando resíduos com $\text{RMSF} \geq 3\text{\AA}$ como flexíveis, podemos considerar aproximadamente as mesmas regiões como as mais flexíveis, com destaque para as regiões iniciais e finais. Os valores do $Z - score$ das distâncias κ/τ máximas (Figura 30) também permite chegar a conclusões gerais semelhantes, no entanto discordam quanto a magnitude das variações, especialmente nas regiões iniciais e final. Considerando resíduos mais flexíveis apenas aqueles com $Z - score > 1.0$, o DBL (22-33) apresenta a maior região contínua com resíduos mais flexíveis. Os descritores de GD e o RMSF concordam quanto as regiões mais flexíveis e ambos sugerem um DBL flexível como uma delas.

4.2.4 Identificação de estados conformacionais

Identificar estados conformacionais energeticamente favorecidos é útil para investigar aspectos funcionais da proteína. Considerando o funil de enovelamento, essas conformações serão observadas com mais frequência em um *ensemble* representativo do equilíbrio conformacional em solução. Logo, podemos identificar esses estados como as regiões mais populadas (ou de maior densidade) em um espaço definido por descritores

Figura 27 – Valores de κ e τ por resíduos observados no *ensemble* obtido por simulações dos modelos do E1DBD do HPV18.

Fonte: Elaborada pelo autor.

que reflitam apropriadamente os graus de liberdade do sistema.

Existem inúmeros esquemas de partição possíveis (ou soluções de *clustering*) que podem ser definidas para um conjunto de conformações, serão consideradas corretas aquelas que refletirem alguma informação útil para o problema em questão. Por essa razão optamos por propor uma solução de *clustering* local para cada resíduo. Utilizando o HDBScan (ver [seção 3.3](#)) identificamos apenas os resíduos A4, G13, D27, T30, K50, T51, I53, Q54, R121, G123, I127, L141 com mais de um *cluster* e cada um captura aspectos diferentes da dinâmica da proteína. Por exemplo, o resíduo A4 faz parte de uma hélice a qual porção inicial de 4 resíduos desenovela depois de algum tempo nas simulações e os dois estados identificados pelo HDBScan capturam esse aspecto ([Figura 31A](#)). Podemos propor diferentes grupos para as conformações *ensemble* baseado apenas nesse resíduo e estudar sua relação com outras regiões da proteína. O caso do resíduo A4 permite identificar que a hélice enovelada é observada em todos os três estados do ângulo do DBL, enquanto o estado desenovelado é uma característica mais comum para os estados "Fechado" e "Semi-fechado" ([Figura 31B](#)). Os únicos resíduos do DBL que apresentam mais de um estado são D27 ([Figura 32A](#)) e T30 ([Figura 34A](#)). Apesar dos estados serem amplos demais para serem representados por uma única conformação geral do DBL, um

Figura 28 – Gráfico de "violino" das distribuições dos valores de κ e τ por resíduos observados no *ensemble* obtido por simulações dos modelos do E1DBD do HPV18. Resíduos com $Z - score > 1.0$ estão marcados em vermelho

Fonte: Elaborada pelo autor.

dos *clusters* identificados para este resíduo é predominante em conformações do estado "Aberto" (Figura 32B). Os resultados de *clustering* do resíduo T30 sugere uma relação direta com o ângulo de abertura, pois um dos *clusters* é predominante para o estado "Fechado", parte dos estados de transição pertencem ao estado "Semi-fechado" e todas as conformações do estado "Aberto" pertencem ao mesmo *cluster*, que também inclui parte do estado "Semi-fechado" (Figura 34B). Os resultados permitem concluir que existe uma relação direta entre os estados dos resíduos D27 e T30 com o ângulo de abertura do DBL. Portanto, sem a necessidade de pressupor nenhuma informação sobre o espaço conformacional, seria possível identificar estados semelhantes aos apontados pela variável coletiva definida (o ângulo do DBL) baseado apenas nos resultados em valores de $\kappa - \tau$ dos resíduos.

4.2.5 Medidas de Informação

Calculamos os valores de informação associada aos resíduo, R_{res} com o objetivo de investigar se seria possível identificar regiões importantes para a função da proteína através desses valores. A distribuição dos valores de R_{res} apresenta média de 2.98 bits e desvio padrão de 1.07 bits. Poucos resíduos apresentam $R_{res} \geq 4.06$ bits (Figura 35A), são

Figura 29 – *Root Mean Square Flucuation* do *ensemble* obtido por simulações dos modelos do E1DBD do HPV18.

Fonte: Elaborada pelo autor.

Figura 30 – O *Z - score* das distâncias κ/τ máximas por resíduo do *ensemble* obtido por simulações dos modelos do E1DBD do HPV18.

Fonte: Elaborada pelo autor.

eles T12, V41, N42, P43, G70, L74, L76, C81, S84, R85, L86, L96, Q106, P107, P108, D133 (Figura 35B). A distribuição desses resíduos na superfície revela um fenômeno semelhante ao observado no *ensemble* da Ubiquitina (ver subseção 4.1.3), os resíduos que apresentam maiores valores de R_{res} encontram-se próximos a regiões envolvida com a interação da E1DBD com outras moléculas (Figura 36). A maior região contínua exposta na superfície é formada pelos resíduos S84, R85, L86 e Q106 e estão em contato direto com resíduos do DBL, os resíduos P107 e P108 também estão próximos do DBL mas não possuem átomos expostos na superfície. Nenhum deles está diretamente envolvido com as interações. A proximidade desses resíduo com regiões importantes tornar plausível alguma influência nos mecanismos de ligação. No entanto, assim como para o *ensemble* da Ubq, também não encontramos estudos de mutação sobre esses resíduos e seu real significado permanece incerto.

Figura 31 – A) Resultados de *Clustering* pelo HDBScan dos valores de curvatura e torção observados para o resíduo A4 (amarelo) no *ensemble* conformacional do E1DBD do HPV18. Dois estados foram identificados, no primeiro (azul) a maior parte das conformações mantém uma α -hélice e no segundo (verde) a maioria apresenta a mesma hélice com os quatro primeiros resíduos desenovelados. B) Valores de curvatura e torção das conformações do resíduo A4 segundo os estados "Fechado"(roxo), "Semi-fechado"(azul), "Aberto"(amarelo) e o da conformação ligada ao DNA (vermelho)

Fonte: Elaborada pelo autor.

Uma das hipóteses investigadas é se o valor de R_{conf} permitiria identificar características do espaço conformacional, como estados conformacionais e regiões de transição. A distribuição dos valores de R_{conf} calculados para o *ensemble* do E1DBD do HPV18

Figura 32 – A) *Clusters* encontrados pelo HDBSCAN dos valores de curvatura e torção observados para o resíduo D27 no *ensemble* conformacional do E1DBD do HPV18. B) Valores de curvatura e torção das conformações do resíduo D27 segundo os estados "Fechado"(roxo), "Semi-fechado"(azul), "Aberto"(amarelo) e o da conformação ligada ao DNA (vermelho).

Fonte: Elaborada pelo autor.

Figura 33 – Resultados de *Clustering* por HDBScan dos valores de curvatura e torção observados para o resíduo T30 do *ensemble* conformacional do E1DBD do HPV18. B) Valores de curvatura e torção das conformações do resíduo T30 segundo os estados "Fechado"(roxo), "Semi-fechado"(azul), "Aberto"(amarelo) e o da conformação ligada ao DNA (vermelho).

Fonte: Elaborada pelo autor.

apresenta média de 1285.9 bits e desvio padrão de 1.58 bits. O número de bits por conformação varia no máximo em 12 bits. Os valores de R_{conf} não apresenta nenhuma relação direta com os estados definidos de acordo com o ângulo DBL (Figura 37), apesar de uma

Figura 34 – Resultados de *Clustering* por HDBScan dos valores de curvatura e torção observados para o resíduo 4 (amarelo) do *ensemble* conformacional do E1DBD do HPV18. Dois estados foram identificados, no primeiro (azul) a maior parte das conformações mantém uma α -hélice e no segundo (verde) a maioria apresenta a mesma hélice com somente os quatro primeiros resíduos desenovelados.

Fonte: Elaborada pelo autor.

discreta diminuição para os ângulos mais abertos. Uma observação curiosa é que as duas conformações que apresentam os menores valores de R_{conf} (1278.4 e 1278.8) encontram-se exatamente nas regiões de transição entre o estado "Aberto-"Semi-fechado"(56.4°) e "Semi-fechado-"Aberto"(31.6°), respectivamente. Apesar disso, não é possível identificar os estados ou suas transições utilizando apenas os valores de R_{conf} . O valor de R_{conf} é definido pela contribuição igualitária da incerteza, H_n , de todos os resíduos de uma conformação. De modo que as conformações que apresentam muitos resíduos em estados menos prováveis (H_n maior) apresentarão valores de R_{conf} maiores. O ângulo do DBL reflete uma característica local com pequenas barreiras de transição no *ensemble*, de modo que a contribuição individual dos resíduos que influenciam o ângulo do DBL não são suficientes para que sejam significativamente refletidas em R_{conf} .

4.2.6 Análise de acessibilidade da conformação do DBL ligada ao DNA

Um aspecto fundamental da ligação do E1-DBD com o DNA é a forma que o *backbone* do DBL assume e os valores de κ e τ caracterizam diretamente esta característica. Portanto, podemos utilizá-los para fazer comparações dos estados “fechado” e “semi-fechado” e com o DBL ligado ao DNA. Os estados apresentam perfis de torção e curvatura semelhantes, mas com discretas reduções na variância de nos resíduos R22-T30(DBL), Y59-G70 e T122-V130 (Figura 38). Essa redução sugere que a transição entre os estados pode ser interpretada como um processo de ajuste fino envolvendo esses resíduos. Uma inspeção mais próximas dos valores de curvatura e torção do DBL revela que os estados apresentam médias diferentes e desvio padrão parcialmente sobrepostos (Figura 39). O resíduo F24 é o que apresenta a menor variação dos valores de κ e τ entre os resíduos do DBL, indicando uma posição estável. K25 e S26 apresentam valores maiores de curvatura no estado “fechado” e há uma tendência no sentido inverso para D27, K28 e T29. Ao comparar os valores de κ e τ característico do *backbone* do DBL quando ligado ao DNA com os apresentados no estado “fechado”, R22, N23, S26 e K28 apresentam valores próximos aos limites do desvio padrão e somente K25 apresenta valor igual à média. F24 e D27 são os únicos que apresentam valores fora do desvio padrão de ambos os estados analisados. A acessibilidade de valores de κ e τ em solução próximos aos valores observados para o forma do DBL quando ligado ao DNA para a maioria dos resíduos sugere que a ligação da perspectiva dos *C α s* segue um mecanismo de *Conformational Shift* seguido de encaixe induzido. Esse mecanismo de ligação sugerido é compatível com o modelo de ligação em três etapas (difusão, seleção de conformação e reenovelamento) proposto por Raik Grunberg e colegas.¹⁵⁵

Os resultados apresentados sugerem algumas características do mecanismo de ligação ao DNA do E1DBD, pois a maioria dos resíduos apresenta formas próximas à forma do DBL ligada ao sítio específico da E1. É provável que essas formas permitam a ligação do E1DBD a sítios não específicos do DNA, uma vez que a maior parte da

Figura 35 – A) Distribuição dos valores de R_{res} e B) os valores de R_{res} por resíduo do *ensemble* do E1DBD do HPV18.

Fonte: Elaborada pelo autor.

interação E1DBD-DNA se baseia em complementaridade de forma entre o *backbone* do DNA e o *backbone* do DBL. A ligação não específica ao DNA é um mecanismo prevalente em proteínas que reconhecem sítios em um genoma¹⁵⁶ e o E1DBD não é exceção.¹³⁷ A principal vantagem dessa característica é que a proteína ao ser capaz de se ligar a uma região qualquer do genoma e deslizar por algumas bases, ela reduz a busca em três dimensões pelo sítio específico para apenas uma dimensão.¹⁵⁷ Sendo assim, o primeiro evento antes do

Figura 36 – Superfície do E1DBD do HPV18 com o *loop* de ligação ao DNA, DBL (azul), a hélice de ligação ao DNA, DBH (laranja), a hélice de dimerização (verde) e resíduos que apresentaram $R_{res} \geq 4.06bits$ (vermelho). A maioria dos resíduos com $R_{res} \geq 4.06bits$ encontram-se próximas à regiões diretamente envolvidas com a interação com outras moléculas.

Fonte: Elaborada pelo autor.

reconhecimento do sítio de ligação da E1 é a ligação a uma região não específica do DNA. O resíduo F24 faz contato direto com os fosfatos do *backbone* do DNA e o posicionamento do resíduo D27 é necessário para a forma característica do DBL quando ligado ao DNA, apesar de não interagir diretamente com o DNA. É provável que posicionamentos semelhantes aos observados no complexo E1-DBD-*ori* sejam acessíveis para o E1DBD ao se ligar a sítios não específicos do DNA e, possivelmente, o posicionamento final observado para os resíduos F24 e D27 sejam favorecidos energeticamente somente pelo sítio específico.

Sendo a ligação não específica ao DNA importante para a função da E1, alguma plasticidade do DBL se faz necessária e entender os fatores que determinam esse aspecto pode auxiliar a explicar as diferenças do E1DBD entre HPVs. Apesar das simulações terem sido realizadas somente com a proteína em solução, as conformações observadas trazem informações intrínsecas ao espaço conformacional e podemos utilizá-los para inferir aspectos determinantes da dinâmica do DBL. Considerando que o deslizamento do E1DBD em uma fita de DNA requer necessariamente um DBL flexível, a forma e o ângulo de abertura precisam tolerar algum grau de mudança. A ligação não específica ao DNA certamente muda o equilíbrio conformacional e uma mudança no sentido do estado “fechado” observado nas simulações deve ser favorecida, considerando o posicionamento entre o DBL e DBH

Figura 37 – Gráfico dos valores de R_{conf} e ângulo do DBL para cada conformação do *ensemble* do E1DBD do HPV18 obtido por simulação. Em destaque, as duas conformações com menores valores de R_{conf}

Fonte: Elaborada pelo autor.

para ligação ao DNA ser otimizado nesse estado. Em um complexo E1DBD deslizando em uma fita de DNA, transições no sentido do estado “semi-fechado” tornariam o complexo mais tolerante a mudanças locais significativas na forma do DNA sem desfazer toda a interface. Infelizmente, é difícil prever qual o efeito exato da ligação não específica ao DNA nas populações desses estados, pois simulamos apenas a proteína em água. No entanto, mesmo que a distribuição mude completamente no complexo E1DBD-DNA, os estados observados e as transições entre eles permitem investigar características estruturais intrínsecas ao DBL. De modo que os estados e transições observados nas simulações podem ser utilizados para investigar o papel de resíduos na determinação da forma e do ângulo de abertura do DBL.

Antes de prosseguir, podemos fazer alguns comentários sobre as outras duas regiões e suas possíveis relações com a interação do DBD e outros domínios da E1. A E1 é composta por quatro domínios e o DBD se posiciona entre o domínio regulatório N-terminal e o domínio de oligomerização.¹³⁷ Os resíduos Q54-G70 fazem parte de uma fita posterior a hélice de dimerização (DH) e os resíduos R121-I127 são próximos a região C-terminal do DBD, ambas espacialmente próximas (Figura 40). A proximidade espacial dessas regiões

Figura 38 – Variância dos valores de κ e τ entre os estados "fechado"(laranja) e "semi-fechado"(azul).

Fonte: Elaborada pelo autor.

conecta a DH a região C-terminal próxima do domínio de oligomerização. A dimerização do DBD ocorre somente no sítio de ligação da E1 e é um processo importante para o papel da E1 na manutenção da replicação viral, mas não para a ligação do DBD ao sítio.¹⁴¹ O processo de oligomerização da E1 ocorre apenas depois da dimerização do DBD, o que está diretamente relacionado a natureza palindrômica do sítio da E1. Portanto, uma relação estrutural envolvendo essas duas regiões é um candidato razoável à via de comunicação entre o estado de dimerização da E1 e o domínio de oligomerização. Infelizmente, a falta de informação estrutural sobre como esses domínios se conectam fisicamente inviabiliza a formulação de hipóteses sobre como mudanças estruturais no DBL afetam ou não o domínio de oligomerização baseada nos *ensembles* obtidos.

4.2.7 Resíduos chaves para a transição entre os estados “fechado” e “semi-fechado” do DBL

Os valores de κ e τ dos resíduos F24 e D27 não serem acessíveis em solução sugere que essas posições possuem algum papel na dinâmica de interação entre o DBL e o DNA.

Figura 39 – Média e desvio padrão de κ e τ observados para a região do *DNA Binding loop* entre os estados "fechado"(laranja) e "semi-fechado"(azul claro) e os valores de κ e τ da conformação do DBL observada quando ligada ao DNA.

Fonte: Elaborada pelo autor.

A inspeção da vizinhança desses resíduos e seu comportamento nas transições entre os estados permite a proposição de dois grupos que influenciam diretamente a dinâmica do DBL (Figura 41): 1) uma região hidrofóbica, formada por F24, W34 e P107, e 2) junção DBL-DBH, composta por uma rede de ligações de hidrogênio mediada por R85, D27 e Q106. Esse tipo de classificação é feita para facilitar a discussão sobre a contribuição dos resíduos para a estabilidade do DBL ou ligação ao DNA não está inteiramente restrita aos papéis dos grupos propostos. O posicionamento do F24 na direção do P107 e W107 forma um bloco central que permite o espaçamento interno necessário para a forma característica do DBL quando ligado ao DNA, portanto, favorece o posicionamento dos resíduos que interagem diretamente com o DNA. A R85 atua como peça central da junção DBL-DBH ao interagir simultaneamente com D27 (DBL) e Q106 (fita- β central). A transição entre os estados “fechado” e “semi-fechado” dependem necessariamente da R85, pois somente com

Figura 40 – Posição na superfície do E1DBD do DBL(azul), do DBH(laranja), 59-70 (vermelho) e 121-127(amarelo) e a posição do domínio de oligomerização.

Fonte: Elaborada pelo autor.

a quebra dessa rede de ligações de hidrogênio o DBL é capaz acessar maiores ângulos de abertura. Precisamos ressaltar que a configuração da junção DBL-DBH não foi modelada em nenhum dos modelos utilizados como ponto de partida das simulações e ainda assim se tornaram características dominantes nos estados “fechado” observados no *ensemble*.

Se o modelo proposto estiver correto e for determinante para a função do E1DBD, esses resíduos precisam apresentar algum grau de conservação entre os HPVs. Podemos utilizar a medida de incerteza, H_x proposta por Shannon para avaliar a conservação de um resíduo baseado em um alinhamento, quando $H_x = 0bits$ um único resíduo é observado naquela posição e o valor de H_x será máximo somente quando todos os resíduos forem igualmente prováveis para aquela posição ($H_{x,max} = 4.3bits$). A média de H_x para as sequências do E1DBD de HPVs é de 0.99 e desvio padrão de 0.6. Cinco dos sete resíduos equivalentes aos dos dois grupos discutidos possuem $H_x \approx 0$ (Tabela 2). Os resíduos da região hidrofóbica são os mais conservados. A posição 24 apresenta apenas uma fenilalanina ou uma tirosina, sugerindo que a presença do anel aromático é importante. O resíduo 34 é um triptofano em 98,26% das sequências de HPVs, as exceções são duas

Figura 41 – Mecanismo proposto envolvendo resíduos determinantes para abertura e dinâmica do DBL. A R85 forma ligações de hidrogênio entre a D27 e a Q106 (azul) e os resíduos F24, P107 e W34 formam uma região hidrofóbica interna que permite que o DBL mantenha sua forma característica. O HPV18 possui uma alanina na posição 89 (vermelho), enquanto os demais HPVs apresentam outros resíduos capazes de interagir com a arginina central.

Fonte: Elaborada pelo autor.

sequências VE1_HPVR7 e R9QBG6_HPVR16 que apresentam uma arginina e uma serina, respectivamente. Somente, prolinas são observadas na posição 107, possivelmente devido a um papel estrutural associado à mudança de direção no *backbone* ao final da fita- β e não apenas devido ao papel proposto na dinâmica do DBL. Os resíduos da junção DBL-DBH apresentam uma tolerância maior a mudanças, com exceção da arginina central que é conservada em todos os HPVs. As alternativas mais comuns (87,81%) na posição 27 são aspartato, asparagina e serina, todos capazes de formar ligações de hidrogênio com a arginina central. Um resultado inesperado foi a tirosina ser uma alternativa viável para 7,82% dos HPVs (HPV12, HPVR7, HPV47, HPV21, HPV14, HPV20, HPV25, HPV19 e HPV8). Aparentemente não há relação entre esses HPVs e não encontramos dados experimentais publicados sobre o E1DBD desses HPVs. A posição 89 é a menos conservada e tolera uma grande variedade de resíduos, o que é compatível com um papel de ajuste fino da dinâmica de transição entre os estados “fechado” e “semi-fechado”.

Resíduos da região hidrofóbica e da junção DBL-DBH são bem conservados entre os HPVs, no entanto, há alguns desvios que podem ajudar a explicar o porquê da baixa especificidade associada ao DBDE1 do HPV18. Por exemplo, o BPV apresenta resíduos e disposição similares para a maioria dos resíduos em posições equivalentes (P265, W192,

Tabela 2 – Análise de conservação dos resíduos propostos como determinantes para abertura do ângulo do DBL.

Resíduo	Aminoácido	%	Incerteza (H_x)
24	Y	12.17	0.20
	F	87.83	
27	C	1.73	1.25
	D	53.91	
	H	1.73	
	N	26.95	
	S	6.95	
	T	0.87	
	Y	7.82	
34	S	0.87	0.63
	R	0.87	
	W	98.26	
85	R	100.00	0.00
89	A	26.09	2.19
	C	0.87	
	E	20.00	
	G	4.35	
	F	0.87	
	I	2.61	
	H	10.43	
	K	1.74	
	M	1.74	
	L	2.61	
	Q	5.21	
	S	13.91	
	T	2.61	
	V	6.09	
	Y	0.87	
106	E	78.26	0.73
	D	4.35	
	G	0.87	
	N	0.87	
	Q	14.78	
	R	0.87	
107	P	100.00	0.00

Fonte – Elaborada pelo autor.

F182, R243 e D185) e as maiores mudanças em relação ao HPV18 são 1) uma leucina (L191) no lugar da P107 e 2) uma arginina (R294) que permite a formação de uma ligação

de hidrogênio extra com a arginina central, o que não é possível com A89 do HPV18. O HPV16 também apresenta os mesmos resíduos em posições equivalentes, mas na posição 89 há um glutamato que também permite ligações de hidrogênio extras com a arginina central. Essa ligação a mais com a arginina central pode aumentar o custo de transição entre os estados “Fechado” e “Semi-fechado”. A ausência dessa interação no E1DBD no HPV18 implica em um custo mais baixo de transição entre esses estados, o que pode ajudar a explicar a baixa especificidade característica do E1DBD do HPV18. Nós aplicamos a mesma estratégia discutida nesta seção ao E1DBD do HPV16 e BPV1, além de mutações nos resíduos dos grupos propostos e os resultados são coerentes com as previsões do modelo. As implicações biológicas para biologia do HPV estão fora do escopo do presente texto, mas um artigo sobre essas implicações foi escrito e submetido para o *Journal of Biomolecular Structure and Dynamics*.

4.3 c-Myb-KIX

Os casos mais extremos dentro do espectro de flexibilidade são os das proteínas não estruturadas e utilizamos os *ensembles* conformacionais das etapas de ligação da c-Myb com a KIX³ para avaliar a aplicação dos métodos propostos nesse cenário. Esses *ensembles* foram determinados por Metadinâmica restringida por dados de RMN e representam o primeiro caso de ligação de uma proteína não estruturada via *templated folding* a ser caracterizada com esse nível de detalhe.³ A c-Myb forma uma α -hélice ao interagir com a KIX e as conformações da c-Myb refletem o processo de enovelamento de uma proteína não estruturada em uma hélice. Nesta seção, exploraremos o uso dos descritores de geometria diferencial e teoria da informação no estudo de proteínas não-estruturada, portanto, nos ateremos majoritariamente a análise das conformações da c-Myb.

4.3.1 Medidas de flexibilidade

Os valores de curvatura e torção das conformações dos *ensembles* da c-Myb apresentam uma variação maior que os observados para os casos anteriores, especialmente para os valores de torção (Figura 42). Também podemos observar que resíduos vizinhos apresentam valores próximos de curvatura e torção para um grande número de conformações, compatível com a formação de hélices. A variação maior nos valores de torção sugere que mudanças significativas no direcionamento em várias regiões da proteína são observadas nas conformações, o que é esperado para uma proteína não estruturada. Por outro lado, a distribuição desses valores (Figura 43) permite identificar regiões com variação comparável a observada na cauda C-terminal da Ubq (ver subseção 4.1.1). O que sugere que a alta flexibilidade observada em proteínas como a c-Myb é de certa forma similar em magnitude ao que se observa em regiões terminais de proteína rígidas. As distribuições também permitem identificar alguns resíduos que apresentam mais de uma região com população significativa de conformações, especialmente para os valores de torção.

Figura 42 – Valores de curvatura (Azul) e torção (Vermelho) das conformações da c-Myb pertencentes aos *ensembles* das etapas de ligação à KIX.

Fonte: Elaborada pelo autor.

Um dos obstáculos para realizar medidas de flexibilidade de proteínas não estruturadas é a necessidade de se obter um bom alinhamento estrutural das conformações. Os valores de RMSF são especialmente sensíveis ao alinhamento estrutural, o que dificulta seu uso nesses casos mais extremos de flexibilidade. No entanto, a c-Myb é uma proteína pequena e as conformações analisadas tendem a formar uma hélice, de modo que não é possível obter uma sobreposição estrutural razoável. Os valores de RMSF dos C α s (Figura 44) apresentam média de 5.7 Å e sugerem uma proteína muito flexível, sendo a região S101-Q109 a que apresenta os menores valores de RMSF. A análise do Z – score das distâncias máximas apontam o oposto (Figura 45), S101, K107 e Q110 os únicos com Z – score ≥ 1.0 . A região M100-L106 é a primeira a enovelar, formando uma volta de hélice a partir da segunda etapa da ligação. Por ser a primeira a enovelar, a maior parte das conformações apresenta um posicionamento estável, o que reduz a média da flutuação do C α desses resíduos. Já os valores do Z – score das distâncias máximas $\kappa - \tau$ são determinados de acordo com os valores dos extremos inferiores e superiores da distribuição de κ e τ observados, após a remoção de *outliers* (definidas como regiões com menos de 2% das conformações). Isto é, a menos que os valores desses extremos sejam populados por menos de 2% da população, eles serão utilizados no cálculo da distância máxima $\kappa - \tau$. Se mais de 2% das conformações desses resíduos visitam outros estados diferentes de uma volta de hélice, essa informação não é perdida nos valores de Z – score

Figura 43 – Gráficos em "violino" da distribuição dos valores de curvatura e torção observados nas conformações da c-Myb pertencentes aos *ensembles* das etapas de ligação à KIX.

Fonte: Elaborada pelo autor.

da distâncias máxima κ/τ . Apesar de ser a primeira a enovelar, essa região apresenta metade dos resíduos mais flexíveis.

4.3.2 Identificação de estados conformacionais

As conformações da c-Myb obtidas compõem quatro etapas bem definidas do processo de ligação com a KIX por *templated folding*. O conjunto dessas conformações reflete o processo de *folding* e permite avaliar se seria possível identificar essas etapas utilizando apenas os descritores de geometria diferencial. Além disso, os resultados de *clustering* e os valores de κ e τ de cada resíduo auxiliam a visualizar o seu papel no processo de enovelamento da c-Myb. Por exemplo, os resíduos E89, K90, I92, K93, L95, E96 e M100-E105 apresentam mais de um *cluster* identificados via HDBScan. No entanto, somente os *clusters* dos resíduos E96 e S101 permitem identificar praticamente todas as etapas da ligação. Todos os resíduos refletem uma característica local do processo de enovelamento, mas vamos nos ater aos resíduos E96, S101 e K107 apenas para ilustrar como o método proposto pode ser útil em análises de casos como esse.

O resultado de *clustering* do resíduo E96 e S101 indicam que esses resíduos possuem um papel de destaque nas duas primeiras etapas da formação do complexo c-Myb-KIX. S101 é um dos mais flexíveis segundo os valores do Z – *score* da distância máxima κ/τ e

Figura 44 – *Root Mean Square Flucuation* dos carbonos α das conformações da c-Myb pertencentes aos *ensembles* das etapas de ligação à KIX.

Fonte: Elaborada pelo autor.

Figura 45 – O *Z - score* das distâncias κ/τ máximas por resíduo das conformações da c-Myb pertencentes aos *ensembles* das etapas de ligação à KIX.

Fonte: Elaborada pelo autor.

apresenta três *clusters* que permitem identificar as duas primeiras etapas do enovelamento e considera as duas última como pertencentes a um único *cluster* (Figura 46). S101 apresenta uma maior variação geral nos valores de curvatura e torção na primeira etapa de ligação. Nesta etapa, a região da c-Myb que interage diretamente com a KIX (E96-M100) e os resíduos anteriores (K88-L95) permanecem numa conformação de "alça"relativamente estável quando comparada aos demais resíduos (Figura 47). Essa estabilidade é mediada

principalmente pelo T11 da KIX que interage com L96 e L97 da c-Myb³ e os resíduos posteriores (S101-L112) variam seu posicionamento interagindo com outras regiões da KIX sem formar nenhuma interação estável. S101 viabiliza essa variação e isso é refletido na ampla distribuição dos seus valores de torção. Na segunda etapa (cMyb-KIX(I26V)), S101 apresenta uma variação significativa apenas nos valores de curvatura. Nesta etapa, o resíduo S101 é a peça central de uma rede de ligações de hidrogênio estável envolvendo os resíduos N104, E105 e T102 da c-Myb e o K21 da KIX (Figura 48). O K21 da c-Myb forma uma ligação de hidrogênio com o oxigênio da cadeia principal do S101, impedindo movimentos de rotação e, conseqüentemente, limitando as mudanças nos valores de torção. Nas duas ultimas etapas, já temos uma α -hélice padrão formada e ambos as etapas podem ser considerados como pertencentes ao mesmo estado conformacional da perspectiva do S101. Já o resíduo E96 faz parte da região que interage diretamente com a KIX logo na primeira etapa de enovelamento e os resultados de *clustering* são semelhantes aos do S101 (Figura 49). Assim como no S101 mas em menor escala, na primeira etapa o E96 apresenta uma distribuição mais ampla nos valores de curvatura e torção. Esse resíduo tem sua posição estabilizada pela interação com R61 da KIX (Figura 50), de modo que além da contribuição do T11, também podemos incluir a R61 como uma interação fora do sítio de ligação final que é importante para a formação do complexo c-Myb-KIX. Na segunda etapa, a E96 apresenta uma ampla variação nos valores de torção, e uma redução nos de curvatura. O que reflete uma inversão de papéis entre os resíduos nesta etapa, pois a S101 encontra-se na região estabilizada, enquanto a E96 permite a mudança de posicionamento dos resíduos anteriores. Nas ultimas etapas, ambos já formam hélices e os resultados de *clustering* também considera essas etapas como pertencentes ao mesmo *cluster*.

Os valores de κ e τ dos resíduos E96 e S101 não distinguem as duas últimas etapas de ligação, enquanto o resíduo K107 só permite distinguir as duas últimas etapas (Figura 51). Esses *clusters* do resíduo K107 refletem algo que só ocorre nessas etapas finais. O K107 é um dos resíduos mais flexíveis segundo os valores de $Z - score$ da distância máxima κ/τ e apresenta três *clusters*. Um *cluster* corresponde à penúltima etapa da ligação com a KIX e os demais dividem a etapa final (KIXwt) em dois grupos. Nas duas últimas etapas a hélice já está praticamente enovelada, com exceção da última volta de hélice. O processo de enovelamento dessa última volta é refletido nos valores de κ e τ desse resíduo e a subdivisão da etapa final também reflete estágios diferentes de enovelamento dessa volta de hélice no *ensemble* da etapa final.

Os resultados de *clustering* desses resíduos permitiram identificar de forma intuitiva a contribuição individual dos resíduos para o processo de enovelamento da c-Myb. Essas características locais podem trazer informações úteis sobre cenário global do *ensemble* e são facilmente identificadas usando os valores de κ e τ . A estratégia utilizada ilustra como os descritores de geometria diferencial podem ser usados para identificar características locais da dinâmica do *backbone* de proteínas.

Figura 46 – A) Resultados de *clustering* e B) distribuição das 4 etapas de ligação dos valores de curvatura e torção do resíduo 101 observados nas conformações da c-Myb pertencentes aos *ensembles* das etapas de ligação à KIX. Há quatro *ensembles* referidos de acordo a mutação da KIX utilizada para determiná-los. Os resultados de *clustering* permitem identificar quase todos os estados individualmente.

Fonte: Elaborada pelo autor.

Figura 47 – *Ensemble* conformacional da c-Myb-KIX(I72V). Os resíduos 88-100 da c-Myb forma uma alça relativamente estável (*rainbow*) devido ao contato com a T11 da KIX, enquanto os demais resíduos, 101-112, permanecem com posicionamento instável (cinza).

Fonte: Elaborada pelo autor.

Figura 48 – Rede de ligação de hidrogênio envolvendo a S101, T102, N104 e E105 da c-Myb e K21 da KIX. Essa rede só é formada na segunda etapa de enovelamento (cMyb-KIX-I26V).

Fonte: Elaborada pelo autor.

Figura 49 – A) Resultados de *clustering* e B) distribuição das 4 etapas de ligação dos valores de curvatura e torção do resíduo E96 (azul) observados nas conformações da c-Myb pertencentes aos *ensembles* das etapas de ligação à KIX. Há quatro *ensembles* referidos de acordo a mutação da KIX utilizada para determiná-los. Os resultados de *clustering* permitem identificar quase todos os estados individualmente. A região 88-95 (cinza) e o resíduo 101 (vermelho) encontram-se destacadas.

Fonte: Elaborada pelo autor.

Figura 50 – Interação entre a R61 e a E96 auxilia na estabilização da c-Myb na primeira etapa da ligação com a KIX.

Fonte: Elaborada pelo autor.

4.3.3 Medidas de Informação

Por definição, o processo de enovelamento de uma proteína implica em redução na incerteza da posição dos seus átomos. Esse conceito de redução de incerteza pode ser medido quantitativamente pelos valores de R_{conf} e R_{res} e podemos supor que esses valores devem aumentar à medida que a proteína se enovela. Para investigar essa hipótese, podemos analisar as etapas individuais da ligação da c-Myb-KIX da perspectiva global do processo de enovelamento. Para este fim, definimos um alfabeto baseado na distribuição dos valores de κ e τ (ver [subseção 3.4.1](#)) observados em todas as etapas da c-Myb, geramos sequências conformacionais para as conformações de cada etapa de ligação e calculamos R_{conf} e R_{res} . Da perspectiva da c-Myb, o processo de enovelamento se reflete em aumento de R_{res} ([Figura 52A](#)) e R_{conf} ([Figura 54A](#)) para as três primeiras etapas, com exceção do estágio final. Isto é, a incerteza quanto à posição dos resíduos em geral tende a diminuir à medida que a proteína se enovela em uma α -hélice e sendo a penúltima etapa (c-Myb-KIX(L43A)) com maior informação associada aos resíduos ($\langle R_{res} \rangle = 5.26$ bits, $\sigma_{R_{res}} = 0.64$; $\langle R_{conf} \rangle = 197.29$, $\sigma_{R_{conf}} = 0.54$). Há um aumento de 4.35 bits no valor médio de R_{conf} entre a primeira ($\langle R_{conf} \rangle = 192.94$ bits) e a penúltima etapa. Entre a penúltima e a última etapa há uma diminuição de 3.16 bits no valor médio de R_{conf} . O *ensemble* da c-Myb-KIXwt apresenta uma variação maior de posição que a da c-Myb-KIX(L43A) (**FigX**) o que explica essa diminuição nos valores de R_{conf} somente na última etapa. De modo que, os valores de R_{conf} e R_{res} de fato tendem a aumentar durante o processo de enovelamento e os resultados sugerem que a penúltima etapa de ligação da c-Myb é a que

Figura 51 – A) Resultados de *clustering* e B) distribuição das 4 etapas de ligação dos valores de curvatura e torção do resíduo K107 (amarelo) observados nas conformações da c-Myb pertencentes aos *ensembles* das etapas de ligação à KIX. Há quatro *ensembles*, todos referidos de acordo a mutação da KIX utilizada para determiná-los. O resultado de *clustering* permite identificar apenas as duas últimas etapas de ligação e reflete, o processo de enovelamento da última volta de hélice. A região 88-95 (cinza) e os resíduos S101(vermelho) e E96(azul) encontram-se destacados.

Fonte: Elaborada pelo autor.

possui um posicionamento mais estável.

Figura 52 – Distribuição dos valores de R_{conf} da das quatro etapas de ligação da A)c-Myb com a B)KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4).

Fonte: Elaborada pelo autor.

Os valores de R_{conf} e R_{res} refletem a incerteza sobre os valores de κ e τ , que por sua vez refletem a posição dos C α s do *backbone* da proteína. Logo, mudanças nos valores de R_{conf} refletem mudanças de entropia conformacional da proteína. Portanto,

podemos interpretar essas mudanças entre as etapas como uma estimativa da mudança de entropia conformacional entre as etapas de ligação. Aplicando a mesma metodologia utilizada para a c-Myb, o processo de ligação se traduz em perda de informação para a KIX somente entre a primeira ($\langle R_{conf} \rangle = 748.19, \sigma_{R_{conf}} = 1.29$ bits) e a segunda etapa ($\langle R_{conf} \rangle = 743.19, \sigma_{R_{conf}} = 1.14$) da ligação. Para todas as demais, a distribuição de R_{conf} e R_{res} são semelhantes (Figura 54B e Figura 52B). A transição da c-Myb da região de interação com o resíduo T11 para o sítio de ligação final ocorre entre a primeira e a segunda etapa. Isto é, o posicionamento da c-Myb no sítio de ligação final aumenta a incerteza quanto as posições dos C α s da KIX, e essa incerteza permanece praticamente a mesma nas etapas seguintes. O valor de $\langle R_{conf} \rangle$ da KIX diminui 4.52 bits entre a primeira e a terceira etapa, que é aproximadamente o mesmo valor que a c-Myb ganha entre essas etapas ($\delta \langle R_{conf} \rangle = 4.35$ bits). No entanto, a etapa final apresenta diminuição de $\langle R_{conf} \rangle$ para a c-Myb e KIX de -3.5 e -1.18 , respectivamente. O balanço final da ligação com a c-Myb para a KIX é de -5.72 bits no valor de $\langle R_{conf} \rangle$ e de $+1.19$ bits para a c-Myb. Os dados sugerem que a KIX fornece a maior contribuição de entropia conformacional para o processo de ligação e compensa a perda de entropia associada ao processo de enovelamento da c-Myb. Poderíamos utilizar a Equação 1.30 para estimar o valor da mudança de entropia e a Equação 1.36 para estimar a energia mínima necessária para que a ligação ocorra, no entanto, sem dados experimentais para comparar os valores calculados não há como julgar a precisão dessas estimativas. De qualquer forma, os dados sugerem que a KIX apresenta a maior contribuição para mudança de entropia conformacional e ao comparar os valores de R_{conf} com a constante de associação k_{on} reportadas³ observamos uma relação inversamente proporcional entre essas duas grandezas (Figura 53).

Ao examinar os valores de R_{res} da KIX, observamos um fenômeno semelhante ao encontrado nos sistemas discutidos anteriormente. A KIX possui dois sítios de ligação, um para a c-Myb e outro que interage com domínio do *mixed lineage leukemia (MLL) transcription factor*. Apesar da perda de informação geral da KIX, os resíduos com valores de R_{res} ($R_{res} > 5$ bits) permanecem praticamente os mesmos para todas as etapas. Além dos resíduos nas proximidades dos sítios de ligação da MLL e da c-Myb, todos os resíduos que interagem com a c-Myb na primeira etapa da ligação também permanecem com os maiores valores de R_{res} (Figura 55). Entretanto, sem dados do efeito de mutações nesses resíduos na interação entre a KIX e outros ligantes é difícil fazer afirmações sobre qual é o papel desses resíduos na função da proteína.

4.4 Vantagens, limitações e perspectivas para os métodos propostos

4.4.1 Representação do espaço conformacional de proteínas por geometria diferencial

A representação do espaço conformacional do *backbone* de proteínas utilizando descritores de geometria diferencial apresenta várias conveniências para análise de *ensembles*

Figura 53 – Valores de $\langle R_{conf} \rangle$ dos *ensembles* da KIX correspondente e valores da constante de associação reportada para cada etapa de ligação com a KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4). Valores de k_{on} divulgados no material suplementar em³

Fonte: Elaborada pelo autor.

conformacionais de proteínas. A representação do *backbone* como uma curva espacial é válida para qualquer tipo de proteína, os valores de curvatura e torção são características intrínsecas do forma do *backbone* e independem de sua posição no espaço. Sendo assim, esses descritores são gerais o suficiente para serem aplicados a todas as proteínas e evitam a necessidade de resolver o problema da sobreposição estrutural, assim como os ângulos $\phi - \psi$ mas evitando problemas associados a periodicidade desses ângulos. Evitar o problema de sobreposição diminui o custo computacional ao comparar estruturas, o que é especialmente vantajoso para realizar medidas de flexibilidade em regiões ou proteínas muito flexíveis. Uma das maiores limitações para esse tipo de análise é a dificuldade de se obter uma boa solução de sobreposição estruturais, pois essa solução afeta diretamente o valor da uma medida baseada na posição dos átomos. Esse fenômeno é exemplificado nas comparações realizadas entre os resultados de *Z - score* da distância máxima κ/τ e o RMSF. Além disso, o significado de mudanças de curvatura e torção são conceitos intuitivos. O que torna a fácil a interpretação do significado desses valores, o que mantém uma das principais vantagens da representação do espaço conformacional por variáveis coletivas mas sem a

Figura 54 – Distribuição dos valores de R_{res} da das quatro etapas de ligação da A)c-Myb com a B)KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4).

Fonte: Elaborada pelo autor.

necessidade de assumir *a priori* características sobre os graus de liberdade do sistema analisado. Consequentemente, a identificação de estados conformacionais locais e sua interpretação também é facilitada. A solução de *clustering* implementada baseada em densidade é compatível com o conceito de estado nativo de uma proteína segundo o funil de

Figura 55 – A) R_{res} calculados para as conformações da KIX do *ensemble* da primeira etapa de ligação da c-Myb com a KIX (c-Myb-KIXI72V). B) Distribuição dos resíduos $R_{res} \geq 5.0$ (rosa), $R_{res} \geq 6.0$ (vermelho), do sítio de ligação a c-Myb (azul) e sítio de ligação ao MLL (amarelo) na superfície da KIX. Uma conformação da c-Myb está representada na forma de *ribbon*.

Fonte: Elaborada pelo autor.

enovelamento e se mostrou muito eficaz especialmente para monitoramento de hélices. Essa facilidade se deve a necessidade de manter valores similares de curvatura e torção ao longo

dos resíduos para que uma hélice seja formada e mantida. Por fim, a representação por geometria diferencial possui várias vantagens associada a métodos popularmente utilizados e evita a maior parte das limitações associadas a esses métodos, o que a torna uma solução mais adequada para análise de *ensembles* conformacionais.

Um outro aspecto interessante da representação por geometria diferencial é poder resolver um problema relativamente negligenciado: a falta de uma linguagem computacionalmente conveniente e genérica para representar estruturas de proteínas. Como comentado por Goriely e colaboradores em uma revisão publicada em 2008,⁸⁵ a maior parte das áreas identificadas como as atuais fronteiras do conhecimento da biologia estrutural estão ligadas a limitações experimentais, o que é evidente para proteínas de membrana, proteínas intrinsecamente não estruturadas, complexos macromoleculares, dentre outras. No entanto, a falta de uma linguagem matemática adequada para representar estruturas proteicas usando variáveis contínuas é uma das fronteiras intelectuais da biologia estrutural pouco apreciada. A representação de estruturas de proteínas através das coordenadas atômicas é natural, porem é uma representação discreta e impede o uso de métodos contínuos de análise. Ao mapear a representação por coordenadas atômicas com uma representação contínua, como a por geometria diferencial, podemos utilizar técnicas analíticas que permitiriam investigar características mecânicas gerais de proteínas. Neste mesmo trabalho, Goriely e colaboradores argumentam e discutem sobre aplicações desses métodos baseadas em representações de geometria diferencial para 1) determinação de estruturas baseada em dados experimentais, 2) estudo e predição do contínuo de *folds* de proteínas e 3) estudo de mecânica de estruturas proteicas e modelagem de deformações dessas estruturas. O método proposto nesta tese exemplifica as vantagens desse tipo de representação para o estudo de mudanças conformacionais em relação aos métodos popularmente utilizados. Outras áreas poderiam ser beneficiadas de modo semelhante. Por exemplo, da perspectiva do *backbone* representado como uma curva espacial, o problema de identificar todos os enovelamentos possíveis para proteínas consiste em identificar o subconjunto de todas as curvas espaciais possíveis que é compatível com o esperado para esse tipo de estruturas. Sendo possível definir uma função matemática que capture essa qualidade das curvas, identificar esse subconjunto de possíveis enovelamentos seria relativamente simples. Isso permitiria explorar todo o espaço de enovelamentos possíveis para proteínas. Existem alguns trabalhos já publicados explorando aplicações de representação contínua, algumas comentadas por Goriely e colaboradores.⁸⁵ Estabelecer uma representação universal de proteínas baseada em variáveis contínuas permite lidar com problemas difíceis de serem abordados por uma representação discreta, como a das coordenadas atômicas, e a representação baseada em geometria diferencial é uma boa candidata a esse posto.

Apesar de ainda em fase de protótipo, os programas desenvolvidos para realizar as análises propostas neste trabalho apresentam um custo computacional baixo. Uma boa performance geral é obtida em um computador pessoal comum para o ano de 2017 (intel

core i7-4770 3.40 GHz e 16.0Gb de memória RAM). A etapa que apresenta o maior custo computacional é o cálculo dos descritores de geometria diferencial das conformações e a implementação atual permite paralelizar de acordo com o número de *cores* ou *threads* disponíveis. O *ensemble* do E1-DBD do HPV18 (145 resíduos e 9001 conformações) é o maior utilizado neste estudo e o cálculo dos descritores de geometria diferencial leva menos de 4 minutos a partir de 3 *cores* em um intel core i7-4770 3.40GHz (Figura 56) com consumo de memória RAM de ≈ 2.2 Gb. O programa que calcula os valores de curvatura e torção foi escrito em D, utiliza uma solução relativamente simples de paralelização e há espaço para melhorias de performance. Uma nova implementação voltada otimizada para tirar proveito de servidores de alta performance viabilizaria a análise de centenas de milhares de conformações com um custo computacional ordens de magnitude menor que as alternativas atualmente disponíveis. Pretendemos continuar desenvolvendo o programa, mas o atual protótipo desenvolvido neste trabalho será disponibilizado gratuitamente para fins não comerciais. Atualmente estamos trabalhando no desenvolvimento de uma API (*Application Programming Interface* em R e em Python para facilitar a implementação dos métodos propostos em diferentes rotinas de análise por parte de terceiros.

Figura 56 – Tempo em segundos para processar 9006 conformações do *ensembles* do E1-DBD do HPV18 (145 resíduos) por *core* utilizado pelo protótipo desenvolvido para esta tese. Performance em um intel core i7-4770 3.40GHz, 16Gb de memória RAM, Ubuntu 16.04.

Fonte: Elaborada pelo autor.

Apesar das vantagens comentadas, a maior limitação da representação proposta neste trabalho é se basear apenas nos Cas das conformações. Portanto, qualquer informação

associada a interação envolvendo cadeias laterais que não afete significativamente a posição dos seus respectivos $C\alpha$, não será refletida pelos descritores de geometria diferencial. A adição de uma representação das conformações de cadeias laterais ao método proposto se faz necessária para que o método se torne mais sensível a esse aspecto. A análise de *ensembles* conformacionais de proteínas relativamente rígidas seria especialmente beneficiada com a inclusão de informação ligada ao posicionamento de cadeias laterais.

4.4.2 Teoria da Informação aplicada a análise conformacional

Além da representação espacial, um outro aspecto relevante do espaço conformacional de proteínas é a relação probabilística entre os estados conformacionais. Existe uma relação entre a probabilidade associada à observação de cada conformação acessível para uma proteína e suas características funcionais (ver [seção 1.1](#)). Uma ferramenta matemática apropriada para avaliar esse aspecto poderia se basear em conceitos de teoria da informação. Por esta razão, decidimos explorar a aplicação dos conceitos de incerteza e informação calculados a partir dos descritores de geometria diferencial (ver [seção 3.4](#)). Para isso desenvolvemos um sistema de codificação desses descritores em um número finito de símbolos e calculamos a informação associada aos resíduos, R_{res} , e às conformações, R_{conf} , baseada nas sequências resultantes da codificação. Investigamos se haveria alguma relação entre os resíduos ligados diretamente a função da proteína e o valor de R_{res} , e se o valores de R_{conf} poderiam ser utilizado para identificar características estruturais em um *ensemble* conformacional de proteínas.

De modo geral, os valores maiores (ou menores) de R_{res} não estão associados a resíduos envolvidos diretamente com regiões funcionais da proteína nos sistemas teste investigados. No entanto, observamos que os resíduos que apresentaram maiores valores de R_{res} tendem a estar próximos de regiões envolvidas diretamente com a função da proteína. Além disso, esses resíduos encontram-se posicionados de tal forma que conectam diferentes regiões funcionais. Devido a proximidade com esses sítios de ligação, é plausível que esses resíduos também possam estar envolvidos com a ligação a outras moléculas. Maiores valores de R_{res} implicam uma redução maior da incerteza sobre qual estado aquele resíduo assume. Da perspectiva de uma molécula que pudesse interagir com uma proteína, um resíduo cujo o estado é previsível na distribuição de conformações acessíveis poderia servir como base para uma primeira interação. Considerando a probabilidade de um encontro aleatório, uma molécula que fosse capaz de interagir com um resíduo em um determinado estado e esse estado fosse presente em todas as conformações, esse resíduo seria um bom alvo para maximizar a chance de formação de um complexo. É possível que esses resíduos atuem aumentando as chances de um encontro com uma determinada molécula resultar no seu posicionamento final no sítio de ligação apropriado. O que seria compatível com o conceito de *binding funnel*,³⁶ pois permitiria que diferentes pontos iniciais de contato resultassem na formação de complexo. No entanto, não encontramos dados experimentais

na literatura que permitissem confirmar ou rejeitar essa hipótese. Desse modo, o significado de resíduos que apresentam maiores valores de R_{res} é incerto.

Similar ao R_{res} , os valores de R_{conf} refletem a redução da incerteza ao observar uma determinada conformação e se mostraram úteis para analisar as etapas de enovelamento. Os resultados observado para a c-Myb sugerem que R_{conf} tende a aumentar a medida que a proteína se enovela. O processo de enovelamento consiste em, a partir de uma conformação instável, a proteína assume conformações progressivamente mais estáveis até atingir o mínimo global do funil de enovelamento, que é chamado de estado nativo da proteína. Uma consequência desse processo é que no estado final, os resíduos apresentarão um subconjunto de estados menos diverso que os observados fora desse mínimo global. Para esse tipo de análise, a noção de redução da incerteza de observar uma conformação final é intuitiva, o que torna R_{conf} uma medida atrativa. Um alfabeto definido a partir dos descritores de geometria diferencial de um *ensemble* conformacional representativo do equilíbrio em solução poderia ser utilizado para identificar esses estados independente de qual seja o enovelamento final. Seria possível identificar o estado final de enovelamento apenas baseado nas probabilidades associadas a cada conformação, evitando assim a necessidade de estabelecer métricas entre as conformações e um estado de referência. No contexto de *ensembles* conformacionais representativos do equilíbrio em solução, uma possível aplicação para R_{conf} é a sua utilização para identificação de estados de transição. Sendo possível identificar estados de transição, os valores de R_{conf} podem ser úteis para otimização de protocolos de *ensemble docking* e identificação de conformações alvo para desenvolvimento de fármacos.

Um aspecto do método proposto pouco explorado no presente trabalho é a representação conformacional por sequências de símbolos. Por exemplo, essas sequências permitem julgar as semelhanças entre duas conformações utilizando a identidade entre sequências. Esse tipo de abordagem pode ser utilizada para análise de trajetórias e identificar estados visitados que são semelhantes entre si. Por exemplo, ao definir um conjunto de conformações utilizando todas as conformações das etapas de ligação da c-Myb com a KIX e plotar uma matriz de identidade ([Figura 57](#)) podemos observar que cada etapa tende a apresentar conformações mais semelhante entre si, do que entre as demais etapas. A mesma estratégia de análise pode ser aplicada usando apenas um subconjunto de resíduos, o que permitiria focar a análise apenas em uma região de interesse. Outra possível aplicação dessa representação é a definição de um alfabeto "universal" baseado em todas os valores de curvatura e torção observados no *Protein Data Bank*. Com esse alfabeto seria possível incluir informação estrutural em *pipelines* de análise de sequências de proteínas diferentes utilizando métodos de alinhamento de sequência, o que pode ser útil para, por exemplo, estudos da evolução e da relação entre sequência e estrutura em famílias de proteínas.

Apesar dos resultados interessantes, a estratégia de análise baseada em teoria

Figura 57 – Matriz de identidade das sequências conformacionais das etapas da ligação da c-Myb com a KIX. As etapas estão indicadas segundo a ordem em que ocorrem, isto é, cmyb-KIXI72V(1), cmyb-KIXI26V(2), cmyb-KIXL43A(3) e cmyb-KIXwt(4)

Fonte: Elaborada pelo autor.

da informação precisa de melhorias em diferentes aspectos. Por exemplo, a estratégia herda a principal limitação da representação por geometria diferencial, isto é, só reflete características associadas ao *backbone* da proteína. Portanto, qualquer informação que não afete a posição dos $C\alpha$ não será refletida nas medidas de informação. Outras representações podem ser utilizadas como base para o cálculo de R_{res} e $R + conf$. Possivelmente, uma representação baseada em superfícies de moléculas pode ser mais adequada, especialmente para análise de eventos de ligação. Além disso, outros aspectos ainda precisam de uma melhor fundamentação teórica. Por exemplo, o número de símbolos utilizados ainda é arbitrário e os valores de R_{conf} são proporcionais ao número de conformações em um *ensemble*. Diferentes números de símbolos foram testados e as mesmas conclusões poderiam ser obtidas para alfabetos a partir de ≈ 15 símbolos. No entanto, um critério objetivo para essa decisão precisa ser estabelecido e bem fundamentado. Para este fim, pretendemos explorar possíveis justificativas para o número de símbolos baseadas em conceitos estabelecidos em teoria da codificação, especificamente os aplicados a compressão de dados. A dependência do número de conformações para os valores de R_{conf} limita a

comparação entre *ensembles* que possuam números próximos de conformações, um método de normalização poderia evitar essa limitação.

Apesar das limitações e significado ainda incerto, a abordagem proposta abre uma nova perspectiva para análise conformacional. Uma vez que um sistema de codificação seja validado como adequado, utilizando a representação por sequências de símbolos, seria possível aplicar diretamente métodos de teoria da comunicação a análise conformacional. Além de permitir explorar o uso desses métodos para solução de problemas já conhecidos, novas perguntas poderiam ser formuladas a partir dessa perspectiva. Por exemplo, existe algum mecanismo em proteínas equivalente aos sistemas de correção de erros em transmissão de dados? Existe um limite de bits de informação que podem ser transmitidos em um evento de ligação? Se sim, qual é a relação desse limite com a estrutura e com a função da proteína? Existe algum paralelo entre a dinâmica de proteínas e as soluções utilizadas rotineiramente em dispositivos de comunicação do nosso cotidiano? Um sistema de codificação adequado e aplicável a todas as proteínas permitirá buscar respostas para perguntas como essas e, apesar das limitações, o método proposto nesta tese poderá ser um primeiro passo nessa direção.

5 CONCLUSÃO

Neste trabalho exploramos a aplicação de uma representação do espaço conformacional de proteínas baseada em descritores de geometria diferencial e de conceitos de teoria da informação a problemas de análise conformacional. Os descritores de geometria diferencial se mostraram eficientes em identificar estados conformacionais, realizar medidas de flexibilidade e representar a dinâmica conformacional de forma intuitiva. Essa representação evita a necessidade de sobreposição de estruturas e as medidas propostas apresentaram resultados igualmente eficientes ou superiores a métodos comumente utilizados, com destaque para aplicação em proteínas não estruturadas e monitoramento de hélices. A principal limitação é não capturar nenhuma informação da dinâmica que não afete a posição dos $C\alpha$. No entanto, a representação é genérica o suficiente para ser aplicável a todos os casos no espectro de flexibilidade de proteínas, apresenta baixo custo computacional associado e permite identificar estados conformacionais de forma fácil e intuitiva. Além de manter as principais vantagens e evitar as principais limitações dos métodos popularmente utilizados.

A abordagem proposta baseada em teoria da informação abre a possibilidade de lidar com a informação espacial de proteínas por uma nova perspectiva. Por exemplo, o sistema de codificação proposto permite tratar a informação espacial das conformações como sequências de símbolos, o que abre a possibilidade de aplicar métodos de análise de sequência para análise de *ensembles* conformacionais de proteínas. Além disso, as duas grandezas de informação definidas capturam características relevantes do espaço conformacional. Nos conjuntos testes, os resíduos que apresentam maiores valores de R_{res} tendem a estar próximos ou entre regiões envolvidas diretamente com interações com outras moléculas associadas a função da proteína. Uma hipótese sobre o papel desses resíduos foi levantada e estudos de mutação podem esclarecer se há e, existindo, qual é o real papel desses resíduos. R_{conf} se mostrou útil para análise de processo de enovelamento, pois permite identificar o estado final enovelado baseado apenas nas probabilidades dos estados observados para cada resíduo. Dessa forma é possível identificar o estado final de enovelamento e a ordem com que as regiões enovelam sem a necessidade de estabelecer métricas ou de comparar com algum estado de referência.

Por fim, a estratégia de análise baseada em teoria da informação apresenta resultados interessantes, mas precisa ser melhor desenvolvida e somente após comparação das previsões com dados experimentais será possível esclarecer o real significado dos resultados obtidos. Por outro lado, a representação baseada em geometria diferencial é adequada para resolver problemas de análises conformacionais e será uma ferramenta útil para biólogos estruturais. Pretendemos disponibilizar o *software* protótipo desenvolvido neste trabalho gratuitamente

para fins não comerciais.

REFERÊNCIAS

- 1 MU, Y.; NGUYEN, P. H.; STOCK, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. **Proteins**, v. 58, n. 1, p. 45–52, 2004.
- 2 KAMADURAI, H. B. et al. Insights into ubiquitin transfer cascades from a structure of a ubch5bubiquitin-hectnedd4l complex. **Molecular Cell**, v. 36, n. 6, p. 1095–1102, 2009.
- 3 TOTO, A. et al. Molecular recognition by templated folding of an intrinsically disordered protein. **Scientific Reports**, v. 6, n. 1, p. 21994, 2016.
- 4 BOWNE-ANDERSON, H. et al. Microtubule dynamic instability: a new model with coupled gtp hydrolysis and multistep catastrophe. **Bioessays**, v. 35, n. 5, p. 452–461, 2013.
- 5 SCHMEING, T. M.; RAMAKRISHNAN, V. What recent ribosome structures have revealed about the mechanism of translation. **Nature**, v. 461, n. 7268, p. 1234–1242, 2009.
- 6 TRITSCH, N. X.; GRANGER, A. J.; SABATINI, B. L. Mechanisms and functions of gaba co-release. **Nature Reviews Neuroscience**, v. 17, n. 3, p. 139–45, 2016.
- 7 GOLDSTEIN, J. L.; BROWN, M. S. A century of cholesterol and coronaries: from plaques to genes to statins. **Cell**, v. 161, n. 1, p. 161–172, 2015.
- 8 ROSSJOHN, J. et al. T cell antigen receptor recognition of antigen-presenting molecules. **Annual Review of Immunology**, v. 33, n. 1, p. 169–200, 2015.
- 9 PAGADALA, N. S.; SYED, K.; TUSZYNSKI, J. Software for molecular docking: a review. **Biophysical Reviews**, v. 9, n. 2, p. 91–102, 2017.
- 10 MONTALVAO, R. W. et al. Structure determination of protein-protein complexes using nmr chemical shifts: case of an endonuclease colicin-immunity protein complex. **Journal of the American Chemical Society**, v. 130, n. 47, p. 15990–6, 2008.
- 11 MONTALVAO, R. W.; SIMONE, A. D.; VENDRUSCOLO, M. Determination of structural fluctuations of proteins from structure-based calculations of residual dipolar couplings. **Journal of Biomolecular NMR**, v. 53, n. 4, p. 281–92, 2012.
- 12 SIMONE, A. D. et al. Characterization of the interdomain motions in hen lysozyme using residual dipolar couplings as replica-averaged structural restraints in molecular dynamics simulations. **Biochemistry**, v. 52, n. 37, p. 6480–6486, 2013.
- 13 SIMONE, A. D.; MONTALVAO, R. W.; VENDRUSCOLO, M. Determination of conformational equilibria in proteins using residual dipolar couplings. **Journal of Chemical Theory and Computation**, v. 7, n. 12, p. 4189–4195, 2011.
- 14 BORKAR, A. N. et al. A method of determining rna conformational ensembles using structure-based calculations of residual dipolar couplings. **Journal of Chemical Physics**, v. 138, n. 21, p. 215103, 2013.

- 15 CAVALLI, A.; MONTALVAO, R. W.; VENDRUSCOLO, M. Using chemical shifts to determine structural changes in proteins upon complex formation. **Journal of Physical Chemistry B**, v. 115, n. 30, p. 9491–4, 2011.
- 16 JANIN, J.; STERNBERG, M. J. Protein flexibility, not disorder, is intrinsic to molecular recognition. **F1000 Biology Reports**, v. 5, p. 2, 2013. Disponível em: <https://f1000.com/prime/reports/b/5/2>. Acesso em: 13 de novembro de 2017.
- 17 KENDREW, J. C. et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. **Nature**, v. 181, n. 4610, p. 662–666, 1958.
- 18 PELLEGRINI, L. et al. Insights into dna recombination from the structure of a rad51–brca2 complex. **Nature**, v. 420, n. 6913, p. 287–293, 2002.
- 19 SCOTT, D. E. et al. Small-molecule inhibitors that target protein-protein interactions in the rad51 family of recombinases. **Chemmedchem**, v. 10, n. 2, p. 296–303, 2015.
- 20 JUBB, H. C. et al. Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. **Progress in Biophysics and Molecular Biology**, v. 128, p. 3–13, 2017. Disponível em: <https://doi.org/10.1016/j.pbiomolbio.2016.10.002>. Acesso em: 13 de novembro de 2017.
- 21 PALAMINI, M.; CANCIANI, A.; FORNERIS, F. Identifying and visualizing macromolecular flexibility in structural biology. **Frontiers in Molecular Biosciences**, v. 3, p. 3–17, 2016. Disponível em: <https://www.frontiersin.org/articles/10.3389/fmolb.2016.00047/full>. Acesso em: 13 de novembro de 2017.
- 22 ADAMS, M. J. et al. Structure-function relationships in lactate dehydrogenase. **Proceedings of the National Academy of Sciences of the United States of America**, v. 70, n. 7, p. 1968–72, 1973.
- 23 WHITE, J. L. et al. A comparison of the structures of apo dogfish m4 lactate dehydrogenase and its ternary complexes. **Journal of Molecular Biology**, v. 102, n. 4, p. 759–79, 1976.
- 24 HUBER, R.; BENNETT, W. S. Functional significance of flexibility in proteins. **Biopolymers**, v. 22, n. 1, p. 261–279, 1983.
- 25 QIN, H.; LIM, L.; SONG, J. Protein dynamics at eph receptor-ligand interfaces as revealed by crystallography, nmr and md simulations. **BMC Biophysics**, v. 5, n. 1, p. 2, 2012.
- 26 WEI, G. et al. Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. **Chemical Reviews**, v. 116, n. 11, p. 6516–6551, 2016.
- 27 MIRSKY, A. E.; PAULIN, L. On the structure of native, denatured, and coagulated proteins. **Proceedings of the National Academy of Sciences of the United States of America**, v. 22, n. 1892, p. 439–447, 1936.
- 28 KARUSH, F. Heterogeneity of the binding sites of bovine serum albumin 1. **Journal of the American Chemical Society**, v. 72, n. 6, p. 2705–2713, 1950.

-
- 29 KOSHLAND, D. E.; RAY, W. J.; ERWIN, M. J. Protein structure and enzyme action. **Federation Proceedings**, v. 17, n. 4, p. 1145–50, 1958.
- 30 MONOD, J.; WYMAN, J.; CHANGEUX, J. P. On the nature of allosteric transitions: a plausible model. **Journal of Molecular Biology**, v. 12, n. 1, p. 88–118, 1965.
- 31 WEBER, G. Energetics of ligand binding to proteins. **Advances in Protein Chemistry**, v. 29, p. 1–83, 1975.
- 32 ENGLANDER, S. W.; KALLENBACH, N. R. Hydrogen exchange and structural dynamics of proteins and nucleic acids. **Quarterly Reviews of Biophysics**, v. 16, n. 04, p. 521, 1983.
- 33 FRAUENFELDER, H.; SLIGAR, S. G.; WOLYNES, P. G. The energy landscapes and motions of proteins. **Science**, v. 254, n. 5038, p. 1598–603, 1991.
- 34 MARTÍNEZ, L. Introducing the levinthal's protein folding paradox and its solution. **Journal of Chemical Education**, v. 91, n. 11, p. 1918–1923, 2014.
- 35 WOLYNES, P. G.; EATON, W. A.; FERSHT, A. R. Chemical physics of protein folding. **Proceedings of the National Academy of Sciences**, v. 109, n. 44, p. 17770–17771, 2012.
- 36 TSAI, C.-J. et al. Folding funnels, binding funnels, and protein function. **Protein Science**, v. 8, n. 6, p. 1181–1190, 1999.
- 37 CSERMELY, P.; PALOTAI, R.; NUSSINOV, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. **Trends in Biochemical Sciences**, v. 35, n. 10, p. 539–546, 2010.
- 38 TOTO, A. et al. The mechanism of binding of the kix domain to the mixed lineage leukemia protein and its allosteric role in the recognition of c-myc. **Protein Science**, v. 23, n. 7, p. 962–969, 2014.
- 39 DILL, K. A.; MACCALLUM, J. L. The protein-folding problem, 50 years on. **Science**, v. 338, n. 6110, p. 1042–1046, 2012.
- 40 PENG, Z.; MIZIANTY, M. J.; KURGAN, L. Genome-scale prediction of proteins with long intrinsically disordered regions. **Proteins**, v. 82, n. 1, p. 145–158, 2014.
- 41 JAVIER ZEA, D. et al. Protein conformational diversity correlates with evolutionary rate. **Molecular Biology and Evolution**, v. 30, n. 7, p. 1500–1503, 2013.
- 42 PARISI, G. et al. Conformational diversity and the emergence of sequence signatures during evolution. **Current Opinion in Structural Biology**, v. 32, p. 58–65, 2015.
Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0959440X15000147?via%3Dihub>>. Acesso em: 13 de novembro de 2017.
- 43 SALDAÑO, T. E. et al. Evolutionary conserved positions define protein conformational diversity. **Plos Computational Biology**, v. 12, n. 3, p. e1004775, 2016.
- 44 CAMPBELL, E. et al. The role of protein dynamics in the evolution of new enzyme function. **Nature Chemical Biology**, v. 12, n. 11, p. 944–950, 2016.

- 45 MA, B.; NUSSINOV, R. Protein dynamics: conformational footprints. **Nature Chemical Biology**, v. 12, n. 11, p. 890–891, 2016.
- 46 WHITNEY, D. S.; VOLKMAN, B. F.; PREHODA, K. E. Evolution of a protein interaction domain family by tuning conformational flexibility. **Journal of the American Chemical Society**, v. 138, n. 46, p. 15150–15156, 2016. Disponível em: <<http://pubs.acs.org/doi/10.1021/jacs.6b05954>>. Acesso em: 13 de novembro de 2017.
- 47 UVERSKY, V. N. The triple power of d³: protein intrinsic disorder in degenerative diseases. **Frontiers in Bioscience**, v. 19, n. 1, p. 181–258, 2014.
- 48 HIGUERUELO, A. P.; JUBB, H.; BLUNDELL, T. L. Protein-protein interactions as druggable targets: recent technological advances. **Current Opinion in Pharmacology**, v. 13, n. 5, p. 791–796, 2013.
- 49 WILSON, C. et al. Using ancient protein kinases to unravel a modern cancer drug's mechanism. **Science**, v. 347, n. 6224, p. 882–886, 2015.
- 50 AGAFONOV, R. V.; WILSON, C.; KERN, D. Evolution and intelligent design in drug development. **Frontiers in Molecular Biosciences**, v. 2, n. 5, p. 27, 2015.
- 51 HOSPITAL, A. et al. Molecular dynamics simulations: advances and applications. **Advances and Applications in Bioinformatics and Chemistry**, v. 8, p. 37–47, 2015. Disponível em: <<https://doi.org/10.2147/AABC.S70333>>. Acesso em: 13 de novembro de 2017.
- 52 MORI, T. et al. Molecular dynamics simulations of biological membranes and membrane proteins using enhanced conformational sampling algorithms. **Biochimica et Biophysica Acta**, v. 1858, n. 7, p. 1635–1651, 2016.
- 53 HERTIG, S.; LATORRACA, N. R.; DROR, R. O. Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. **Plos Computational Biology**, v. 12, n. 6, p. e1004746, 2016.
- 54 GU, R.-X. et al. Ganglioside-Lipid and ganglioside-protein interactions revealed by coarse-grained and atomistic molecular dynamics simulations. **Journal of Physical Chemistry B**, v. 121, n. 15, p. 3262–3275, 2017.
- 55 KOVERMANN, M.; ROGNE, P.; WOLF-WATZ, M. Protein dynamics and function from solution state nmr spectroscopy. **Quarterly Reviews of Biophysics**, v. 49, n. 5, p. e6, 2016.
- 56 MAISURADZE, G. G.; LIWO, A.; SCHERAGA, H. A. Principal component analysis for protein folding dynamics. **Journal of Molecular Biology**, v. 385, n. 1, p. 312–329, 2009.
- 57 SITTEL, F.; JAIN, A.; STOCK, G. Principal component analysis of molecular dynamics: on the use of cartesian vs. internal coordinates. **Journal of Chemical Physics**, v. 141, n. 1, 2014. Disponível em: <<http://aip.scitation.org/doi/10.1063/1.4885338>>. Acesso em: 13 de novembro de 2017.
- 58 GENDOO, D. M. A.; HARRISON, P. M. The landscape of the prion protein's structural response to mutation revealed by principal component analysis of multiple nmr ensembles. **Plos Computational Biology**, v. 8, n. 8, p. e1002646, 2012.

-
- 59 OKAZAKI, K. I.; TAKADA, S. Structural comparison of f1-atpase: interplay among enzyme structures, catalysis, and rotations. **Structure**, v. 19, n. 4, p. 588–598, 2011.
- 60 KOLODNY, R.; KOEHL, P.; LEVITT, M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. **Journal of Molecular Biology**, v. 346, n. 4, p. 1173–1188, 2005.
- 61 HASEGAWA, H.; HOLM, L. Advances and pitfalls of protein structural alignment. **Current Opinion in Structural Biology**, v. 19, n. 3, p. 341–348, 2009.
- 62 ZHAO, Y. et al. A fast parallel clustering algorithm for molecular simulation trajectories. **Journal of Computational Chemistry**, v. 34, n. 2, p. 95–104, 2013.
- 63 ASLAM, N. et al. The accuracy of protein structure alignment servers. **Electronic Journal of Biotechnology**, v. 20, p. 9–13, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0717345816000208?via%3Dihub>. Acesso em: 13 de novembro de 2017.
- 64 RITCHIE, D. W. Calculating and scoring high quality multiple flexible protein structure alignments. **Bioinformatics**, v. 32, n. 17, p. 2650–2658, 2016.
- 65 SIEW, N. et al. MaxSub: an automated measure for the assessment of protein structure prediction quality. **Bioinformatics**, v. 16, n. 9, p. 776–85, 2000.
- 66 ZHANG, Y.; SKOLNICK, J. TM-align: a protein structure alignment algorithm based on the tm-score. **Nucleic Acids Research**, v. 33, n. 7, p. 2302–9, 2005.
- 67 XU, J.; ZHANG, Y. How significant is a protein structure similarity with tm-score = 0.5? **Bioinformatics**, v. 26, n. 7, p. 889–895, 2010.
- 68 LAIO, A.; GERVASIO, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. **Reports on Progress in Physics**, v. 71, n. 12, p. 126601, 2008.
- 69 FIORIN, G.; KLEIN, M. L.; HÉNIN, J. Using collective variables to drive molecular dynamics simulations. **Molecular Physics**, v. 111, n. 22-23, p. 3345–3362, 2013.
- 70 RAMACHANDRAN, G. N.; SASISEKHARAN, V. Conformation of polypeptides and proteins. **Advances in Protein Chemistry**, v. 23, p. 283–438, 1968. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/4882249>. Acesso em: 13 de novembro de 2017.
- 71 BYSTROFF, C.; BAKER, D. Prediction of local structure in proteins using a library of sequence-structure motifs. **Journal of Molecular Biology**, v. 281, n. 3, p. 565–577, 1998.
- 72 HINSEN, K. Comment on: “energy landscape of a small peptide revealed by dihedral angle principal component analysis”. **Proteins**, v. 64, n. 3, p. 795–797, 2006.
- 73 RACKOVSKY, S.; SCHERAGA, H. A. Differential geometry and polymer conformation. 1. comparison of protein conformations 1a,b. **Macromolecules**, v. 11, n. 6, p. 1168–1174, 1978.
- 74 HAGLER, A. T.; HONIG, B. On the formation of protein tertiary structure on a computer. **Proceedings of the National Academy of Sciences of the United States of America**, v. 75, n. 2, p. 554–8, 1978.

- 75 LOUIE, A.; SOMORJAI, R.; KLUG, A. Differential geometry of proteins. **Journal of Molecular Biology**, v. 168, n. 1, p. 143–162, 1983.
- 76 LOUIE, A. H.; SOMORJAI, R. L. Differential geometry of proteins: a structural and dynamical representation of patterns. **Journal of Theoretical Biology**, v. 98, n. 2, p. 189–209, 1982.
- 77 RACKOVSKY, S.; GOLDSTEIN, D. A. Differential geometry and protein conformation. v. medium-range conformational influence of the individual amino acids. **Biopolymers**, v. 26, n. 7, p. 1163–87, 1987.
- 78 DA-FU, D.; JIANG, Q.; ZU-KANG, F. A differential geometric treatment of protein structure comparison. **Bulletin of Mathematical Biology**, v. 56, n. 5, p. 923–943, 1994.
- 79 MONTALVÃO, R. W. et al. CHORAL: a differential geometry approach to the prediction of the cores of protein structures. **Bioinformatics**, v. 21, n. 19, p. 3719–25, 2005.
- 80 RAGANATHAN, S. et al. **Automated and accurate protein structure description**, 2008. Disponível em: <<http://arxiv.org/abs/0811.3587>>. Acesso em: 13 de novembro de 2017.
- 81 HAUSRATH, A. C.; GORIELY, A. Repeat protein architectures predicted by a continuum representation of fold space. **Protein Science**, v. 15, n. 4, p. 753–760, 2006.
- 82 _____. Continuous representations of proteins: construction of coordinate models from curvature profiles. **Journal of Structural Biology**, v. 158, n. 3, p. 267–281, 2007.
- 83 FEOLI, A.; NESTERENKO, V. V.; SCARPETTA, G. Functionals linear in curvature and statistics of helical proteins. **Nuclear Physics B**, v. 705, n. 3, p. 577–592, 2005.
- 84 THAMWATTANA, N.; MCCOY, J. A.; HILL, J. M. Energy density functions for protein structures. **Quarterly Journal of Mechanics and Applied Mathematics**, v. 61, n. 3, p. 431–451, 2008.
- 85 GORIELY, A.; HAUSRATH, A.; NEUKIRCH, S. The differential geometry of proteins and its applications to structure determination. **Biophysical Reviews and Letters**, v. 03, n. 01n02, p. 77–101, 2008.
- 86 LEUNG, H. et al. ARABESQUE: a tool for protein structural comparison using differential geometry and knot theory. **World Research Journal of Peptide and Protein**, v. 1, n. 1, p. 33–40, 2012.
- 87 PITT, W. R.; MONTALVÃO, R. W.; BLUNDELL, T. L. Polyphony: superposition independent methods for ensemble-based drug discovery. **BMC Bioinformatics**, v. 15, n. 1, p. 324, 2014.
- 88 SHANNON, C. E. A mathematical theory of communication. **Bell System Technical Journal**, v. 27, n. 1, p. 379–423, 1948.
- 89 TRIBUS, M.; MCIRVINE, E. C. Energy and information. **Scientific American**, v. 225, n. 3, p. 179–188, 1971.

- 90 SCHNEIDER, T. D. A brief review of molecular information theory. **Nano Communication Networks**, v. 1, n. 3, p. 173–180, 2010.
- 91 SHETH, N. et al. Comprehensive splice-site analysis using comparative genomics. **Nucleic Acids Research**, v. 34, n. 14, p. 3955–3967, 2006.
- 92 MACHTA, J. Entropy, information, and computation. **American Journal of Physics**, v. 67, n. 12, p. 1074–1077, 1999.
- 93 SARKAR, S. Decoding "coding": information and dna. **Bioscience**, v. 46, n. 11, p. 857–864, 1996.
- 94 SMITH, J. M. The concept of information in biology. **Philosophy of Science**, v. 67, n. 2, p. 177–194, 2000.
- 95 BARBIERI, M. What is information? **Philosophical Transactions of the Royal Society A**, v. 374, n. 2063, p. 20150060, 2016.
- 96 RASHEVSKY, N. Life, information theory, and topology. **Bulletin of Mathematical Biophysics**, v. 17, n. 3, p. 229–235, 1955.
- 97 _____. Life, information theory, probability, and physics. **Bulletin of Mathematical Biophysics**, v. 22, n. 4, p. 351–364, 1960.
- 98 ASTLER, H. Q. U. Information theory in radiobiology. **Annual Review of Nuclear Science**, v. 8, n. 1, p. 387–400, 1958.
- 99 YOCKEY, H. P. An application of information theory to the central dogma and the sequence hypothesis. **Journal of Theoretical Biology**, v. 46, n. 2, p. 369–406, 1974.
- 100 _____. Self organization origin of life scenarios and information theory. **Journal of Theoretical Biology**, v. 91, n. 1, p. 13–31, 1981.
- 101 _____. An application of information theory to the physics of tissue damage. **Radiation Research**, v. 5, n. 2, p. 146, 1956.
- 102 SCHNEIDER, T. D. et al. Information content of binding sites on nucleotide sequences. **Journal of Molecular Biology**, v. 188, n. 3, p. 415–431, 1986.
- 103 BENOS, P. V.; BULYK, M. L.; STORMO, G. D. Additivity in protein-dna interactions: how good an approximation is it? **Nucleic Acids Research**, v. 30, n. 20, p. 4442–51, 2002.
- 104 SCHNEIDER, T. D.; STEPHENS, R. M. Sequence logos: a new way to display consensus sequences. **Nucleic Acids Research**, v. 18, n. 20, p. 6097–6100, 1990.
- 105 LYAKHOV, I. G.; KRISHNAMACHARI, A.; SCHNEIDER, T. D. Discovery of novel tumor suppressor p53 response elements using information theory. **Nucleic Acids Research**, v. 36, n. 11, p. 3828–3833, 2008.
- 106 SCHNEIDER, T. D. 70% efficiency of bistate molecular machines explained by information theory, high dimensional geometry and evolutionary convergence. **Nucleic Acids Research**, v. 38, n. 18, p. 5995–6006, 2010.

- 107 _____. Evolution of biological information. **Nucleic Acids Research**, v. 28, n. 14, p. 2794–2799, 2000.
- 108 _____. Theory of molecular machines. ii. energy dissipation from molecular machines. **Journal of Theoretical Biology**, v. 148, n. 1, p. 125–137, 1991.
- 109 _____. Theory of molecular machines. i. channel capacity of molecular machines. **Journal of Theoretical Biology**, v. 148, n. 1, p. 83–123, 1991.
- 110 HARTE, J.; NEWMAN, E. A. Maximum information entropy: a foundation for ecological theory. **Trends in Ecology and Evolution**, v. 29, n. 7, p. 384–389, 2014.
- 111 MOUSAVIAN, Z.; DÍAZ, J.; MASOUDI-NEJAD, A. Information theory in systems biology. part ii: protein–protein interaction and signaling networks. **Seminars in Cell & Developmental Biology**, v. 51, p. 14–23, 2016. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1084952115300203?via%3Dihub>.
- 112 MOUSAVIAN, Z.; KAVOUSI, K.; MASOUDI-NEJAD, A. Information theory in systems biology. part i: gene regulatory and metabolic networks. **Seminars in Cell & Developmental Biology**, v. 51, p. 3–13, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1084952115300197?via%3Dihub>. Acesso em: 13 de novembro de 2017.
- 113 BYEON, J.-Y.; LEE, J. Protein contact prediction by using information theory. **Journal of the Korean Physical Society**, v. 70, n. 9, p. 876–879, 2017.
- 114 COLLIER, J. H. et al. Statistical inference of protein structural alignments using information and compression. **Bioinformatics**, v. 33, n. January, p. btw757, 2017.
- 115 SCHNEIDER, T. D. Information content of individual genetic sequences. **Journal of Theoretical Biology**, v. 189, n. 4, p. 427–441, 1997.
- 116 OTU, H. H.; SAYOOD, K. A divide-and-conquer approach to fragment assembly. **Bioinformatics**, v. 19, n. 1, p. 22–29, 2003.
- 117 KURUOGLU, E. E.; ARNDT, P. F. The information capacity of the genetic code: is the natural code optimal? **Journal of Theoretical Biology**, v. 419, p. 227–237, 2017. Disponível em: <https://doi.org/10.1016/j.jtbi.2017.01.046>. Acesso em: 13 de novembro de 2017.
- 118 GIBRAT, J.-F.; GAMIER, J.; ROBSONT, B. Further developments of protein secondary structure prediction using information theory new parameters and consideration of residue pairs. **Journal of Molecular Biology**, v. 198, p. 425–443, 1987. Disponível em: <https://www.sciencedirect.com/science/article/pii/0022283687902920?via%3Dihub>. Acesso em: 13 de novembro de 2017.
- 119 ALTSCHUL, S. F. Amino acid substitution matrices from an information theoretic perspective. **Journal of Molecular Biology**, v. 219, n. 3, p. 555–565, 1991.
- 120 KORBER, B. T. et al. Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. **Proceedings of the National Academy of Sciences**, v. 90, n. 15, p. 7176–7180, 1993.

-
- 121 KANNAN, N.; SCHNEIDER, T. D.; VISHVESHWARA, S. Logos for amino-acid preferences in different backbone packing density regions of protein structural classes. **Acta Crystallographica Section D**, v. 56, n. 9, p. 1156–1165, 2000.
- 122 ROMERO, P. et al. Sequence complexity of disordered protein. **Proteins**, v. 42, n. 1, p. 38–48, 2001.
- 123 LIU, Y.; BAHAR, I. Sequence evolution correlates with structural dynamics. **Molecular Biology and Evolution**, v. 29, n. 9, p. 2253–2263, 2012.
- 124 MORCOS, F. et al. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. **Proceedings of the National Academy of Sciences**, v. 111, n. 34, p. 12408–12413, 2014.
- 125 NOEL, J. K.; MORCOS, F.; ONUCHIC, J. N. Sequence co-evolutionary information is a natural partner to minimally-frustrated models of biomolecular dynamics. **F1000Research**, v. 5, n. 0, p. 1–7, 2016. Disponível em: <<https://f1000research.com/articles/5-106/v1>>. Acesso em: 13 de novembro de 2017.
- 126 STEFANI, A. G. et al. Application of methods from information theory in protein-interaction analysis. In: _____. **Information- and communication theory in molecular biology**. Cham: Springer International Publishing, 2017. (Lecture Notes in Bioengineering), cap. 13, p. 293–313.
- 127 LENAERTS, T. et al. Quantifying information transfer by protein domains: analysis of the fyn sh2 domain structure. **BMC Structural Biology**, v. 8, n. 1, p. 43, 2008.
- 128 LEVINE, M. V.; WEINSTEIN, H. NbIT - a new information theory-based analysis of allosteric mechanisms reveals residues that underlie function in the leucine transporter leut. **Plos Computational Biology**, v. 10, n. 5, p. e1003603, 2014.
- 129 BISHOP, R. L. There is more than one way to frame a curve. **American Mathematical Monthly**, v. 82, n. 3, p. 246–251, 1975.
- 130 THEODORIDIS, S.; KOUTROUMBAS, K. Proximity measures. In: _____. **Pattern recognition**. 4th. ed. Burlington: Elsevier Inc., 2009. cap. 11, p. 602.
- 131 SHIMAZAKI, H.; SHINOMOTO, S. Kernel bandwidth optimization in spike rate estimation. **Journal of Computational Neuroscience**, v. 29, n. 1-2, p. 171–182, 2010.
- 132 CAMPELLO, R. J. G. B. et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. **ACM Transactions on Knowledge Discovery From Data**, v. 10, n. 1, p. 1–51, 2015.
- 133 CHANG, P. L.; RINNE, A. W.; DEWEY, T. G. Structure alignment based on coding of local geometric measures. **BMC Bioinformatics**, v. 7, p. 346, 2006. Disponível em: <<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-346>>. Acesso em: 13 de novembro de 2017.
- 134 YAU, R.; RAPE, M. The increasing complexity of the ubiquitin code. **Nature Cell Biology**, v. 18, n. 6, p. 579–586, 2016.

- 135 KERSCHER, O.; FELBERBAUM, R.; HOCHSTRASSER, M. Modification of proteins by ubiquitin and ubiquitin-like proteins. **Annual Review of Cell and Developmental Biology**, v. 22, n. 1, p. 159–180, 2006.
- 136 BERNDSEN, C. E.; WOLBERGER, C. New insights into ubiquitin e3 ligase mechanism. **Nature Structural & Molecular Biology**, v. 21, n. 4, p. 301–307, 2014.
- 137 BERGVALL, M.; MELENDY, T.; ARCHAMBAULT, J. The e1 proteins. **Virology**, v. 445, n. 1-2, p. 35–56, 2013.
- 138 ENEMARK, E. J.; STENLUND, A.; JOSHUA-TOR, L. Crystal structures of two intermediates in the assembly of the papillomavirus replication initiation complex. **EMBO Journal**, v. 21, n. 6, p. 1487–96, 2002.
- 139 CHEN, G.; STENLUND, A. The e1 initiator recognizes multiple overlapping sites in the papillomavirus origin of dna replication. **Journal of Virology**, v. 75, n. 1, p. 292–302, 2001.
- 140 ENEMARK, E. J. et al. Crystal structure of the dna binding domain of the replication initiation protein e1 from papillomavirus. **Molecular Cell**, v. 6, n. 1, p. 149–158, 2000.
- 141 AUSTER, A. S.; JOSHUA-TOR, L. The dna-binding domain of human papillomavirus type 18 e1. crystal structure, dimerization, and dna binding. **Journal of Biological Chemistry**, v. 279, n. 5, p. 3733–3742, 2004.
- 142 ROY, A.; KUCUKURAL, A.; ZHANG, Y. I-TASSER: a unified platform for automated protein structure and function prediction. **Nature Protocols**, v. 5, n. 4, p. 725–38, 2010.
- 143 EMSLEY, P.; COWTAN, K. Coot: model-building tools for molecular graphics. **Acta Crystallographica Section D**, v. 60, n. Pt 1, p. 2126–32, 2004.
- 144 CHEN, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. **Acta Crystallographica Section D**, v. 66, n. Pt 1, p. 12–21, 2010.
- 145 SCHRÖDINGER LLC. The PyMOL molecular graphics system, version 1.8. 2015. Disponível em: <<https://pymol.org/2/support.html>>. Acesso em: 13 de novembro de 2017.
- 146 ABRAHAM, M. J. et al. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. **SoftwareX**, v. 1–2, p. 19 – 25, 2015. ISSN 2352-7110. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2352711015000059>>. Acesso em: 13 de novembro de 2017.
- 147 SCHMID, N. et al. Definition and testing of the gromos force-field versions 54a7 and 54b7. **European Biophysics Journal**, v. 40, n. 7, p. 843–856, 2011.
- 148 BUSSI, G.; DONADIO, D.; PARRINELLO, M. Canonical sampling through velocity rescaling. **Journal of Chemical Physics**, v. 126, n. 1, 2007.
- 149 NOSÉ, S.; KLEIN, M. Constant pressure molecular dynamics for molecular systems. **Molecular Physics**, v. 50, n. 5, p. 1055–1076, 1983.

-
- 150 RADHAKRISHNAN, I. et al. Solution structure of the kix domain of cbp bound to the transactivation domain of creb: a model for activator:coactivator interactions. **Cell**, v. 91, n. 6, p. 741–752, 1997.
- 151 BEAL, R. E. et al. The hydrophobic effect contributes to polyubiquitin chain recognition †. **Biochemistry**, v. 37, n. 9, p. 2925–2934, 1998.
- 152 SLOPER-MOULD, K. E. et al. Distinct functional surface regions on ubiquitin. **Journal of Biological Chemistry**, v. 276, n. 32, p. 30483–30489, 2001.
- 153 ROSCOE, B. P. et al. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. **Journal of Molecular Biology**, v. 425, n. 8, p. 1363–1377, 2013.
- 154 GENHEDEN, S.; RYDE, U. Will molecular dynamics simulations of proteins ever reach equilibrium? **Physical Chemistry Chemical Physics**, v. 14, p. 8662–8677, 2012. Disponível em: <<http://dx.doi.org/10.1039/C2CP23961B>>. Acesso em: 13 de novembro de 2017.
- 155 GRÜNBERG, R.; LECKNER, J.; NILGES, M. Complementarity of structure ensembles in protein-protein binding. **Structure**, v. 12, n. 12, p. 2125–2136, 2004.
- 156 ROHS, R. et al. Origins of specificity in protein-dna recognition. **Annual Review of Biochemistry**, v. 79, n. 1, p. 233–269, 2010. PMID: 20334529. Disponível em: <<http://www.annualreviews.org/doi/10.1146/annurev-biochem-060408-091030>>. Acesso em: 13 de novembro de 2017.
- 157 SNEPPEN, K.; ZOCCHI, G. Physics of genetic regulation: the lambda-phage in e. coli. In: _____. **Physics in molecular biology**. New York: Cambridge University Press, 2005. cap. 7, p. 170–176.
- 158 ABBENA, E.; SALAMON, S.; GRAY, A. Curves in space. In: _____. **Modern differential geometry of curves and surfaces with mathematica**. 3rd. ed. New York: ChapMan and Hall/CRC, 2006. cap. 7, p. 191–212.
- 159 _____. Construction of space curves. In: _____. **Modern differential geometry of curves and surfaces with mathematica**. 3rd. ed. New York: ChapMan and Hall/CRC, 2006. cap. 8, p. 229–249.

Apêndices

APÊNDICE A – CÁLCULO DA CURVATURA EM FUNÇÃO DE UMA CURVA $\gamma(t)$

A demonstração da [Equação 3.12](#) pode ser encontrada em vários livros textos introdutórios sobre geometria diferencial.¹⁵⁸

Teorema A.0.1. *Dada uma curva $\gamma(t) : I \rightarrow R^3$, a curvatura κ em função de um parâmetro é dada por*

$$\kappa(t) = \frac{\|\gamma'(t) \times \gamma''(t)\|}{\|\gamma'(t)\|^3} \quad (\text{A.1})$$

Demonstração. Considerando que

$$\vec{T}(t) = \frac{\gamma'(t)}{\|\gamma'(t)\|} \quad (\text{A.2})$$

$$\|\gamma'(t)\| = \frac{ds}{dt} \quad (\text{A.3})$$

teremos

$$\gamma'(t) = \|\gamma'(t)\| \vec{T}(t) = \frac{ds}{dt} \vec{T}(t) \quad (\text{A.4})$$

$$\gamma''(t) = \frac{ds'}{dt} \vec{T}(t) + \frac{ds}{dt} \vec{T}'(t) \quad (\text{A.5})$$

Logo,

$$\gamma'(t) \times \gamma''(t) = \left(\frac{ds}{dt}\right)^2 (\vec{T} \times \vec{T}'). \quad (\text{A.6})$$

Sendo $\|\vec{T}\| = 1, \forall t, \langle \vec{T}(t), \vec{T}'(t) \rangle = 0$ e $\|\vec{T}(t) \times \vec{T}'(t)\| = \|\vec{T}(t)\| \|\vec{T}'(t)\|$, obteremos

$$\|\gamma'(t) \times \gamma''(t)\| = \left(\frac{ds}{dt}\right)^2 \|\vec{T}(t) \times \vec{T}'(t)\| \quad (\text{A.7})$$

$$\begin{aligned} \|\gamma'(t) \times \gamma''(t)\| &= \left(\frac{ds}{dt}\right)^2 \|\vec{T}(t) \times \vec{T}'(t)\| \\ &= \left(\frac{ds}{dt}\right)^2 \|\vec{T}(t)\| \|\vec{T}'(t)\| \\ &= \left(\frac{ds}{dt}\right)^2 \|\vec{T}'(t)\| \end{aligned} \quad (\text{A.8})$$

Portanto,

$$\|\vec{T}'(t)\| = \frac{\|\gamma'(t) \times \gamma''(t)\|}{\left(\frac{ds}{dt}\right)^2} = \frac{\|\gamma'(t) \times \gamma''(t)\|}{\|\gamma'(t)\|^2} \quad (\text{A.9})$$

Podemos expressar a curvatura em função de um parâmetro t , seguindo o seguinte raciocínio

$$\frac{d\vec{T}}{dt} = \frac{d\vec{T}}{ds} \frac{ds}{dt} \quad (\text{A.10})$$

e

$$\frac{d\vec{T}}{ds} = \frac{\frac{d\vec{T}}{dt}}{\frac{ds}{dt}} = \frac{\frac{d\vec{T}}{ds}}{\left\|\frac{d\gamma}{dt}\right\|} \quad (\text{A.11})$$

Logo,

$$\kappa(t) = \left\|\frac{d\vec{T}}{ds}\right\| = \left\|\frac{\frac{d\vec{T}}{dt}}{\frac{d\gamma}{dt}}\right\| \quad (\text{A.12})$$

$$\kappa(t) = \frac{\|\vec{T}'(t)\|}{\|\gamma'(t)\|} \quad (\text{A.13})$$

Consequentemente, a curvatura κ será

$$\kappa(t) = \frac{\|\vec{T}'(t)\|}{\|\gamma'(t)\|} = \frac{\|\gamma'(t) \times \gamma''(t)\|}{\|\gamma'(t)\|^3} \quad (\text{A.14})$$

□

APÊNDICE B – CÁLCULO DA TORÇÃO EM FUNÇÃO DE UMA CURVA $\gamma(t)$

Demonstração da equação [Equação 3.26](#), também pode ser encontrada em vários livros textos sobre geometria diferencial.¹⁵⁸

Teorema B.0.1. *Dada uma curva $\gamma(t) : I \rightarrow R^3$, a qual $\kappa \neq 0$. A torção em função dessa curva paramétrica será dada por*

$$\tau(t) = \frac{\langle (\gamma'(t) \times \gamma''(t), \gamma'''(t)) \rangle}{\|\gamma'(t) \times \gamma''(t)\|^2} \quad (\text{B.1})$$

Demonstração. Considerando que

$$\vec{T}(s) = \frac{d\gamma}{ds} = \frac{d\gamma}{dt} \frac{dt}{ds} \quad (\text{B.2})$$

Sendo, $\frac{dt}{ds} = \dot{t}$,

$$\vec{T}(s) = \gamma'(t)\dot{t} \quad (\text{B.3})$$

Logo,

$$\vec{T}'(s) = \frac{d}{ds}(\gamma'(t)\dot{t}) = \gamma'(t)\ddot{t} + \gamma''(t)\dot{t}^2 \quad (\text{B.4})$$

$$\vec{T}''(s) = \frac{d}{ds}(\gamma'(t)\ddot{t} + \gamma''(t)\dot{t}^2) = \gamma'(t)\dddot{t} + 3\gamma''(t)\dot{t}\ddot{t} + \gamma'''(t)\dot{t}^3 \quad (\text{B.5})$$

O produto interno $\vec{T}(s) \times \vec{T}'(s)$ será

$$\begin{aligned} \vec{T}(s) \times \vec{T}'(s) &= (\gamma'(t)\dot{t}) \times (\gamma'(t)\ddot{t} + \gamma''(t)\dot{t}^2) \\ &= \dot{t}\ddot{t}(\gamma'(t) \times \gamma'(t)) + \dot{t}^3(\gamma'(t) \times \gamma''(t)) \\ &= \dot{t}^3(\gamma'(t) \times \gamma''(t)) \end{aligned} \quad (\text{B.6})$$

Portanto,

$$\begin{aligned} \langle (\vec{T}(t) \times \vec{T}'(t)), \vec{T}''(t) \rangle &= \langle \dot{t}^3(\gamma'(t) \times \gamma''(t)) \times (\gamma'(t)\ddot{t} + 3\gamma''(t)\dot{t}\ddot{t} + \gamma'''(t)\dot{t}^3) \rangle \\ &= \ddot{t}\dot{t}^3 \langle (\gamma'(t) \times \gamma''(t)), \gamma'(t) \rangle + 3\ddot{t}\dot{t}^4 \langle (\gamma'(t) \times \gamma''(t)), \gamma''(t) \rangle \\ &\quad + \dot{t}^6 \langle (\gamma'(t) \times \gamma''(t)), \gamma'''(t) \rangle \end{aligned} \quad (\text{B.7})$$

Considerando que

$$\langle (\gamma'(t) \times \gamma''(t)), \gamma'(t) \rangle = \langle (\gamma'(t) \times \gamma''(t)), \gamma''(t) \rangle = 0 \quad (\text{B.8})$$

a [Equação B.7](#) será

$$\langle (\vec{T}(s) \times \vec{T}'(s)), \vec{T}''(s) \rangle = \dot{t}^6 \langle (\gamma'(t) \times \gamma''(t)), \gamma'''(t) \rangle \quad (\text{B.9})$$

Observe que

$$\dot{t} = \frac{dt}{ds} = \frac{1}{\frac{ds}{dt}} = \frac{1}{\|\gamma'(t)\|} \quad (\text{B.10})$$

Portanto,

$$\langle (\vec{T}(s) \times \vec{T}'(s)), \vec{T}''(s) \rangle = \frac{\langle (\gamma'(t) \times \gamma''(t)), \gamma'''(t) \rangle}{\|\gamma'(t)\|^6} \quad (\text{B.11})$$

Antes de prosseguir, observe que

$$\begin{aligned} \vec{T} \times \vec{T}' &= \vec{T} \times \kappa \vec{N} = \kappa (\vec{T} \times \vec{N}) \\ \langle (\vec{T} \times \vec{T}'), \vec{T}'' \rangle &= \langle \kappa (\vec{T} \times \vec{N}), -\kappa \vec{T} + \kappa' \vec{N} + \tau \vec{B} \rangle \\ &= -\kappa^3 \langle (\vec{T} \times \vec{T}), \vec{N} \rangle + \kappa^2 \tau \langle (\vec{T} \times \vec{N}), \vec{B} \rangle + \kappa' \langle (\vec{T} \times \vec{N}), \vec{N} \rangle \\ &= \kappa^2 \tau \langle \vec{B}, \vec{B} \rangle \\ \langle (\vec{T} \times \vec{T}'), \vec{T}'' \rangle &= \kappa^2 \tau \end{aligned} \quad (\text{B.12})$$

Rearranjando a [Equação B.12](#), teremos

$$\tau(t) = \frac{\langle (\vec{T} \times \vec{T}'), \vec{T}'' \rangle}{\kappa^2(t)} \quad (\text{B.13})$$

De acordo com a [Equação B.11](#) e com o teorema demonstrado no [Apêndice A](#) ($\kappa(t) = \frac{\|\gamma'(t) \times \gamma''(t)\|}{\|\gamma'(t)\|^3}$),

$$\tau(t) = \frac{\langle (\vec{T}(t) \times \vec{T}'(t)), \vec{T}''(t) \rangle}{\kappa^2(t)} \quad (\text{B.14})$$

$$\tau(t) = \frac{\langle (\gamma'(t) \times \gamma''(t)), \gamma'''(t) \rangle}{\kappa^2(t) \|\gamma'(t)\|^6} \quad (\text{B.15})$$

$$\tau(t) = \frac{\langle (\gamma'(t) \times \gamma''(t)), \gamma'''(t) \rangle}{\|\gamma'(t) \times \gamma''(t)\|^2} \quad (\text{B.16})$$

□

APÊNDICE C – PROVA DO TEOREMA FUNDAMENTAL DAS CURVAS ESPACIAIS

O *Teorema Fundamental das Curvas Espaciais*(TFCE) é o princípio que fundamenta o uso dos valores de curvatura, κ , e torção, τ , para comparar *backbones* de conformações proteínas. A prova do TFCE pode ser encontrada em diversos livros texto sobre geometria diferencial, para este apêndice reproduzimos a demonstração conforme desenvolvida no livro a *Modern Differential Geometry of Curves and Surfaces with Mathematica*.¹⁵⁹ O TFCE fundamenta as seguintes afirmações feitas neste trabalho: 1) podemos caracterizar uma conformação de um *backbone* usando apenas os valores de κ e τ 2) sem a necessita de sobreposição estrutural. Para que a afirmação 2) seja verdade, precisamos demonstrar que κ e τ não variam com movimentos euclidianos em R^3 . Para essa demonstração precisaremos inicialmente de um fato sobre produto vetorial triplo para poder determinar o efeito do movimento euclidiano em κ , τ e no comprimento de arco, s .

Lemma C.0.1. *Seja $A : R^3 \rightarrow R^3$ um mapeamento linear e $\mathbf{a}, \mathbf{b}, \mathbf{c} \in R^3$. Logo,*

$$[(A\mathbf{a})(A\mathbf{b})(A\mathbf{c})] = \det(A)[\mathbf{abc}] \quad (\text{C.1})$$

Demonstração. Considerando $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$ como vetores coluna, A será uma matriz que multiplica esses vetores. Considere $(\mathbf{a}|\mathbf{b}|\mathbf{c})$ sendo a matriz 3×3 cujas colunas são $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$. Logo,

$$[(A\mathbf{a})(A\mathbf{b})(A\mathbf{c})] = A\mathbf{a} \cdot (A\mathbf{b} \times A\mathbf{c}) = \det(A\mathbf{a}|A\mathbf{b}|A\mathbf{c}) \quad (\text{C.2})$$

$$\det(A\mathbf{a}|A\mathbf{b}|A\mathbf{c}) = \det(A(\mathbf{a}|\mathbf{b}|\mathbf{c})) = \det(A)[\mathbf{abc}] \quad (\text{C.3})$$

□

Teorema C.0.2. *A curvatura e o valor absoluto da torção são invariantes em relação a movimentos Euclidianos em R^3 , mas a torção pode mudar de sinal caso a orientação da curva seja revertida.*

Demonstração. Considere $\alpha : (a, b) \rightarrow R^3$ uma curva e $F : R^3 \rightarrow R^3$ um movimento Euclidiano. Sendo A a parte linear de F , de modo que $\forall \mathbf{p} \in R^3$ teremos $F(\mathbf{p}) = A\mathbf{p} + F(0)$. Nós definimos uma curva $\gamma : (a, b) \rightarrow R^3$ por $\gamma = F \circ \alpha$; então para $a < t < b$ teremos

$$\gamma(t) = A\alpha(t) + F(0) \quad (\text{C.4})$$

Logo, $\gamma'(t) = A\alpha'(t)$, $\gamma''(t) = A\alpha''(t)$ e $\gamma'''(t) = A\alpha'''(t)$. Considere s_α e s_γ o comprimento de arco de α e γ , respectivamente. Como A será uma transformação ortogonal,

teremos

$$s'_\gamma(t) = \|\gamma'(t)\| = \|A\alpha'(t)\| = \|\alpha'(t)\| = s'_\alpha(t) \quad (\text{C.5})$$

Podemos computar a curvatura de γ usando a [Equação 3.12](#) e a *identidade de Lagrange* (i. e., $\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\|^2\|\mathbf{b}\|^2 - (\mathbf{a} \cdot \mathbf{b})^2$).

$$\kappa_\gamma(t) = \frac{\|\gamma'(t) \times \gamma''(t)\|}{\|\gamma'(t)\|^3} = \frac{\|A\alpha'(t) \times A\alpha''(t)\|}{\|A\alpha'(t)\|^3} \quad (\text{C.6})$$

$$\kappa_\gamma(t) = \frac{\sqrt{\|A\alpha'(t)\|^2\|A\alpha''(t)\|^2 - (A\alpha'(t) \cdot A\alpha''(t))^2}}{\|A\alpha'(t)\|^3} \quad (\text{C.7})$$

$$\kappa_\gamma(t) = \frac{\sqrt{\|\alpha'(t)\|^2\|\alpha''(t)\|^2 - (\alpha'(t) \cdot \alpha''(t))^2}}{\|\alpha'(t)\|^3} \quad (\text{C.8})$$

$$\kappa_\gamma(t) = \frac{\|\alpha'(t) \times \alpha''(t)\|}{\|\alpha'(t)\|^3} = \kappa_\alpha(t) \quad (\text{C.9})$$

Similarmente, considerando a [Equação 3.26](#), [Teorema C.0.1](#) e $\det(A) = \pm 1$, o mesmo raciocínio se aplica a τ , pois

$$\tau_\gamma(t) = \frac{[\gamma'(t)\gamma''(t)\gamma'''(t)]}{\|\gamma'(t) \times \gamma''(t)\|^2} = \frac{[(A\alpha'(t))(A\alpha''(t))(A\alpha'''(t))]}{\|A\alpha'(t) \times A\alpha''(t)\|^2} \quad (\text{C.10})$$

$$\tau_\gamma(t) = \frac{\det(A)[\alpha'(t)\alpha''(t)\alpha'''(t)]}{\|\alpha'(t) \times \alpha''(t)\|^2} = \quad (\text{C.11})$$

$$\tau_\gamma(t) = \det(A)\tau_\alpha(t) = \pm\tau_\alpha(t) \quad (\text{C.12})$$

□

Sendo s , τ e κ invariantes, podemos comparar curvas utilizando esses valores independente de sua posição no espaço. Assim, justificamos a afirmação 2). Para fundamentar a afirmação 1), precisamos demonstrar que duas curvas só serão iguais, se e somente se, seus valores de κ e τ forem iguais. Uma forma de demonstrar isso é provar que duas curvas que possuam os mesmos valores de κ e τ só diferem por movimentos de corpo rígido F e que essa condição só será satisfeita se as duas curvas forem iguais.

Teorema C.0.3. *Seendo $\kappa : (a, b) \rightarrow \mathbb{R}$ e $\tau : (a, b) \rightarrow \mathbb{R}$ funções diferenciáveis e contínuas e $\kappa > 0$, existe uma curva $\alpha(s) : I \rightarrow \mathbb{R}^3$ cuja curvatura é $\kappa(s)$ e a torção é $\tau(s)$. Se existe uma segunda curva $\gamma(s) : I \rightarrow \mathbb{R}^3$ com os mesmos valores de $\kappa(s)$ e $\tau(s)$, então existe um movimento de corpo rígido $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ tal que $\tilde{\gamma}(s) = F \circ \tilde{\alpha}(s)$.*

Demonstração. Como $\vec{\alpha}$ e $\vec{\gamma}$ possuem $\kappa > 0$, ambos *Frenet-Serret frames* $(\{\vec{T}_\alpha, \vec{N}_\alpha, \vec{B}_\alpha\})$ e $(\{\vec{T}_\gamma, \vec{N}_\gamma, \vec{B}_\gamma\})$ serão bem definidos. Fixando s_0 como $a < s_0 < b$, podemos escolher um movimento de rotação A que mapeia $\vec{\alpha}'(s_0)$ em $\vec{\gamma}'(s_0)$. Ao rotacionar novamente no sentido do plano perpendicular a esse vetor, também mapearemos $\vec{N}_\alpha(s_0)$ em $\vec{N}_\gamma(s_0)$. O [Teorema C.0.1](#) garante que A mapeia $\vec{B}_\alpha(s_0)$ em $\vec{B}_\gamma(s_0)$. Então, existe um movimento Euclidiano F em R^3 tal que

$$F(\vec{\alpha}(s_0)) = \vec{\gamma}(s_0) \quad (\text{C.13})$$

$$A(\vec{T}_\alpha(s_0)) = \vec{T}_\gamma(s_0) \quad (\text{C.14})$$

$$A(\vec{N}_\alpha(s_0)) = \vec{T}_\gamma(s_0) \quad (\text{C.15})$$

$$A(\vec{B}_\alpha(s_0)) = \vec{V}_\gamma(s_0) \quad (\text{C.16})$$

Primeiro precisamos demonstrar a existência dessa curva. Isto é, se $\kappa(s) = \kappa$ e $\tau(s) = \tau$ são funções diferenciáveis e contínuas, existe uma curva parametrizada pelo comprimento de arco $\vec{\alpha}(s) : I \rightarrow R^3$ cuja curvatura é κ e a torção é τ . Observe que para $a < s < b$ o valor de $\vec{\alpha}(s_0)$ pode ser arbitrário. De forma semelhante, os valores de $\vec{T}(s_0)$ e $\vec{N}(s_0)$ podem ser definidos desde que $\|\vec{T}(s_0)\| = \|\vec{N}(s_0)\| = 1$ e $\vec{T}(s_0) \cdot \vec{N}(s_0) = 0$. Agora considere o seguinte sistema de 12 equações diferenciais:

$$\begin{cases} x'_i(s) = t_i(s) \\ t'_i(s) = \kappa(s)n_i(s) \\ n'_i(s) = -\kappa(s)t_i(s) + \tau(s)b_i(s) \\ b'_i(s) = -\tau(s)n_i(s) \end{cases}$$

todos para $1 \leq i \leq 3$, junto com as condições iniciais

$$\begin{cases} x_i(s_0) = p_i \\ t_i(s_0) = q_i \\ n_i(s_0) = r_i \\ b_1(s_0) = q_2 r_3 - q_3 r_2 \\ b_2(s_0) = q_3 r_1 - q_1 r_3 \\ b_3(s_0) = q_1 r_2 - q_2 r_1 \end{cases}$$

Como $\|\vec{T}(s_0)\| = \|\vec{N}(s_0)\| = 1$ e $\vec{T}(s_0) \cdot \vec{N}(s_0) = 0$, logo

$$\sum_{i=1}^3 q_i^2 = 1 = \sum_{i=1}^3 r_i^2 \quad (\text{C.17})$$

$$\sum_{i=1}^3 q_i r_i = 0 \quad (\text{C.18})$$

Se $\vec{\gamma} = (x_1, x_2, x_3)$, $\vec{T} = (t_1, t_2, t_3)$ e $\vec{B} = (b_1, b_2, b_3)$, podemos expressar o primeiro sistema de equação de acordo com $\gamma' = \vec{T}$ e obteremos a equação do *Frenet-Serret frame*.

$$\begin{cases} \vec{T}'_i(s) = \kappa(s)\vec{N}_i(s) \\ \vec{N}'_i(s) = -\kappa(s)\vec{T}_i(s) + \tau(s)\vec{B}_i(s) \\ \vec{B}'_i(s) = -\tau(s)\vec{N}_i(s) \end{cases}$$

A partir do teorema de existência e unicidade da solução de sistemas de equações diferenciais, sabemos que esse sistema linear possui uma única solução. Considerando um triedro ortonormal $\{T_0, N_0, B_0\}$ com orientação positiva e $s_0 \in I$, existe uma família de triedros $\{\vec{T}(s), \vec{N}(s), \vec{B}(s)\}$, $s \in I$, com $\vec{T}(s_0) = \vec{T}_0$, $\vec{N}(s_0) = \vec{N}_0$, $\vec{B}(s_0) = \vec{B}_0$. E esta família mantém-se ortonormal para todo $s \in I$. Observe que

$$\frac{d}{ds}(\langle \vec{T}(s), \vec{N}(s) \rangle) = 0 \quad (\text{C.19})$$

Portanto, $\langle \vec{T}(s), \vec{N}(s) \rangle$ é constante e

$$\langle \vec{T}(s), \vec{N}(s) \rangle = \langle \vec{T}(s_0), \vec{N}(s_0) \rangle = 0 \quad (\text{C.20})$$

Usando um raciocínio semelhante, obteremos

$$\langle \vec{T}(s), \vec{B}(s) \rangle = \langle \vec{B}(s), \vec{N}(s) \rangle = 0 \quad (\text{C.21})$$

$$\|\vec{T}(s)\| = \|\vec{N}(s)\| = \|\vec{B}(s)\| = 1 \quad (\text{C.22})$$

Isto é, o triedro $\vec{T}, \vec{N}, \vec{B}$ é ortonormal para todo $s \in I$. Podemos definir $\vec{\alpha}(s) = \int_{s_0}^s \vec{T}(s) ds$. Consequentemente, $\vec{\alpha}'(s) = \vec{T}(s)$ e $\vec{\alpha}''(s) = \kappa \vec{N}(s)$, portanto a curvatura de $\vec{\alpha}(s)$ será κ . Além disso sabemos que (ver [Apêndice B](#)):

$$\vec{\gamma}''' = -\kappa^2 \vec{T} + \kappa' \vec{N} + \tau \kappa \vec{B} \quad (\text{C.23})$$

Logo, a torção será

$$\tau = \frac{\langle \vec{\gamma}' \times \vec{\gamma}'', \vec{\gamma}''' \rangle}{\kappa^2} \quad (\text{C.24})$$

$$\tau = -\frac{\langle \vec{T} \times \kappa \vec{N}, (-\kappa^2 \vec{T} + \kappa' \vec{N} + \tau \kappa \vec{B}) \rangle}{\kappa^2} \quad (\text{C.25})$$

Logo, a curva $\vec{\alpha}$ é a curva que corresponde a curva especificada pelos valores κ e τ

Agora precisamos demonstrar que $\vec{\alpha}$ é a única solução ao descontar movimentos de translação e rotação em R^3 . Considere $\vec{\gamma} : I \rightarrow R^3$ uma segunda curva que $\kappa_\gamma(s) = \kappa_\alpha(s)$, $\tau_\gamma(s) = \tau(s)$, $s \in I$, e $\{\vec{T}_{\gamma,0}, \vec{N}_{\gamma,0}, \vec{B}_{\gamma,0}\}$ sendo o *Frenet-Serret frame* de $\vec{\gamma}(s_0)$. Um movimento F permite levar o diedro $\{\vec{T}_{\gamma,0}, \vec{N}_{\gamma,0}, \vec{B}_{\gamma,0}\}$ a uma posição que coincida com $\{\vec{T}_{\alpha,0}, \vec{N}_{\alpha,0}, \vec{B}_{\alpha,0}\}$. Aplicando o teorema de existência e unicidade da solução de sistemas de equações diferenciais e com um raciocínio análogo ao desenvolvido anteriormente ambos os correspondem a mesma família $\{\vec{T}_{\alpha,0}, \vec{N}_{\alpha,0}, \vec{B}_{\alpha,0}\}$ de triedro ortonormal em R^3 . Como $\vec{\alpha}(s) = \int_{s_0}^s \vec{T}_\alpha(s) ds$, $\vec{\gamma}(s) = \int_{s_0}^s \vec{T}_\gamma(s) ds$ e $F \circ \vec{T}_\gamma(s) = \vec{T}_\alpha(s)$, podemos concluir que

$$(F \circ \vec{\alpha})(s) = \vec{\gamma}(s) + \vec{q}, \forall s \quad (\text{C.26})$$

onde, $\vec{q} \in R^3$ Pela escolha de F , $\vec{q} = 0$ e F mapeia $\vec{\alpha}$ em $\vec{\gamma}$.

□