1

MÉTODOS DE ATRIBUIÇÃO DE ESTRUTURA SECUNDÁRIA

1.1 INTRODUÇÃO

Os métodos computacionais de atribuição de estruturas secundárias surgiram com o objetivo de automatizar e reduzir a subjetividade da atribuição feita por cristalógrafos. Em geral, tais métodos buscam por um conjunto de padrões estruturais capazes de representar a estrutura secundária atribuída por especialistas.

Entretanto, como a atribuição feita por diferentes especialistas pode diferir devido ao uso de padrões subjetivos e, sobretudo, pela dificuldade de selecionar parâmetros objetivos que reproduzam a atribuição feita por especialistas, os métodos computacionais apresentam variações entre si.

Diferentes autores sugerem a impossibilidade de se eleger qual o melhor método, uma vez que todos estão corretos segundo os princípios adotados por eles [**kabsch_dictionary_1983**, **colloch_comparison_1993**, **frishman_knowledge-based_1995**, **zhang_secondary_2015**]. Assim, métodos que utilizam a informação da estrutura secundária de proteínas, como os métodos de predição de estruturas secundárias, acabaram optando, possivelmente de forma natural, por escolherem um método de atribuição como padrão. Consequentemente, podemos observar o método DSSP como o mais utilizado como padrão de referência.

Nesse trabalho, nós decidimos comparar quatro métodos de atribuição que utilizam diferentes parâmetros para classificar elementos de estrutura secundária. O objetivo disso foi entender como a atribuição varia entre os métodos e assim, avaliar qual informação deveríamos utilizar e analisar ao longo do trabalho.

1.1.1 *DSSP*

Em 1983, Kabsch e Sander publicaram o algoritmo de atribuição de estruturas secundárias de proteínas que viria a ser o mais utilizado até os dias atuais, o DSSP (*Dictionary of Protein Secondary Structure*)[**kabsch_dictionary_1983**].

No trabalho, os autores afirmam que a atribuição de estruturas secundárias a partir das coordenadas atômicas de estruturas proteicas é um problema de reconhecimento de padrões. Nesse contexto, eles optaram por identificar esses padrões através de ligações de hidrogênio entre átomos da cadeia principal.

A utilização das ligações de hidrogênio da cadeia principal ao invés de ângulos Φ e Ψ ou de posições relativas de C_α foi justificada pela simplicidade. A presença, ou ausência, de ligações de hidrogênio poderiam ser avaliadas por um simples critério energético, enquanto que outras características precisariam do ajuste numérico de um número maior de parâmetros.

Na época havia um pouco mais de 100 estruturas depositadas no Protein Data Bank.

O DSSP define as ligações de hidrogênio utilizando um modelo eletrostático. Nesse modelo, uma ligação de hidrogênio HB ocorrerá se, e somente se, a energia E for menor que $-0.5kcal/mol$. Para o cálculo são utilizadas as cargas parciais $+q_1, -q_1$ nos átomos C e O, e $-q_2, +q_2$ nos átomos N e H, onde $q_1 = 0.42e$ e $q_2 = 0.20e$.

$$E < -0.5kcal/mol \implies HB = \text{Verdade} \tag{1.1}$$

onde

$$E = q_1 q_2 (1/r(ON) + 1/r(CH) - 1/r(OH) - 1/r(CN)) * f \tag{1.2}$$

Na equação 1.2, $r(AB)$ é a distância interatômica entre A e B em ångström e o fator dimensional $f = 332$.

Os autores afirmam que, por este modelo, uma boa ligação de hidrogênio teria aproximadamente $-3kcal/mol$. Assim, a escolha do limiar como$-0.5kcal/mol$ torna o modelo mais tolerante a erros nas coordenadas atômicas e a ligações de hidrogênios bifurcadas **[kabsch_dictionary_1983]**.

Uma vez definido o modelo para identificar ligações de hidrogênio, elas são procuradas e anotadas na cadeia polipeptídica em duas classes ou padrões elementares: (1) padrão *n-Turn* e (2) padrão *Bridge*.

O padrão *n-Turn*, onde $n \in \{3, 4, 5\}$, apresentam uma ligação de hidrogênio entre o CO do resíduo i e o NH do resíduo $i + n$.

$$n\text{-}Turn \impliedby HB(i, i+n),\ n \in \{3,\ 4,\ 5\} \tag{1.3}$$

O padrão *Bridge* pode ocorrer de duas formas, a paralela e a antiparalela.

$$Paralela \impliedby [HB(i-1, j) \wedge HB(j, i+1)] \vee [HB(j-1, i) \wedge HB(i, j+1)] \tag{1.4}$$

$$Antiparalela \impliedby [HB(i, j) \wedge HB(j, i)] \vee [HB(i-1, j+1) \wedge HB(j-1, i+1)] \tag{1.5}$$

Sendo que as sequências $i-1, i, i+1$ e $j-1, j, j+1$ não apresentam sobreposição (*overlap*) de resíduos entre si.

As ocorrências dos padrões elementares *n-Turn* e *Bridge* ao longo da cadeia polipeptídica são utilizadas para a atribuição dos elementos

de estrutura secundária. Como exemplos, repetições consecutivas do padrão *4-Turn* indicam a ocorrência de uma hélice α, enquanto resíduos consecutivos com padrão *Bridge* formam uma fita de uma folha β.

No trabalho, os autores mencionam que o algoritmo proposto produz hélices mais curtas, com um resíduo a menos em cada extremidade da hélice, em relação a anotação seguindo as regras da IUPAC. Outra característica do algoritmo é que hélices que apresentem algumas ligações de hidrogênio ausentes, são mantidas como uma hélice única ao invés de múltiplas hélices com *kinks*. O mesmo ocorre com resíduos que formam *bulges* em fitas, sendo os mesmos anotados como parte integrante da fita.

1.1.2 *STRIDE*

O método de atribuição de estrutura secundárias STRIDE foi desenvolvido com o objetivo de reproduzir, com o máximo de acurácia possível, a anotação humana feita por cristalógrafos [**frishman_knowledge-based_1995**]. No método são utilizados como critérios tanto as energias de ligações de hidrogênio quanto a propensão dos ângulos torcionais Φ e Ψ da cadeia polipeptídica.

A energia das ligações de hidrogênio são calculadas de acordo com a seguinte função:

$$E_{hb} = E_r * E_t * E_p \tag{1.6}$$

Onde E_r (1.7) é o termo de distância, E_p (1.8) e E_t (1.9) descrevem as propriedades direcionais da ligação de hidrogênio.

$$E_r = \frac{-3E_m {r_m}^8}{r^8} + \frac{-4E_m {r_m}^6}{r^6} \tag{1.7}$$

Onde r é a distância entre o nitrogênio N e o oxigênio O da cadeia principal de resíduos diferentes, $E_m = -2,8\,kcal/mol$ e $r_m = 3,0\,\text{Å}$.

Os termos angulares são representados pelas seguintes funções:

$$E_p = \cos^2 \mathrm{p} \tag{1.8}$$

e

$$E_t = \begin{cases} (0.9 + 0.1 \sin 2t_i) \cos t_0 & 0 < t_i < 90° \\ K_1(K_2 - \cos^2 t_i)^3 \cos t_0 & 90° < t_i < 110° \\ 0 & t_i > 110° \end{cases} \tag{1.9}$$

onde $K_1 = 0,9/cos^6 110°$, $K_2 = cos^2 110°$e os ângulos t_i e t_0 são, respectivamente, os desvios angulares da ligação de hidrogênio.

A propensão dos ângulos torcionais para hélices α e fitas β são calculadas respectivamente como:

$$P_i^\alpha = \begin{cases} \frac{N_i^\alpha}{N_i^{total}} & se\ -180° < \Psi < 10°\ e\ -120° < \Phi < 45° \\ 0 & caso\ contrário \end{cases} \tag{1.10}$$

e

$$P_i^\beta = \begin{cases} \frac{N_i^\beta}{N_i^{total}} & se\ -180° < \Psi < 0°\ \text{e -180°<}\Phi\text{<-120°}\ ou\ 45° < \Phi < 180° \\ 0 & caso\ contrário \end{cases} \tag{1.11}$$

onde N_i^α e N_i^β são, respectivamente, os números de resíduos definidos como hélice e fita em um quadrante Φ e Ψ de dimensões 20°×20°, e N_i^{total} é o número total de resíduos nesse quadrante.

A hélice mínima é definida por duas ligações de hidrogênio consecutivas, ou seja, entre resíduos $k, k+4$ e $k+1, k+5$, utilizando a função:

$$E_{hb}^{k,k+4}\left(1 + W_1^\alpha + W_2^\alpha \frac{P_k^\alpha + P_{k+4}^\alpha}{2}\right) < T_1^\alpha \tag{1.12}$$

Caso a condição seja verdadeira, os resíduos centrais $k+1, k+2, k+3\, e\, k+4$ são classificados como hélice. Já os resíduos $k, k+5$ serão classificados como hélice se respeitarem as condições adicionais:

$$P_k^\alpha < T_2^\alpha \tag{1.13}$$

$$P_{k+5}^\alpha < T_3^\alpha \tag{1.14}$$

Nas funções 1.12,1.13 e 1.14, W_1^α, W_2^α, T_1^α, T_2^α e T_3^α são pesos e limites empíricos que foram otimizados.

A folha β mínima é definida por duas ligações de hidrogênio consecutivas através das seguintes funções:

$$\begin{cases} E_{hb1}\left(1 + W_1^\beta + W_2^\beta . CONF_{Antiparalela}\right) < T_{Antiparalela}^\beta \\ E_{hb2}\left(1 + W_1^\beta + W_2^\beta . CONF_{Antiparalela}\right) < T_{Antiparalela}^\beta \end{cases} \tag{1.15}$$

$$\begin{cases} E_{hb1}\left(1 + W_1^\beta + W_2^\beta . CONF_{Paralela}\right) < T_{Paralela}^\beta \\ E_{hb2}\left(1 + W_1^\beta + W_2^\beta . CONF_{Paralela}\right) < T_{Paralela}^\beta \end{cases} \tag{1.16}$$

onde

$$CONF = \frac{P_{Int1}^\beta + P_{Int2}^\beta}{2} \tag{1.17}$$

W_1^β,W_2^β, $T_{Paralela}^\beta$, $T_{Antiparalela}^\beta$ são pesos e limites empíricos que foram otimizados.

A otimização dos pesos foi realizada de forma a aumentar a acurácia (Q_3) entre a atribuição feita pelo método e a atribuição feita por cristalógrafos. Utilizando 223 proteínas, os autores definiram os seguintes valores: $W_1^\alpha = W_2^\alpha = 1$, $T_1^\alpha = 230.0$, T_2^α e $T_3^\alpha = 0.06$, $W_1^\beta = W_2^\beta = 0.2$, $T_{Paralela}^\beta = -240.0$ e $T_{Antiparalela}^\beta = -310.0$.

1.1.3 *KAKSI*

KAKSI é um método de atribuição de estruturas secundárias proposto por Martin e colaboradores [**martin_protein_2005**]. Esse método foi desenvolvido utilizando padrões de distâncias entre carbonos alfa (C_α) e de ângulos Φ e Ψ.

A heurística de atribuição de estrutura secundárias busca primeiramente por hélices, sendo que um resíduo é classificado como hélice se respeitar os critérios de distâncias entre C_α **ou** os critérios de ângulos Φ e Ψ. Em seguida, é feito a classificação dos resíduos em fitas. Somente os resíduos não-hélice podem ser classificados em fitas, e para tal, eles precisam repeitar os critérios de distância entre C_α **e** os critérios de ângulos Φ e Ψ.

1. Critérios para classificação de hélices:
   a) Distância entre C_α

      Todas as distâncias entre C_α em uma janela de seis resíduos [i, i+5] precisam estar dentro do intervalo$[M_\alpha - \varepsilon_H \times SD_\alpha, M_\alpha + \varepsilon_H \times SD_\alpha]$, onde M_α e SD_α são respectivamente a média e o desvio padrão observado em hélices α.

   b) Ângulos Φ e Ψ

      Todos os pares de ângulos Φ e Ψ em uma janela de quatro resíduos precisam satisfazer as condições: $\Phi < 0°$ e $-90° < \Psi < 60°$. Além disso, ao menos um par de ângulos precisa estar em uma região densamente povoada, com densidade$> \sigma_H$.

2. Critérios para classificação de folhas:
   a) Distância entre C_α

      Todos as distâncias entre C_α em duas janelas de três resíduos precisam estar no intervalo $[M_\beta - \varepsilon_b \times SD_\beta, M_\beta + \varepsilon_b \times SD_\beta]$, onde M_β e SD_β são respectivamente a média e o desvio padrão observado em folhas β.

   b) Ângulos Φ e Ψ

      Cada par de ângulos Φ e Ψ presente na zona povoada de resíduos em folhasβ incrementa um contador em 1. Quando um resíduo central da janela apresenta $-120° < \Psi < 50°$ (fora da região de fitas), o contador é reiniciado em zero. Esse critério é satisfeito se o contator $\geq \sigma_b$.

Além dos critérios acima, há outros para detecção de *kink* em hélices e um critério de correção de segmentos, que altera um resíduo para o estado de coil quando ocorre continuidade de segmentos hélice-fita, ou fita-hélice, tornando as hélices 1 resíduo menor.

Os vários parâmetros necessários ao método foram ajustados empiricamente utilizando um conjunto de 2880 domínios estruturais, com identidade sequencial inferior à 40%, resolvidos por cristalografia e com resolução superior a 2.25Å.

1.1.4 *PROSS*

PROSS é um método de atribuição de estruturas secundárias que utiliza somente os ângulos torcionais Φ e Ψ da cadeia principal. Originalmente, cada resíduo era classificado em um mesoestado (*mesostate*) de acordo com o valor de seus ângulos torcionais [**srinivasan_physical_1999**, **gong_does_2005**]. Cada mesoestado equivale a um quadrante de dimensões $60° \times 60°$, totalizando 36 mesoestados possíveis (Figura 1.1). Em seguida, cada resíduo é classificado em hélice α, fita β, volta β, P_{II} ou coil de acordo com um conjunto de regras. São elas:

1. Hélicesα: Uma região é identificada como hélice se houverem 5 ou mais resíduos contínuos no conjunto de mesoestados {O, P}.

2. Fita β: Uma região é definida como fita se houverem 3 ou mais resíduos contínuos no conjunto de mesoestados {L, G, F, A, R, M}.

3. Volta β: São definidos como voltas β todos os pares de dipeptídeos com combinações no conjunto {OO, OP, OJ, PO, PP, PJ, JO, JP, JJ, Mo, Mp, Mj, Ro, Rp, Rj, oo, op, oj, po, pp, pj, jo, jp, jj, mO, mP, mJ, rO, rP, rJ}.

4. P_{II} : Resíduos que não foram classificados como fita, mas que estão no conjunto {M, R} são categorizados como poliprolina II.

5. Coil: Todos os outros resíduos não classificados como hélices α, fitas β, voltas β ou P_{II} são classificados como coil.

Posteriormente, outra versão do método PROSS passou a utilizar 144 mesoestados com área $30° \times 30°$ (Figura 1.2). As regras continuaram sendo estabelecidas de forma semelhante a feita anteriormente, mas utilizando estados mais específicos. Por exemplo:

1. Hélicesα: Uma região é identificada como hélice se houverem 5 ou mais resíduos contínuos no conjunto de mesoestados {De, Df, Ed, Ee, Ef, Fe}.

2. Fita β: Uma região é definida como fita se houverem 3 ou mais resíduos contínuos no conjunto de mesoestados {Bj, Bk, Bl, Cj, Ck, Cl, Dj, Dk, Dl}.

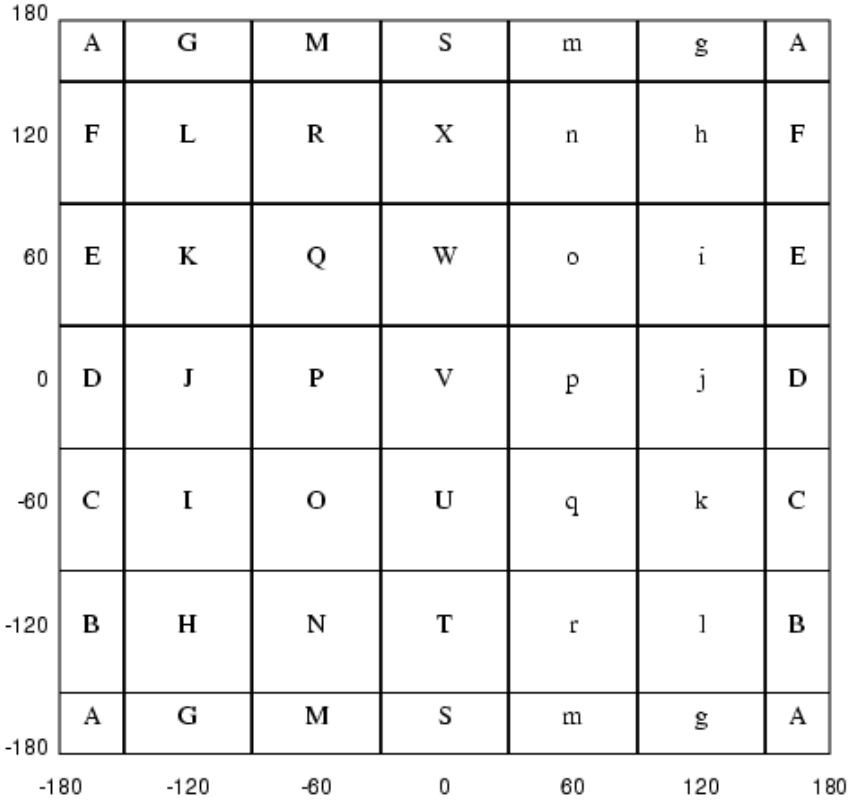

Figura 1.1: Mesoestados 60°×60° para descrever os resíduos de acordo com os ângulos torcionais Φ e Ψ.

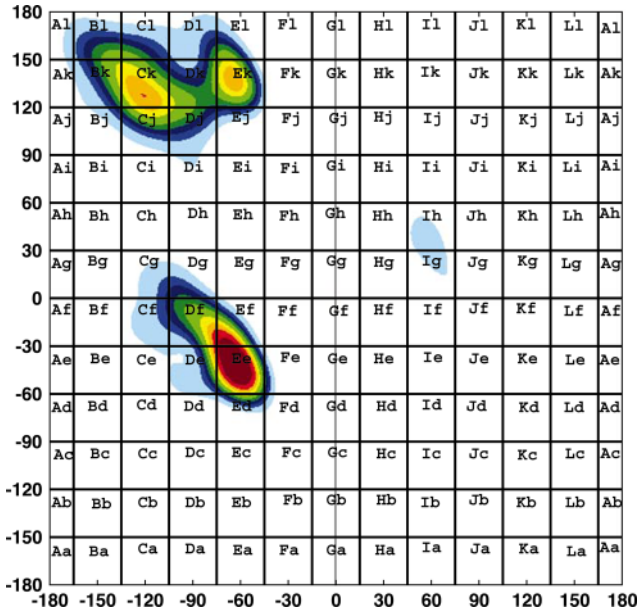

Figura 1.2: Mesoestados 30°×30° para descrever os resíduos de acordo com os ângulos torcionais Φ e Ψ.

Segundo os autores, a utilização somente dos ângulos torcionais da cadeia principal apresenta a vantagem de representar, corretamente, as fitas β como estruturas secundárias [**srinivasan_physical_1999**]. Métodos como o DSSP, ao utilizarem ligações de hidrogênio, acabam definindo as folhas β como estruturas secundárias, as quais, no contexto físico, seriam mais apropriadamente classificadas como estruturas terciárias.

1.2 MATERIAIS E MÉTODOS

1.2.1 *Conjunto de dados*

O conjunto de proteínas utilizado ao longo deste trabalho foi obtido do banco de dados "TOP8000-best-hom50" (atualizado em 2015). Esse banco de dados é organizado pelo Richardson Lab da Universidade de Duke e está disponível em github.com/rlabduke/reference_data. A estruturas proteicas do banco de dados estão separadas por cadeias e atendem aos seguintes critérios:

- Resolução $<$ 2,0Å
- MolProbity score $<$ 2,0
- $\leq$5% dos resíduos apresentando comprimentos de ligação anormais ($> 4\sigma$)
- $\leq$5% dos resíduos apresentando ângulos de ligação anormais ($> 4\sigma$)
- $\leq$5% dos resíduos com desvios anormais do C_β ($> 0,25$Å)
- identidade sequencial entre elas $<$50% (HOM50)

A estrutura secundária dessas proteínas foram atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS. Proteínas que possuíam resíduos indeterminados na sequência de aminoácidos ou que apresentaram algum erro de execução durante a atribuição da estrutura secundária por algum dos quatro métodos foram removidas do conjunto.

O conjunto de dados final é composto por 6749 cadeias polipeptídicas das 7233 presentes no banco de dados "TOP8000-best-hom50".

1.3 RESULTADOS

A comparação das estruturas secundárias atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS demonstrou que há diferenças significativas entre eles. Aproximadamente 24% dos 1,7 milhão de resíduos que presentes no conjunto de dados não apresentaram consenso na estrutura secundária atribuída pelos quatro métodos (Figura 1.3).

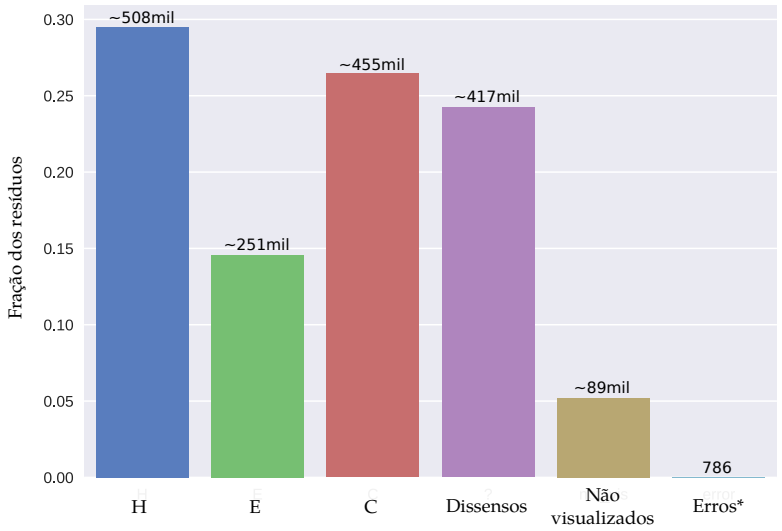

Figura 1.3: Estrutura secundária atribuída aos resíduos das proteínas do conjunto de dados. Entre os resíduos que apresentaram consenso entre os métodos, aproximadamente 29,5% foram classificados como hélices (H); 14,5% como fitas (E); 26,5% como coils (C). 24,2% não apresentam consenso entre os quatro métodos de atribuição (Dissenso); 5,2% são resíduos não visualizados na estrutura atômica resolvida e 0,04% dos resíduos não tiveram a estrutura secundária atribuída por todos os quatro métodos (Erro*).

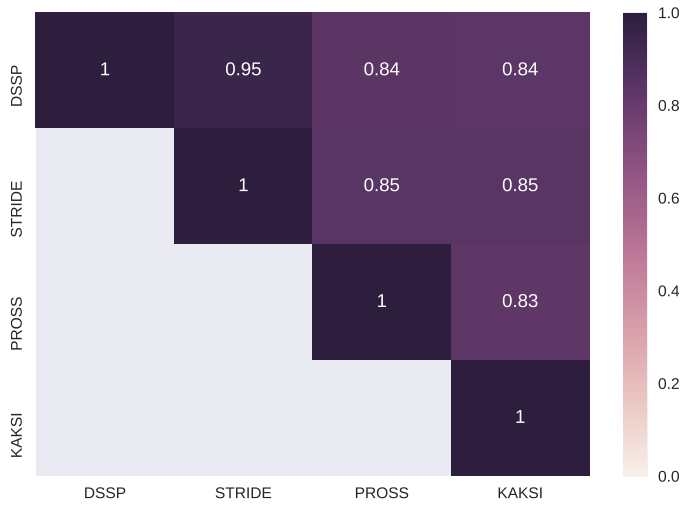

Figura 1.4: Similaridade entre as estruturas secundárias atribuídas para cada resíduo entre os quatro métodos de atribuição de estruturas secundárias: DSSP, STRIDE, KAKSI e PROSS.

As similaridades obtidas pela comparação pareada dos métodos de atribuição indicam que, com exceção do DSSP x STRIDE, aproximadamente 85% dos resíduos apresentam consenso entre dois métodos (Figura 1.4). Considerando a comparação conjunta dos quatro métodos que demonstra consenso em 74,4% dos resíduos podemos inferir que as posições de dissenso variam entre os métodos.

1.3.1 *Características das estruturas secundárias atribuídas*

Em relação ao número de estruturas secundárias atribuídas, o DSSP atribuiu o maior número de hélices, ~67 mil. Esse valor é apenas 5,7% maior que o número atribuído pelo STRIDE, mas 31,6% maior que o KAKSI e 42% maior que o atribuído pelo PROSS. Quanto as fitas, o método que atribuiu o maior número foi o STRIDE ~68mil. O número de fitas atribuídas por cada método apresentou uma variação menor que as hélices. O número de fitas atribuídas pelo DSSP foi apenas 0,03% menor, pelo PROSS, 0,3% menor e pelo KAKSI, 0,5% menor (Figura 1.5).

A distribuição do comprimento das estruturas secundárias também difere entre os métodos de atribuição. A mediana do comprimento das hélices atribuídas pelo DSSP é de 8 resíduos, 9 no STRIDE, 10 no PROSS e de 12 resíduos no KAKSI (Figura 1.6). Hélices mais curtas são mais frequentes tanto no DSSP quanto no STRIDE. Caso duas ou mais hélices curtas sejam unidas e anotadas como uma hélice longa por outros métodos, teríamos um menor número de hélices.

Figura 1.5: Número de estruturas secundárias atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS.

Isso explicaria, ao menos parcialmente, o fato dos métodos KAKSI e PROSS anotarem um menor número de hélices (Figura 1.5).

As distribuições dos comprimentos das fitas apresentam mediana de 5 resíduos para os quatro métodos de atribuição (Figura 1.7).

As distribuições dos comprimentos dos coils apresentam mediana de 4 resíduos para os quatro métodos de atribuição. O comprimento dos coils para o DSSP e o STRIDE são muito similares, o que também foi observado em fitas.

1.3.2 *Características dos dissensos*

O número de dissensos observados nas proteínas apresenta uma relação linear com o comprimento da cadeia polipeptídica. Assim, em média, aproximadamente 25% dos resíduos não apresentam consenso entre os métodos de atribuição analisados (Figura 1.9).

Na comparação entre as estruturas secundárias atribuídas para cada resíduo, quatro tipos de dissensos podem ocorrer. O tipo $H \leftrightarrow C$, onde o resíduo é anotado como hélice por alguns métodos e fita por outros, e da mesma forma os tipos $E \leftrightarrow C$, $H \leftrightarrow E$ e $H \leftrightarrow C \leftrightarrow E$. A análise dos dissensos demonstrou que os dois tipos dominantes são $H \leftrightarrow C$, com 48,7%, e $E \leftrightarrow C$, com 50,5%, totalizando 99,2% dos casos de dissenso. Dos demais tipos, $H \leftrightarrow E$ ocorre 0,1% e $H \leftrightarrow C \leftrightarrow E$ ocorre em 0,7% dos casos (Figura 1.10).

Outra característica analisada é a região relativa onde as regiões de dissenso ocorrem. Para isso, definimos 9 tipos de regiões que diferem pelo tipo de estrutura secundária que precede e sucede o

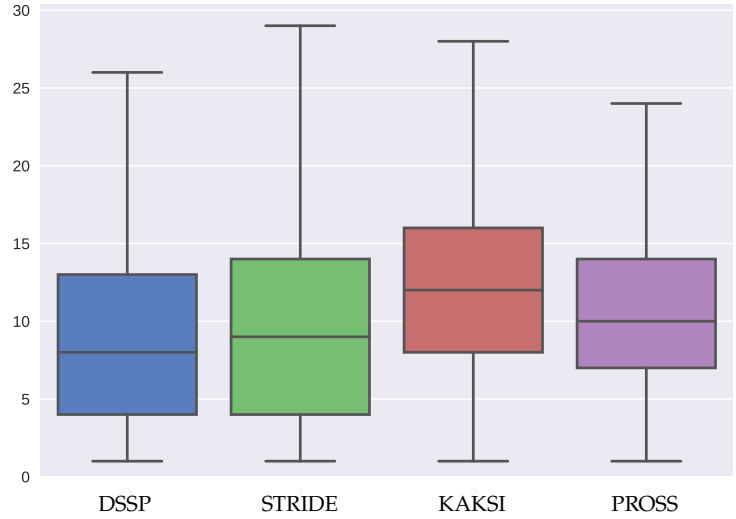

Figura 1.6: Distribuição do comprimento de hélices por método de atribuição de estrutura secundária.

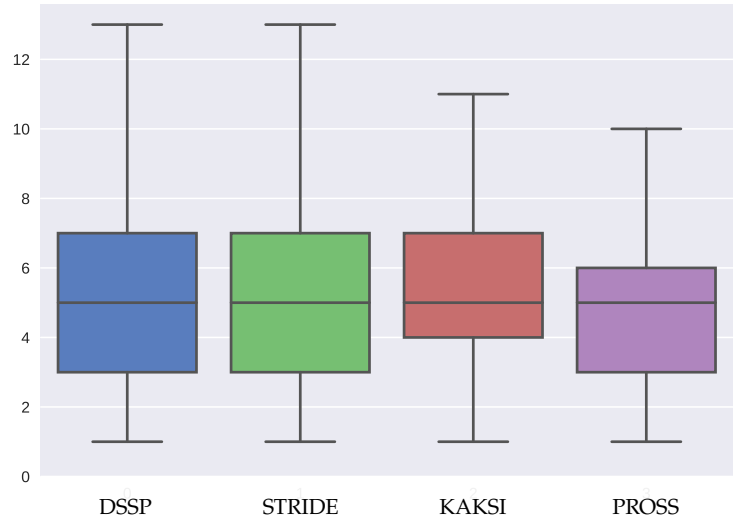

Figura 1.7: Distribuição do comprimento de fitas por método de atribuição de estrutura secundária.

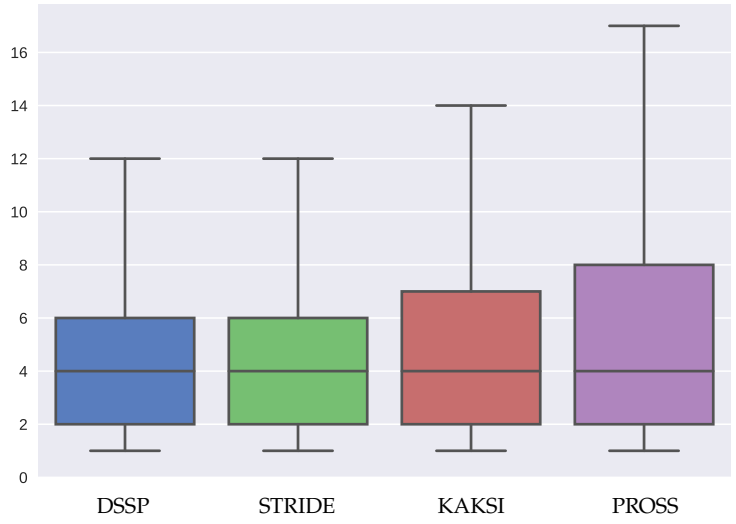

Figura 1.8: Distribuição do comprimento dos coils por método de atribuição de estrutura secundária.

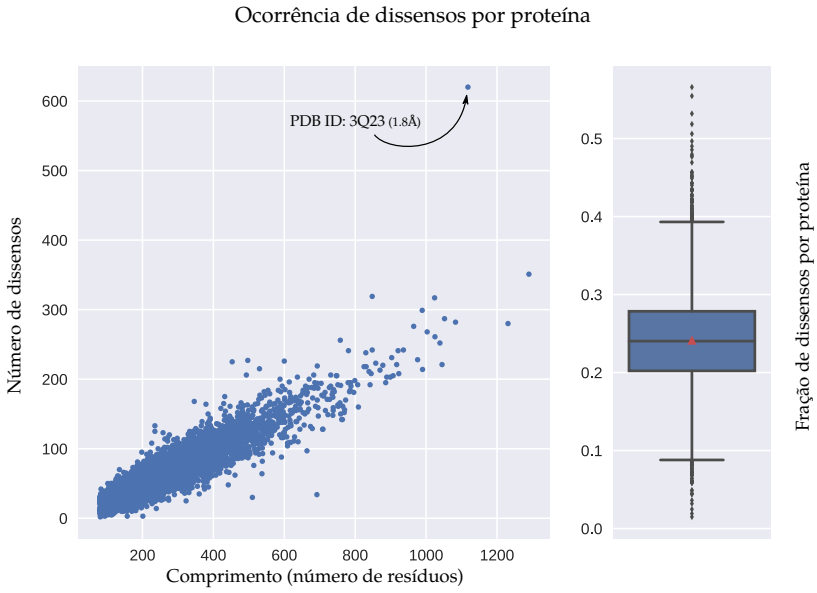

Figura 1.9: Relação entre o número de dissensos e o comprimento da cadeia polipeptídica. Distribuição da fração de resíduos que não apresentaram consenso (dissensos) por proteína.

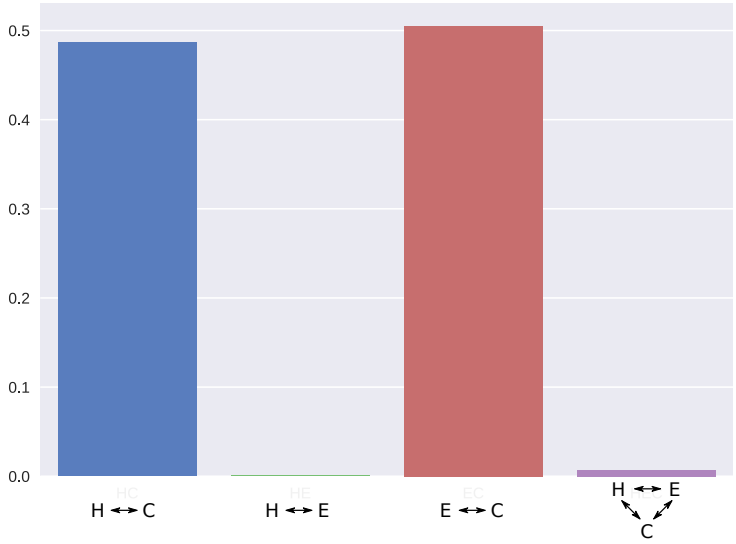

Figura 1.10: Proporção dos quatro tipos de dissensos que os resíduos podem apresentar.

dissenso. Por exemplo, $C?H$ é um dissenso antecedido por um coil e sucedido por uma hélice, o que indicaria que há variações em relação ao resíduo que inicia a hélice.

Notamos que a maioria dos dissensos ocorre em regiões de transição entre coil e as estruturas secundárias hélice ou fita. Assim, os tipos mais frequentes foram $C?E$, $E?C$, $H?C$ e $C?H$ (Figura 1.11). Isso indica que os métodos de atribuição tem pouca precisão em anotar onde começam e terminam as estruturas secundárias.

Os tipos $E?H$ e $H?E$ foram pouco frequentes, o que era esperado uma vez que a transição entre esses dois tipos de estruturas costumam ocorrer pelo intermédio de regiões de coil.

O tipo $C?C$ também apresentou um grande número de ocorrências. Isso indica que alguns métodos de atribuição podem falhar na atribuição de regiões inteiras de estruturas secundárias, por exemplo, ignorando a presença de uma hélice inteira ou de uma fita.

A frequência relativa dos aminoácidos em regiões de dissenso mostrou que o resíduo mais comum nessas regiões é a glicina. Acreditamos que um dos motivos seja a maior variação dos ângulos Φ e Ψ que esse resíduo pode apresentar, fato que poderia dificultar a atribuição da estrutura secundária por métodos que utilizam tais ângulos, como o PROSS, ou a posição relativa de carbonos alfa, como o KAKSI.

Em relação aos aminoácidos hidrofóbicos, parece haver uma menor tendência deles ocorrerem em regiões de dissenso. É possível que a maior frequência desses aminoácidos em regiões internas da estrutura proteica limite a variabilidade dos estados conformacionais e apresente ligações de hidrogênio mais próximas ao padrão ideal.

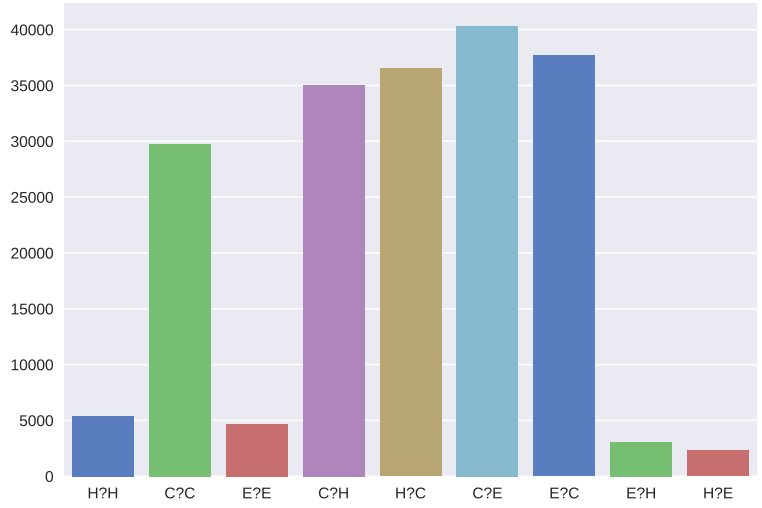

Figura 1.11: Ocorrências de dissensos por região relativa à estrutura secundária precedente e sucedente. Ex. C?H indica uma região de dissenso imediatamente posterior a um coil e anterior a uma hélice.

Isso poderia explicar um maior consenso na atribuição da estrutura secundária por diferentes métodos.

Observamos ainda uma maior frequência do aspartato (D) em relação ao glutamato (E) e da asparagina (N) em relação à glutamina (Q). Essas diferenças podem estar relacionadas à maior propensão do aspartato e da asparagina em relação ao glutamato e à glutamina nos inícios e finais de hélice (*helix capping*) [**aurora_helix_1998**, **doig_n-_1995**] e inícios e finais de fitas (*beta-sheet capping*) [**farzadfard_beta-sheet_2008**].

1.4 CONCLUSÃO

Uma das etapas iniciais no desenvolvimento de métodos de predição de estrutura secundária é a preparação dos dados de referência. Parte dessa preparação consiste na atribuição da estrutura secundária em proteínas com estrutura atômica resolvida. Em geral, trabalhos da literatura costumam utilizar o método DSSP como padrão. Entretanto, a diferença observada entre as atribuições de estrutura secundária efetuadas por diferentes métodos computacionais indica que essa etapa pode ser importante para o desenvolvimento de preditores com maior acurácia.

Ao longo deste trabalho nós exploramos a utilização dos dados de cada um dos quatro métodos citados e sua influência no treinamento de preditores. Como veremos adiante, a utilização apenas de resíduos que apresentam consenso na atribuição resulta em preditores com

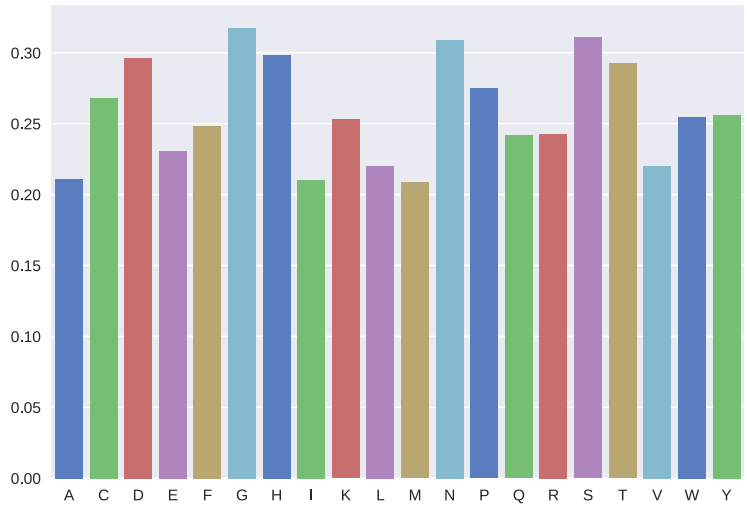

Figura 1.12: Frequência dos aminoácidos em regiões de dissenso. A frequência foi normalizada pelo número total de ocorrências do aminoácido nas proteínas do conjunto de dados.

acurácia maior ou semelhante ao melhor método de atribuição. Neste contexto, melhor método de atribuição significa o que produziu os dados que resultaram no preditor com maior acurácia.