1 Introdução

Hello (**kendrew1958three**)

2 Revisão de literatura

2.1 Métodos de atribuição de estruturas secundárias

2.1.1 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*).

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 ao invés de ângulos Φ e Ψ ou de posições relativas de C_α. A justificativa utilizada foi que a presença ou ausência de ligações de hidrogênio poderiam ser avaliadas por um simples critério energético, enquanto que outras características precisariam do ajuste de um número maior de parâmetros.

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

$$E < -0.5kcal/mol \implies HB = \text{Verdade} \tag{2.1}$$

onde

$$E = q_1 q_2 (1/r(ON) + 1/r(CH) - 1/r(OH) - 1/r(CN)) * f \tag{2.2}$$

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

Os autores afirmam que, por este modelo, uma boa ligação de hidrogênio teria aproximadamente -3 kcal/mol. Assim, a escolha de um limiar em -0.5 kcal/mol torna o modelo mais tolerante à erros nas coordenadas atômicas e à ligações de hidrogênios bifurcadas (**Kabsch1983**).

Uma vez definido o modelo para identificar ligações de hidrogênio, essas são testadas 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 $i + n$.

$$n\text{-}Turn \impliedby Hbond(i, i+n), n \in \{3, 4, 5\} \tag{2.3}$$

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

$$Parallel\ Bridge \impliedby [Hbond(i-1, j) \wedge Hbond(j, i+1)] \vee [Hbond(j-1, i) \wedge Hbond(i, j+1)] \tag{2.4}$$

$$Antiparallel\ Bridge \impliedby [Hbond(i, j) \wedge Hbond(j, i)] \vee [Hbond(i-1, j+1) \wedge Hbond(j-1, i+1)] \tag{2.5}$$

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 é utilizada para a atribuição dos elementos de estrutura secundária. Como exemplos, a repetição de padrões *4-Turn* consecutivos indica a ocorrência de uma hélice α, enquanto que resíduos consecutivos que apresentem padrões *Bridge* formam uma fita de uma folha β.

No trabalho, os autores mencionam que esse algoritmo proposto produz hélices mais curtas, com um resíduo a menos em cada extremidade, 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 multiplas hélices com *kinks*. O mesmo ocorre com resíduos que formam *bulges* em fitas, sendo os mesmo também anotados como parte da fita.

2.1.2 Stride

stride

2.1.3 KAKSI

KAKSI é um método de atribuição de estruturas secundárias proposto por Martin e colaboradores **Martin2005** Esse método foi desenvolvido utilizando padrões de distâncias entre C_α e de ângulos Φ e Ψ.

Resumidamente, a heurística de atribuição de estrutura secundárias busca primeiramente por hélices, sendo que um resíduo é classificado como hélice se atender aos critérios de distâncias entre C_α ou aos 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 atender aos critérios de disntâncias entre C_α e aos critérios de ângulos Φ e Ψ.

1. Critérios para classificação de hélices:

   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.

Ângulos Φ e Ψ

Todos os pares de ângulos Φ e Ψ em uma janela de quatro resíduos precisam satifazer 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:

Distância entre C_α

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

Ângulos Φ e Ψ

Cada par de ângulos Φ e Ψ presente na zona povoada de resíduos em folhas β incrementa um contador em 1. Quando um resíduo central da janela apresenta $-120° < \Psi < 50°$, o contador é reiniciado em zero. Esse critério é satifeito se o contator $>= \sigma_b$.

Além destes critérios, há critérios 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 cristalogria e com resolução superior a 2.25Å.

2.1.4 PROSS

pross

2.2 Métodos de predição de estruturas secundárias

2.2.1 Primeira geração (1957-1978)

Dentre os métodos de predição de estrutura secundária da primeira geração, o mais reconhecido foi desenvolvido por Chou e Fasman **key**

Esse método foi desenvolvido utilizando a frequência de cada aminoácidos em cada tipo de estrutura secundária, hélices α ou folhas β. A análise descrita no trabalho **key** foi utilizada para determinar os parâmetros P_α e P_β usados na predição.

A partir dos parâmetros P_α e P_β eles determinaram as seguintes regras para predição:

A. Hélices 1. Nucleação de hélices: Regiões de seis resíduos com ao menos quatro deles sendo h_α ou H_α. A formação de hélice é desfavorável se o segmento contiver um terço ou mais de hélices breakers (b_α ou B_α) ou menos da metade de hélices formers.

Table 2.1: My caption

P_α			P_β			$P_\alpha - P_\beta$	
E	1,53		M	1,67		E	1,27
A	1,45	H_α	V	1,65	H_β	H	0,53
L	1,34		I	1,60		A	0,48
H	1,24		C	1,30		K	0,33
M	1,20		Y	1,29		D	0,18
Q	1,17		F	1,28		L	0,12
V	1,14	h_α	Q	1,23	h_β	N	0,08
W	1,14		L	1,22		S	0,07
F	1,12		T	1,20		P	-0,03
K	1,07	I_α	W	1,19		W	-0,05
I	1,00		A	0,97	I_β	Q	-0,06
D	0,98		R	0,90		R	-0,11
T	0,82		G	0,81	i_β	F	-0,16
R	0,79	i_α	D	0,80		G	-0,28
S	0,79		K	0,74		T	-0,38
C	0,77		S	0,72		M	-0,47
N	0,73	b_α	H	0,71	b_β	V	-0,51
Y	0,61		N	0,65		C	-0,53
P	0,59	B_α	P	0,62		I	-0,6
G	0,53		E	0,26	B_β	Y	-0,68

2. Terminação de hélices: As hélices iniciadas pela regra A.1. são estendidas até a presença de um tetrapeptídeo com $\langle P_\alpha \rangle < 1.0$. Regiões de folha β também podem terminar as regiões de hélice.

3. Prolinas não podem ocorrer na região central de hélices ou na extremidade C-terminal da mesma.

4. Bordas das hélices: Prolinas, aspartatos e glutamatos preferem a extremidade N-terminal das hélices enquanto histidinas, lisinas e argininas preferem a extremidade C-terminal.

Regra 1: qualquer segmento de seis resíduos ou mais com $\langle P_\alpha \rangle > 1.03$, $\langle P_\alpha \rangle > \langle P_\beta \rangle$ e satisfazendo as condições A.1. até A.4. é predito como hélice.

B. Folhas β 1. Nucleação: Regiões de 5 resíduos com ao menos 3 deles sendo h_β ou H_β. A formação de fita β é desfavorável se o segmento contiver um terço ou mais de fita breakers (b_β ou B_β) ou menos da metade de fita formers. 2. Terminação: As fitas iniciadas pela regra B.1. são estendidas até a presença de um tetrapeptídeo com $\langle P_\beta \rangle < 1.0$. Regiões de hélice α também podem terminar as regiões de folhas β. 3. Glutamatos ocorrem raramente em regiões de folha β. Prolinas ocorrem raramente na região central das mesmas. 4. Bordas das folhas β: Residuos carregados ocorrem raramente na extremidade N-terminal da fita e são pouco frequentes nas regiões central e C-terminal. Triptofano ocorre com maior frequência na região N-terminal e raramente na C-terminal da fita β.

Regra 2: qualquer segmento de cinco resíduos ou mais com $\langle P_\beta \rangle > 1.05$, $\langle P_\beta \rangle > \langle P_\alpha \rangle$ e satisfazendo as condições B.1. até B.4. é predito como fita β.

Os parâmetros foram calculados utilizando 15 estruturas proteicas. A aplicação das regras a essas estruturas apresentou acurácia de 77% na classificação dos resíduos entre hélices, fitas β e coil. Entretanto, outros trabalhos sugerem que a acurácia do método esteja entre 50-60%. Possivelmente, o uso do mesmo conjunto de proteínas para treinamento e teste contribuiu para que a acurácia fosse superestimada.

2.2.2 Segunda geração (1983-1992)

segunda geração - GORIII

2.2.3 Terceira geração

Qian e Sejnowski **key** foram possivelmente os primeiros a publicarem um trabalho utilizando redes neurais artificiais para a predição da estrutura secundária de proteínas. Segundo os autores, a inspiração surgiu de um trabalho que utilizava redes neurais artificiais para converter texto em fonemas (*text-to-speech*), cujo primeiro autor era Sejnowski (ref NETtalk 1986).

A melhor arquitetura testada por eles consistia de duas redes neurais em sequência. A primeira recebia como entrada 13 resíduos de uma sequência proteíca e como saída emitia 3 valores no intervalo entre 0 e 1 correspondentes aos elementos de estrutura secundária coil, hélice e fita. Cada resíduo era codificado em um vetor binário com 21 elementos representando os 20 aminoácidos e um elemento indicando a ausência de

um aminoácido. A segunda rede neural tinha como entrada os valores de sáida da rede anterior para uma janela de 13 resíduos e emitia novamente 3 valores no intervalo entre 0 e 1 representando a pontuação para os 3 elementos de estrutura secundária. O elemento com o maior valor correspondia a estrutura secundária predita.

Os autores exploraram também a utilização de uma camada oculta de neurônios, mas esta, apesar de diminuir o erro de classificação durante o treinamento, foi incapaz de reduzir o erro no conjunto de teste.

Segundo os autores, a performance de 64,3% (Q3) atingida pela rede neural artifical, apesar de superior aos métodos anteriores, foi decepcionante. Dentre as possíveis explicações, eles minimizaram o impacto de estarem utilizando apenas 106 estruturas (85 para treinamento) com o argumento de que um maior número de estruturas não irá melhorar a predição para proteínas não homólogas as do conjunto de treinamento. Assim, a explicação mais plausível segundo eles era que a influência de resíduos mais distantes na sequência, e que portanto não estão presentes na janela utilizada como entrada, precisaria ser considerada na predição. Caso isso fosse confirmado, eles acreditavam que seriam necessários novos métodos para considerar esses efeitos. Eles concluem dizendo que um banco de dados maior de proteínas homólogas poderia permitir uma rede neural artificial aprender a equivalência dos aminoácidos em diferentes contextos na proteína.

Outro modelo de rede neural aplicado à predição de estruturas secundárias foi proposto por Holley e Karplus **key** Este modelo também utilizou uma janela, neste caso de 17 aminoácidos. Nesta rede, os autores utilizaram uma camada de neurônios oculta, sendo a camada com dois neurônios a que demonstrou maior acurácia no conjunto de teste. A acurácia do método foi de aproximadamente 63%. Os autores testaram ainda uma codificação de aminoácidos com características como hidrofobicidade, carga e flexibilidade da cadeia principal, entretanto, a acurácia observada foi de 61%, inferior a codificação binária.

2.2.4 Quarta geração

quarta geração - aprendizado de máquina (NN) + dados evolutivos (PSSM) - PSIPred

3 Materiais e métodos

3.1 Conjunto de dados

Estruturas

3.2 Rede Neural similar ao modelo de Holley e Karplus

O modelo original de Holley e Karplus **key** utilizava dois neurônios na camada oculta. Posteriormente, Chandonia e Karplus **10.1002/pro.5560050422** aumentaram o conjunto de dados de 62 para 318 proteínas e observaram que um aumento da camada oculta de 2 para 8 neurônios produzia um aumento da acurácia de 63% para 67%.

Para avaliarmos qual seria o desempenho de uma rede neural com topologia similar a proposta por Holley e Karplus **key** nós implementamos algumas redes neurais utilizando o framework Pytorch e treinamos com o conjunto de proteínas utilizado ao longo desse trabalho.

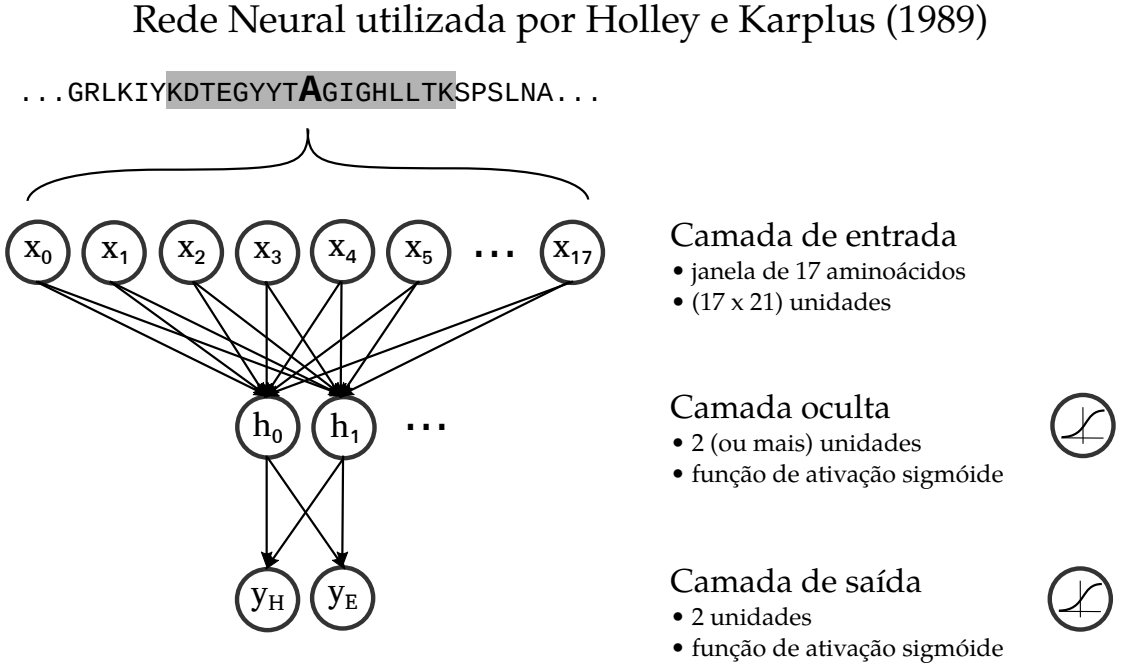


Figure 3.1: Rede neural utilizada nos trabalhos de Holley e Karplus **key** e Chandonia e Karplus **10.1002/pro.5560050422** A rede neural utiliza como entrada uma janela de 17 aminoácidos da proteína, cada um codificado em um vetor de tamanho 21 (20 aa + 1 posição que indica ausência de aminoácidos). Na camada oculta foram testadas várias configurações, diferindo entre si pelo número de neurônios. A camada de saída possuía 2 neurônios, um representando a saída para hélice e outro para fitas. Todos os neurônios possuiam funções de ativação sigmóide. Os rótulos de treinamento foram $(1,0) \rightarrow hélice$, $(0,1) \rightarrow fita$, $(0,0) \rightarrow coil$.

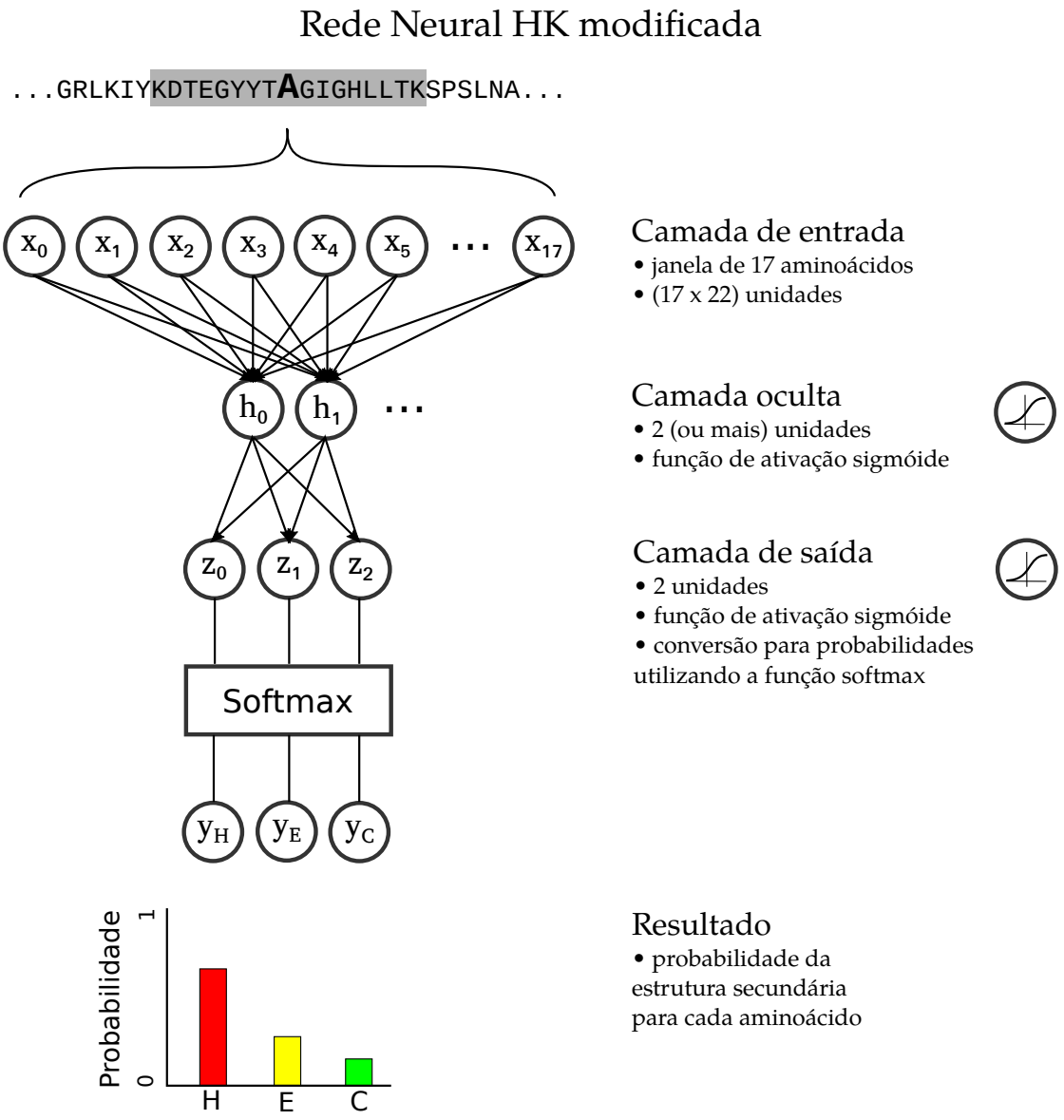

Rede Neural HK modificada
...GRLKIYKDTEGYYTAGIGHLLTKSPSLNA...
x0 x1 x2 x3 x4 x5 ... x17
h0 h1 ...
z0 z1 z2
Softmax
yH yE yC
Probabilidade
1
0
H E C
Camada de entrada
• janela de 17 aminoácidos
• (17 x 22) unidades
Camada oculta
• 2 (ou mais) unidades
• função de ativação sigmóide
Camada de saída
• 2 unidades
• função de ativação sigmóide
• conversão para probabilidades utilizando a função softmax
Resultado
• probabilidade da estrutura secundária para cada aminoácido

Figure 3.2

4 Resultados

4.1 Análise de métodos de atribuição de estrutura secundária

Diferentemente de vários métodos publicados de predição de estrutura secundária que utilizaram os resultados de atribuição provenientes de apenas um método, comumente o DSSP, nós optamos por utilizar os resultados de um consenso entre quatro métodos. Consequentemente, é importante analisar como esses métodos diferem entre si.

As diferenças entre as metologias de atribuição foram previamente discutidas na seção **??**. Nesta seção, analisaremos as diferenças nas estruturas secundárias atribuídas aos resíduos.

As análises a seguir foram feitas entre os resíduos de 6749 proteínas do banco de dados Top8000-HOM50 que tiveram estrutura secundária atribuída por todos os métodos. Isso excluí 5,19% dos resíduos que não tiveram a estrutura secundária atribuída por não terem sido visualizados na estrutura experimental e 0,05% dos resíduos que não tiveram sua estrutura atribuída por um ou mais programas devido à possíveis falhas dos programas.

Notamos que, entre os quatro métodos de atribuição analisados, há a ocorrência de uma variação considerável entre as estruturas secundárias atribuídas para cada resíduo. A figura **??** demonstra a similaridade observada entre os métodos. Como esperado, a maior similaridade é observada entre os métodos DSSP e STRIDE, uma vez que o último é inspirado no primeiro. Entretanto, a similaridade entre os demais métodos encontra-se na faixa entre 83% e 85%. A porcentagem de resíduos que apresentaram consenso entre os quatro métodos foi de apenas 74,43%.

4.2 Resultados de predição com rede neurais artificiais

O método de predição proposto por Holley e Karplus em **key** foi treinado utilizando 48 estruturas proteicas resolvidas, de um conjunto de 62 estruturas selecionadas. Nessas condições, o método demonstrou uma acurácia de aproximadamente 63%. Posteriormente, Chandonia e Karplus **10.1002/pro.5560050422** demonstraram que o aumento no número de estruturas do conjunto de treinamento poderia aumentar a acurácia da predição utilizando redes neurais. Entretanto, como observado por eles, tal aumento pode requerer modificações na topologia da rede neural. Assim, com um conjunto de 318 estruturas proteicas e aumentando o número de neurônios da camada oculta de 2 para 8, eles conseguiram uma acurácia de 67%.

Tabela

A acurácia máxima observada de 74% para o conjunto de teste evidencia que a quantidade de dados e informação ainda é um dos fatores que a serem explorados para aumentar

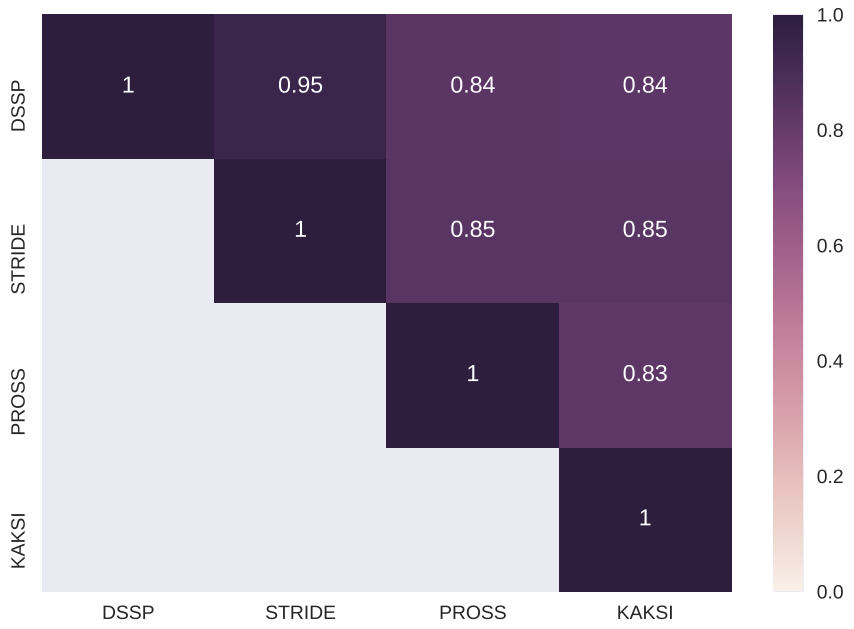


Figure 4.1: Similaridade entre as estruturas secundárias atribuídas para cada resíduo entre os quatro métodos de atribuição de estruturas secundárias: DSSP, STRIDE, KAKSI e PROSS.

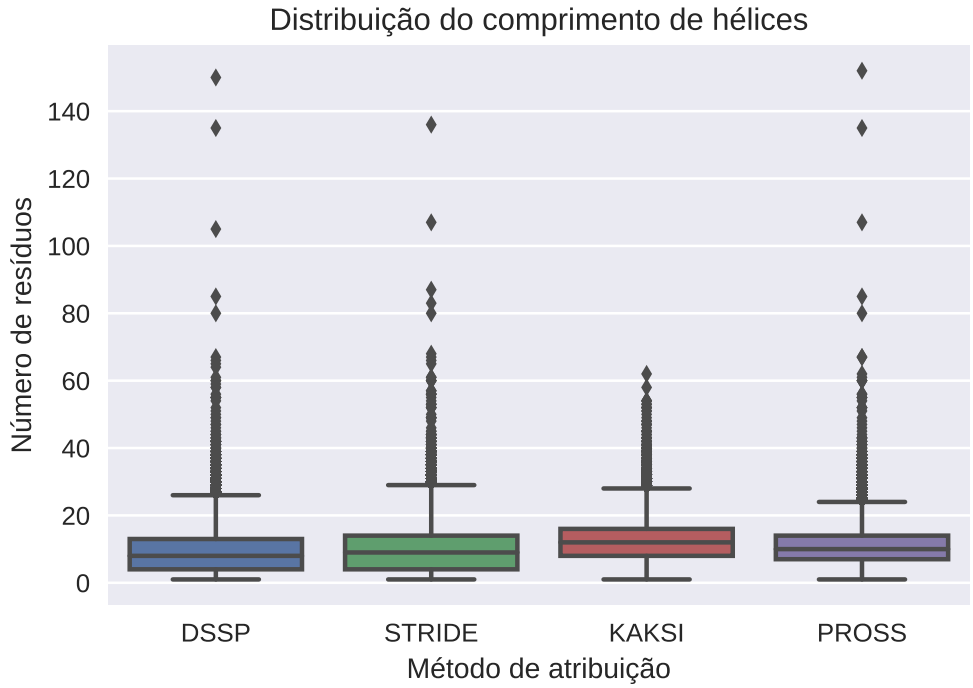


Figure 4.2: Distribuição do comprimento de hélices por método de atribuição de estrutura secundária.

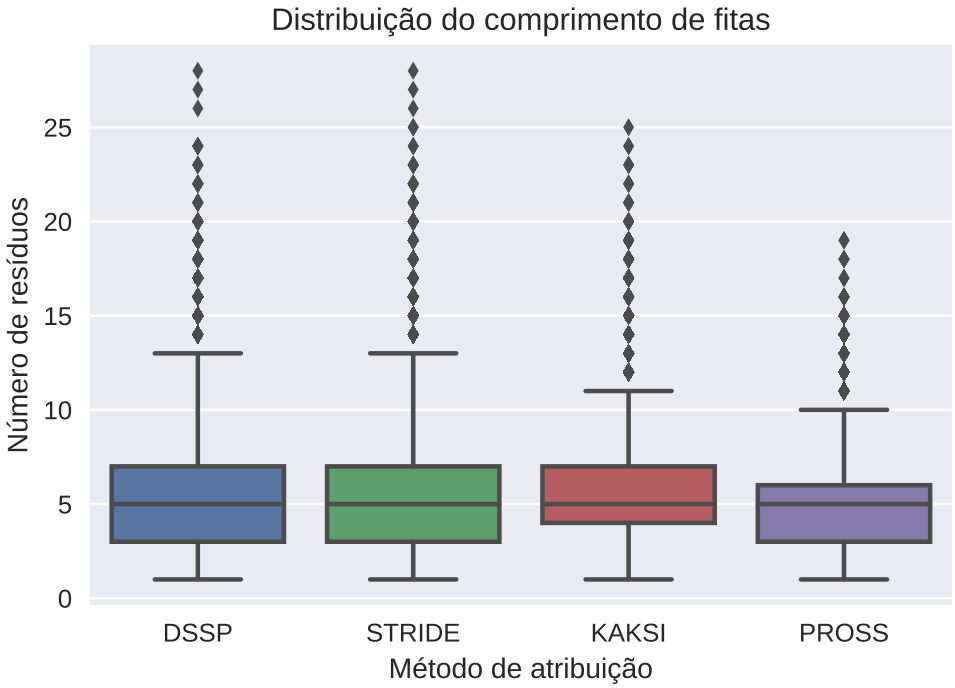


Figure 4.3: Distribuição do comprimento de fitas por método de atribuição de estrutura secundária.

a acurácia. Assim, o platô dos métodos de predição baseados em redes neurais artificiais e que utilizam apenas a sequência de aminoácidos como entrada pode não ter sido atingido ainda.

4.3 Resultados de predição com o PSIPRED

O PSIPRED é um método de predição de estruturas secundárias que utiliza redes neurais artificias em conjunto com PSSM (**10.1006/jmbi.1999.3091**) (ver **??**). O método foi originalmente treinado com informações da estrutura secundária atribuídas pelo DSSP.

Devido as variações observadas entre os métodos de atribuição de estrutura secundária, nós realizamos a comparação dos resultados de predição com a estrutura secundária atribuída por diferentes métodos, assim como com o consenso entre os métodos de atribuição.

No conjunto de proteínas utilizado nesse trabalho, o PSIPRED comparado à atribuição pelo DSSP, demonstrou uma acurácia média (Q3) de 86%, superior aos 78% descrito na literatura.

A acurácia média para a predição de fitas β, como esperado, foi inferior a predição de hélices e coils.

Um resultado interessante foi a acurácia média observada entre a predição e o consenso dos métodos de atribuição estrutura secundária. Tanto a acurácia geral (Q3) quantos a acurácia por classe (Qh, Qe, Qc) demonstraram aumentos significativo em comparação com os métodos de atribuição individuais.

Isso indica que nas regiões onde não há consenso entre os métodos de atribuição, a acurácia média (Q3) é próxima ou inferior a 68%.

$Q3_{\text{consenso}} * P_{\text{consenso}} + Q3_{\text{não consenso}} * P_{\text{não consenso}} = Q3_{\text{total}}$
$Q3_{\text{não consenso}} = \frac{Q3_{\text{total}} - Q3_{\text{consenso}} * P_{\text{consenso}}}{P_{\text{não consenso}}}$
$Q3_{\text{não consenso}} = \frac{0.86 - 0.92 * 0.75}{0.25}$
$Q3_{\text{não consenso}} = 0.68$

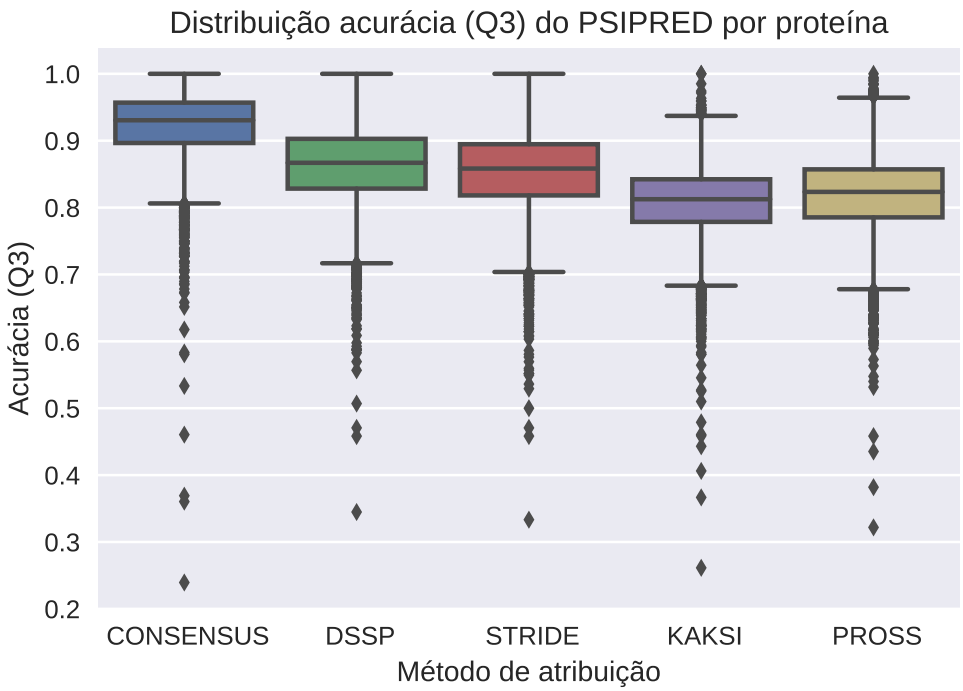

Distribuição acurácia (Q3) do PSIPRED por proteína
Acurácia (Q3)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
CONSENSUS
DSSP
STRIDE
KAKSI
PROSS
Método de atribuição

Figure 4.4

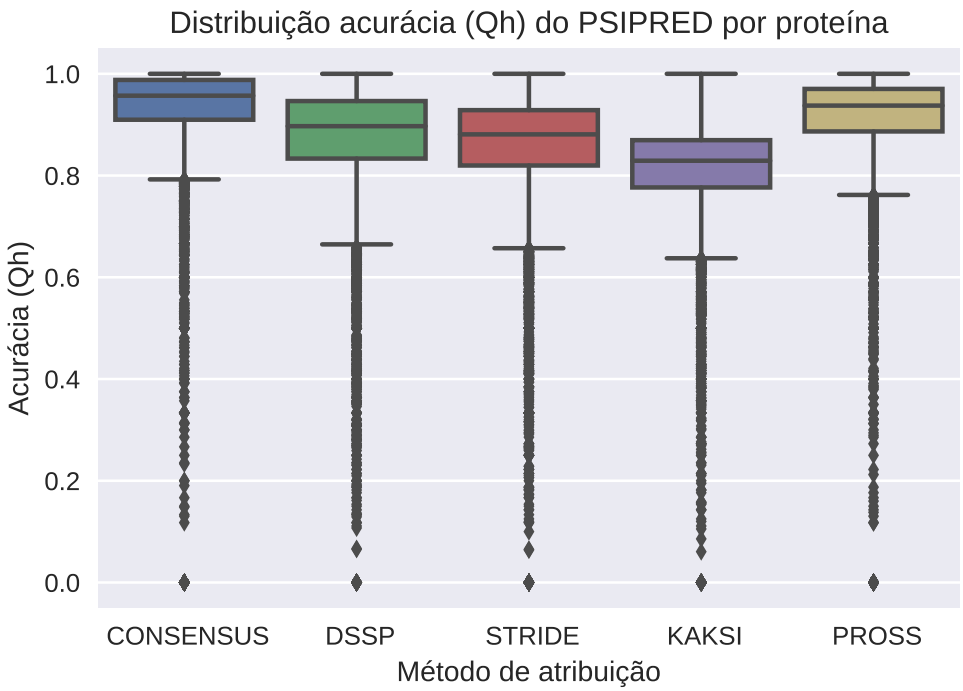

Distribuição acurácia (Qh) do PSIPRED por proteína
Acurácia (Qh)
1.0
0.8
0.6
0.4
0.2
0.0
CONSENSUS
DSSP
STRIDE
KAKSI
PROSS
Método de atribuição

Figure 4.5

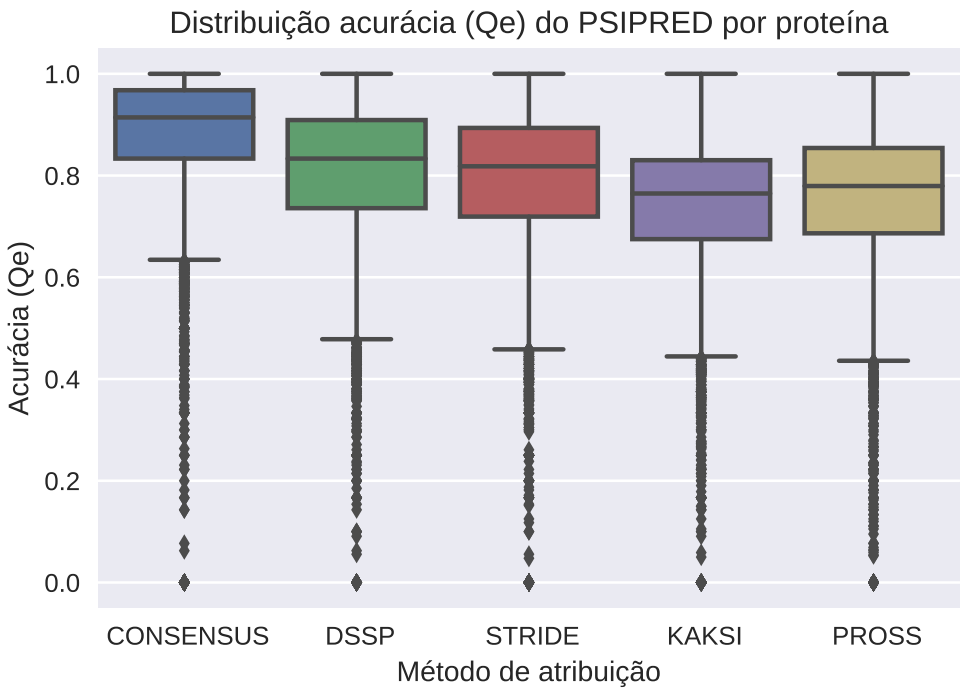

Distribuição acurácia (Qe) do PSIPRED por proteína
Acurácia (Qe)
1.0
0.8
0.6
0.4
0.2
0.0
CONSENSUS
DSSP
STRIDE
KAKSI
PROSS
Método de atribuição

Figure 4.6

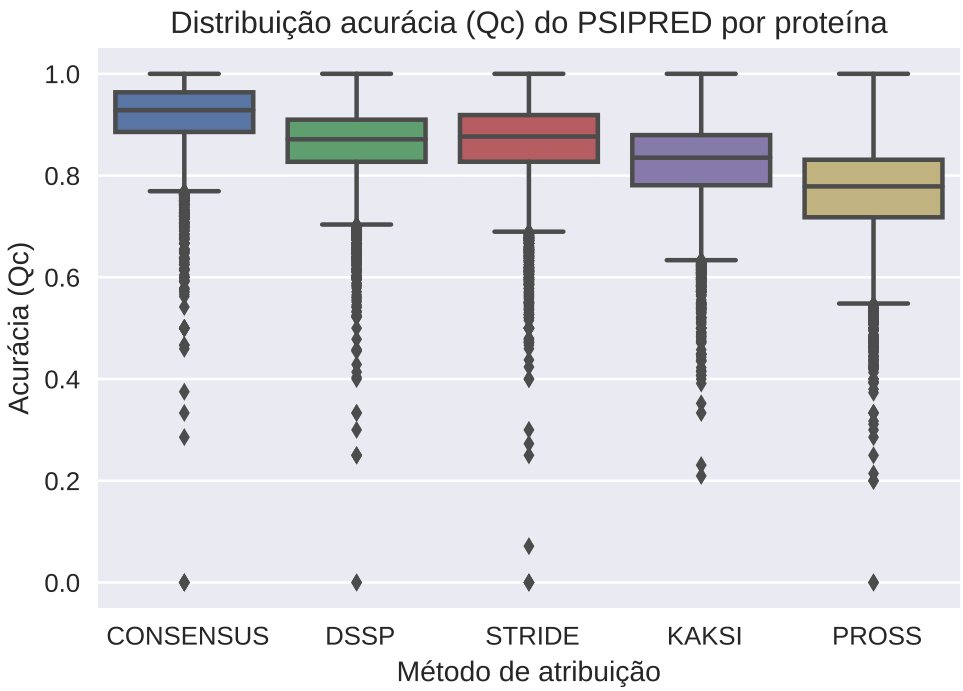

Distribuição acurácia (Qc) do PSIPRED por proteína
Acurácia (Qc)
1.0
0.8
0.6
0.4
0.2
0.0
CONSENSUS
DSSP
STRIDE
KAKSI
PROSS
Método de atribuição

Figure 4.7

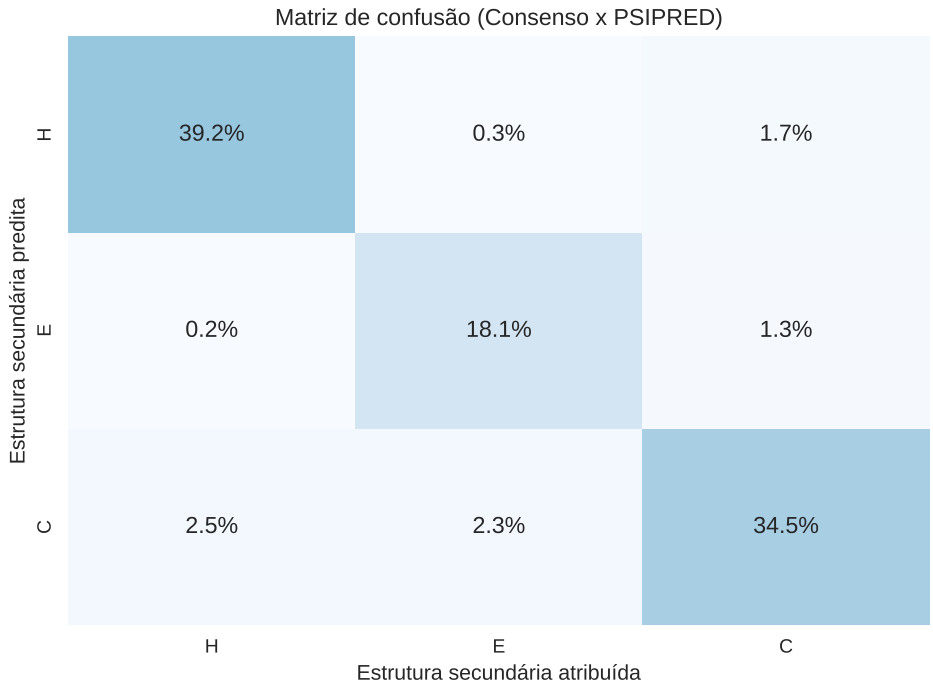
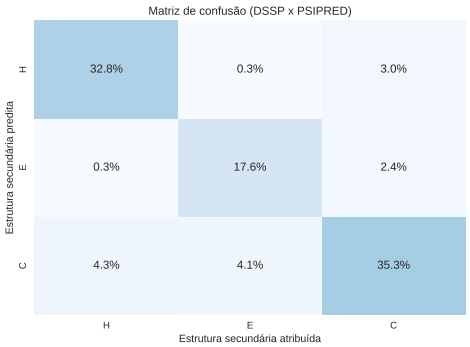
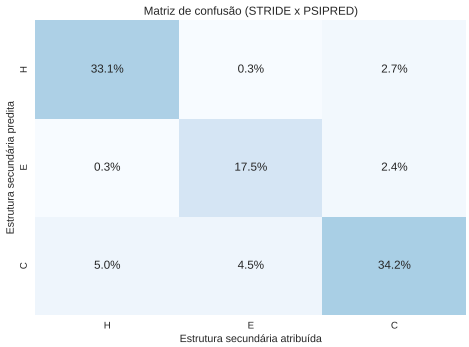

Matriz de confusão (Consenso x PSIPRED)
Estrutura secundária predita
H
E
C
39.2%
0.3%
1.7%
0.2%
18.1%
1.3%
2.5%
2.3%
34.5%
H
E
C
Estrutura secundária atribuída

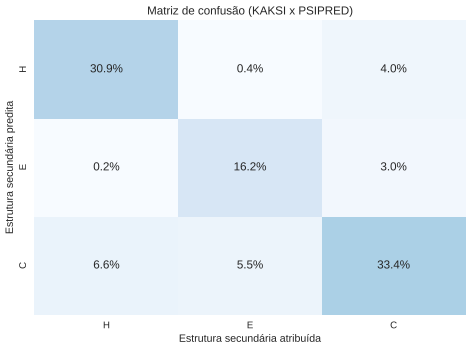
Figure 4.8



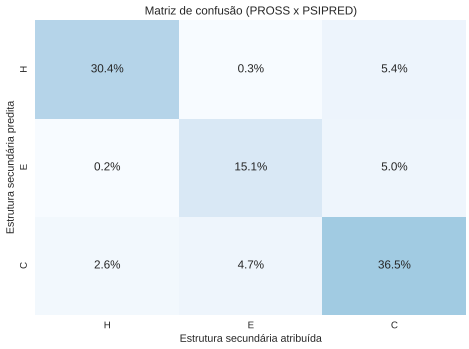
(a) Initial condition



(b) Rupture



(c) DFT, Initial condition



(d) DFT, rupture

Figure 4.9: Illustration of various images