

Proposta de otimização do método CSM de extração de features de proteínas via três métricas de distância

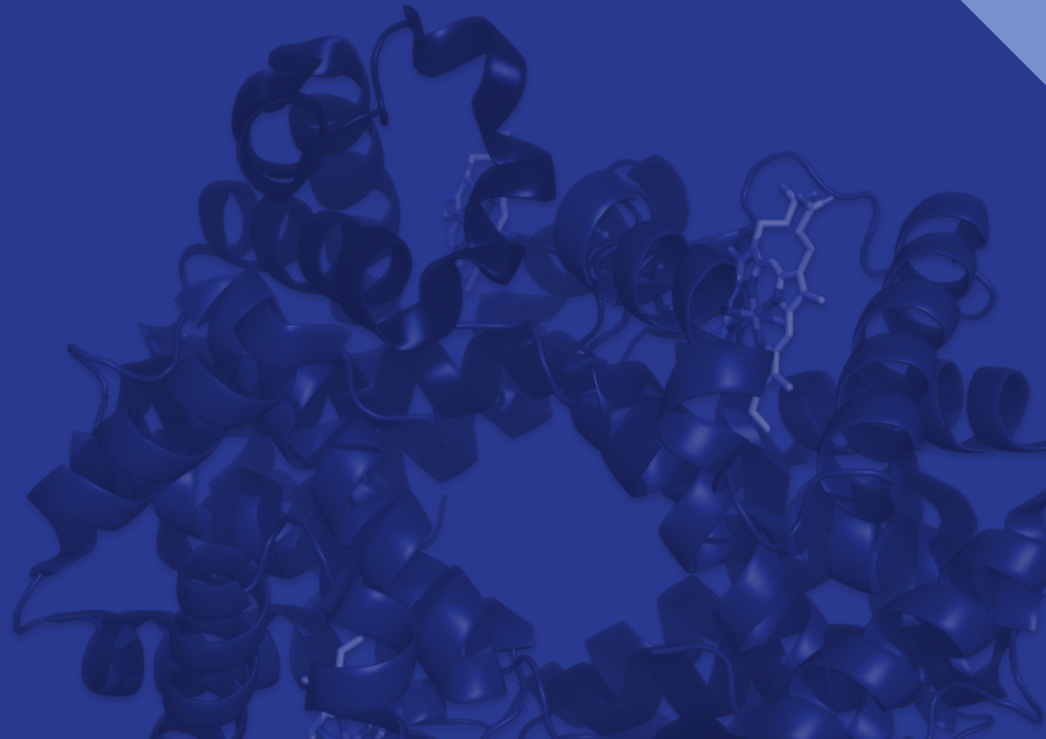
André Luiz Moreira Dutra
Ruhann Carlos Pereira de Almeida



Sumário

- Motivação
- Hipótese e Proposta de Otimização
- Metodologia
- Dados Utilizados
- Resultados
- Problemas encontrados
- Perspectivas futuras

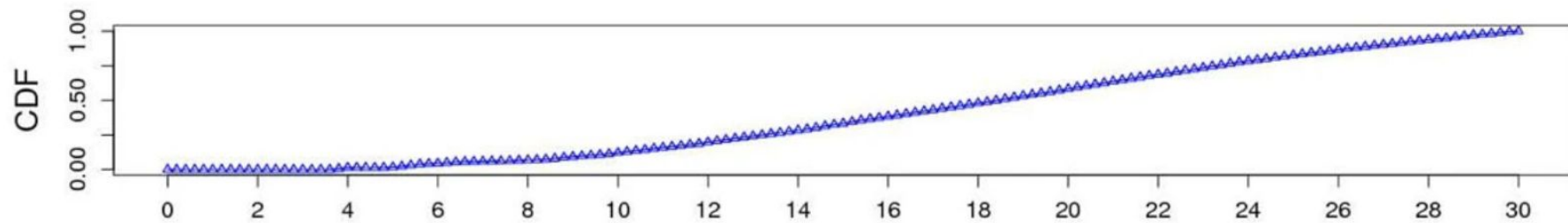
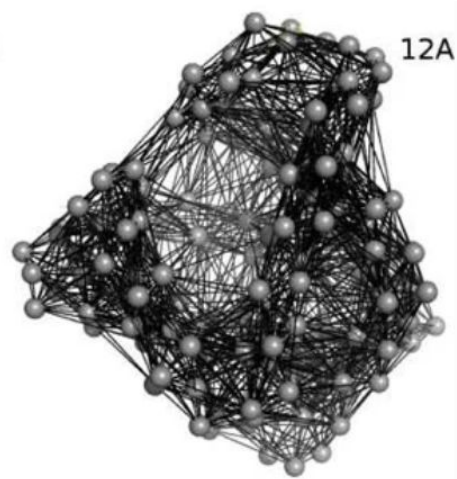
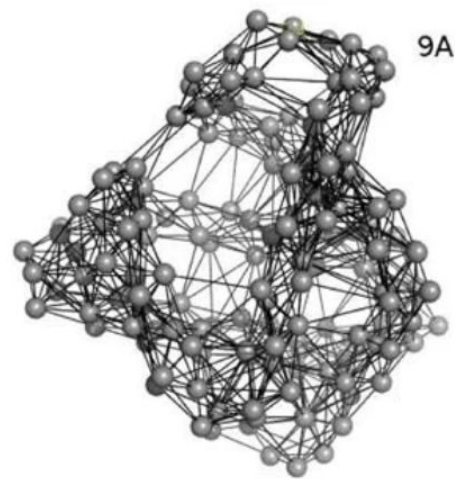
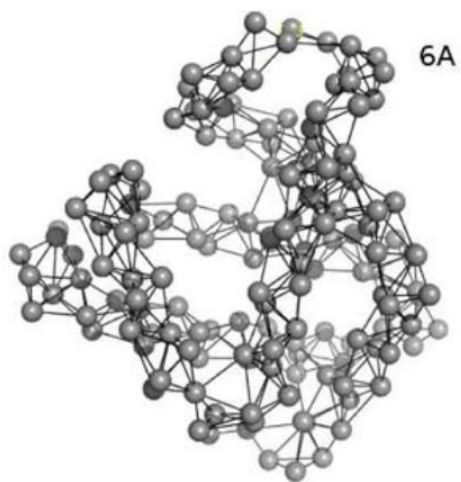
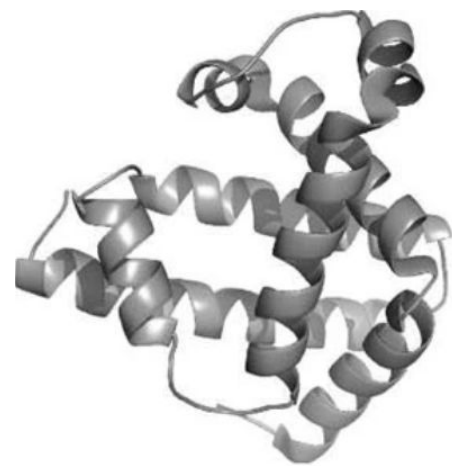
Motivação



O método CSM

- Método de extração de features de moléculas para fins de classificação.
- Se baseia na matriz de distâncias entre os átomos avaliados.
- Realiza uma varredura por intervalos quantizados na matriz de distâncias, fazendo a contagem de pares de átomos com distância contida em cada intervalo avaliado.
- O conjunto das contagens constitui o vetor de features da molécula, e o conjunto de vetores constitui a matriz CSM.



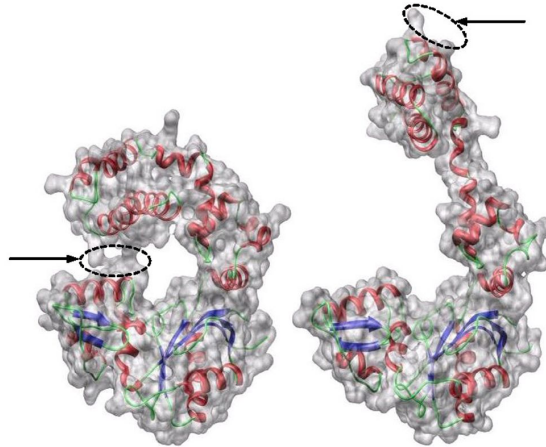


Algorithm 1 Cutoff Scanning Matrix calculation

```
function GENERATECSM(ProteinSet, CSM, DistanceMIN, DistanceMAX, DistanceSTEP)  
  for all protein i  $\in$  (ProteinSet) do  
    j = 0  
    Calculate the distances between all pairs of  $C_{\alpha}$   
    for dist  $\leftarrow$  DistanceMIN; to DistanceMAX; step DistanceSTEP do  
      CSM[i][j]  $\leftarrow$  Get frequency of pairs of  $C_{\alpha}$  within a distance dist  
      j ++  
  return CSM
```

Flexibilidade das proteínas

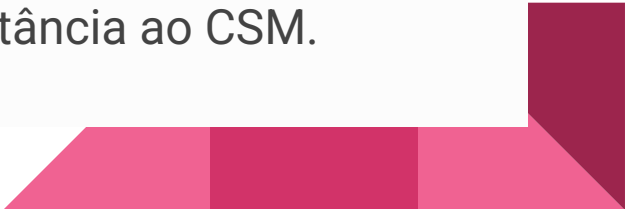
- A estrutura da proteína é flexível, podendo se apresentar em múltiplas disposições espaciais diferentes.
- No entanto, o CSM se baseia em um frame estático da molécula, classificando-a segundo apenas uma de suas múltiplas disposições.



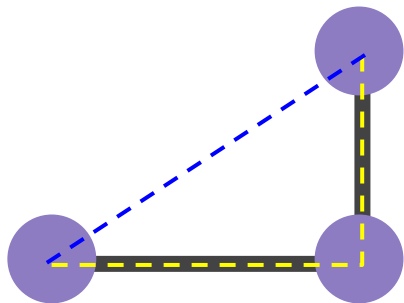
Hipótese e Proposta de Otimização



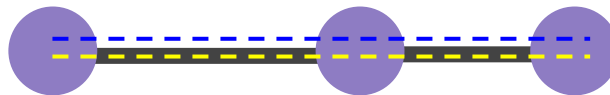
Hipótese

- O CSM original utiliza a distância euclidiana entre os átomos para o cálculo das features.
 - “Será possível capturar múltiplas disposições de uma proteína flexível em um mesmo descritor utilizando métricas de distância menos restritivas?”
 - Se sim, uma otimização simples ao CSM poderia ser feita utilizando a métrica de distância ideal para a geração de features.
 - Essencialmente, objetivamos observar o desempenho da classificação de features geradas aplicando diferentes métricas de distância ao CSM.
- 

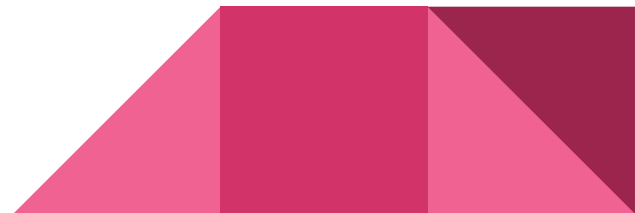
Hipótese



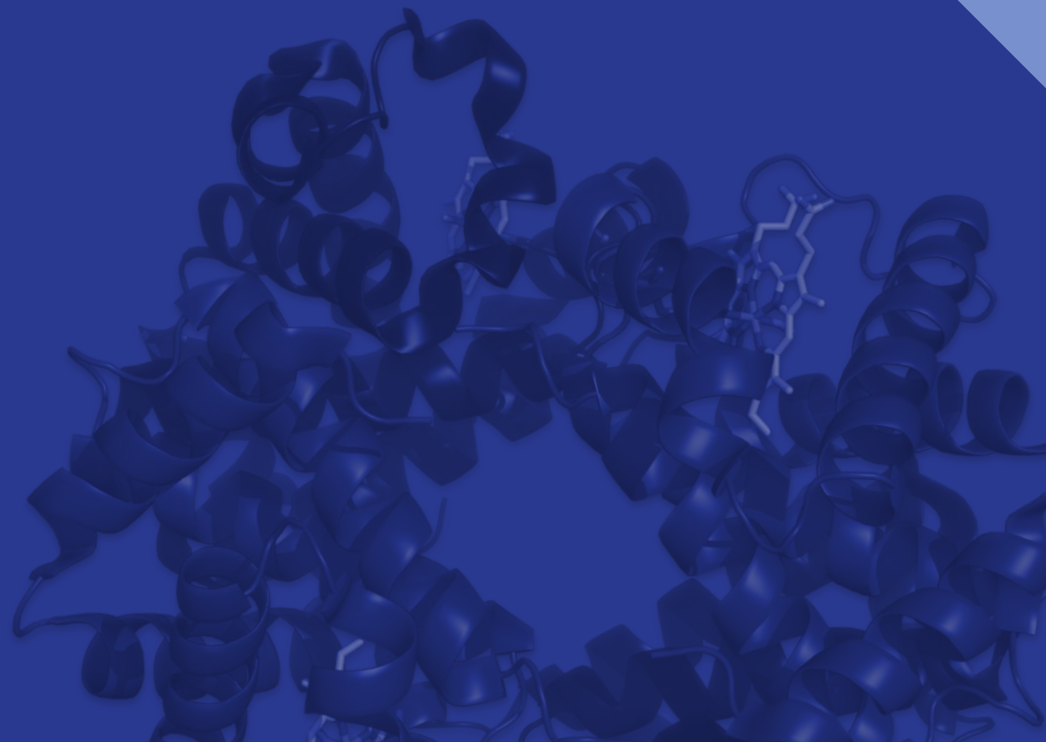
euc = 5
manh = 7



euc = 7
manh = 7



Metodologia

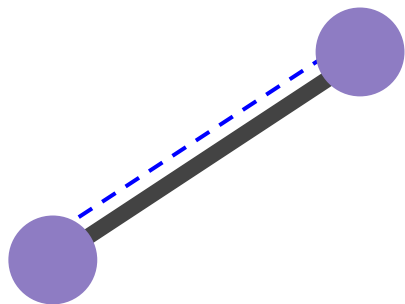


Modelos utilizados

- Três métricas de distância:
 - Euclidiana
 - Distância de Manhattan
 - Distância de Chebyshev
- Três classificadores:
 - Support Vector Machine-based classifier (núcleo RBF)
 - Decision Tree (altura máxima = 5)
 - Multi-Layer Perceptron Classifier (Rede Neural, uma camada de 100 neurônios)

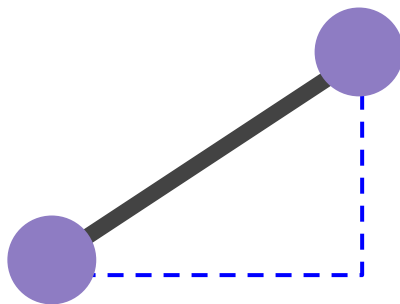


Métricas de distância



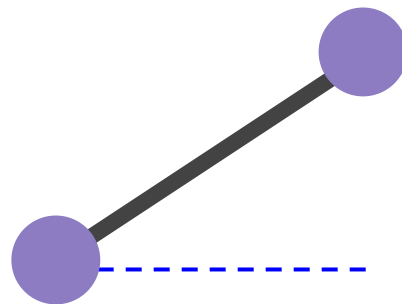
Euclidiana

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Manhattan

$$d = |x_2 - x_1| + |y_2 - y_1|$$



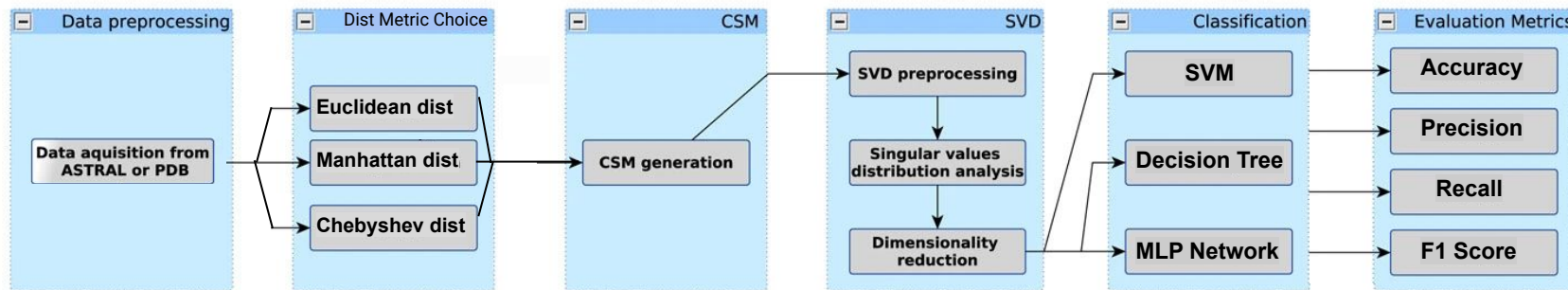
Chebyshev

$$d = \max(|x_2 - x_1|, |y_2 - y_1|)$$

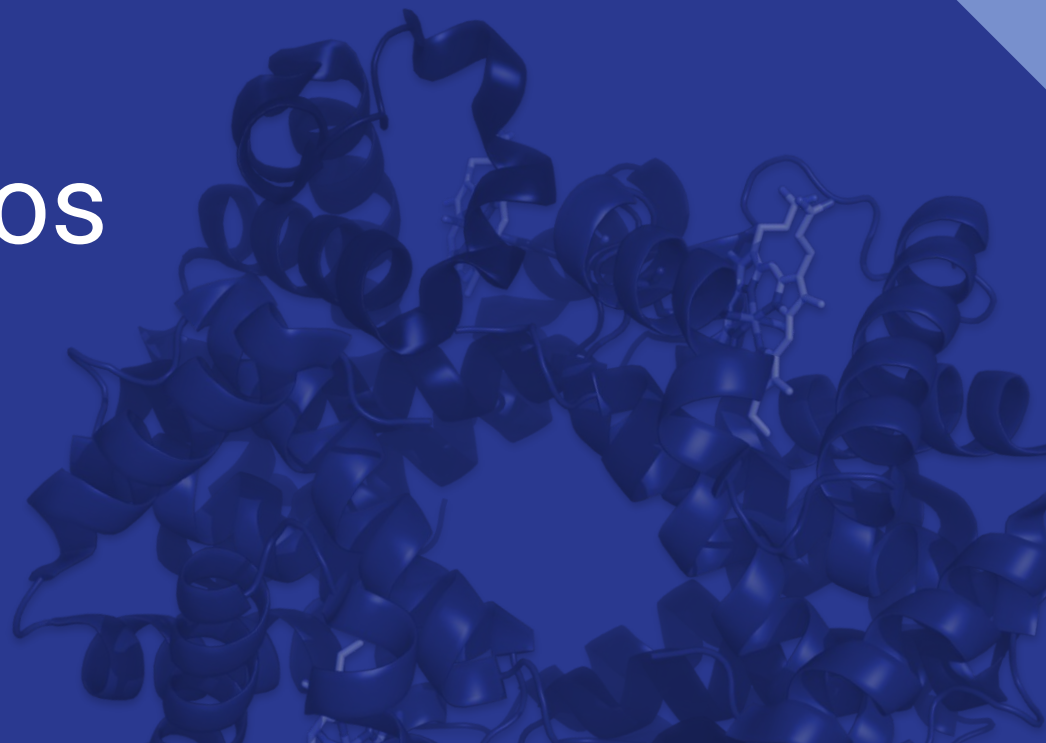


Pipeline de dados

O processo utilizado é semelhante ao apresentado em [1], alterando apenas os métodos de classificação e adicionando outras métricas de avaliação.



Dados Utilizados



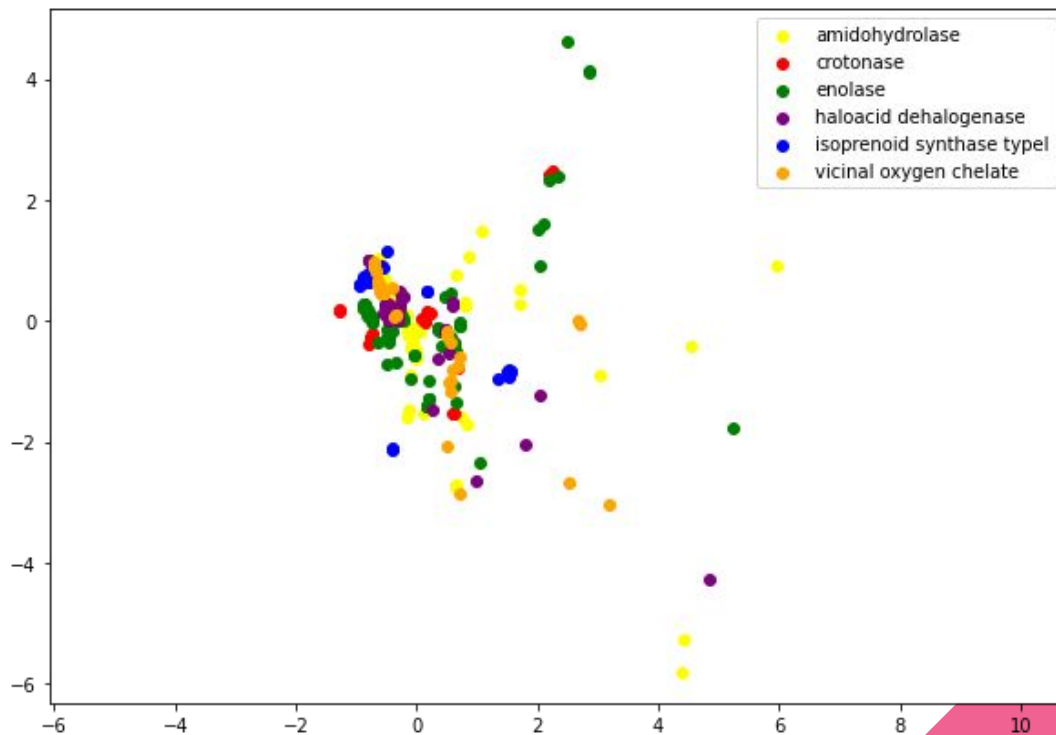
Base de Dados Utilizada

- Conjunto padrão-ouro de superfamílias de enzimas mecanisticamente diversas, disponível em [3].
- Foram consideradas seis superfamílias (amidohidrolase, crotonase, haloácido desalogenase, isoprenóide sintase tipo I e oxigênio quelato vicinal).
- 47 famílias distribuídas em 365 cadeias diferentes.
- Classificação feita apenas para as superfamílias.

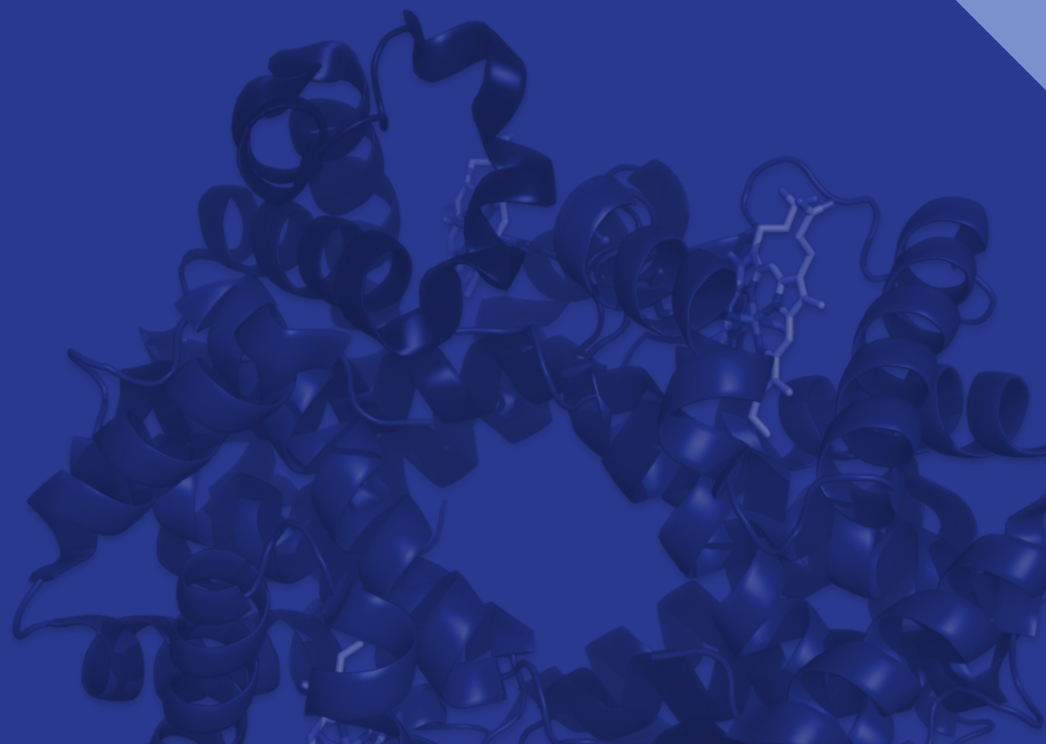


Visualização dos Dados

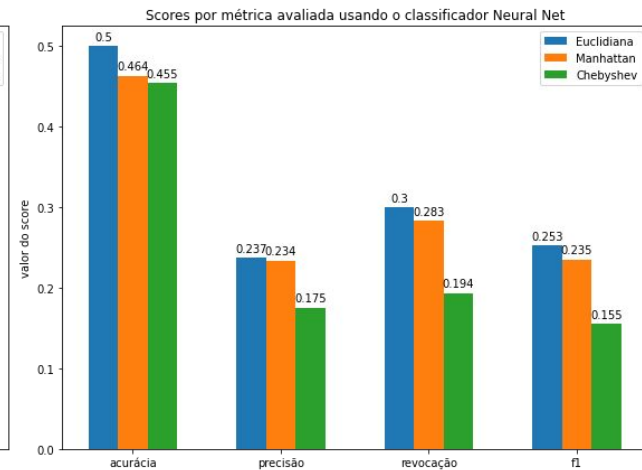
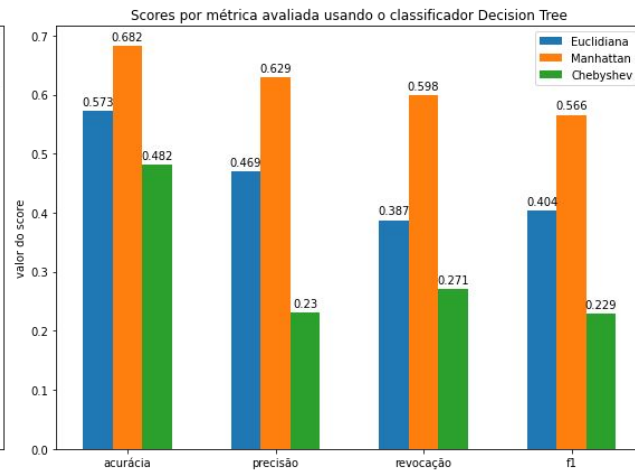
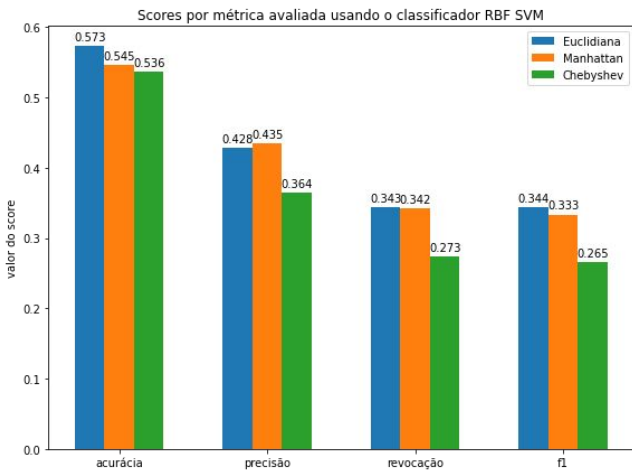
Distribuição pós PCA normalizada para a base de distância euclidiana(CA)



Resultados

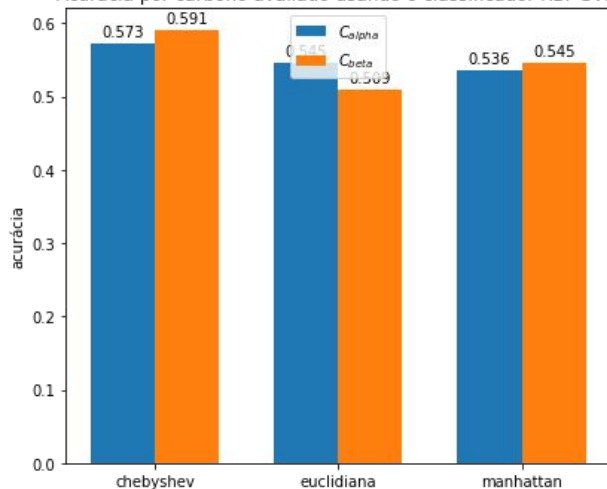


Scores de Classificação para cada métrica

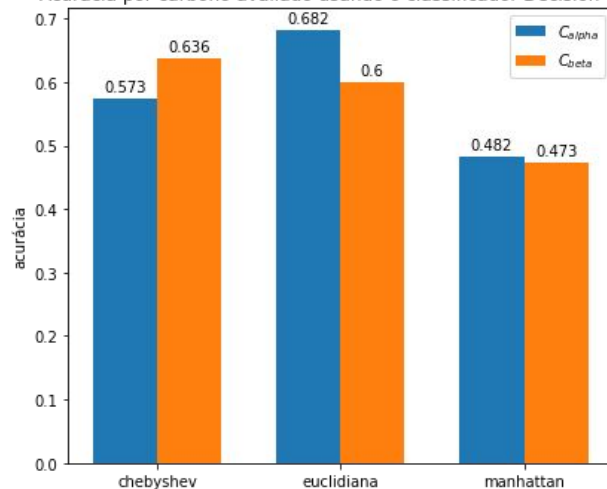


Scores de cada métrica para os carbonos alfa e beta

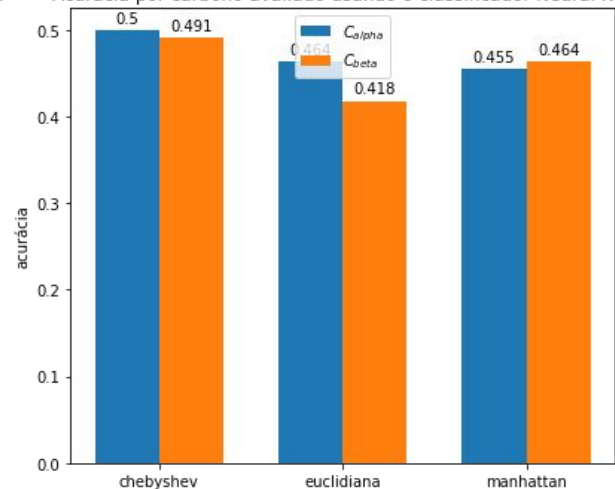
Acurácia por carbono avaliado usando o classificador RBF SVM



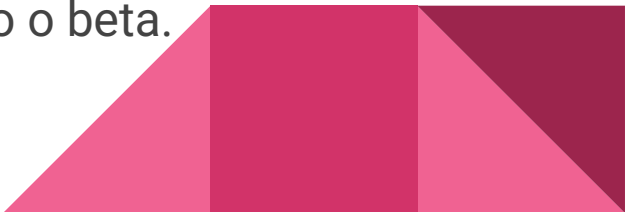
Acurácia por carbono avaliado usando o classificador Decision Tree



Acurácia por carbono avaliado usando o classificador Neural Net



Discussão

- O score de todas as classificações foi relativamente baixo, muito provavelmente devido à base de dados pequena.
 - A distância de Manhattan apresentou um desempenho consideravelmente alto em relação às demais, porém condicionado à escolha do classificador.
 - A escolha do carbono alfa em relação ao beta só trouxe diferenças significativas utilizando a distância euclidiana, corroborando o artigo original.
 - Para outras métricas, o desempenho utilizando o carbono alfa foi ou pouco significativamente maior ou até menor que utilizando o beta.
- 

Problemas encontrados



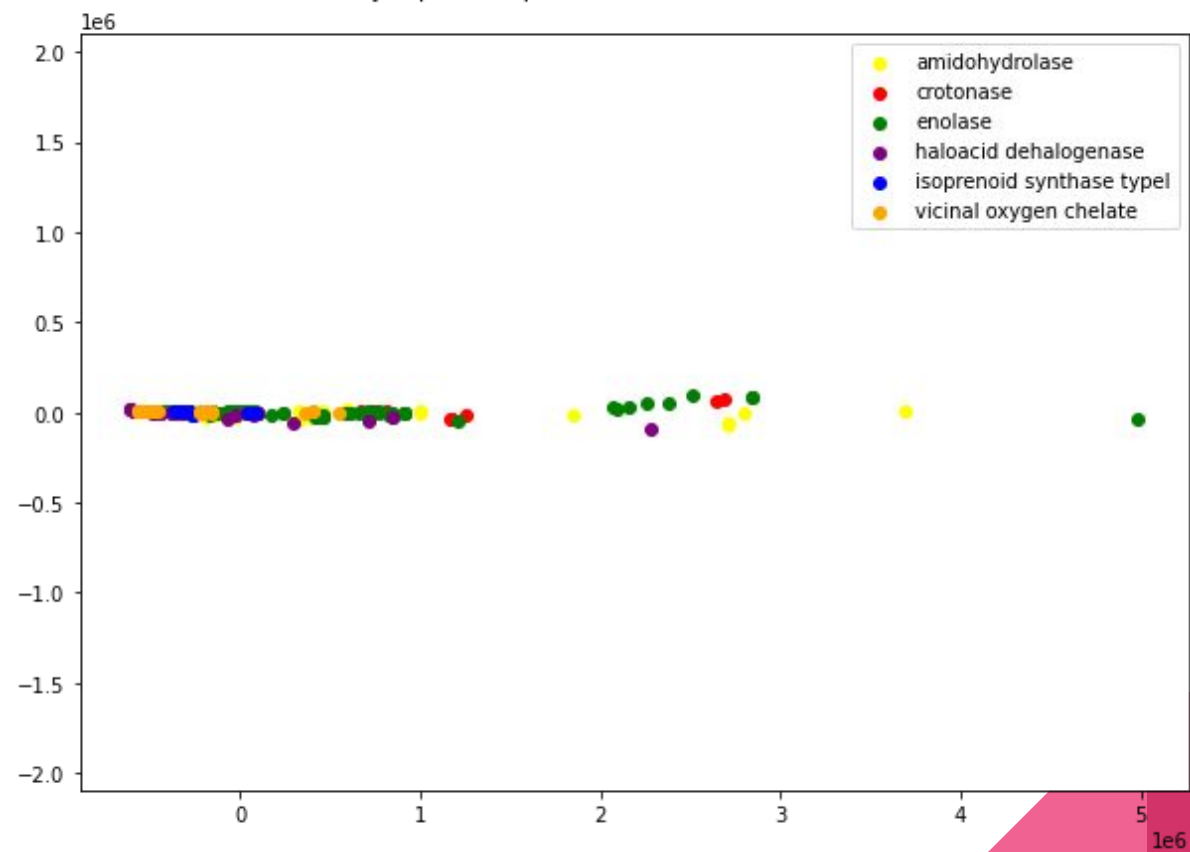
Base de dados muito pequena

- Espaço de treinamento muito pouco esparsos e de dimensionalidade baixa.
- Métricas de avaliação com baixo score.
- Não fomos capazes de reproduzir os resultados do artigo original para a distância euclidiana.

Porcentagem da variância total explicada pelos dois primeiros componentes do PCA para:

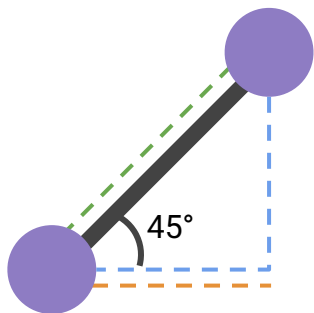
```
Dist Euclidiana (CA): 100.0%
Dist Manhattan (CA): 100.0%
Dist Chebyshev (CA): 99.99999999999999%
Dist Euclidiana (CB): 100.0%
Dist Manhattan (CB): 99.99999999999999%
Dist Chebyshev (CB): 100.0%
```

Distribuição pós PCA para a base de distância euclidiana(CA)



Métricas dependem da disposição da molécula

- As métricas de Manhattan e Chebyshev dependem da rotação da molécula.
- Uma mesma molécula rotacionada pode gerar dois vetores de features diferentes.



euc = 1
manh = $\sqrt{2} \approx 1.414$
cheb = $\sqrt{2}/2 \approx 0.707$



euc = 1
manh = 1
cheb = 1

Perspectivas Futuras



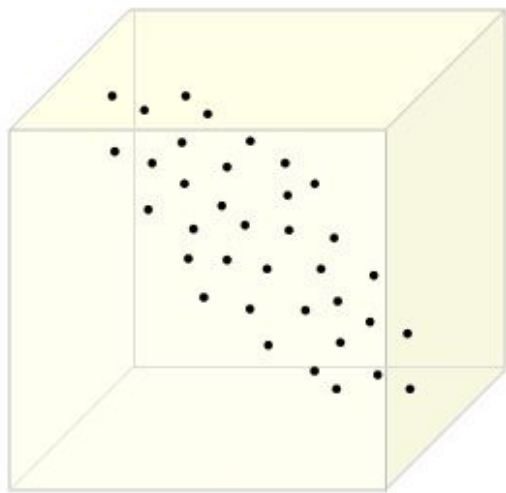
Propostas de melhoria do experimento

- Aumento da base de dados de treinamento.
- Expectativa de aumento da dimensionalidade dos dados.
- Realizar uma validação cruzada dos classificadores para o ajuste fino dos parâmetros.
- Aplicar o PCA (SVD) nos pontos da proteína antes de coletar as features, mantendo todos os componentes.

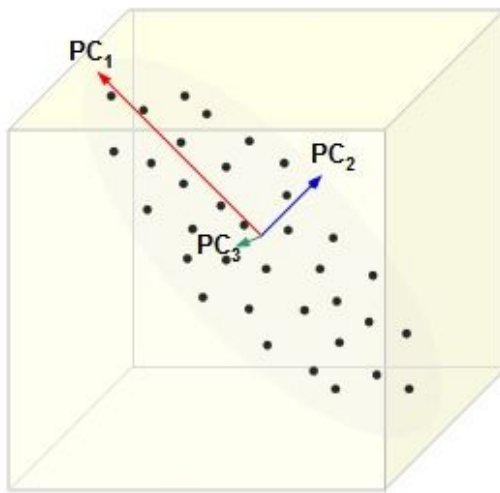


Aplicação do PCA Diretamente na Proteína

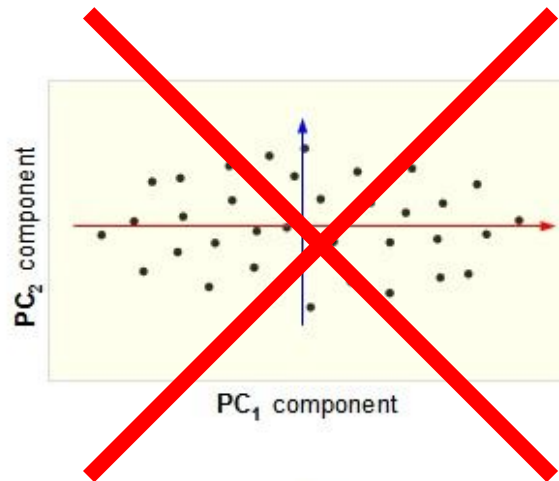
- A SVD define uma base ortonormal nas direções de maior variância e menor covariância.
- Além de padronizar a base das moléculas, a disposição dos vetores da base pode ser significativa para as distâncias que dependem da base de deslocamento.



a



b



c

Referências

- [1]<https://link.springer.com/article/10.1186/1471-2164-12-S4-S12>
- [2]<https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa>
- [3]<https://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-1-r8>
- [4]https://engineering.purdue.edu/precisetest/wp/wp-content/uploads/2012/05/Fang_localdia2009.pdf



Obrigado!

