




Análise Exploratória e Clusterização de Tecidos Humanos Usando Séries Temporais



Eduardo R. Rodrigues 13696679
Guilherme V. O. Taratá 10817476
Katiely F. de Lacerda 12777100
Wellington M. Amaral 11315054



Introdução

- Investigar dataset de expressão genética de tecidos humanos
- Avaliar a possibilidade da identificação de clusters significativos de genes
- Analisar a dinâmica temporal das expressões genéticas (RNASeq X Idade)

Dataset projeto GTEX

- Dados de coletas de tecidos humanos para análise genética
- Contém 17382 amostras com 56200 genes cada
- Possui a idade, tecido, sexo e expressão genética de cada amostra
- Conteúdo anonimizado sobre dados pessoais

PANTHER Classification System

- The PANTHER (**P**rotein **A**nalysis **T**hrough **E**volutionary **R**elationships)
- The mission of the PANTHER knowledgebase is to support biomedical and other research by providing comprehensive information about the evolution of protein-coding gene families, particularly protein phylogeny, function and genetic variation impacting that function.

Results ②

	Reference list	genes_total.txt
Uniquely Mapped IDs:	20592 out of 20592	304 out of 321
Unmapped IDs:	0	3
Multiple mapping information:	0	15

Export [Table](#) [XML with user input ids](#) [JSON with user input ids](#)Displaying only results for FDR P < 0.05. [click here to display all results](#)

	Homo sapiens (REF)	genes_total.txt (▼ Hierarchy NEW! ②)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
B cell negative selection	2	2	.03	64.15	+	1.38E-03	1.03E-02
↳B cell selection	4	2	.06	32.07	+	3.38E-03	2.17E-02
↳immune system process	2260	95	35.23	2.70	+	2.11E-19	1.07E-17
↳B cell differentiation	132	17	2.06	8.26	+	1.47E-10	3.52E-09
↳lymphocyte differentiation	290	31	4.52	6.86	+	3.56E-16	1.42E-14
↳mononuclear cell differentiation	335	38	5.22	7.28	+	2.03E-20	1.10E-18
↳leukocyte differentiation	424	49	6.61	7.41	+	1.74E-26	1.28E-24
↳cell differentiation	3581	153	55.82	2.74	+	5.67E-35	6.12E-33
↳cellular developmental process	3605	153	56.20	2.72	+	1.22E-34	1.28E-32
↳cellular process	14613	313	227.80	1.37	+	1.40E-35	1.54E-33
↳developmental process	5689	213	88.68	2.40	+	3.00E-46	6.38E-44
↳hemopoiesis	685	69	10.68	6.46	+	5.63E-34	5.75E-32
↳cell development	2181	117	34.00	3.44	+	9.30E-34	9.38E-32
↳anatomical structure development	5189	203	80.89	2.51	+	7.38E-46	1.42E-43
↳lymphocyte activation	478	41	7.45	5.50	+	6.58E-18	3.00E-16
↳leukocyte activation	601	50	9.37	5.34	+	3.54E-21	1.99E-19
↳cell activation	724	57	11.29	5.05	+	5.09E-23	3.12E-21
↳multicellular organismal process	6692	226	104.32	2.17	+	4.60E-43	7.37E-41
↳B cell activation	191	24	2.98	8.06	+	3.58E-14	1.17E-12
glomerular mesangial cell proliferation	2	2	.03	64.15	+	1.38E-03	1.03E-02
↳kidney development	303	22	4.72	4.66	+	7.17E-09	1.42E-07

Preparação dos dados

- Z-Normalização
- Remoção de variantes
- Seleção de genes senescentes (GenAge)
- Escolha de apenas um tipo de tecido

Metodologias

- Classificação utilizando Extremely Random Trees
- Classificação por idade + tipo de tecido
- DTW vs Spectral Clusters em Dados Tabulares
- Clusterização utilizando Spectral Clustering
- Clusterização DTW de tecidos utilizando idades das amostras

Classificação utilizando Extremely Random Trees

Premissa:

- A expressão genética possui aparência similar a uma série temporal;
- Tratar a expressão genética como um vetor de características e classificar o tipo de tecido.

Problemas enfrentados:

- Funciona, mas perde o sentido de série temporal que é o objetivo do trabalho.



Resultado

Neste caso foram classificados apenas tipos de tecido sem considerar idade

```
from sklearn.metrics import accuracy_score
```

```
y_pred = clf.predict(x_test)  
accuracy_score(y_test, y_pred)
```

```
✓ 129 ms (2023-12-04T00:41:47/2023-12-04T00:41:47)  
0.9010925819436457
```

Classificação por idade + tipo de tecido

Premissa:

- Criar uma prova de conceito para aplicar clusterizadores de séries temporais.
- Caso um classificador consiga classificar tipo de tecido + idade, um clusterizador pode funcionar também

Problemas enfrentados:

- Apesar da baixa acurácia, o classificador funcionou.
- Falta de tempo.

Resultado

Visto que há mais de 100 classes, a acurácia de 31% é uma boa pista

```
from sklearn.metrics import accuracy_score
```

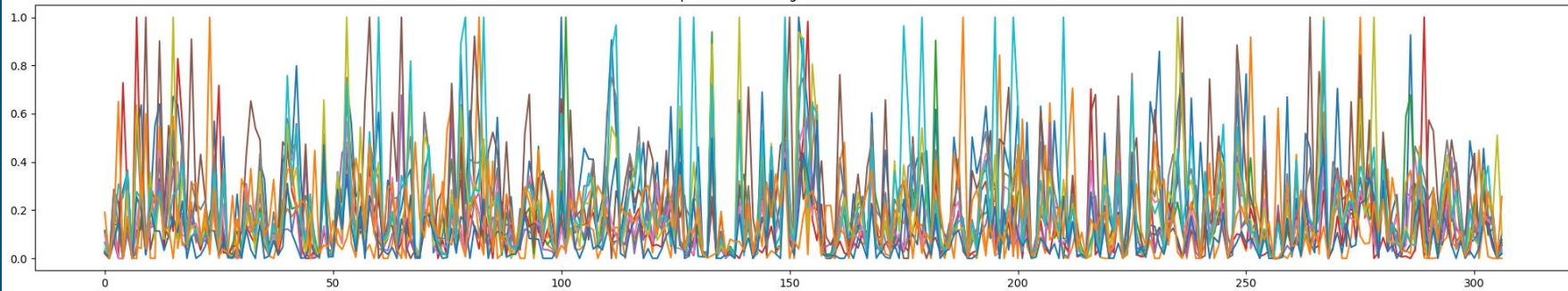
```
y_pred = clf.predict(x_test)  
accuracy_score(y_test, y_pred)
```

```
✓ 305 ms (2023-12-03T23:50:42/2023-12-03T23:50:43)  
0.3139735480161012
```

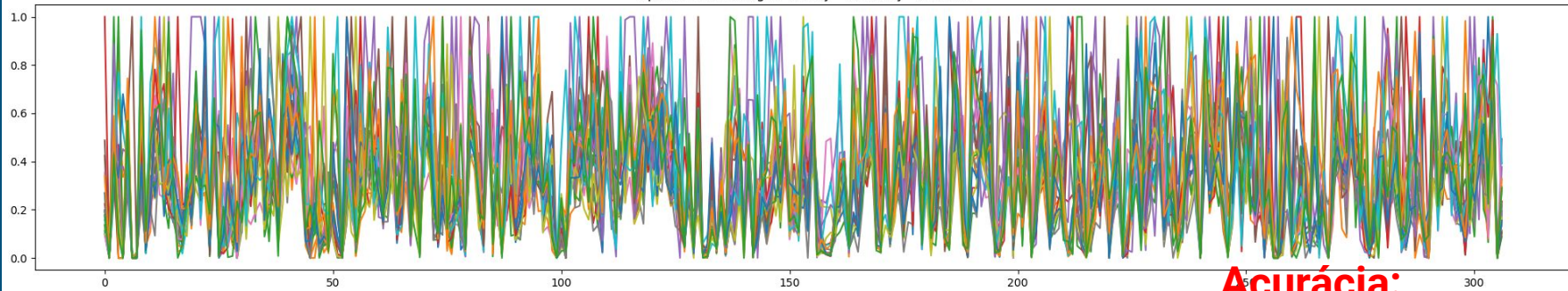
DTW vs Spectral Clusters em Dados Tabulares

A comparação dos métodos para ajudar a esclarecer o funcionamento da DTW e sua relação com o eixo do tempo.

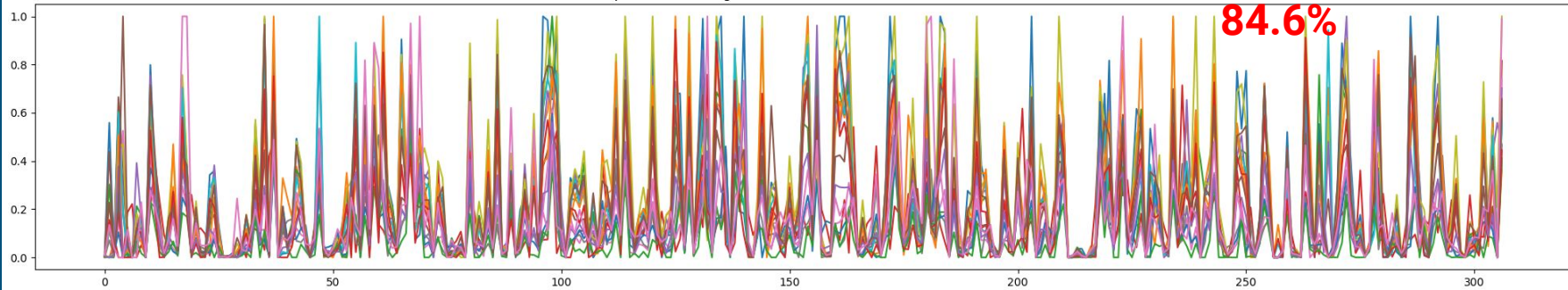
Spectral Clustering 1 - Stomach - 60.0%



Spectral Clustering 2 - Artery - Coronary - 100.0%

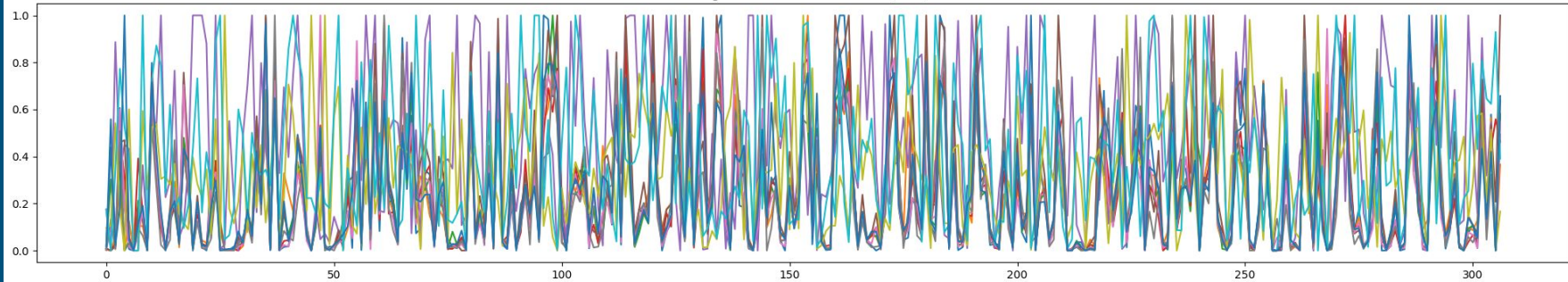


Spectral Clustering 3 - Brain - Frontal Cortex (BA9) - 100.0%

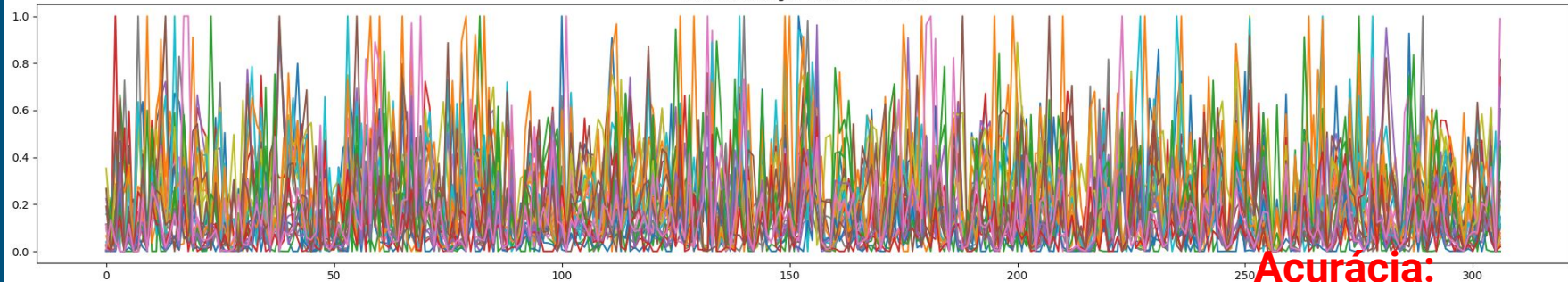


Acurácia:
84.6%

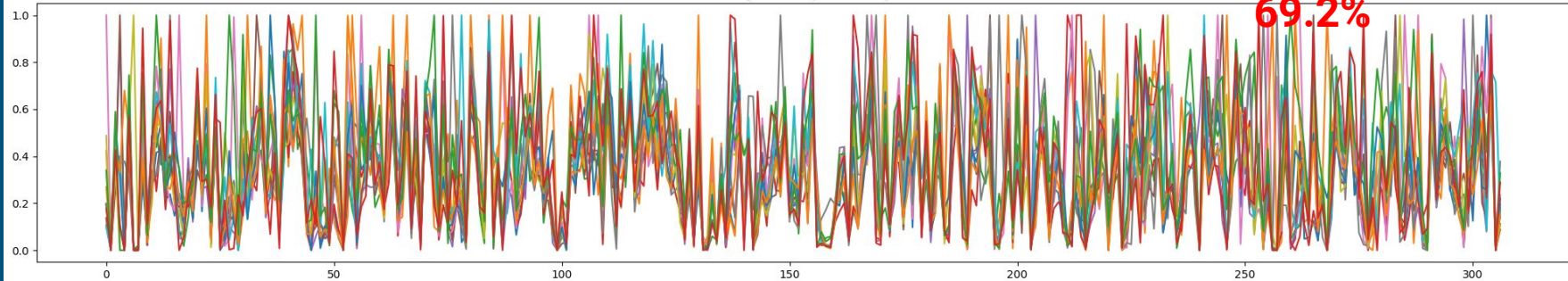
DTW Clustering 1 - Brain - Frontal Cortex (BA9) - 47.06%



DTW Clustering 2 - Stomach - 80.0%



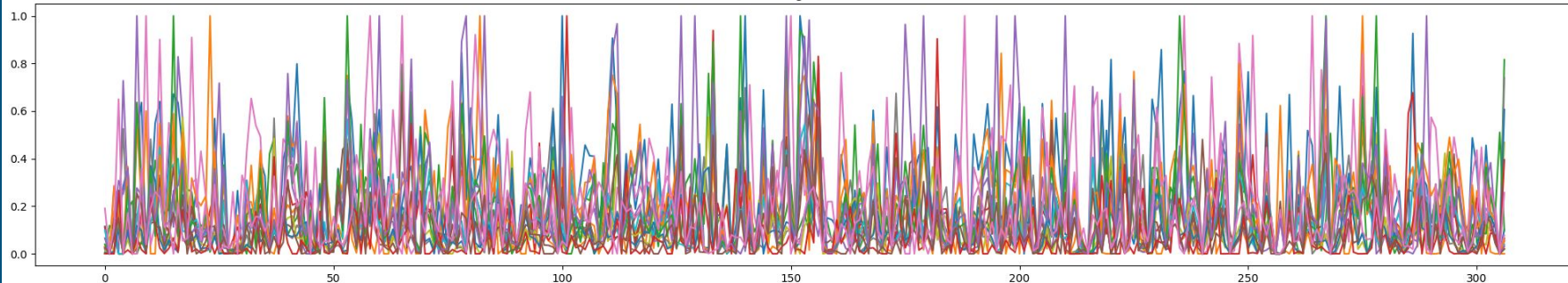
DTW Clustering 3 - Artery - Coronary - 73.33%



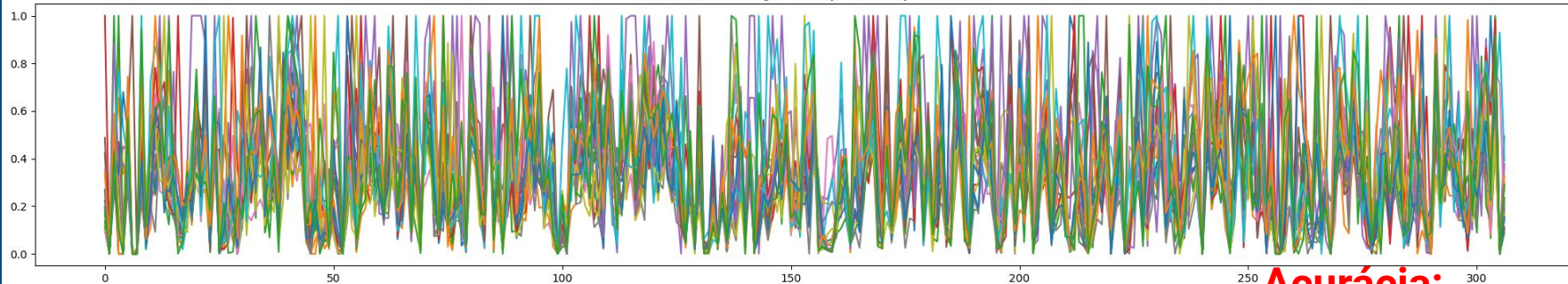
Acurácia:

69.2%

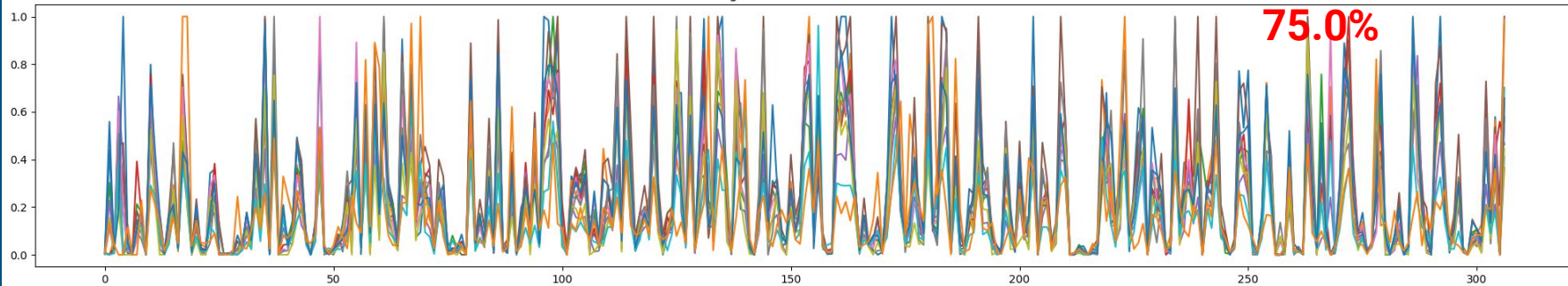
Euclidean Clustering 1 - Stomach - 60.0%



Euclidean Clustering 2 - Artery - Coronary - 100.0%



Euclidean Clustering 3 - Brain - Frontal Cortex (BA9) - 70.59%



Acurácia:
75.0%

Criação da “Série Temporal”

- Separação dos tecidos
- Ordenação por idade das amostras
- Z-Normalização + Min-Max
- Avaliação pelos Termos GO

Clusterização utilizando DTW (idade + tipo de tecido)

Premissa:

- Encontrar clusters que coincidem com a literatura (Termos GO).
- Utilizar pré-seleção de genes senescentes (GenAge).

Problemas enfrentados:

- “Clusters significativos negativos.”
- Viés dos dados
- Demora na execução.
- Dataset anonimizado

Clusterização utilizando DTW (idade + tipo de tecido)

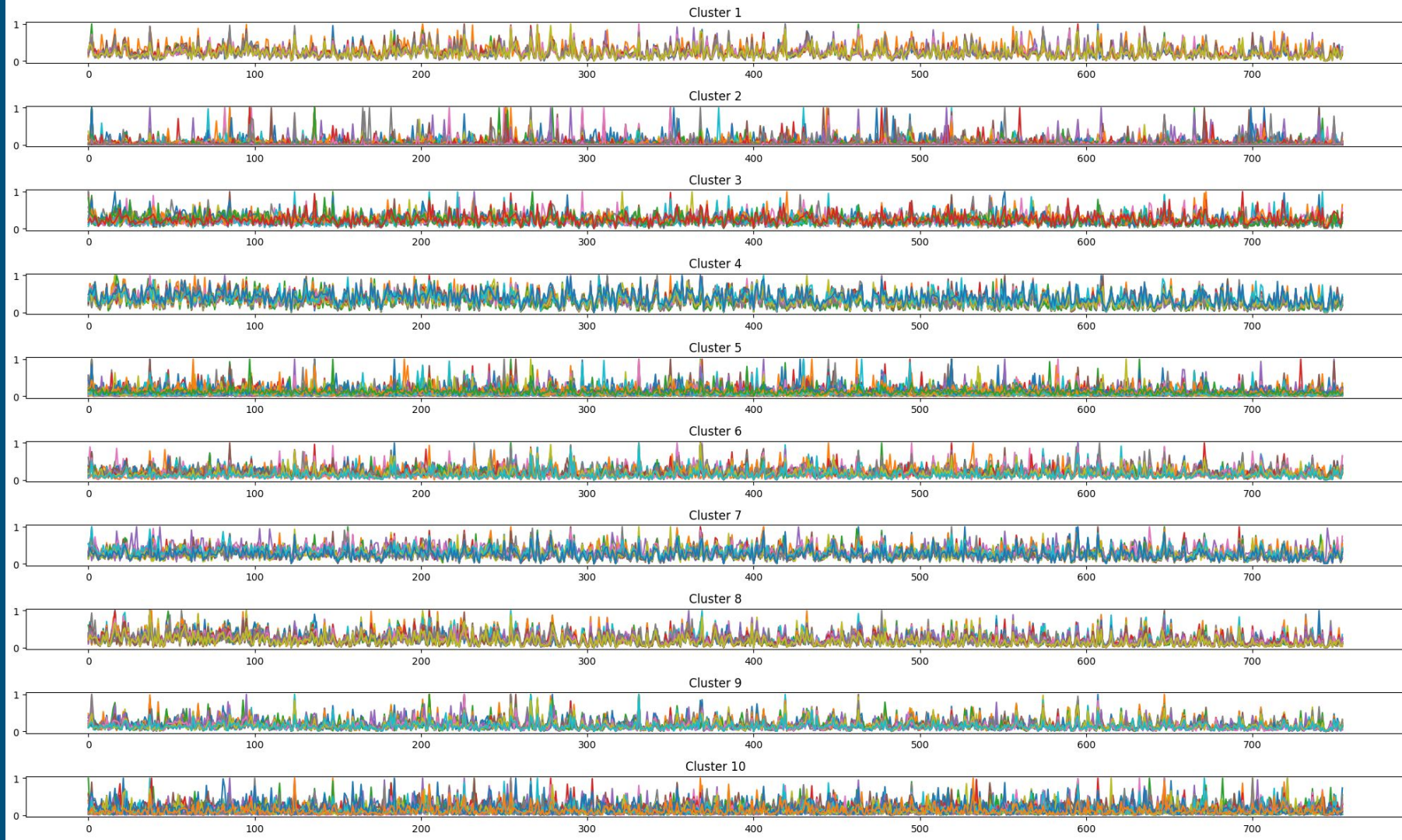
Pontos positivos:

- Encontrou clusters com alguma significância.

Displaying only results for FDR P < 0.05, [click here to display all results](#)

	genes_total.txt (REF)	GENES_CLUSTERS_4.txt (▼ Hierarchy NEW! ?)					
GO molecular function complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
organic cyclic compound binding	204	17	34.95	.49	-	1.09E-05	6.49E-03
heterocyclic compound binding	201	16	34.44	.46	-	5.70E-06	6.76E-03

PANTHER Classification System



Spectral Clustering

Premissa:

- Servir de comparação para os resultados da DTW.

Problemas enfrentados:

- Não criou clusters significativos coincidentes com os Termos GO.

Conclusão:

Dados Tabulares:

Apesar da DTW conseguir classificar dados tabulares sem relação temporal, o desempenho em relação aos outros métodos não foi satisfatório. E o tempo de execução foi significativamente maior.

Tecidos + Idades:

- Neste dataset a DTW parece ser capaz de detectar grupos de genes com relações significativas, porém não por semelhanças, mas pela exclusão delas.
- Talvez alguns ajustes no dataset possam levar a conclusões mais promissoras:
 - Como o ajuste no eixo das idades pelo balanceamento das categorias.
 - Filtro pelo sexo das amostras
 - Remoção de outliers e amostras doentes.

Bibliografia

PANTHER Classification System

<https://www.pantherdb.org/>

GTEx Portal

<https://www.gtexportal.org/home/>

SCIKIT-LEARN

<https://scikit-learn.org/>

TS-LEARN

tslearn.readthedocs.io/