EXERCÍCIO I2A2 – ANÁLISE GENÔMICA PARA DETECÇÃO DE PARENTESCO

05 DE MARÇO DE 2021

ELIO NAKNAO

I. INTRODUÇÃO

O exercício em questão envolve a análise genômica de 65.215 (sessenta e cinco mil, duzentos e quinze) genes mapeados referentes a 48 (quarenta e oito pessoas) para a identificação da existência de parentesco entre as referidas 48 (quarentas e oito pessoas).

Os dados foram fornecidos em uma tabela Excel “readcount.xls” sendo que:

> As 48 (quarente e oito) pessoas são identificadas da seguinte forma: H223H224 H225 H226 H227 H228 H229 H230 H231 H232 H233 H234 H235 H236 H237 H238 H239 H240 H241 H242 H243 H244 H245 H246 H247 H248 H249 H250 H251 H252 H253 H254 H255 H256 H257 H258 H259 H260 H261 H262 H263 H264 H265 H266 H267 H268 H269 H270.
>
> Os 65.215 (sessenta e cinco mil, duzentos e quinze) genes são identificados com o seguinte formato ENSG00000XXXXXX, onde X são números que representam os respectivos genes, como por exemplo: ENSG00000281922, ENSG00000000003, ENSG00000000003, entre outros.

II. ANÁLISE PRELIMINAR

Inicialmente, ressalta-se que se interpretou os valores da tabela “readcount.xls” como a “contagem do número de ocorrências” (*readcount*) de determinado gene em determinada amostra (pessoa). Adicionalmente, utilizou-se da hipótese de que pessoas (amostras) com *readcounts* similares para os mesmos genes de forma consistente ao longo da tabela devem ter algum grau de parentesco, considerando a hereditariedade dos genes.

Para fins de visualização apenas, foram selecionadas aleatoriamente 48 (quarenta e oito) genes e um gráfico do tipo *heatmap* foi gerado contra as 48 (quarenta e oito).

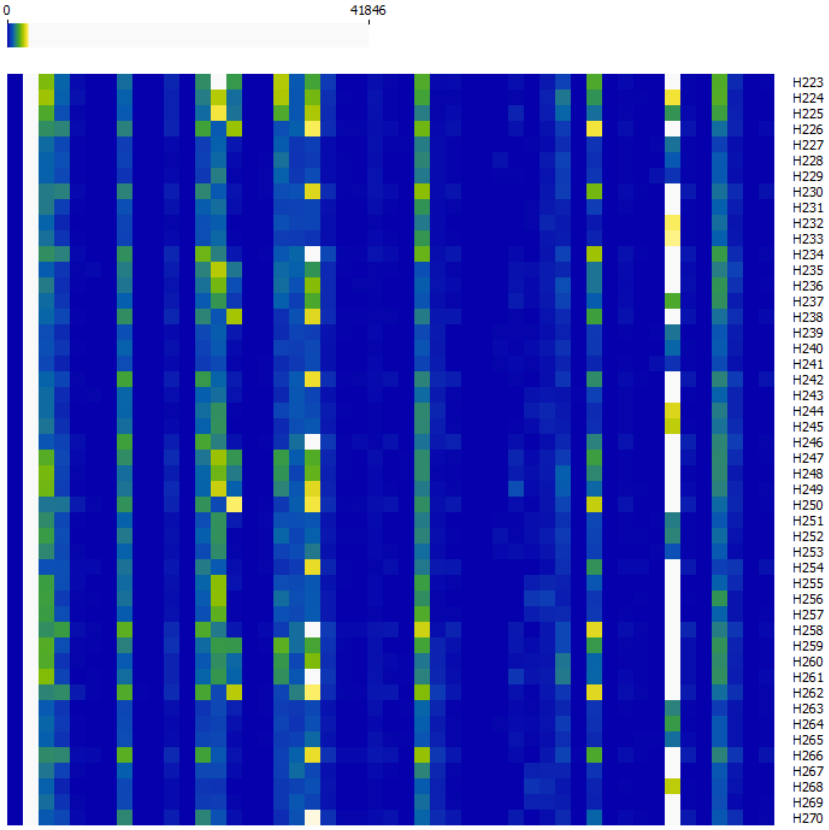

Gráfico 1 (genes x pessoas)

A partir de uma análise visual preliminar do Gráfico 1 nota-se que existem inúmero genes cujo *readcount* é baixo ou mesmo inexistente assim como muito alto para a totalidade das amostras, ou seja, referido gene não teria efeito na diferenciação das pessoas para fins de determinação de parente.

De forma complementar, pode-se notar também por meio de análise visual que algumas amostras contêm *readcounts* similares para determinados genes com a mesma denominação o que poderia ser um potencial parentesco.

Como base nas análises preliminares as premissas inicialmente parecem se mostrar consistentes portanto, uma análise mais generalista foi realizada para a análise da totalidade dos dados.

A metodologia adotada foi a “clusterização” das diversas amostras com base nos *readcounts* dos respectivos genes por meio da metodologia denominada “K-Means” em especial por se tratar de uma

metodologia de aprendizado não-supervisionado, visto que não foi fornecido, para este exercício, um conjunto de dados de treinamento já categorizados ou classificados.

.

Para a efetiva realização das análises, utilizou-se o *framework* Orange 3 em sua versão 3.26.0 junto ao ambiente Ananconda.

III. VETORIZAÇÃO DOS DADOS

A fim de se realizar o tratamento inicial dos dados realizou-se a "vetorização" dos dados, de forma que cada gene representaria um eixo num gráfico de coordenadas, e o valor dos *readcounts* representaria o tamanho do vetor no referido eixo. Desta forma, cada pessoas seria representada por um vetor que seria equivalente à soma de todos os vetores dos respectivos genes. Essa operação acaba por gerar um espaço de vetores de tamanho 65.215 (sessenta e cinco mil, duzentos e quinze) dimensões.

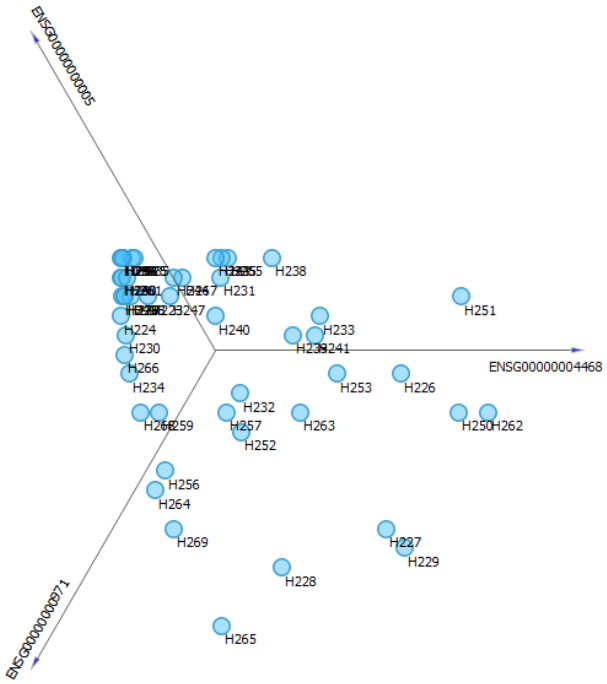

Gráfico 2 (Projeção 3D)

De forma ilustrativa, o Gráfico 2 mostra a projeção do espaço de vetores de 65.215 (sessenta e cinco mil, duzentos e quinze) dimensões em 3 (três dimensões), que auxilia na visualização de possível existência de algum padrão de agrupamento das pessoas em relação aos 3 (três) genes nos quais os demais foram projetados (ENSG00000000971, ENSG00000004468 e ENSG00000000005)

Levando em conta o grande número de dimensões do espaço de vetores assim com as evidencias preliminares de que existem genes que não teriam grande influência na operação de "clusterização" das pessoas tendo em vista o seu parentesco pois não propiciam uma diferenciação entre elas, uma vez que o *readcounts* são similares de forma consistente para todas as amostras, como por exemplo no caso em que todos os *readcounts* de determinado gene para todas as pessoas são iguais a zero.

Para tratar do problema acima relatado, optou-se por utilizar a técnica de redução de dimensionalidade denominada "Principal Componente Analysis" ou PCA. Este método consiste em identificar determinados eixos que tenham baixo poder explicativo no âmbito conjunto de dados e eliminá-los por meio da projeção dos dados nestes eixos ao longo dos demais eixos restantes. Com esta eliminação a dimensão total dos dados é reduzida com uma perda mínima de informação.

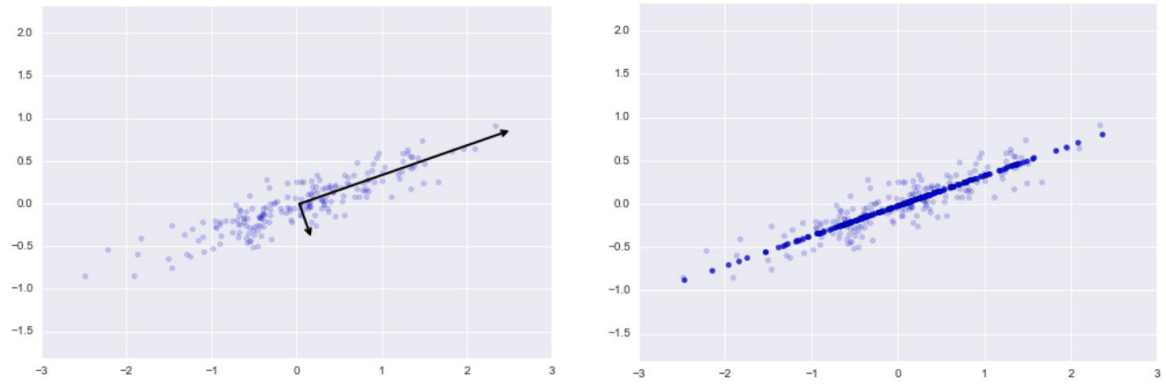

Gráfico 3 e 4 – Exemplo -PCA[1]

IV. PRINCIPAL COMPONENT ANALYSIS - PCA

Utilizou-se Orange 3 para realizar a análise PCA com base na eliminação de eixos com as menores variâncias mantendo os eixos (Componentes Principais) de forma que o Gráfico 5, abaixo, foi gerado. Observa-se que o próprio Orange 3 faz a normalização dos dados.

O Gráfico 5 demonstra que mantendo-se apenas 3 (três) Componentes Principais, a variância explicada é 31% (trinta e um por cento). Com 29 (vinte e nove) Componentes Principais, tem-se uma variância de 80% (oitenta por cento). Curiosamente, conforme a análise PCA, 100% da variância pode ser explicada por apenas 47 (quarenta e sete) Componentes Principais, ou seja, houve uma redução de dimensionalidade de 65.215 (sessenta e cinco mil, duzentos e quinze) para apenas 47 (quarenta e sete) dimensões.

[1] VanderPlas, Jake, Python Data Science Handbook, O'Reilly Media, 2017, páginas 435 e436.

Destaca-se também que a partir da 3ª (terceira) Componente Principal, a variância explicada se torna bastante pequena e cada Componente Principal adicional tem a mudança de sua variância explicada muito pequena em relação à Componente Principal anterior.

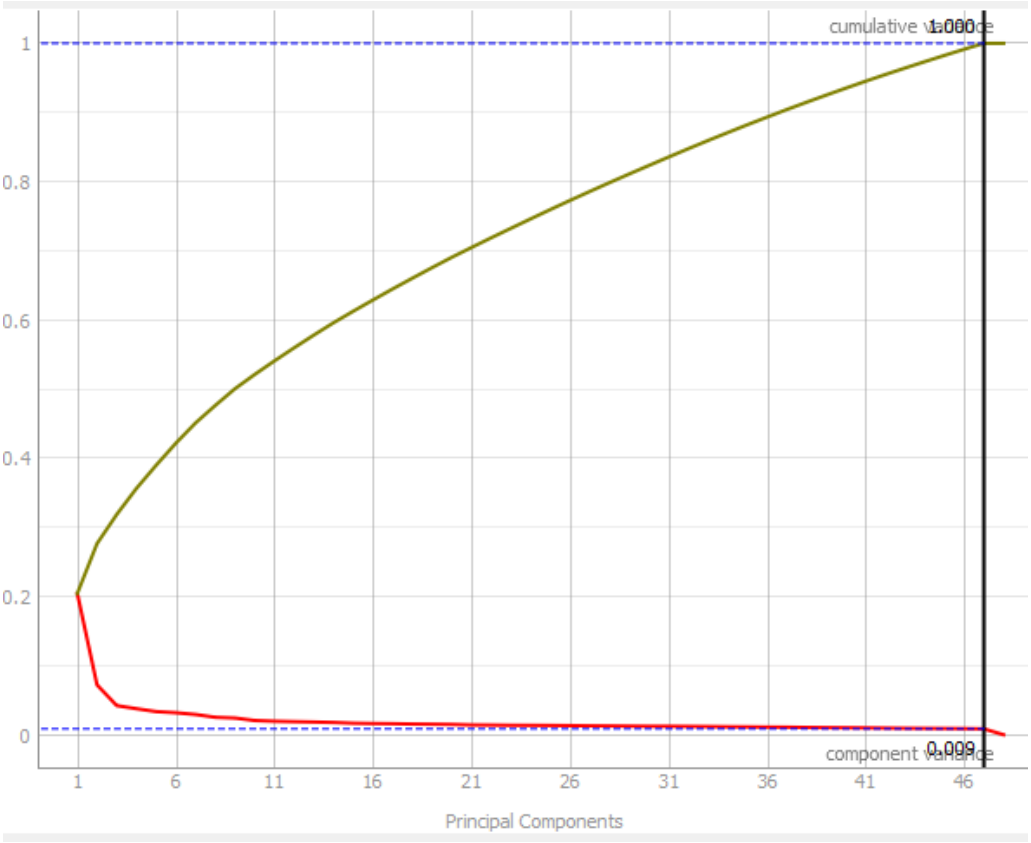

Gráfico 5 – no. de PCAs x Variância Explicada

Desta maneira, optando-se por utilizar as 47 (quarenta e sete) dimensões, com 100% da variância explicada, obtemos um novo conjunto de dados, no qual os 65.215 (sessenta e cinco mil, duzentos e quinze) genes foram reduzidos a 47 (quarenta e sente) Componentes Principais.

Com base na nova conformação de dados gerou-se novamente um Gráfico 6 do tipo *heatmap* que mostra os valores de cada das novas 47 (quarenta e sete) Componentes Principais em relação a cada uma das 48 (quarenta e oito) pessoas.

Ressalta-se que após a utilização do método PCA e realizada a redução de dimensionalidade, as Componentes Principais, em geral, não representam diretamente os genes originais, mas sim um novo conjunto de características geradas a partir dos genes originais.

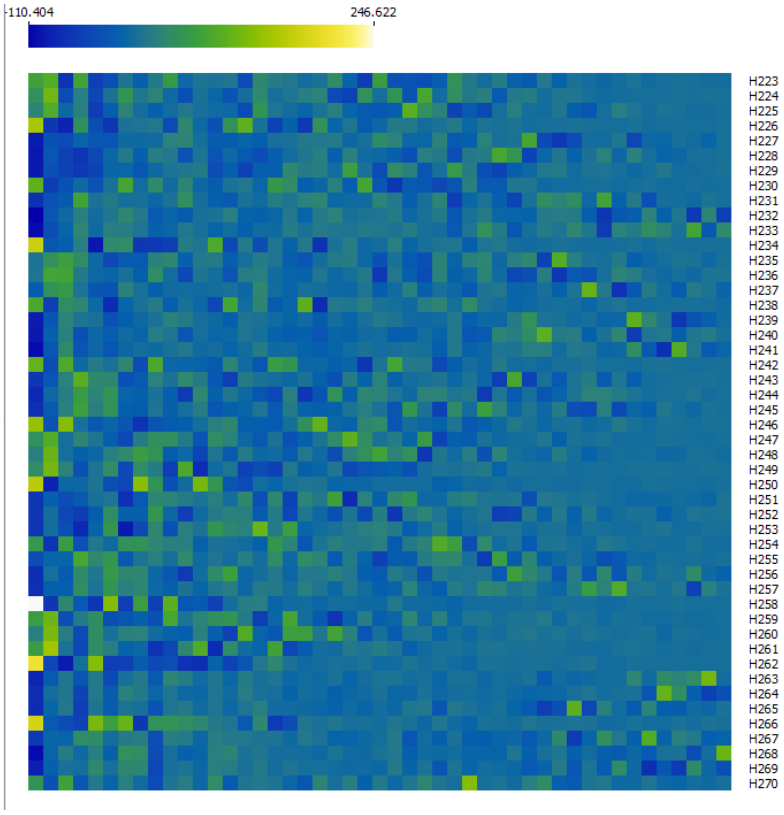

Gráfico 6 – *Heatmap* após PCA

Realizada esta redução de dimensionalidade, passa-se ao próximo passo da análise que é o K-Means para a "clusterização" das pessoas de acordo com as características das novas PCAs encontradas.

V. K-MEANS

O conceito por trás do K-Means é que os dados se agrupam de alguma forma nos respectivos espaços dimensionais. Neste caso, cada pessoa é indicada e localizada no espaço dimensional pela composição das respectivas PCA.

Desta forma, o K-Means é um algoritmo que inicia diversos centroides no espaço dimensional e calcula as distâncias entre cada um dos centroides e cada dos pontos representativos de cada uma das pessoas.

A partir disso o K-Means "agrupa" junto ao centroide aqueles pontos que têm a menor distância até ele conforme pode ser visto no exemplo do Gráfico 7.

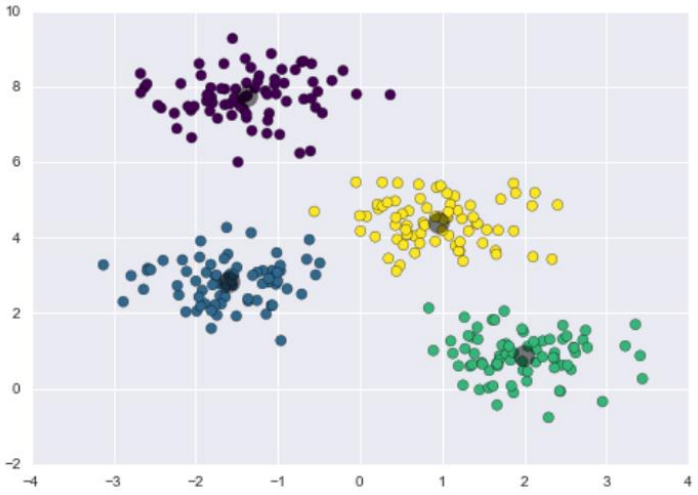

Gráfico 7 – Exemplo - K-Means[2].

Para a determinação do número de centroides do K-Means (ou seja, o valor K) o Orange fornece uma ferramenta que roda o algoritmo e fornece os resultados com base nos “Silhouette Scores” que para os dados em análise determinou que o número de K = 4 sendo o melhor, ou seja, as pessoas poderiam ser categorizadas em 4 “clusters”

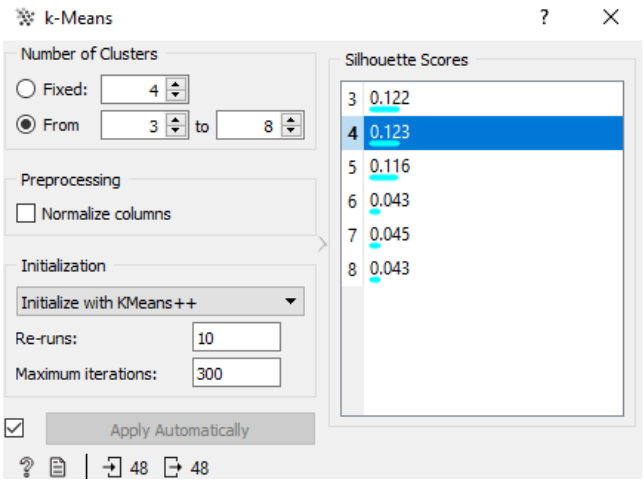

Figura 1 – Silhouette Scores – K-Means – Orange 3

O Gráfico 8 mostra a projeção das pessoas e suas respectivas “clusterizações” em 2 (duas) dimensões, projetadas sobre as Componentes Principais 1 e 2.

[2] VanderPlas, Jake, Python Data Science Handbook, O’Reilly Media, 2017, páginas 464.

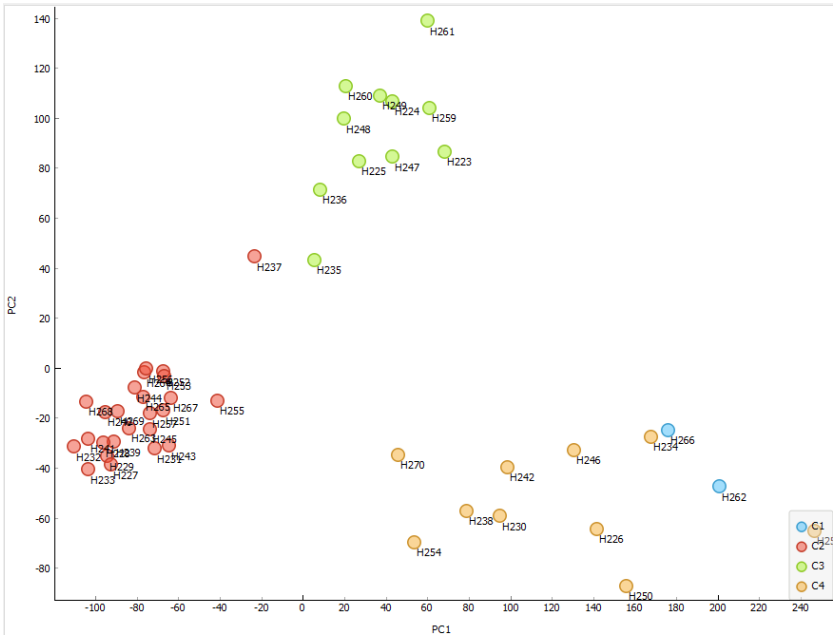

Gráfico 8 – Clusterização (Projeção 2D)

Deste resultado, pode-se inferir que as pessoas:

a) H262, H266: pertencem ao grupo C1

b) H227 H228 H229 H231 H232 H233 H237 H239 H240 H241 H243 H244 H245 H251 H252 H253 H255 H256 H257 H263 H264 H265 H267 H268 H269: pertencem ao grupo C2

c) H223 H224 H225 H235 H236 H247 H248 H249 H259 H260 H261 pertencem ao grupo C3:

d) H226 H230 H234 H238 H242 H246 H250 H254 H258: pertencem ao grupo C4.

E que cada grupo seria uma família.

Adicionalmente, o Orange contém uma ferramenta que permite visualizar a similaridade dos dados de acordo com os diversos “k” utilizados na “clusterização”.

O Gráfico 9 mostra os dados sem a aplicação do K-Means para “clusterização” apenas com base na similaridade dos dados. Referido gráfico reorganiza os dados e mostra uma árvore de similaridade que poderia indicar o parentesco entre as pessoas assim como os respectivos graus de parentesco.

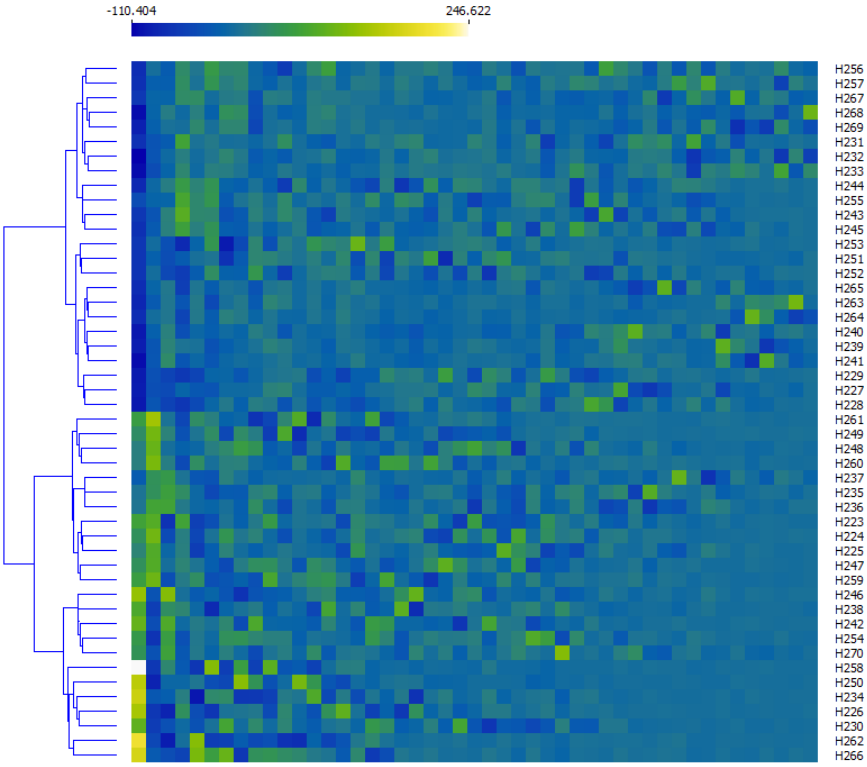

Gráfico 9 – Gráfico de Similaridade

Ao utilizarmos referida ferramenta, mas utilizando a "clusterização" com K-Mean com 5 (cinco) centroides (a ferramenta não permite a visualização abaixo de 5 (cinco) centroides) obtemos o Gráfico 10, abaixo. Nota-se que devido ao fato de serem utilizados 5 (cinco) centroides ao invés de 4 (quatro) os resultados diferem um pouco entre si.

Gráfico 10 – Gráfico de Similaridade – 5 Clusters

VI. CONCLUSÃO.

Com base nas análises realizadas chegou-se à conclusão de que as pessoas poderiam ser divididas em 4 grupos de parentes, compostos da seguinte forma:

a) H262, H266: pertencem ao grupo C1

b) H227 H228 H229 H231 H232 H233 H237 H239 H240 H241 H243 H244 H245 H251 H252 H253 H255 H256 H257 H263 H264 H265 H267 H268 H269: pertencem ao grupo C2

c) H223 H224 H225 H235 H236 H247 H248 H249 H259 H260 H261 pertencem ao grupo C3:

d) H226 H230 H234 H238 H242 H246 H250 H254 H258: pertencem ao grupo C4.

Além disso, como não foi definido o grau de parentesco, a análise pode ser ampliada de acordo com essa variável, conforme mostrado nos Gráficos 9 e 10.

*** * * ***