Supplementary Material: Equations used for the mitoDB website

K. Scheibye-Alsing, M Scheibye-Knudsen

1 Similarity Measures

In the following sections different methods for calculating the similarity between two diseases as represented by their signs-and-symptom frequency vectors (shortened to symptom vectors in the following) are presented. The vectors of the diseases are written in short form as:

$$X = (x_1, x_2, \ldots, x_n) \tag{1}$$

and

$$Y = (y_1, y_2, \ldots, y_n) \tag{2}$$

Where the indices represent the different symptoms. For example:

$$\begin{aligned} X_{Coenzyme\ Q10\ deficiency} = (&x_{Ataxia} = 0.72, \\ &x_{Cerebellar\ atrophy} = 0.61, \\ &\ldots) \end{aligned}$$

1.1 Note about Correlations

In the correlation based measures the correlation lies between -1 and 1, where a correlation coefficient of one for diseases with identical symptoms. This is converted to a “distance” measure going from 0 and up, where 0 is for diseases with identical symptom frequencies, therefore the “distance” and hence the length of the branches in the hierarchical clustering tree are a symbolic representation of the correlations.

2 Correlation Coefficient

The Pearson (centered) correlation coefficient is calculated as:

$$CC_{correlation}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

The coefficient will be affected by the large number of symptoms that are not present in the disease, *ie.* the disease vectors are sparse.

3 Uncentered Correlation

The uncentered correlation coefficient is

$$CC_{Uncentered\ Correlation}(X,Y) = \frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sqrt{(\sum_{i=1}^{n} x_i)^2(\sum_{i=1}^{n} y_i)^2}} \tag{4}$$

The uncentered correlation is unaffected by the sparse nature of the disease vectors, as only symptoms which are present in both diseases will contribute to the correlation coefficient.

4 Mixed Correlation

The mixed correlation is a mix of the centered and the uncentered correlation, *ie.*:

$$CC_{Mixed\ Correlation}(X,Y) = (1-\Theta)\frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sqrt{(\sum_{i=1}^{n} x_i)^2(\sum_{i=1}^{n} y_i)^2}} + \Theta\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{5}$$

Where Θ is the weight of the centered correlation.

The mixed correlation coefficient is a compromise between the previous two. Where the the centered correlation coefficient can be too heavily influenced by the majority of non-present symptoms and the uncentered correlation can focus too specifically on the co-present symptoms, the mixed correlation is a compromise between both both and should therefore moderate the non-ideal features of each method.

5 Euclidean Distance

The Euclidean distance is calculated as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

6 Manhattan Distance

$$D(X,Y) = \sum_{i=1}^{n}|x_i - y_i| \tag{7}$$

7 Complete Linkage

Given two clusters of diseases $G = \{X_1, X_2, \ldots\}$ and $H = \{Y_1, Y_2, \ldots\}$, where each element is a diseases with associated symptom vector $(X = (x_1, x_2, \ldots, x_n))$ etc. Each cluster could potentially consist of only one disease, in the case of a disease that has not (yet) been assigned a cluster.

The distance between each cluster is then:

$$D(G,H) = \max_{X_i \in G, Y_j \in H} D(X_i, Y_j) \tag{8}$$

ie. the maximum distance between any pairs of diseases in the two groups.

8 Single Linkage

Given two clusters of diseases $G = \{X_1, X_2, \ldots\}$ and $H = \{Y_1, Y_2, \ldots\}$, where each element is a diseases with associated symptom vector ($X = (x_1, x_2, \ldots, x_n)$) etc. Each cluster could potentially consist of only one disease, in the case of a disease that has not (yet) been assigned a cluster.

The distance between each cluster is then:

$$D(G, H) = \min_{X_i \in G, Y_j \in H} D(X_i, Y_j) \tag{9}$$

ie. the minimum distance between any pairs of diseases in the two groups.

9 Average Linkage

Given two clusters of diseases $G = \{X_1, X_2, \ldots\}$ and $H = \{Y_1, Y_2, \ldots\}$, where each element is a diseases with associated symptom vector ($X = (x_1, x_2, \ldots, x_n)$) etc. Each cluster could potentially consist of only one disease, in the case of a disease that has not (yet) been assigned a cluster.

The distance between each cluster is then:

$$D(G, H) = \frac{\sum_{X_i \in G, Y_j \in H} D(X_i, Y_j)}{N} \tag{10}$$

where N is the total number of disease pairs. This is the average of the distance between any pairs of diseases in the two groups.

10 Centroid Linkage

Given two clusters of diseases $G = \{X_1, X_2, \ldots, X_K\}$ and $H = \{Y_1, Y_2, \ldots\}$, where each element is a diseases with associated symptom vector ($X_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$) etc. Each cluster could potentially consist of only one disease, in the case of a disease that has not (yet) been assigned a cluster.

The centroid-vector of each cluster is then defined as (for the cluster G):

$$Centroid_G = (\frac{\sum_{i=1,n\ j=1,K} x_{j1}}{K}, \frac{\sum_{i=1,n\ j=1,K} x_{j2}}{K}, \ldots, \frac{\sum_{i=1,n\ j=1,K} x_{jn}}{K}) \tag{11}$$

where N is the total number of disease pairs, *ie.* centroid vector consists of the average of the symptom frequencies for the constituent diseases.

The distance between each cluster is then:

$$D(G, H) = D(Centroid_G, Centroid_H) \tag{12}$$

ie. the distance between the centroid vectors of the two clusters.

11 SVM discussion

The Support Vector Machine (SVM) was implemented using linear kernel, eps-regression. Briefly, the SVM classifies the disease by taking the symptom vector (ie. the frequency of the different symptoms for that disease), mapping the symptoms to a (potentially) higher dimensional space and determining whether the disease lies on one or the other side of a “separating hyperplane”. The hyperplane is generated from a training set, here the set of symptom-vectors for the known mitochondrial diseases and the set of vector for the known non-mitochondrial diseases. The hyperplane is optimised such that it best separates the mitochondrial from the non-mitochondrial diseases.

The linear kernel allows for an easy interpretation of the results. The linearity basically means that the symptoms of the disease are directly used in the classification without any non-linear mapping, which allows the different symptoms and their contribution the score to be assessed individually. *Eg.* The symptom with the highest affinity in mitochondrial diseases (for one of the filtered SVM) is “Optic atrophy” with a weight of 1.000827, which means that a disease having 100% afflicted patients with this symptoms will get +1.000827 contribution to the score.

“Eps regression” allows for a continuum of scores, where the other (typical) choice “C-classification” is a binary “either/or” classification. *Ie.* with “C-classification” a disease would get either a -1 or a +1 depending on the symptoms present, while “eps-regression” provides a more nuanced view where diseases could possibly be “marginally mitochondrial” (having a score just above 0, *eg.* 0.2).

12 Mitochondrialness Score

The mito-score was computed using the following equation

$$M_s(x) = 100 \frac{\sum_{i=1}^{n} x_i \overline{x_i}}{\sum_{i=1}^{n} x_i \overline{x_i} + \sum_{i=1}^{n} x_i \overline{y_i}} \tag{13}$$

Where $x = (x_1, x_2, \ldots, x_n$ is the symptom vector for the tested disease, $\overline{x} = (\overline{x_1}, \overline{x_2}, \ldots, \overline{x_n}$ is the average symptom vector of the mitochondrial diseases, and $\overline{y} = (\overline{y_1}, \overline{y_2}, \ldots, \overline{y_n}$ is the average symptom vector of the non-mitochondrial diseases.

This equation can be reformulated as (using the standard vector dot-product and its geometric interpretation).

$$\begin{aligned} M_s(x) &= 100 \frac{x \cdot \overline{x}}{x \cdot \overline{x} + x \cdot \overline{y}} & (14) \\ &= 100 \frac{|x||\overline{x}| \cos\theta_{x\overline{x}}}{|x||\overline{x}| \cos\theta_{x\overline{x}} + |x||\overline{x}| \cos\theta_{x\overline{y}}} & (15) \\ &= 100 \frac{|\overline{x}| \cos\theta_{x\overline{x}}}{|\overline{x}| \cos\theta_{x\overline{x}} + |\overline{y}| \cos\theta_{x\overline{y}}} & (16) \end{aligned}$$

where $\theta_{x\overline{x}}$ and $\theta_{x\overline{y}}$ are the “angles” between the disease symptom vector and the average symptom vectors. *Ie.* the mito-score can be thought of as the projection of the symptom vector of the disease onto the average mitochondrial symptom vector normalised by the mitochondrial projection and the non-mitochondrial projection.

13 Network

The link strength of the connections between two diseases in the network is calculated as:

$$L(x, y) = \sum_{i=1}^{n} x_i y_i = x \cdot y \tag{17}$$

ie. two diseases are linked if they share a common symptom. In the visualisation of the network the distances between pairs of nodes is linearly dependent on the value, the thickness of the line connecting them is dependent on the square root of the score, and the opacity linearly dependent on the score.

The score will go from 0 and up, and will, for two diseases sharing a symptom with 100% prevalence be one, for two diseases sharing two symptoms with 100% prevalence be two, *etc.*

When building the network a cut-off can be chosen so that diseases are only shown as linked, if their link strength is above the chosen cut-off threshold (defaults to 0, and set under advanced options).

14 Cophenetic Correlation Coefficient

A method for evaluating is the cophenetic correlation coefficient, given by the following equation.

$$CCC(X, Y) = \frac{\sum_{i,j=1}^{n}(D(x_i, x_j) - \overline{D})(D_c(x_i, x_j) - \overline{D_c})}{\sqrt{\sum_{i,j=1}^{n}(D(x_i, x_j) - \overline{D})^2 \sum_{i,j=1}^{n}(D_c(x_i, x_j) - \overline{D_c})^2}} \tag{18}$$

Where $D(x_j, x_i)$ is the distance between disease vectors x_i and x_j and $D_c(x_j, x_i)$ is the cophenetic distance between the vectors, *ie.* the distance between the clusters containing x_i and x_j when they are merged.

Since the cophenetic correlation intrensically uses the calculated pairwise distance between disease vectors, it is dependent on the quality of these distances, hence if the distance measure choosen does not reflect the biological relations well, a hierarchical clustering can have a high cophenetic correlation coefficient, even though the clustering does not reflect relevant biological relations, *ie.* junk in, junk out.

Therefore the main use of the cophenetic correlation coefficient is to evaluate different linkage schemes, and can be used as a guideline for which linkage scheme best reproduces a clustering refleting the pairwise distances.