Supplementary Material

Incorporation of biological knowledge into distance for clustering genes

Grzegorz M Boratyn[1] , Susmita Datta[2], and Somnath Datta[3]

[1]Clinical Proteomics Center, University of Louisville, Louisville, KY 40202; [2,3]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202

Selection of functional distance scaling parameters: The parameters $\boldsymbol{\alpha}$, β, and γ of the functional distance can be found by considering the following constraints:

Goal 1: Distance between genes with similar functions is smaller or equal to distance between genes with different functions.

Goal 2: Distance between a pair of annotated and unannotated genes is smaller or equal to distance between two unannotated genes.

Goal 3: Distance between genes with different functions is larger or equal to distance between unannotated genes.

In order to satisfy goal 1, the new distance metric should satisfy the following set of inequalities:

$$\begin{cases} \underset{i,j\in\mathcal{F}_1}{E}(d_{ij}^M + d_{ij}^{F1}) \le \underset{i\in\mathcal{F}_1, j\in\mathfrak{F}-\mathcal{F}_1}{E}(d_{ij}^M + d_{ij}^{F1}) \\ \underset{i,j\in\mathcal{F}_2}{E}(d_{ij}^M + d_{ij}^{F1}) \le \underset{i\in\mathcal{F}_2, j\in\mathfrak{F}-\mathcal{F}_2}{E}(d_{ij}^M + d_{ij}^{F1}) \\ \dots \\ \underset{i,j\in\mathcal{F}_f}{E}(d_{ij}^M + d_{ij}^{F1}) \le \underset{i\in\mathcal{F}_f, j\in\mathfrak{F}-\mathcal{F}_f}{E}(d_{ij}^M + d_{ij}^{F1}) \\ d_{ij}^{F1} \le 0, \text{ for } i,j \in \mathfrak{F}, \end{cases} \tag{1}$$

where $E(\cdot)$ denotes expectation of a random variable. In some cases the linear system (1) may not have a solution because of data set properties. Hence addition parameter $C \ge 0$ is added to the right hand side of the inequalities, and (1) is expressed more generally as:

$$\begin{cases} \underset{i,j\in\mathcal{F}_1}{E}(d_{ij}^M + d_{ij}^{F1}) \le \underset{i\in\mathcal{F}_1, j\in\mathfrak{F}-\mathcal{F}_1}{E}(d_{ij}^M + d_{ij}^{F1}) + C \\ \underset{i,j\in\mathcal{F}_2}{E}(d_{ij}^M + d_{ij}^{F1}) \le \underset{i\in\mathcal{F}_2, j\in\mathfrak{F}-\mathcal{F}_2}{E}(d_{ij}^M + d_{ij}^{F1}) + C \\ \dots \\ \underset{i,j\in\mathcal{F}_f}{E}(d_{ij}^M + d_{ij}^{F1}) \le \underset{i\in\mathcal{F}_f, j\in\mathfrak{F}-\mathcal{F}_f}{E}(d_{ij}^M + d_{ij}^{F1}) + C \\ d_{ij}^{F1} \le 0, \text{ for } i,j \in \mathfrak{F}, \end{cases} \tag{2}$$

The parameter C is specified by a user and should be $C > 0$ only if (2) are not satisfied for $C = 0$. Often there are infinitely many values of $\boldsymbol{\alpha}$ that satisfy (2). We select the one that minimizes L1 norm. Hence:

$$\begin{aligned} \boldsymbol{\alpha} &= \arg\min_{\boldsymbol{\alpha}} \sum_{k=1}^{f} |\alpha_k| \\ &\quad \text{subject to:} \\ \underset{i,j\in\mathcal{F}_k}{E}(d_{ij}^M + d_{ij}^{F1}) &\le \underset{i\in\mathcal{F}_k, j\in\mathfrak{F}-\mathcal{F}_k}{E}(d_{ij}^M + d_{ij}^{F1}) + C \\ &\quad \text{for } k = 1, 2, \dots, f \\ d_{ij}^{F1} &\le 0, \text{ for } i,j \in \mathfrak{F}, \end{aligned} \tag{3}$$

Due to (4) $d_{ij}^{F1} = -\mathbf{F}_i \boldsymbol{\alpha} \mathbf{F}_j^T$, hence (3) becomes:

$$\begin{aligned}
&\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \sum_{k=1}^{f} \alpha_k \\
&\quad \text{subject to:} \\
&\sum_{l=1}^{f} \alpha_l \left(\sum_{i \in \mathcal{F}_k, j \in \mathfrak{F} - \mathcal{F}_k} \frac{F_{il} F_{jl}}{N_0^k} - \sum_{i,j \in \mathcal{F}_k, i \neq j} \frac{F_{il} F_{jl}}{N_1^k} \right) \leq \\
&\quad \leq \frac{1}{N_0^k} \sum_{i \in \mathcal{F}_k, j \in \mathfrak{F} - \mathcal{F}_k} d_{ij}^M - \frac{1}{N_1^k} \sum_{i,j \in \mathcal{F}_k, i \neq j} d_{ij}^M + C \\
&\alpha_k \geq 0 \\
&\text{for } k = 1, 2, \ldots, f
\end{aligned} \tag{4}$$

where $N_0^k = |\{(i,j) | i \in \mathcal{F}_k, j \in \mathfrak{F} - \mathcal{F}_k\}|$, $N_1^k = |\{(i,j) | i, j \in \mathcal{F}_k, i \neq j\}|$, $|\cdot|$ indicates cardinality of a set.

Let us now consider goals 2 and 3. Let

$$\begin{aligned}
(p,q) &= \arg\max_{i \in \mathfrak{F}, j \notin \mathfrak{F}} \quad d_{ij}^M + d_{ij}^{F1} \\
(r,s) &= \arg\min_{i,j \notin \mathfrak{F}} \quad d_{ij}^M + d_{ij}^{F1} \\
(x,y) &= \arg\max_{i,j \notin \mathfrak{F}} \quad d_{ij}^M + d_{ij}^{F1} \\
(v,z) &= \arg\min_{i,j \in \mathfrak{F}, \text{ s.t. } \mathbf{F}_i \mathbf{F}_j^T = 0} \quad d_{ij}^M + d_{ij}^{F1}
\end{aligned} \tag{5}$$

The following must hold in order to satisfy these goals:

$$\begin{cases}
D_{pq} \leq D_{rs} \\
D_{xy} \leq D_{vz} \\
d_{ij}^{F2} \geq 0 \\
d_{ij}^{F3} \geq 0 \\
\text{for } i, j \in \mathcal{G}
\end{cases} \tag{6}$$

Due to (2), (4), (5), and (7), (6) becomes:

$$\begin{cases}
d_{pq}^M + d_{pq}^{F1} + \beta u_p u_q + \gamma(1 - \text{sign}(\mathbf{F}_p \mathbf{F}_q^T)) \leq \\
\leq d_{rs}^M + d_{rs}^{F1} + \beta u_r u_s + \gamma(1 - \text{sign}(\mathbf{F}_r \mathbf{F}_s^T)) \\
d_{xy}^M + d_{xy}^{F1} + \beta u_x u_y + \gamma(1 - \text{sign}(\mathbf{F}_x \mathbf{F}_y^T)) \leq \\
\leq d_{vz}^M + d_{vz}^{F1} + \beta u_v u_z + \gamma(1 - \text{sign}(\mathbf{F}_v \mathbf{F}_z^T)) \\
d_{ij}^{F2} \geq 0 \\
d_{ij}^{F3} \geq 0 \\
\text{for } i, j \in \mathcal{G}
\end{cases} \tag{7}$$

Due to (5), $u_p u_q = 0$, $\mathbf{F}_p \mathbf{F}_q^T > 0$, $u_r u_s = 1$, $\mathbf{F}_r \mathbf{F}_s^T = f$, $u_x u_y = 1$, $\mathbf{F}_x \mathbf{F}_y^T = f$, $u_v u_z = 0$, $\mathbf{F}_v \mathbf{F}_z = 0$, $u_i \geq 0$, and $F_{ik} \geq 0$, for $i \in \mathcal{G}$, $k = 1, 2, \ldots, f$. Hence (7) becomes:

$$\begin{cases}
\beta \geq d_{pq}^M + d_{pq}^{F1} - (d_{rs}^M + d_{rs}^{F1}) \\
\gamma \geq \beta + d_{xy}^M + d_{xy}^{F1} - (d_{vz}^M + d_{vz}^{F1}) \\
\beta \geq 0, \gamma \geq 0
\end{cases} \tag{8}$$

There can be infinitely many values of β and γ that satisfy (8). We select the equality solution or zero, hence:

$$\begin{aligned} \beta &= \max\{\beta', 0\} \\ \gamma &= \max\{\gamma', 0\}, \end{aligned} \tag{9}$$

where β' and γ' solve:

$$\begin{cases} \beta' = d_{pq}^{M} + d_{pq}^{F1} - (d_{rs}^{M} + d_{rs}^{F1}) \\ \gamma' = \beta + d_{xy}^{M} + d_{xy}^{F1} - (d_{vz}^{M} + d_{vz}^{F1}). \end{cases} \tag{10}$$