Matrix Representation

Four kinds of matrices (order k, k = 0-12) can be used in **QuBiLS-MIDAS** software, *namely* Non-stochastic (NS), Simple stochastic (SS), Double stochastic (DS) and Mutual Probabilistic (MP) matrices.

1. Non-stochastic (NS) matrix

The codification of 3D information of the non-covalent interactions of the molecular structure is fulfilled through rules among two, three and four atoms, and the values of these rules are represented in the *two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrix*, $\mathbb{G}$, $\mathbb{GT}$ and $\mathbb{GQ}$, respectively. The generalized expressions of these matrices are the k^{th} *two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices*, denoted by $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$, where superscript k indicates the power to which $\mathbb{G}$, $\mathbb{GT}$ and $\mathbb{GQ}$ are raised. Thus, for k = 0, all coefficients g_{ij}^0, gt_{ijl}^0 and gq_{ijlh}^0 corresponding to the matrices $\mathbb{G}^0$, $\mathbb{GT}^0$ and $\mathbb{GQ}^0$ have value 1; and for k = 1, the coefficients g_{ij}^1 of the matrix $\mathbb{G}^1$, gt_{ijl}^1 of the matrix $\mathbb{GT}^1$ and gq_{ijlh}^1 of the matrix $\mathbb{GQ}^1$, represent the information of the interactions among two, three and four atoms respectively. The formal definitions of these elements is shown below:

$$\begin{aligned} g_{ij}^1 &= D_{ij} \quad if\ atoms\ i\ and\ j\ are\ not\ equal \\ &= L_{ij} \quad i = j \wedge lone\text{-}pairs\ are\ considered\ \left(or\ D_{io}\right) \\ &= 0 \quad otherwise \end{aligned} \tag{1}$$

$$\begin{aligned} gt_{ijl}^1 &= T_{ijl} \quad if\ atoms\ i,\ j\ and\ l\ are\ not\ equal \\ &= L_{ijl} \quad i = j = l \wedge lone\text{-}pairs\ are\ considered\ \left(or\ D_{io}\right) \\ &= 0 \quad otherwise \end{aligned} \tag{2}$$

$$\begin{aligned} gq_{ijlh}^1 &= Q_{ijlh}\ if\ atoms\ i,\ j,\ l\ and\ h\ are\ not\ equal \\ &= L_{ijlh} \quad i = j = l = h \wedge lone\text{-}pairs\ are\ considered\ (or\ D_{io}) \\ &= 0 \quad otherwise \end{aligned} \tag{3}$$

where, D_{ij} is a (dis-)similarity metric between two atoms, T_{ijl} is an measure for ternary relations of atoms, Q_{ijlh} is an measure for quaternary relations of atoms, while L_{ij}, L_{ijl} and L_{ijlh} are the diagonal entries, which could have assigned two different values to achieve greater discrimination of the molecular structures: 1) representing the number of lone-pairs electrons for atoms, or 2) the spatial distance, D_{io} for each atom *i* and center of the molecule, *o*.

2. Simple-stochastic (SS) matrix

The k^{th} *simple-stochastic two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices*, ${}_{ss}\mathbb{G}^k$, ${}_{ss}\mathbb{GT}^k$ and ${}_{ss}\mathbb{GQ}^k$, can be directly obtained from $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$, respectively. The coefficients of these matrices are computed as follows:

$$ {}_{ss}g_{ij}^{k} = \frac{{}_{ns}g_{ij}^{k}}{S_j} = \frac{{}_{ns}g_{ij}^{k}}{\sum_{j=1}^{n} {}_{ns}g_{ij}^{k}} \quad \textbf{(4)} $$

$$ {}_{ss}gt_{ijl}^{k} = \frac{{}_{ns}gt_{ijl}^{k}}{S_{jl}} = \frac{{}_{ns}gt_{ijl}^{k}}{\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt_{ijl}^{k}} \quad \textbf{(5)} $$

$$ {}_{ss}gq_{ijlh}^{k} = \frac{{}_{ns}gq_{ijlh}^{k}}{S_{jlh}} = \frac{{}_{ns}gq_{ijlh}^{k}}{\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq_{ijlh}^{k}} \quad \textbf{(6)} $$

where, ${}_{ns}g_{ij}^{k}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gt_{ijl}^{k}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gq_{ijlh}^{k}$ are the elements of the k^{th} power of the matrix $\mathbb{GQ}$.

3. Double-stochastic (DS) matrix

The k^{th} *double-stochastic two-tuples spatial-(dis)similarity matrices*, ${}_{ds}\mathbb{G}^k$, can also be directly obtained from $\mathbb{G}^k$. Here, ${}_{ds}\mathbb{G}^k$ = [${}_{ds}g_{ij}^{k}$], is a square matrix of order *n* (*n* = number of atomic nuclei). It should be remarked that the matrix ${}_{ds}\mathbb{G}^k$ has the property that *the sum of the elements in each row or in each column* is 1. Notice that ${}_{ss}\mathbb{G}^k$ matrix (simple stochastic) is not symmetric, therefore, with the aim of equalize the probabilities in both senses is employed a *double-stochastic scaling*. This scaling is performed by using Sinkhorn-Knopp algorithm.

4. Mutual probability (MP) matrix

The k^{th} *mutual probability two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices*, ${}_{mp}\mathbb{G}^k$, ${}_{mp}\mathbb{GT}^k$ and ${}_{mp}\mathbb{GQ}^k$, can be directly obtained from $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$, respectively. The coefficients of these matrices are computed as follows:

$$ {}_{mp}g_{ij}^{k} = \frac{{}_{ns}g_{ij}^{k}}{S_{ij}} = \frac{{}_{ns}g_{ij}^{k}}{\sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns}g_{ij}^{k}} \tag{7} $$

$$ {}_{mp}gt_{ijl}^{k} = \frac{{}_{ns}gt_{ijl}^{k}}{S_{ijl}} = \frac{{}_{ns}gt_{ijl}^{k}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt_{ijl}^{k}} \tag{8} $$

$$ {}_{mp}gq_{ijlh}^{k} = \frac{{}_{ns}gq_{ijlh}^{k}}{S_{ijlh}} = \frac{{}_{ns}gq_{ijlh}^{k}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq_{ijlh}^{k}} \tag{9} $$

where, ${}_{ns}g_{ij}^{k}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gt_{ijl}^{k}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gq_{ijlh}^{k}$ are the elements of the k^{th} power of the matrix $\mathbb{GQ}$.