

## README: completely synthetic data

The completely synthetic data sets used in our empirical study are available. Each set is available as a compressed archive of *tab-delimited text* files.

To uncompress archive “Xrep\_low\_noise.tar.gz”:

- On linux: tar -xzvf Xrep\_low\_noise.tar.gz
- On windows: winzip would extract the archive

Each archive contains 5 synthetic data sets for a given number of repeated measurements and noise level. Each data set contains 400 genes, 20 experiments. The non-sporadic datasets consists of 6 classes in which classes 1 – 4 are periodic sine functions, while classes 5-6 are linear. The sporadic datasets consists of 7 classes in which class 7 consists of the sporadic genes. The third column contains the class numbers starting from 1.

### How did we generate these synthetic data sets?

Let  $\mu(i,j)$  be the artificial pattern of gene  $i$  and experiment  $j$  before error is added, and suppose gene  $i$  belongs to class  $k$ , where  $i = 1, 2, 3, \dots, 400, j = 1, 2, 3, \dots, 20, k=1,2,\dots, 6$ . When  $k = 1, 2, 3, 4$ , we set  $\mu(i,j) = \sin(2\pi j/10 - \pi k/4)$ . The size of each of these periodic classes is 67. When  $k = 5$ ,  $\mu(i,j) = j/20$ . When  $k = 6$ ,  $\mu(i,j) = -j/20$ . The size of class 5 and class 6 is 66 each. Let  $X(i,j,r)$  be the error-added value for gene  $i$ , experiment  $j$  and repeated measurement  $r$ . Let the randomly sampled error from the yeast galactose data be  $\epsilon_{ij}$  for gene  $i$  and experiment  $j$ . Let  $\sigma$  be the multiplicative factor that controls the noise level. When  $\sigma=1$ , we call it the “low noise level. When  $\sigma = 6$ , we call it the “high noise level”.  $X(i,j,r)$  is generated from a random normal distribution with mean equal to  $\mu(i,j)$ , and standard deviation equal to  $\sigma\epsilon_{ij}$ .

For the sporadic datasets, we generated random numbers from Uniform  $[-1, 1]$  for class 7, which consists of 40 sporadic genes. Classes 1 to 6 consist of 60 genes each.