

**Introduction**

**Problem** Given a set of six real breast cancer microarray datasets (including more than 1000 patients) coming from different populations, hospitals and microarray platforms (see [1]): how can we infer a "consensus" transcriptional network in order to discover new genetic interactions, potentially revealing novel therapeutic targets or prognostic genes?

**MRNET**

MRNET is a transcriptional network inference method particularly adapted for large microarray datasets [2]. MRNET infers a network using the MRMR feature selection method (in a forward selection strategy) where each gene in turn plays the role of the target output $X_y$.

Given a set $X_S$ of selected variables, the MRMR criterion updates $X_S$ by choosing the variable

$$X_i^{MRMR} = \arg \max_{X_i \in X_S} (u_i - r_i)$$

where $u_i = I(X_i; X_y)$ is a relevance term and $r_i = \frac{1}{|X_S|} \sum_{X_i \in X_S} I(X_i; X_y)$ is a redundancy term. Hence, MRNET can infer a network from the matrix of pairwise mutual information (MI matrix).

This fast inference method is freely available in the open-source R and Bioconductor package MINET [3].

**Meta-network**

**Methods**

- **Method A**: Aggregate datasets using standard normalization: $\frac{X_i}{\sigma_i}$
- **Method B**: Aggregate matrices of pairwise mutual information using a weighted average based on $m_j$, the number of samples in dataset $j$:

$$W^\text{comp} = \sum_j \frac{m_j W^j}{m_j}$$

- **Method C**: Aggregate networks using sum of weights of each arc in each network:

$$W = W_1 + W_2 + \ldots + W_n$$

> Which meta-network is the best (A, B or C)?

**Synthetic Experiments**

- **Artificial Dataset**: 300 genes, 800 samples generated with SynTren [2].
- **Low heterogeneity configuration**:
  - $n = 4$ datasets
  - normally distributed noise
  - random noise intensity between 5% and 15%
- **High heterogeneity configuration**:
  - $n = 10$ datasets
  - randomly chosen {gaussian, lognormal and gamma} distributed noise
  - with noise intensity between 5% and 30%
  - non-linear transformation of data randomly chosen between $\{\text{none, } x^2, \log(x), x^3\}$

⇒ on 100 runs, method B significantly outperforms (in terms of average F-score) A and C on both configurations.

**Real data**

Let $Y$ be a multiclass survival variable (indicating time before metastasis), does the set of connected nodes to $Y$ form a predictive signature competitive with previously published ones?

- Six breast cancer datasets
- MRNET applied on each of them
- Meta-network built using Method B
- 100 selected genes connected to survival variable (up to two levels)

Using protocols, signatures and data from [1]

- The performance of the new signature in a Dataset-CV setting is competitive with the best published prognostic signatures studied.
- The selected nodes are highly present in published prognostic signatures representing many different biological processes:

<table>
<thead>
<tr>
<th>Function</th>
<th>Genes</th>
<th>Signatures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proliferation</td>
<td>38</td>
<td>AURKA and GDI</td>
</tr>
<tr>
<td>Immune response</td>
<td>4</td>
<td>STAT1 and IR304D</td>
</tr>
<tr>
<td>Stroma</td>
<td>4</td>
<td>SDPP</td>
</tr>
<tr>
<td>Commercial progs</td>
<td>8</td>
<td>GENET0-76 and ONCOTYPE</td>
</tr>
</tbody>
</table>

**References**

