Similarity network fusion for aggregating data types on a genomic scale

Bo Wang, Aziz M Mezlini, Feyyaz Demir, et al.
Nature Methods, Volume 11, pp. 333-337 (Jan 26, 2014)

Presentation by Mykhaylo M. Malakhov, Souradipto Ghosh Dastidar, and Zhiyu Kang

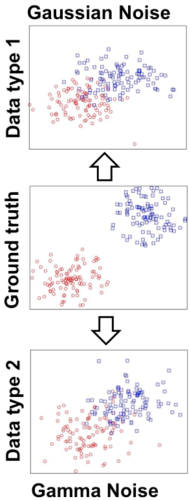
Table of contents

Background

- Easier to collect multiple and diverse genome-scale data sets.
- Data-integrated methods are needed to create a comprehensive view of a given disease or a biological process.
- Three computational challenges:
  - The small number of samples compared to the large number of measurements.
  - The differences in scale, collection bias and noise in each data set.
  - The complementary nature of the information provided by different types of data.

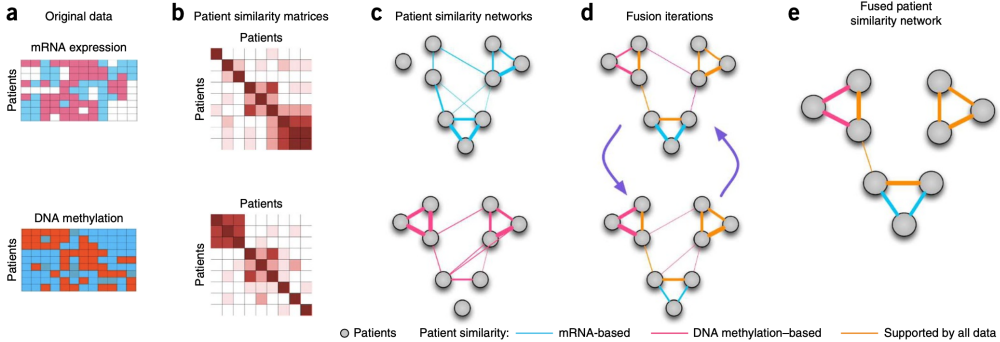
Existing Methods

- The simplest way is to concatenate normalized measurements.
- Independently analyze each data set before combining.
- Preselect a set of important genes from each data source and use Consensus Clustering to combine the data
- iCluster: a joint latent variable model for integrative clustering.

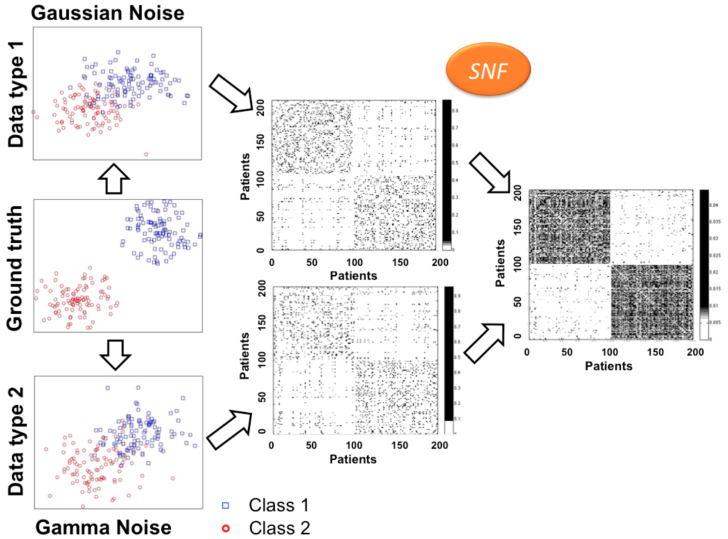


Similarity Network Fusion (SNF)

- SNF uses networks of samples as a basis for integration.

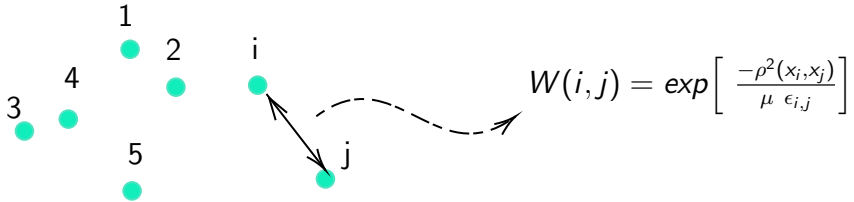


A Simple Simulation



Method

- Graph with n vertices, each vertex representing a patient x_i (each with m measurements).
- Edges weighted by similarity $W(i,j)$



- $\rho(x_i, x_j)$ = Euclidean distance between patients i and j
- μ = empirically set hyperparameter; typically in [0.3,0.8]
- $\epsilon_{i,j}$ = scaling factor := $\frac{mean(\rho(x_i,N_i))+\rho(x_i,x_j)+mean(\rho(x_j,N_j))}{3}$; N_i = neighbours of vertex i

Method (contd.)

- Define Full Kernel **P** s.t.

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\Sigma_{k \neq i} W(i,k)} & j \neq i \\ 1/2 & j = i \end{cases}$$

Note: $\Sigma_j \; P(i,j) = 1 \; \forall \; i$

- Define Local Affinity Kernel based on K nearest neighbours (KNN) **S** s.t.

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\Sigma_{k \in N_i} W(i,k)} & j \in N_i \\ 0 & \text{otherwise} \end{cases}$$

Note: Similarity between non-neighboring points is set to zero

Method (contd.)

Algorithm: Consider m=2. Calculate $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$.

- Initializing with $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$, model $\mathbf{P}_t^{(1)}$ and $\mathbf{P}_t^{(2)}$ like a discrete bivariate markov chain:

$$\mathbf{P}_{t+1}^{(1)} = \mathbf{S}^{(1)} \times \mathbf{P}_t^{(2)} \times (\mathbf{S}^{(1)})^T$$
$$\mathbf{P}_{t+1}^{(2)} = \mathbf{S}^{(2)} \times \mathbf{P}_t^{(1)} \times (\mathbf{S}^{(2)})^T$$

- After t steps,calculate overall status matrix:

$$\mathbf{P}^{(c)} = \frac{\mathbf{P}_t^{(1)} + \mathbf{P}_t^{(2)}}{2}$$

Visualization of Algorithm

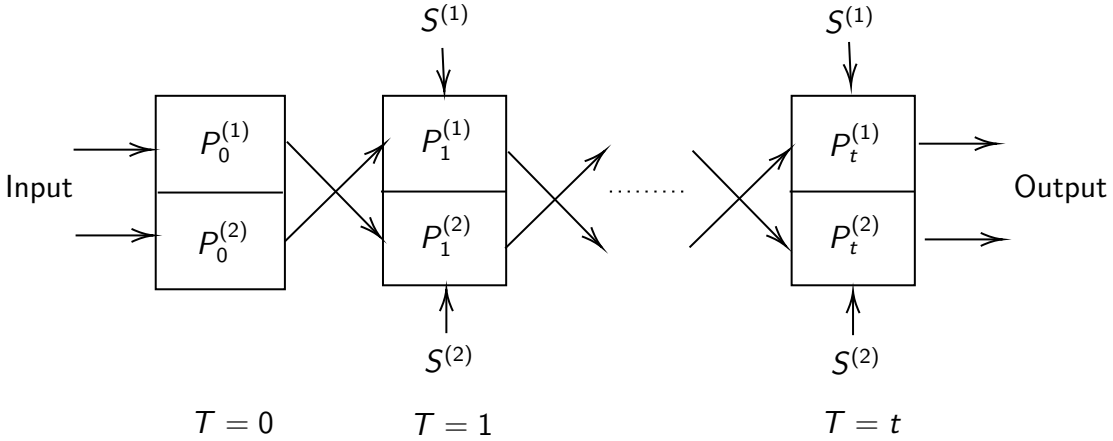


Figure: Discrete Bivariate Markov Chain Model

Method (Contd.)

- **Extension to $m > 2$:**

$$\mathbf{P}^{(\nu)} = \mathbf{S}^{(\nu)} \times \left(\frac{\sum_{k \neq \nu} \mathbf{P}^{(k)}}{m-1} \right) \times (\mathbf{S}^{(\nu)})^T$$

- Given the final similarity matrix, we can identify the clusters using standard network clustering algorithms (eg. spectral clustering)

Case study: glioblastoma multiforme

- Glioblastoma multiforme (GBM) is an aggressive adult brain tumor
- Several previous studies have attempted to identify the subtypes of GBM, but they all used different data types and hence reached different conclusions

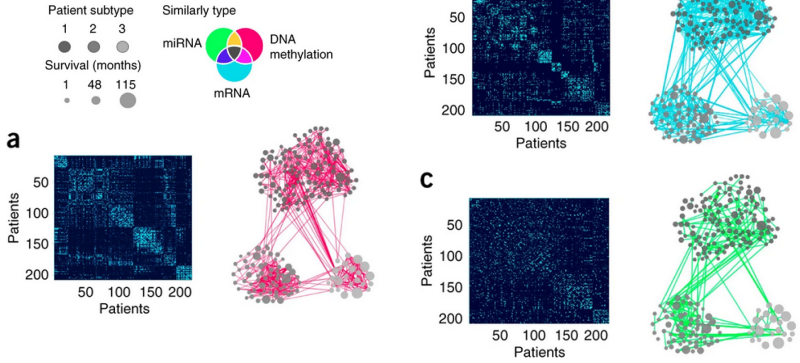
What if we fuse the different data types before clustering?

(A) DNA methylation (1,491 genes)

(B) mRNA expression (12,042 genes)

(C) microRNA expression (534 microRNAs)

Case study: glioblastoma multiforme

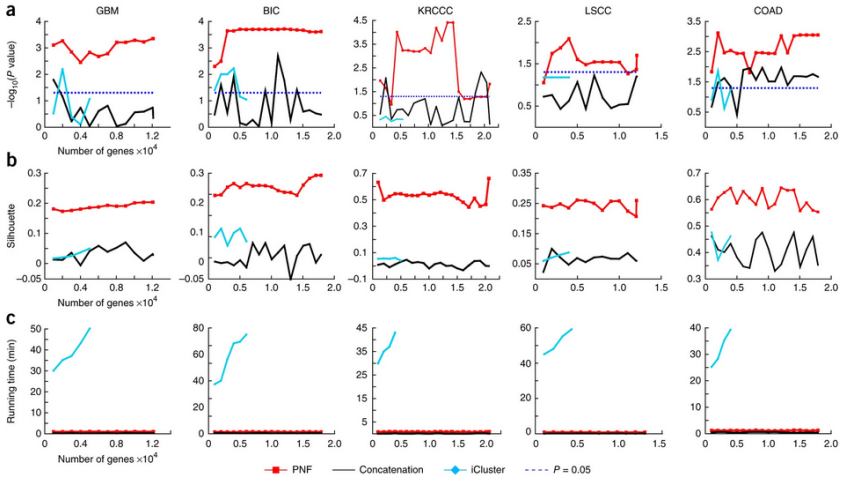


Case study: glioblastoma multiforme

Key result: spectral clustering on the SNF-fused network detected biologically informative clusters.

- The smallest cluster corresponds to a previously identified subtype consisting of younger patients with a better prognosis
- The largest cluster corresponds to patients who had a favorable response to temozolomide, a drug commonly used to treat GBM
- The mid-size cluster is significantly associated with *CTSD* overexpression, which has been found to inhibit treatment

SNF vs concatenation vs iCluster



Questions?