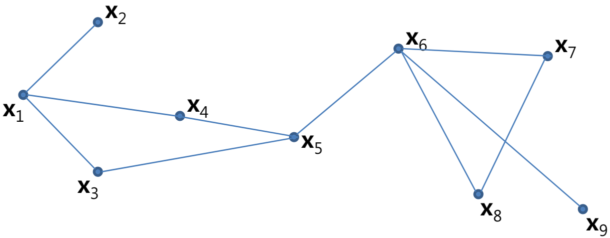
A Unified Approach to Clustering and Ordering via a Graph Theory

Choongrak Kim, Younghee Hong

Department of Statistics,
Pusan National University

Dec. 17, 2011



$\mathbf{x}_i$: vertex (p-vector, $i = 1, \cdots, n$)

a_{ij} : closeness (adjacency, connectivity) between $\mathbf{x}_i$ and $\mathbf{x}_j$

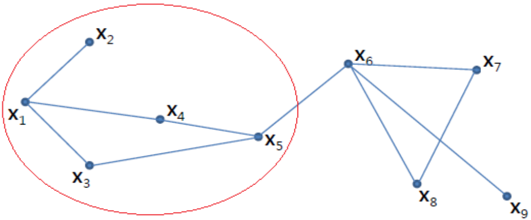
$a_{ij} = a_{ji}$: undirected; $a_{ij} \neq a_{ji}$: directed

$a_{ij} = 0$ or 1 : unweighted; $a'_{ij}s$ are different : weighted

$\mathbf{A} = (a_{ij})$: adjacency matrix

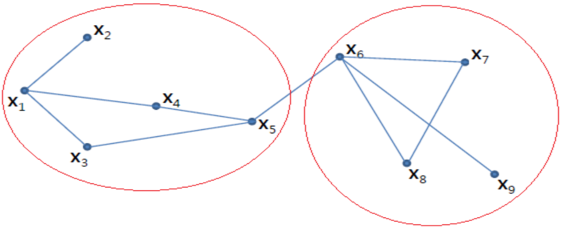
1. Motivation

1.1 Clustering and Hubbing



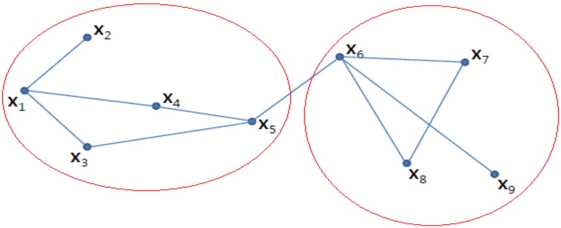
1. Motivation

1.1 Clustering and Hubbing



1. Motivation

1.1 Clustering and Hubbing



$\mathbf{A} = (a_{ij})$: symmetric (undirected) and weighted with $a_{ii} = 0, \ \forall i$

goal : grouping objects with similar characteristics, and finding hub in each group

- z_i : unknown index of the object $\mathbf{x}_i$

  If $z_i = z_j$, then $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same group. Therefore, clustering is finding $\mathbf{z} = (z_1, \cdots, z_n)$.

  (e.g) $z_1 = z_2 = \cdots = z_5, \quad z_6 = z_7 = z_8 = z_9$

- If $\mathbf{x}_i$ is a hub in a group, then z_i must be much different from z_j's, $j \neq i$ in that group.
  Therefore, finding a hub is also finding $\mathbf{z} = (z_1, \cdots, z_n)$.

  (e.g) $z_6 = 0.8, \quad$ other $z_i's$ are almost 0.

1.2 Loci Ordering

$g_i, \; i = 1, \cdots, n$: genes on a chromosome

a_{ij} : recombination fraction between g_i and g_j

$\mathbf{A} = (a_{ij})$: symmetric and weighted with $a_{ii} = 0, \; \forall i$

- goal : order genes on a chromosome

z_i : unknown index of the gene g_i

Loci ordering is ordering of z_i's
(e.g) $z_1 > z_2 > \cdots > z_n$

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, ..., z_n)'$?

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, ..., z_n)'$?

$$(z_i - z_j)^2 a_{ij}$$

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, ..., z_n)'$?

$$Q = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (z_i - z_j)^2 a_{ij}$$

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, \ldots, z_n)'$?

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1)$$

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, \dots, z_n)'$?

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1)$$
$$= \sum_{i=1}^{n} z_i^2 \sum_{j=1}^{n} a_{ij} - \sum_{j=1}^{n}\sum_{i=1}^{n} z_i z_j a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1)$$

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, ..., z_n)'$?

$$
\begin{aligned}
Q &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\
&= \sum_{i=1}^{n} z_i^2 \sum_{j=1}^{n} a_{ij} - \sum_{j=1}^{n}\sum_{i=1}^{n} z_i z_j a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\
&= \sum_{i=1}^{n} z_i^2 d_i - \sum_{j=1}^{n}\sum_{i=1}^{n} z_i z_j a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1), \quad d_i = \sum_{j=1}^{n} a_{ij}
\end{aligned}
$$

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, ..., z_n)'$?

$$
\begin{aligned}
Q &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\
&= \sum_{i=1}^{n} z_i^2 \sum_{j=1}^{n} a_{ij} - \sum_{j=1}^{n}\sum_{i=1}^{n} z_i z_j a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\
&= \sum_{i=1}^{n} z_i^2 d_i - \sum_{j=1}^{n}\sum_{i=1}^{n} z_i z_j a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1), \quad d_i = \sum_{j=1}^{n} a_{ij} \\
&= \mathbf{z}'\mathbf{D}\mathbf{z} - \mathbf{z}'\mathbf{A}\mathbf{z} - \lambda(\mathbf{z}'\mathbf{z} - 1), \quad \mathbf{D} = diag(d_1, \cdots, d_n)
\end{aligned}
$$

2.Estimation of z

2.1 Theory

How to find $\mathbf{z} = (z_1, ..., z_n)'$?

$$
\begin{aligned}
Q &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\
&= \sum_{i=1}^{n} z_i^2 \sum_{j=1}^{n} a_{ij} - \sum_{j=1}^{n}\sum_{i=1}^{n} z_i z_j a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\
&= \sum_{i=1}^{n} z_i^2 d_i - \sum_{j=1}^{n}\sum_{i=1}^{n} z_i z_j a_{ij} - \lambda(\mathbf{z}'\mathbf{z} - 1), \quad d_i = \sum_{j=1}^{n} a_{ij} \\
&= \mathbf{z}'\mathbf{D}\mathbf{z} - \mathbf{z}'\mathbf{A}\mathbf{z} - \lambda(\mathbf{z}'\mathbf{z} - 1), \quad \mathbf{D} = diag(d_1, \cdots, d_n) \\
&= \mathbf{z}'\mathbf{L}\mathbf{z} - \lambda(\mathbf{z}'\mathbf{z} - 1), \quad \mathbf{L} = \mathbf{D} - \mathbf{A} : \textit{Laplacian matrix}
\end{aligned}
$$

2.1 Theory

$$\frac{\partial Q}{\partial \mathbf{z}} = 2\mathbf{L}\mathbf{z} - 2\lambda\mathbf{z} = \mathbf{0}$$

$$\mathbf{L}\mathbf{z} = \lambda\mathbf{z}$$

$$\mathbf{z}'\mathbf{L}\mathbf{z} = \lambda$$

$$\frac{\partial Q}{\partial \mathbf{z}} = 2\mathbf{L}\mathbf{z} - 2\lambda\mathbf{z} = \mathbf{0}$$

$$\mathbf{L}\mathbf{z} = \lambda\mathbf{z}$$

$$\mathbf{z}'\mathbf{L}\mathbf{z} = \lambda$$

$$0 = \lambda_1 < \lambda_2 < \cdots < \lambda_n$$

$\mathbf{z}_2$: Fiedler vector $\Rightarrow$ clustering (Eigenvector corresponding to non-zero smallest eigenvalue)

$$\frac{\partial Q}{\partial \mathbf{z}} = 2\mathbf{L}\mathbf{z} - 2\lambda\mathbf{z} = \mathbf{0}$$

$$\mathbf{L}\mathbf{z} = \lambda\mathbf{z}$$

$$\mathbf{z}'\mathbf{L}\mathbf{z} = \lambda$$

$$0 = \lambda_1 < \lambda_2 < \cdots < \lambda_n$$

$\mathbf{z}_2$: Fiedler vector $\Rightarrow$ clustering (Eigenvector corresponding to non-zero smallest eigenvalue)

$\mathbf{z}_n \Rightarrow$ hubbing

2.2 Relevant Works

- Dempster (1972, Biometrics) : inverse of covariance matrix
- Kim et al. (2008, PNAS) : clustering and classification via graph theory
- Peng et al. (2009, JASA) : variable selection in the inverse of covariance matrix
- Lee and Wasserman (2010, JASA) spectral kernel methods

2.3 Artificial Examples

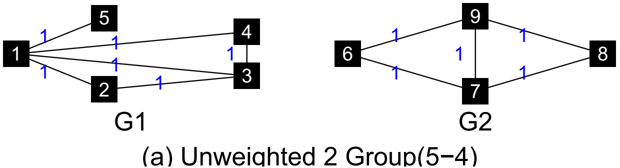


Figure 1. Graph example for matrix representation

2.3 Artificial Examples

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

$$\mathbf{L} = \begin{pmatrix} 4 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 2 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 \end{pmatrix}$$

$$\mathbf{L} = \begin{pmatrix} 4 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 2 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 \end{pmatrix}$$

Table 2.1 Eigenvalues and eigenvectors for the Laplacian matrix of the combined graphs 1(a) in Figure 1. ($noise = 0.0$)

Eigenvalues	5.00	4.00	4.00	4.00	2.00	2.00	1.00	0.00	0.00
1	0.89	0.00	0.00	0.00	0.00	0.00	0.00	-0.45	0.00
2	-0.22	0.00	0.00	0.41	0.00	0.71	-0.29	-0.45	0.00
3	-0.22	0.00	0.00	-0.82	0.00	0.00	-0.29	-0.45	0.00
4	-0.22	0.00	0.00	0.41	0.00	-0.71	-0.29	-0.45	0.00
5	-0.22	0.00	0.00	0.00	0.00	0.00	0.87	-0.45	0.00
6	0.00	0.50	0.00	0.00	0.71	0.00	0.00	0.00	-0.50
7	0.00	-0.50	-0.71	0.00	0.00	0.00	0.00	0.00	-0.50
8	0.00	0.50	0.00	0.00	-0.71	0.00	0.00	0.00	-0.50
9	0.00	-0.50	0.71	0.00	0.00	0.00	0.00	0.00	-0.50

$$\mathbf{L} = \begin{pmatrix} 4.4 & -1 & -1 & -1 & -1 & -0.1 & -0.1 & -0.1 & -0.1 \\ -1 & 2.4 & -1 & 0 & 0 & -0.1 & -0.1 & -0.1 & -0.1 \\ -1 & -1 & 3.4 & -1 & 0 & -0.1 & -0.1 & -0.1 & -0.1 \\ -1 & 0 & -1 & 2.4 & 0 & -0.1 & -0.1 & -0.1 & -0.1 \\ -1 & 0 & 0 & 0 & 1.4 & -0.1 & -0.1 & -0.1 & -0.1 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 & 2.5 & -1 & 0 & -1 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 & -1 & 3.5 & -1 & -1 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 & 0 & -1 & 2.5 & -1 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 & -1 & -1 & -1 & 3.5 \end{pmatrix}$$

Table 2.2 Eigenvalues and eigenvectors for the Laplacian matrix of the combined graphs 1(a) in Figure 1. ($noise = 0.1$)

Eigenvalues	5.40	4.50	4.50	4.40	2.50	2.40	1.40	0.90	0.00
1	0.89	0.00	0.00	0.00	0.00	0.00	0.00	-0.30	-0.33
2	-0.22	0.00	0.00	0.41	0.00	0.71	0.29	-0.30	-0.33
3	-0.22	0.00	0.00	-0.82	0.00	0.00	0.29	-0.30	-0.33
4	-0.22	0.00	0.00	0.41	0.00	-0.71	0.29	-0.30	-0.33
5	-0.22	0.00	0.00	0.00	0.00	0.00	-0.87	-0.30	-0.33
6	0.00	-0.05	-0.50	0.00	0.71	0.00	0.00	0.37	-0.33
7	0.00	-0.65	0.57	0.00	0.00	0.00	0.00	0.37	-0.33
8	0.00	-0.05	-0.50	0.00	-0.71	0.00	0.00	0.37	-0.33
9	0.00	0.76	0.42	0.00	0.00	0.00	0.00	0.37	-0.33

2.3 Artificial Examples

$$\mathbf{L} = \begin{pmatrix} 4.8 & -1 & -1 & -1 & -1 & -0.2 & -0.2 & -0.2 & -0.2 \\ -1 & 2.8 & -1 & 0 & 0 & -0.2 & -0.2 & -0.2 & -0.2 \\ -1 & -1 & 3.8 & -1 & 0 & -0.2 & -0.2 & -0.2 & -0.2 \\ -1 & 0 & -1 & 2.8 & 0 & -0.2 & -0.2 & -0.2 & -0.2 \\ -1 & 0 & 0 & 0 & 1.8 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & 3.0 & -1 & 0 & -1 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & -1 & 4.0 & -1 & -1 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & 0 & -1 & 3.0 & -1 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & -1 & -1 & -1 & 4.0 \end{pmatrix}$$

Table 2.3 Eigenvalues and eigenvectors for the Laplacian matrix of the combined graphs 1(a) in Figure 1. ($noise = 0.2$)

Eigenvalues	5.80	5.00	5.00	4.80	$\cdots$	1.80	0.00
1	0.89	0.00	0.00	0.00	$\cdots$	0.00	-0.33
2	-0.22	0.00	0.00	-0.41	$\cdots$	-0.29	-0.33
3	-0.22	0.00	0.00	0.82	$\cdots$	-0.29	-0.33
4	-0.22	0.00	0.00	-0.41	$\cdots$	-0.29	-0.33
5	-0.22	0.00	0.00	0.00	$\cdots$	0.87	-0.33
6	0.00	-0.23	-0.44	0.00	$\cdots$	0.00	-0.33
7	0.00	-0.40	0.77	0.00	$\cdots$	0.00	-0.33
8	0.00	-0.23	-0.44	0.00	$\cdots$	0.00	-0.33
9	0.00	0.86	0.12	0.00	$\cdots$	0.00	-0.33

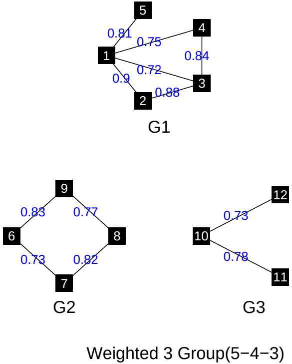


Figure 2. Graph example for matrix representation

Table 2.4 Eigenvalues and eigenvectors for the Laplacian matrix of the combined graphs 2 in Figure 2. ($noise = 0.0$)

Eigenvalues	4.00	3.32	3.15	2.27	...	0.75	0.00	0.00	0.00
1	0.88	0.14	0.00	0.00	...	0.00	0.45	0.00	0.00
2	-0.32	0.40	0.00	0.00	...	0.00	0.45	0.00	0.00
3	-0.10	-0.84	0.00	0.00	...	0.00	0.45	0.00	0.00
4	-0.24	0.35	0.00	0.00	...	0.00	0.45	0.00	0.00
5	-0.22	-0.04	0.00	0.00	...	0.00	0.45	0.00	0.00
6	0.00	0.00	0.49	0.00	...	0.00	0.00	-0.50	0.00
7	0.00	0.00	-0.48	0.00	...	0.00	0.00	-0.50	0.00
8	0.00	0.00	0.51	0.00	...	0.00	0.00	-0.50	0.00
9	0.00	0.00	-0.52	0.00	...	0.00	0.00	-0.50	0.00
10	0.00	0.00	0.00	0.82	...	-0.02	0.00	0.00	0.58
11	0.00	0.00	0.00	-0.43	...	-0.70	0.00	0.00	0.58
12	0.00	0.00	0.00	-0.39	...	0.72	0.00	0.00	0.58

Table 2.5 Eigenvalues and eigenvectors for the Laplacian matrix of the combined graphs 2 in Figure 2. ($noise = 0.1$)

Eigenvalues	4.70	4.02	3.93	3.17	...	1.17	0.00
1	0.88	-0.14	0.00	0.00	...	-0.30	-0.28
2	-0.32	-0.40	0.06	0.00	...	-0.32	-0.27
3	-0.10	0.84	-0.16	0.00	...	-0.30	-0.28
4	-0.24	-0.35	0.08	0.00	...	-0.29	-0.29
5	-0.22	0.04	0.00	0.00	...	-0.27	-0.29
6	0.00	0.00	-0.46	0.00	...	0.39	-0.30
7	0.00	0.00	0.48	0.00	...	0.37	-0.29
8	0.00	0.00	-0.51	0.00	...	0.36	-0.29
9	0.00	0.00	0.51	0.00	...	0.37	-0.29
10	0.00	0.00	0.00	-0.82	...	0.00	-0.29
11	0.00	0.00	0.00	0.43	...	0.00	-0.29
12	0.00	0.00	0.00	0.39	...	0.00	-0.29

Table 2.6 Eigenvalues and eigenvectors for the Laplacian matrix of the combined graphs 2 in Figure 2. ($noise = 0.2$)

Eigenvalues	5.40	4.72	4.71	4.07	...	2.20	0.00
1	0.88	0.14	0.10	0.00	...	0.01	0.29
2	-0.32	0.40	0.35	0.00	...	-0.26	0.27
3	-0.10	-0.84	-0.76	0.00	...	-0.30	0.28
4	-0.24	0.35	0.33	0.00	...	-0.31	0.29
5	-0.22	-0.04	-0.03	0.00	...	0.86	0.29
6	0.00	0.00	-0.19	0.00	...	0.00	0.31
7	0.00	0.00	0.21	0.00	...	0.00	0.29
8	0.00	0.00	-0.22	0.00	...	0.00	0.29
9	0.00	0.00	0.22	0.00	...	0.00	0.29
10	0.00	0.00	0.00	-0.82	...	0.00	0.29
11	0.00	0.00	0.00	0.43	...	0.00	0.29
12	0.00	0.00	0.00	0.39	...	0.00	0.29

2.4 Hard-thresholding

observed value

$= \text{deterministic part} + \text{noise part}$

To remove noise, use the idea of shrankage

Here, we use hard-thresholding functions given by $t_H(x) = x\, I(|x| > \delta)$, where $\delta > 0$ is a thresholding parameter to be estimated.

3. Locus Ordering

3.1 Introduction

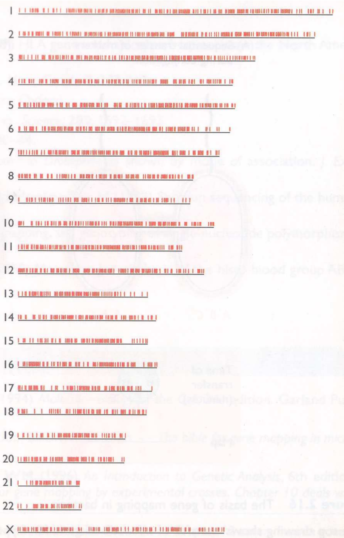
(1) Locus

chromosome location of a gene

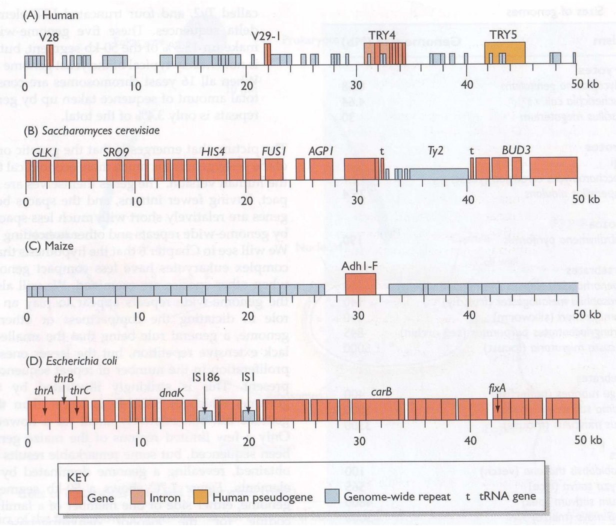
(2) Locus Ordering

- *a linear arrangement of genes or genetic markers in a linkage group*
- *necessary step in constructing genetic map*
- *one of the most important issue in genetic research area*

3.1 Introduction



3.1 Introduction



(3) Crossover

- *As meiosis(cell division leading to the formation of gametes) progresses, crossover might occur at chiasmata.*
- *The chance of crossover is low when two loci are closely located, and it is high when they are apart.*
- *An odd number of crossover between two loci leads to a recombination.*
- *Therefore, recombination fraction, the ratio of recombinant gametes to total gametes, is a measure of distance between two loci.*

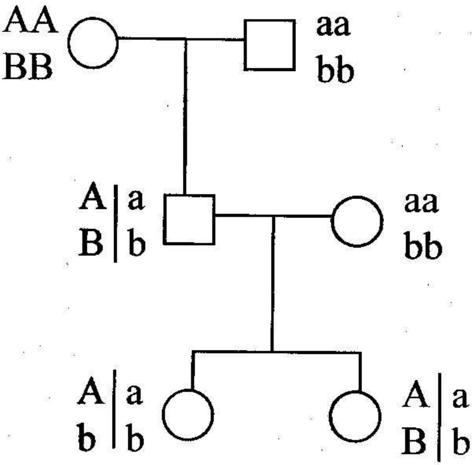
(3) Crossover

$r_{ij}, \ 1 \leq i < j \leq n$

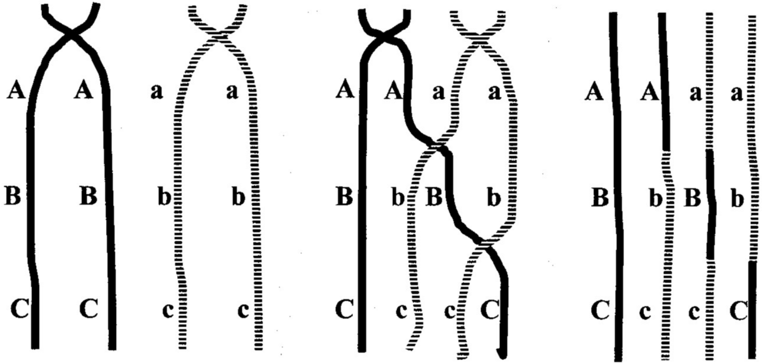
: *two-point recombination fractions for a pair of loci i and j*

If two loci i and j are closely located, i.e., tightly linked, then r_{ij} is close to 0, and if not it is away from 0.

3.1 Introduction



3.1 Introduction



(4) Computational issues

With n loci, there are $n!/2$ possible orderings if the orientation of the orders is ignored.

10 *loci :* $1,814,000$ *possible orders*

If analysis of each order takes 1 second(it might take more than that), evaluating all orders requires 21 days of uninterrupted computation.

(5) Existing methods

- *Falk(1989) : minimum sum of adjacent recombination fractions*
- *Weeks and Lange(1987) : maximum sum of adjacent lod scores*
- *Knapp* et al. *(1989) : minimum sum of the probability of double recombinants*
- *Lander and Green (1987) : maximum likelihood*
- *Thompson (1989) : minimum obligatory crossovers*
- *Kammerer and MacCluer(1988) and Olson and Boehnke(1990) : Comparisons between those methods when the number of loci is* 6 *or* 7*.*
- *Weeks(1991) : overview in locus ordering.*

3.2 Algorithm

(1) Notations

For example,

if $\mathbf{z} = (-1/\sqrt{5}, 0, 1/\sqrt{5}, -\sqrt{2}/\sqrt{5}, \sqrt{2}/\sqrt{5})$,

then the resulting order is $4 - 1 - 2 - 3 - 5$

or $5 - 3 - 2 - 1 - 4$.

(2) Estimation of the thresholding parameter δ

(i) *As δ increases λ_2 decreases*

(ii) *We get more information as λ_2 becomes smaller*

(iii) *If δ is too large, then λ_2 will be zero so that we lose all the information.*

(iv) *To solve this contradicting situation, we consider λ_3. Since the eigenvectors $\mathbf{z}_2$ and $\mathbf{z}_3$ are orthogonal, the information contained in $\mathbf{z}_3$ is independent on the information contained in $\mathbf{z}_2$. Therefore, it is desirable that λ_2 must be relatively small compared to λ_3. Then, most of locus ordering information is contained in the corresponding eigenvector $\mathbf{z}_2$. Choose δ maximizing*

$$\Lambda = \lambda_3/\lambda_2.$$

(2) Accuracy of Locus Ordering

$$\mathbf{t} = (t(1), \cdots, t(n))$$

: $t(i)$ *denote the true order of the* i*th gene*

$$\mathbf{e} = (e(1), \cdots, e(n))$$

: $e(i)$ *denote the estimated order of the* i*th gene*

The accuracy of the estimated order $\mathbf{e}$

$NCL = \sum_{i=1}^{n} I(t(i) = e(i))$: *number of correct loci*

$PIL = \sum_{i=1}^{n} |(t(i) - e(i))|$: *penalized incorrect loci*

3.3 Example

(1) data : 26 loci of barley chromosome IV generated by the North American Barley Genome Mapping Project (NABGMP)

(2) recombination fraction matrix for 26 loci :

(3) Based on the several preliminary locus ordering method it has been known that the best ordering for the 26 loci is $1, 2, \cdots, 26$

3.3 Example

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	.00	.89	.16	.29	.31	.43	.41	.44	.44	.46	.49	.48	.49
2	.89	.00	.95	.81	.79	.69	.71	.66	.65	.63	.63	.62	.62
3	.16	.95	.00	.87	.86	.76	.78	.73	.72	.71	.70	.69	.68
4	.29	.81	.87	.00	.96	.89	.90	.86	.86	.84	.81	.81	.79
5	.31	.79	.86	.96	.00	.90	.89	.83	.82	.83	.82	.80	.80
6	.43	.69	.76	.89	.90	.00	.99	.93	.95	.93	.92	.91	.87
7	.41	.71	.78	.90	.89	.99	.00	.92	.92	.92	.90	.89	.86
8	.44	.66	.73	.86	.83	.93	.92	.00	.99	.98	.93	.94	.91
9	.44	.65	.72	.86	.82	.95	.92	.99	.00	.99	.95	.93	.92
10	.46	.63	.71	.84	.83	.93	.92	.98	.99	.00	.97	.96	.93
11	.49	.63	.70	.81	.82	.92	.90	.93	.95	.97	.00	.98	.96
12	.48	.62	.69	.81	.80	.91	.89	.94	.93	.96	.98	.00	.95
13	.49	.62	.68	.79	.80	.87	.86	.91	.92	.93	.96	.95	.00
14	.47	.65	.70	.82	.81	.90	.89	.94	.94	.96	.98	.98	.96
15	.44	.67	.72	.82	.81	.88	.87	.92	.92	.92	.93	.96	.93
16	.41	.67	.68	.73	.76	.82	.81	.87	.86	.88	.90	.89	.86
17	.45	.60	.63	.68	.67	.74	.72	.78	.77	.79	.78	.79	.77
18	.48	.57	.59	.64	.62	.69	.69	.72	.72	.74	.74	.72	.74
19	.51	.52	.50	.52	.51	.53	.52	.58	.59	.59	.61	.62	.63
20	.52	.51	.51	.52	.53	.57	.56	.59	.60	.60	.63	.63	.64
21	.52	.52	.51	.51	.51	.52	.53	.56	.56	.57	.60	.60	.61
22	.52	.52	.51	.51	.50	.54	.53	.55	.56	.57	.60	.59	.60
23	.52	.53	.52	.52	.51	.53	.54	.56	.57	.57	.60	.59	.61
24	.54	.50	.49	.50	.48	.52	.52	.53	.55	.56	.60	.59	.60
25	.54	.50	.50	.53	.52	.53	.52	.55	.56	.57	.59	.58	.61
26	.57	.48	.47	.49	.47	.50	.51	.53	.54	.52	.57	.57	.60

3.3 Example

	14	15	16	17	18	19	20	21	22	23	24	25	26
1	.47	.44	.41	.45	.48	.51	.52	.52	.52	.52	.54	.54	.57
2	.65	.67	.67	.60	.57	.52	.51	.52	.52	.53	.50	.50	.48
3	.70	.72	.68	.63	.59	.50	.51	.51	.51	.52	.49	.50	.47
4	.82	.82	.73	.68	.64	.52	.52	.51	.51	.52	.50	.53	.49
5	.81	.81	.76	.67	.62	.51	.53	.51	.50	.51	.48	.52	.47
6	.90	.88	.82	.74	.69	.53	.57	.52	.54	.53	.52	.53	.50
7	.89	.87	.81	.72	.69	.52	.56	.53	.53	.54	.52	.52	.51
8	.94	.92	.87	.78	.72	.58	.59	.56	.55	.56	.53	.55	.53
9	.94	.92	.86	.77	.72	.59	.60	.56	.56	.57	.55	.56	.54
10	.96	.92	.88	.79	.74	.59	.60	.57	.57	.57	.56	.57	.52
11	.98	.93	.90	.78	.74	.61	.63	.60	.60	.60	.60	.59	.57
12	.98	.96	.89	.79	.72	.62	.63	.60	.59	.59	.59	.58	.57
13	.96	.93	.86	.77	.74	.63	.64	.61	.60	.61	.60	.61	.60
14	.00	.97	.90	.79	.76	.61	.62	.59	.58	.58	.58	.59	.57
15	.97	.00	.92	.83	.80	.62	.63	.60	.59	.59	.59	.59	.57
16	.90	.92	.00	.88	.82	.67	.69	.66	.66	.62	.63	.62	.62
17	.79	.83	.88	.00	.93	.77	.77	.76	.73	.72	.73	.72	.70
18	.76	.80	.82	.93	.00	.82	.84	.82	.81	.78	.80	.78	.76
19	.61	.62	.67	.77	.82	.00	.97	.92	.93	.93	.95	.91	.89
20	.62	.63	.69	.77	.84	.97	.00	.98	.98	.96	.97	.93	.89
21	.59	.60	.66	.76	.82	.92	.98	.00	1.00	.98	.99	.93	.91
22	.58	.59	.66	.73	.81	.93	.98	1.00	.00	.98	.99	.95	.92
23	.58	.59	.62	.72	.78	.93	.96	.98	.98	.00	.99	.92	.91
24	.58	.59	.63	.73	.80	.95	.97	.99	.99	.99	.00	.96	.93
25	.59	.59	.62	.72	.78	.91	.93	.93	.95	.92	.96	.00	.92
26	.57	.57	.62	.70	.76	.89	.89	.91	.92	.91	.93	.92	.00

3.3 Example

$$\delta = 0, \quad \lambda_2 = 12.16, \quad \lambda_3 = 14.72, \quad \Lambda = 1.21$$

true order	Z	predicted order	$\lvert diff \rvert$	C. I.	
1	0.9695	1	0	1-1	O
2	0.0429	2	0	2-2	O
3	-0.1153	26	23	3-3	X
4	-0.0791	25	21	4-5	X
5	-0.0788	24	19	4-5	X
6	-0.0558	22	16	6-7	X
7	-0.0589	23	16	6-7	X
8	-0.0520	21	13	8-11	X
9	-0.0518	20	11	8-11	X
10	-0.0492	17	7	8-13	X
11	-0.0440	14	3	11-15	O
12	-0.0458	15	3	11-15	O
13	-0.0431	13	0	11-15	O
14	-0.0459	16	2	11-15	X
15	-0.0506	18	3	11-15	X
16	-0.0507	19	3	16-16	X
17	-0.0423	12	5	17-17	X
18	-0.0347	11	7	18-18	X
19	-0.0193	10	9	19-24	X
20	-0.0187	9	11	19-25	X
21	-0.0164	8	13	19-25	X
22	-0.0159	7	15	19-25	X
23	-0.0159	6	17	19-25	X
24	-0.0110	4	20	19-25	X
25	-0.0127	5	20	19-25	X
26	-0.0035	3	23	25-26	X

3.3 Example

$$\delta = 0.1, \quad \lambda_2 = 12.16, \quad \lambda_3 = 14.72, \quad \Lambda = 1.21$$

true order	Z	predicted order	$\lvert diff \rvert$	C. I.	
1	0.9695	1	0	1-1	O
2	0.0429	2	0	2-2	O
3	-0.1153	26	23	3-3	X
4	-0.0791	25	21	4-5	X
5	-0.0788	24	19	4-5	X
6	-0.0558	22	16	6-7	X
7	-0.0589	23	16	6-7	X
8	-0.0520	21	13	8-11	X
9	-0.0518	20	11	8-11	X
10	-0.0492	17	7	8-13	X
11	-0.0440	14	3	11-15	O
12	-0.0458	15	3	11-15	O
13	-0.0431	13	0	11-15	O
14	-0.0459	16	2	11-15	X
15	-0.0506	18	3	11-15	X
16	-0.0507	19	3	16-16	X
17	-0.0423	12	5	17-17	X
18	-0.0347	11	7	18-18	X
19	-0.0193	10	9	19-24	X
20	-0.0187	9	11	19-25	X
21	-0.0164	8	13	19-25	X
22	-0.0159	7	15	19-25	X
23	-0.0159	6	17	19-25	X
24	-0.0110	4	20	19-25	X
25	-0.0127	5	20	19-25	X
26	-0.0035	3	23	25-26	X

3.3 Example

$$\delta = 0.5, \quad \lambda_2 = 5.14, \quad \lambda_3 = 13.01, \quad \Lambda = 2.53$$

true order	Z	predicted order	$\lvert diff \rvert$	C.I.	
1	0.9701	1	0	1-1	O
2	0.0184	2	0	2-2	O
3	-0.0670	25	22	3-3	X
4	-0.0641	24	20	4-5	X
5	-0.0673	26	21	4-5	X
6	-0.0613	23	17	6-7	X
7	-0.0612	22	15	6-7	X
8	-0.0605	21	13	8-11	X
9	-0.0602	20	11	8-11	X
10	-0.0602	19	9	8-13	X
11	-0.0591	16	5	11-15	X
12	-0.0593	17	5	11-15	X
13	-0.0587	14	1	11-15	O
14	-0.0593	18	4	11-15	X
15	-0.0590	15	0	11-15	O
16	-0.0572	13	3	16-16	X
17	-0.0545	12	5	17-17	X
18	-0.0524	11	7	18-18	X
19	-0.0075	10	9	19-24	X
20	-0.0075	9	11	19-25	X
21	-0.0059	8	13	19-25	X
22	-0.0056	7	15	19-25	X
23	-0.0054	6	17	19-25	X
24	0.0012	4	20	19-25	X
25	-0.0039	5	20	19-25	X
26	0.0085	3	23	25-26	X

3.3 Example

$$\delta = 0.58, \quad \lambda_2 = 0.85, \quad \lambda_3 = 5.62, \quad \Lambda = 6.61$$

true order	Z	predicted order	$\|diff\|$	C.I.	
1	0.9766	1	0	1-1	O
2	0.0451	2	0	2-2	O
3	-0.0349	3	0	3-3	O
4	-0.0362	4	0	4-5	O
5	-0.0363	5	0	4-5	O
6	-0.0372	6	0	6-7	O
7	-0.0370	7	0	6-7	O
8	-0.0379	8	0	8-11	O
9	-0.0383	9	0	8-11	O
10	-0.0385	10	0	8-13	O
11	-0.0403	12	1	11-15	O
12	-0.0401	14	2	11-15	O
13	-0.0408	11	2	11-15	O
14	-0.0402	15	1	11-15	O
15	-0.0402	13	2	11-15	O
16	-0.0409	16	0	16-16	O
17	-0.0417	17	0	17-17	O
18	-0.0450	18	0	18-18	O
19	-0.0484	19	0	19-24	O
20	-0.0479	20	0	19-25	O
21	-0.0494	21	0	19-25	O
22	-0.0495	22	0	19-25	O
23	-0.0495	23	0	19-25	O
24	0.0495	24	0	19-25	O
25	-0.0500	25	0	19-25	O
26	0.0520	26	0	25-26	O

3.3 Example

$$\delta = 0.6, \quad \lambda_2 = 0.85, \quad \lambda_3 = 3.77, \quad \Lambda = 4.44$$

true order	Z	predicted order	$\lvert diff \rvert$	C.I.	
1	0.9757	1	0	1-1	O
2	0.0475	2	0	2-2	O
3	-0.0314	3	0	3-3	O
4	-0.0335	4	0	4-5	O
5	-0.0336	5	0	4-5	O
6	-0.0345	7	1	6-7	O
7	-0.0343	6	1	6-7	O
8	-0.0348	8	0	8-11	O
9	-0.0348	9	0	8-11	O
10	-0.0350	10	0	8-13	O
11	-0.0377	14	3	11-15	O
12	-0.0364	13	1	11-15	O
13	-0.0402	15	2	11-15	O
14	-0.0363	12	2	11-15	O
15	-0.0362	11	4	11-15	O
16	-0.0404	16	0	16-16	O
17	-0.0416	17	0	17-17	O
18	-0.0457	18	0	18-18	O
19	-0.0516	20	1	19-24	O
20	-0.0516	19	1	19-25	O
21	-0.0545	21	0	19-25	O
22	-0.0546	22	0	19-25	O
23	-0.0558	24	1	19-25	O
24	-0.0557	23	1	19-25	O
25	-0.0558	25	0	19-25	O
26	-0.0559	26	0	25-26	O

3.3 Example

$$\delta = 0.8, \quad \lambda_2 = 0.23, \quad \lambda_3 = 0.52, \quad \Lambda = 2.26$$

true order	Z	predicted order	$\lvert diff \rvert$	C.I.	
1	0.3229	1	0	1-1	O
2	0.2391	2	0	2-2	O
3	0.1881	3	0	3-3	O
4	0.1391	4	0	4-5	O
5	0.1305	5	0	4-5	O
6	0.1226	7	1	6-7	O
7	0.1227	6	1	6-7	O
8	0.1223	9	1	8-11	O
9	0.1223	8	1	8-11	O
10	0.1222	10	0	8-13	O
11	0.1221	11	0	11-15	O
12	0.1221	12	0	11-15	O
13	0.1209	14	1	11-15	O
14	0.1220	13	1	11-15	O
15	0.0989	15	0	11-15	O
16	0.0917	16	0	16-16	O
17	0.0116	17	0	17-17	O
18	-0.1445	18	0	18-18	O
19	-0.2668	20	1	19-24	O
20	-0.2666	19	1	19-25	O
21	-0.2670	21	0	19-25	O
22	-0.2671	22	0	19-25	O
23	-0.2805	24	1	19-25	O
24	-0.2673	23	1	19-25	O
25	-0.2808	25	0	19-25	O
26	-0.2810	26	0	25-26	O

3.3 Example

$$\delta = 0.83, \quad \lambda_2 = 0.06, \quad \lambda_3 = 1.16, \quad \Lambda = 19.33$$

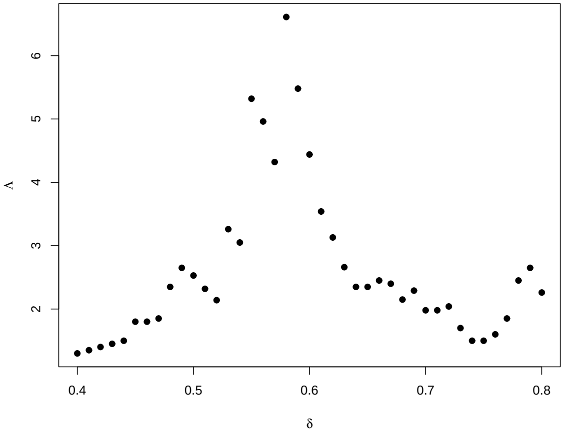
true order	Z	predicted order	$\lvert diff \rvert$	C.I.	
1	0.2856	1	0	1-1	O
2	0.2855	2	0	2-2	O
3	0.2854	5	2	3-3	X
4	0.2854	7	3	4-5	X
5	0.2854	3	2	4-5	X
6	0.2675	8	2	6-7	X
7	0.2854	6	1	6-7	O
8	-0.0461	10	2	8-11	O
9	0.1059	9	0	8-11	O
10	0.2854	4	6	8-13	X
11	-0.1286	12	1	11-15	O
12	-0.1354	13	1	11-15	O
13	-0.1260	11	2	11-15	O
14	-0.1355	15	1	11-15	O
15	-0.1355	14	1	11-15	O
16	-0.1355	15	1	16-16	X
17	-0.1368	19	2	17-17	X
18	-0.1362	18	0	18-18	O
19	-0.1436	22	3	19-24	O
20	-0.1361	17	3	19-25	X
21	-0.1378	20	1	19-25	O
22	-0.1378	21	1	19-25	O
23	-0.1640	24	1	19-25	O
24	-0.1459	23	1	19-25	O
25	-0.1885	25	0	19-25	O
26	-0.2021	26	0	25-26	O

3.3 Example

$$\delta = 0.84, \quad \lambda_2 = 0.0003, \quad \lambda_3 = 0.2410, \quad \Lambda = 926.77$$

true order	Z	predicted order	$\lvert diff \rvert$	C.I.	
1	0.2942	1	0	1-1	O
2	0.2942	1	1	2-2	X
3	0.2942	1	2	3-3	X
4	0.2942	1	3	4-5	X
5	0.2942	1	4	4-5	X
6	0.2942	1	5	6-7	X
7	0.2942	1	6	6-7	X
8	0.2942	1	7	8-11	X
9	-0.1307	9	0	8-11	O
10	-0.1307	9	1	8-13	O
11	-0.1307	9	2	11-15	X
12	-0.1307	9	3	11-15	X
13	-0.1307	9	4	11-15	X
14	-0.1307	9	5	11-15	X
15	-0.1307	9	6	11-15	X
16	-0.1307	9	7	16-16	X
17	-0.1307	9	8	17-17	X
18	-0.1307	9	9	18-18	X
19	-0.1307	9	10	19-24	X

3.3 Example



3.3 Example

δ	λ_2	λ_3	$\Lambda = \lambda_3/\lambda_2$	NCL	NCLI	$\lvert diff \rvert$
0.0	12.16	14.72	1.21	3	5	280
0.1	12.16	14.72	1.21	3	5	280
0.2	11.97	14.71	1.23	3	5	280
0.3	11.63	14.71	1.26	3	5	280
0.4	11.28	14.71	1.30	3	5	280
0.5	5.14	13.01	2.53	3	4	286
0.58	0.85	5.62	6.61	21	26	8
0.6	0.85	3.77	4.44	15	26	18
0.7	0.65	1.29	1.98	13	26	18
0.8	0.23	0.52	2.26	16	26	10
0.82	0.12	0.29	2.42	13	26	20
0.84	0.00026	0.24096	926.77	2	3	181

4. Extensions

4.1 Cryptology (Alphabet scramble)

catssitsit $\Longrightarrow$ statistics

$a_{ij} = P(i$th alphabet coming right before the jth alphabet$)$

$\mathbf{A} = (a_{ij})$: directed and weighted with
$\sum_{i=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ij} = 1, \ \forall i, j$

goal : Unscramble the scrambled word

Laplacian matrix : $\mathbf{I} - \mathbf{A}$

4.2 Protein Structure

base (A, T, G, C) $\Longrightarrow$ amino acid (20 types) $\Longrightarrow$ protein

goal : identification and/or classification of new protein