情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-BIO-22 No.4
2010/7/29

Genetic Network Similarity based on Alignment Score

Hitoshi AFUSO, Takeo OKAZAKI, Morikazu NAKAMURA[†1,†2,†2]

1. Introduction

Various researches about the method to estimate biological networks, such as gene regulatory networks and protein metabolism networks, bring us many important knowledge about life-forms. These estimation methods often had been evaluated its performance by using actual biological data sets. However, to evaluate the performance of methods from various points of view, we need diversified data and it is practically difficult to collect such data sets. To solve the problem, some researches that propose the method to generate artificial data have been done. Particular artificial data generation methods consist of two parts, generation of artificial genetic network and generation of artificial data from generated network. For the generation of artificial genetic network, Afuso *et al.*[1)] used network generation model such as Barabasi-Albert model[2)]. On the other hand, Bulcke *et al.*[3)] and Li *et al.*[4)] used the subnetwork sampling from actual genetic networks. In both cases, it is important whether the methods generated similar artificial network to actual one. So, we need certain measurement that confirms the similarity of generated artificial network to actual. In traditional research above, the authors used the network characters, such as average path length and degree distribution to confirm the similarity of generated network. However, there are some cases that network characters don't reflect the similarity, i.e., although two networks have nearly same value, these networks are not similar and vice versa. To solve that problem, we need the similarity measurement that reflects similarity between two networks more directly. On the other hand, as the similarity measure that is not based on network character, network alignment had been studied. But, this measure is influenced by network size. Then we cannot compare its values that calculated from different networks.

In this research, we proposed new similarity measurement for genetic network. And also, we showed the networks such that network character couldn't catch the similarity but proposal could.

†1 Information Engineering, Graduate School of Engineering and Science, University of the Ryukyus
†2 Faculty of Engineering, University of the Ryukyus

2. Network Alignment with Node Matching

Mohsen et al[5)] had been defined certain problem, called as *Network Alignment Problem*, to propose the network similarity not based on network characters. The problem was led by generalizing the well-known problem in graph theory, which is that *Maximum Common Subgraph Matching Problem*. We show the outline of the network alignment problem in this section. Consider two graphs $A = (V_A, E_A)$ and $B = (V_B, E_B)$ with vertex sets $V_A = \{1, 2, 3, \cdots, n\}$ and $V_V = \{1', 2', 3', \cdots, m'\}$. Let L be a complete bipartite graph between the vertices of A and B, formally $L = \{V_L = V_A \cup V_B, E_L\}$. A *matching* in L is a subset of edges of L. Every matching should satisfies the condition, and no two edges share a common endpoint. Let M be such a matching. For a matching M, we say that an edge $(i, i') \in E_L$ forms a *square* with another edge $(j, j') \in E_L$ if two conditions, edge $(i, j) \in E_A$ and edge $(i', j') \in E_B$ are satisfied. We also call such square as *overlap*. Fig.1 shows the problem setup and definition of squares. In the setup, the edges in E_L have nonnegative weight w, as shown in Fig.1. The weight w represent the similarity among the nodes of A and B. By using terms above, *Network Alignment Problem* is defined as to find the matching M such as maximizes both the sum of weights w of edges in M and the number of overlaps. We can represent this problem mathematically as follows:

Vol.2010-BIO-22 No.4
2010/7/29

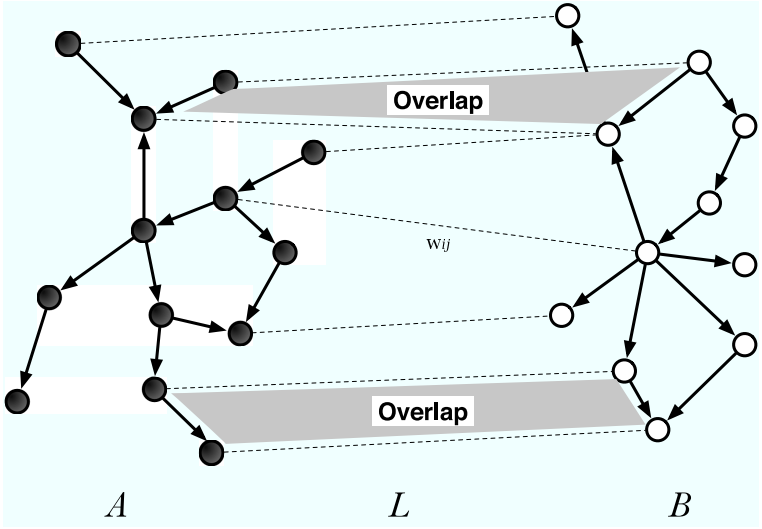

図 1 Setup of Network Alignment with Node Matching.

$$\text{maximize } \mathbf{X} \quad \alpha \mathbf{W} \cdot \mathbf{X} + \beta \mathbf{A}\mathbf{X}\mathbf{B}^T \cdot \mathbf{X} \tag{1}$$

$$\text{subject to} \quad \mathbf{X}\mathbf{1}_m \leq \mathbf{1}_n, \mathbf{X}^T\mathbf{1}_n \leq \mathbf{1}_m \tag{2}$$

In these formulae, $\mathbf{A}$ and $\mathbf{B}$ denote respective adjacency matrices of network A and B. $\mathbf{X}$ is a $|E_A| \times |E_B|$ zero-one matrix where $\mathbf{X}_{i,i'} = 1$ if and only if the edge $(i, i') \in L$ is included in M. $\mathbf{W}$ represent the weight matrix, then $\mathbf{W} = \{w_{i,i'}\}$. Operator "$\cdot$" denotes the operation of inner product among matrices. $\mathbf{1}_n$ denotes the column vector that dimension is n and all elements are 1. α and β are coefficients that represent the weights for similarity and overlaps, respectively. By solving this problem, we can obtain the network alignment score as the value of Formula(1) and also use the value as similarity measurement between given two networks, A and B.

3. Edge Importance and Edge Similarity

In the previous section, we used the similarity among nodes of networks A and B to obtain the similarity among them. However, we can consider the network structure as the connectivity among nodes. So, for the definition of the network similarity that reflects the network structure, it is convenient to use the similarity among edges, not nodes. To do so, we need to define certain measure for edges that indicates the contribution of each edge to the network structure. We call the measure as *Edge Importance*. We proposed two kinds of edge importance, that is, edge importance based on degree of end points, and, edge importance based on shortest path coverage. To apply the edge similarity to our definition of network similarity, we regard the network alignment problem as edge matching problem, not node matching.

3.1 Edge Importance based on Degrees of End Points

This importance is based on two ideas, *amplitude* and *dissimilarity of nodes*. When we think about the importance of edges, we can consider the edge that connects two nodes have high degree, such like **a** in Fig.2, as important edge for the network structure. *Amplitude* Amp_{ij} of edge (i, j) is the measure that reflect this idea and calculated by the formula as follows:

$$Amp_{ij} = |d_j + d_i| \tag{3}$$

In Formula(3), d_i denotes the normalized degree of node i.

And also, we can regard the edge that connects two nodes that have very different degree such like **b** in Fig.2. *Dissimilarity of Nodes* Dis_{ij} reflects this property and is calculated by the formula as follows:

$$Dis_{ij} = |d_j - d_i| \tag{4}$$

By using Formula(3) and (4), we defined the edge importance based on degree $EI^{(d)}_{ij}$ of edge (i, j) as follows:

$$EI^{(d)}_{ij} = \frac{Amp_{ij}}{1 + Dis_{ij}} \tag{5}$$

By using Formula(5), we can calculate the edge importance based on degree.

3.2 Edge Importance based on Degree of Shortest Path Coverage

We have another character of edges that can be used to define the importance. It is the degree of coverage by shortest paths. This importance is based on the idea that if the important edge is removed, then many pairs of connected nodes become unconnected. In Fig.3, the edge $(5, 6)$ is included in four paths, $1 \to 6, 2 \to 6, 3 \to 6, 3 \to 6$, and $4 \to 6$. On the other hand, edge $(7, 8)$ is not included in any other paths. So, if edge $(5, 6)$ is removed, greater number of corruptions of paths will be occurred than the

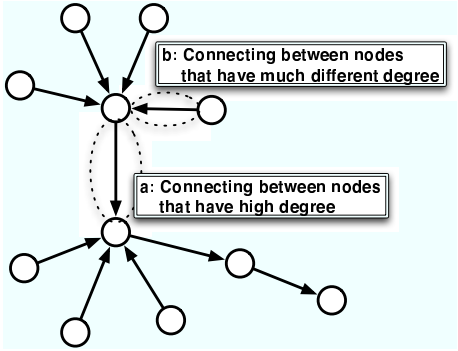

図 2 *Amplitude* and *Dissimilarity*

case of edge $(7, 8)$ is removed. From this consideration, we can consider the edge $(5, 6)$ is important for the network structure than edge $(7, 8)$. Using this idea, we defined the edge importance based on shortest path coverage $EI_{ij}^{(s)}$ of edge (i, j) as follows:

$$EI_{ij}^{(s)} = \frac{C_{ij}}{\sum_{kl} C_{kl}} \tag{6}$$

In this formula, C_{ij} denotes the number of paths that include edge (i, j) and the summation is over all edges in the network.

3.3 Edge Similarity

Using the Formula(5) and (6), we can calculate the edge importance in given one network. Next, we defined the edge similarity between edges that included in different networks, such like A and B in Fig.1. We defined the edge similarity $w_{ii',jj'}^{(e)}$ between edge (i, i') and (j, j') as the difference between edge importances.

$$w_{ii',jj'}^{(e)} = \frac{1}{1 + |EI_{ii'} - EI_{jj'}|} \tag{7}$$

In this formula, $EI_{ii'}$ represents edge importance, based on either degree or shortest coverage degree.

4. Edge Adjacency Penalty

Using the Formula(7), we can search the matching that maximize the sum of the edge similarity. However, simply searching of edge matching that maximize the sum of the

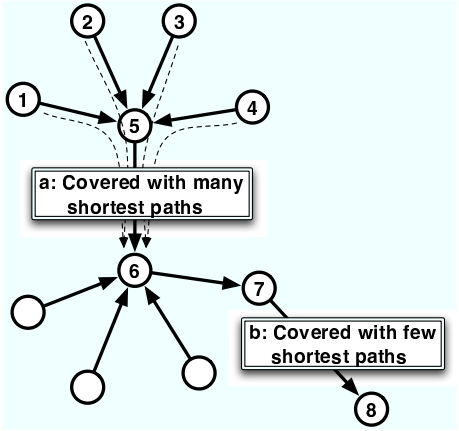

図 3 Shortest path coverage of edges

edge similarity, will lead inappropriate results. As shown in Fig.4, by considering only the maximization of sum of the edge similarity, obtained edge matching that violate the relationship of adjacency among edges. In Fig.4, edge $(3, 2)$ connects to edge $(2, 1)$, but their matched edges $(5, 4)$ and $(6, 5)$ are connected differently. This is the matter for the definition of network similarity that reflects the network structure. To handle this issue, we need to give the penalty to such edge matching that violates the edge adjacency. In general, as shown in Fig.5, there are four pattern of edge adjacency. We represent each pattern as A, B, C and D respectively. Then, to give the penalty, we counted the number of violations of edge adjacency for each combination of edge adjacency in given matching. Let $Mis_{A,B}$ be the number of violations that, in first network, edges are adjacent in pattern A but in second matched edges are adjacent in pattern B. Using this term, we can represent the penalty Pn for violation of edge adjacency mathematically, as follows:

$$Pn_M = \sum_{k \in \{A,B,C,D\}} \sum_{l \in \{A,B,C,D\}} Mis_{k,l} \tag{8}$$

By using Formula(8) as penalty term, we can search the edge matching such that preserve the network structure.

Vol.2010-BIO-22 No.4
2010/7/29

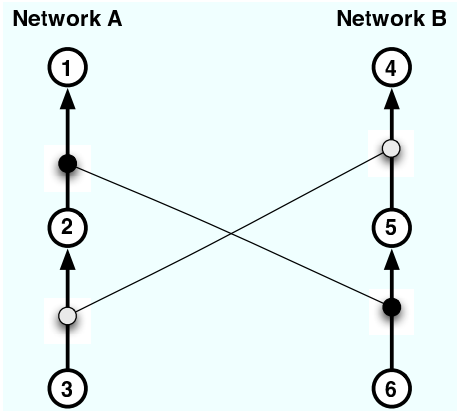

図 4 Violation of Edge Adjacencent Pattern

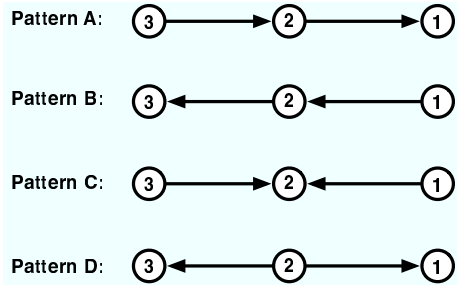

図 5 Patterns of Adjacency of Edges

5. Formulation as Network Alignment with Edge Matching

By using edge similarity and penalty for violation of edge adjacency, we can formulate our edge matching problem mathematically as follows:

$$\text{maximize}_{\mathbf{X}} \quad \mathbf{W} \cdot \mathbf{X} - \alpha P n_M \tag{9}$$

$$\text{subject to} \quad \mathbf{X}\mathbf{1}_m \leq \mathbf{1}_n, \mathbf{X}^T\mathbf{1}_n \leq \mathbf{1}_m \tag{10}$$

In these formulae, $\mathbf{W}$ denotes edge similarity matrix. By solving this problem, we can obtain the optimal edge matching between given two networks. We use the value of Formula(9) corresponding optimal $\mathbf{X}$ as network alignment score.

6. Network Similarity based on Network Alignment

By solving problem above, we can get the edge matching between two networks and also lead the value of Formula(9) that denotes network alignment score. However, this value is influenced by the size of networks that we are focusing. So, we cannot simply compare their values. To this end, we proposed the network similarity $Sim(B|A)$ that can be compared in different pairs of networks, by using the result of network alignment as follows:

$$Sim(B|A) = \frac{\sum_{ij \in M} EI_{ij \in A}}{\sum_{kl} EI_{kl}} \tag{11}$$

In this formula, $Sim(B|A)$ represents the network similarity between given network A and target network B. EI denotes the edge importance. As one can notice in Formula(11), our definition of network similarity is asymmetric.

7. Experiments

To glance the property of our proposal network similarity, we implemented two experiments.

7.1 Comparison to Network Character as Similarity Measure

First, we tried to show the feature of our proposal by comparing with traditional network character. We prepared three networks in advance and calculated its network characters. As network characters to be compared, we used the average path length and average degree, according to traditional research[3),4)]. Prepared networks are shown

表 1 Network Character Values of Given Networks

	Ave. Path Length	Ave. Degree
Network0	1.9387	0.95
Network1	2.5454	0.94
Network2	1.9302	0.95

表 2 Results with Edge Importance based on Degree

	AlignmentScore	NetworkSimilarity
Network1	11.7292	0.5642
Network2	10.4993	0.7052

表 3 Results with Edge Importance based on Degree of Shortest Path Coverage

	AlignmentScore	NetworkSimilarity
Network1	13.1286	0.6866
Network2	11.6479	0.4842

in Fig.6 and in Table.1, their network characters are shown. And next, we searched the optimal network alignment and calculated the network similarity in both cases of edge similarities that we proposed above. Obtained network alignment was shown in Fig.7, and calculated value of network similarities were shown in Table.2 and Table.3, respectively. According to the value of network characters, network0 and network2 are similar and we may regard network0 and netowork1 as not similar each other. However, from the result of network alignment in Fig.7, we can see that three subnetworks in network0 are included in network1. And from the result shown in the Table.3, edge importance based on the degree of shortest path coverage can capture such similarity than traditional network character and edge importance based on the simple degree. From these results, we found that our proposal similarity can capture the similarity corresponding to local structure of network that may be missed by traditional network character.

7.2 Penalty Weight

We tried to take a look the response of the proposal against the change of penalty weight α for each edge importance. In this experiment, we calculated the similarities between one target network and three networks. The target network had 20 nodes and the compared networks had 15, 20 and 30, respectively. The networks that we

表 4 Results from the edge importance based on degree.

	network1	network2	network3
$\alpha = 0.1$	0.5660	0.8188	0.9351
$\alpha = 0.5$	0.6774	0.7068	0.8066
$\alpha = 1.0$	0.5638	0.5196	0.6829
$\alpha = 1.5$	0.6191	0.6485	0.6419
$\alpha = 10.0$	0.5865	0.5605	0.7417

表 5 Results from the edge importance based on shortest path coverage.

	network1	network2	network3
$\alpha = 0.1$	0.7560	0.8170	0.8902
$\alpha = 0.5$	0.6158	0.5305	0.8359
$\alpha = 1.0$	0.6001	0.6404	0.7116
$\alpha = 1.5$	0.6514	0.6404	0.8072
$\alpha = 10.0$	0.3902	0.3902	0.7560

used in the experiment were shown in Fig.8. The calculations of similarity were done in three cases, $\alpha = 0.1, 1.0, 10.0$. The optimal network alignments were searched by using genetic algorithm. In the configuration of genetic algorithm, we represented the matrix $\mathbf{X}$ as simply vector by concatenation of each row of the matrix. And as selection method, roulette and 10% elite preservation was used. We mixed the selected chromosomes by the uniform mixture method. The number of chromosomes was 100, and the number of generations was 2000. As the mutation, we selected a chromosome randomly and executed the deletion or addition of edges into X. In the calculation, we used both types of edge importance. The calculation results were shown in Table.4 and Table.5. From these results, our proposal similarity measurement have few influence from the change of the penalty weight. However, in the case of smaller α, the obtained edge matching didn't preserved the structure of edge adjacency. In other words, many matchings that violate the edge adjacency had been led. In opposite, the obtained edge matching was very strict for local structure in the case of $\alpha = 10$. This result can be simply seen as common subgraph between networks, and it is not enough to capture the similarity among networks. So, it is important to obtain the valid penalty weight α for the appropriate similarity.

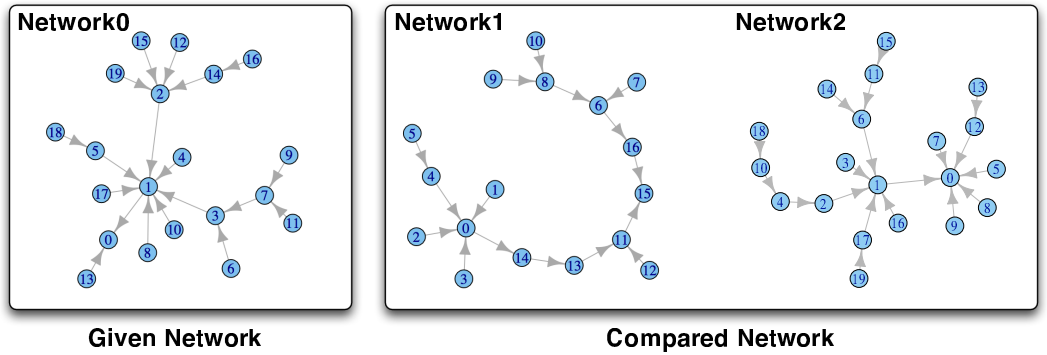

図 6 Networks to be compared each other in the first experiment

8. Conclusion

We proposed the network similarity based on network alignment for the assessment of network generation method. And also we showed the case that our proposal could capture certain similarity that traditional network character may miss. To show the validity of our definition of each edge importance, edge similarity and network similarity, we need more biological discussion, such as the measure or metric of importance of biological relationships between actual genes.

参 考 文 献

1) H.AFUSO, T.OKAZAKI, "PageRank based Score Function for Orientation to Genetic Causal Network", Proceedings of International Technical Conference on Circuits/Systems, Computers and Communications , pp.364-367, 2009
2) A-L.BARABASI, "Linked: The New Science of Networks", American Journal of Physics, Volume 71, Issue 4, pp.409-410, 2003
3) T.V.den BULCKE, K.V.LEEMPUT, et al, "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms", BMC Bioinformatics **7**:43, 2006
4) Y.LI, Y.ZHU, et at, "ReTRN: A retriever of real transcriptional regulatory network and expression data for evaluating structure learning algorithm", Genomics Volume 94, Issue 5, pp.349-354, 2009
5) M.BAYATI, M.GERRITSEN, D.F.GLEICH, et at, "Algorithms for Large, Sparse Network Alignment Problems", Nineth IEEE International Conference on Data Mining, pp.705-710, 2009

Vol.2010-BIO-22 No.4
2010/7/29

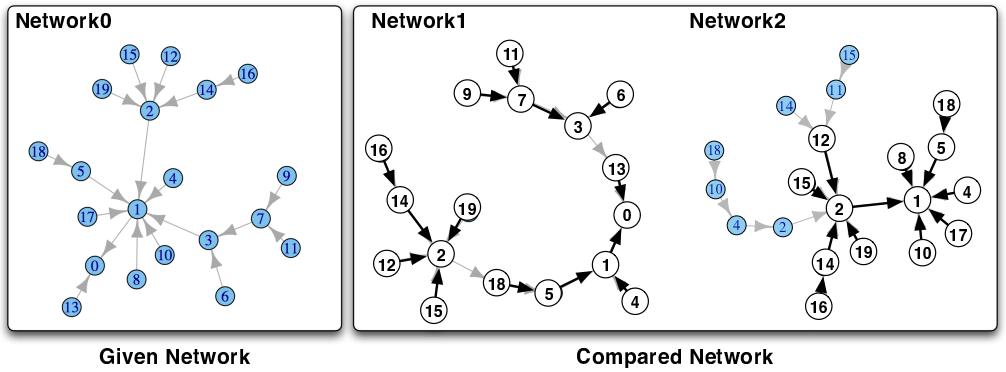

図 7 Obtained Result of Network Alignment

Vol.2010-BIO-22 No.4
2010/7/29

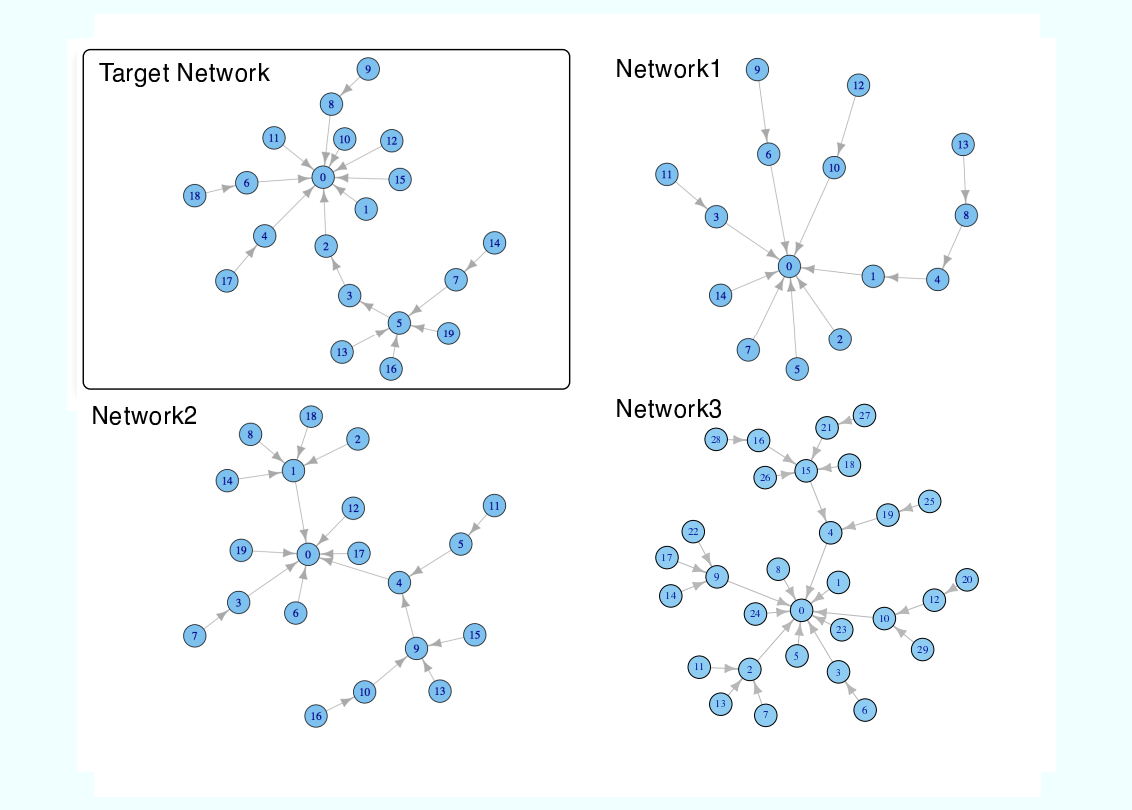

図 8 Networks that used in the experiment.