Abstract

In this paper a complex of algorithms for a factorization of a directed graph, describing a protein network, is constructed. It consists of a sequential extraction of clusters in the graph, a construction of a partial order between clusters and further a construction of acyclic
factor-graph with an ordering of its nodes (clusters) by a principle of a maximal distance from input nodes. Special algorithms are devoted to make a hierarchical classification of cluster nodes and a definition of cluster centres. This complex is applied to a study of key players in the protein network required for connecting ABA signaling and ABA-mediated drought and thermotolerance.

**Mathematics Subject Classification:** 05C80

**Keywords:** a cluster, a partial order, a hierarchical classification, a factorization, a protein network

### Introduction

Besides of free scale networks [1], [2], consisting of a large number of nodes, biologists interest in networks with a sufficiently small number of nodes. These networks are interesting for a visualization of an information about an interaction of proteins in these networks [3]. A characteristic specifics of these representations is a large attention to feedbacks. But detailed investigation of feedbacks needs their formalisation in terms of the graph theory.

Using an information about a protein network it is possible to represent it in a form of a directed graph and to make its factorization by a relation of a cyclic equivalence and to transform the graph into acyclic directed graph. This procedure allows to factorize the network taking into account their feedbacks.

One of main elements of such analysis is a definition of input nodes in the factor-graph and a ranking of other nodes by maximal lengths (numbers of edges) from input nodes to them. After this ranking edges may direct only from nodes with smaller ranks to nodes with larger nodes.

Special interest is in a hierarchical classification of nodes in a graph with cyclically equivalent nodes (clusters) and a definition of their centres.

This complex of algorithms is applied to an analysis of protein network usually used for and investigation of plants stability to droughts and extremal temperatures. Using these algorithms in the network two clusters with sufficiently large numbers of nodes are extracted and their centres are defined.

### 1 Factorization of directed graph by cyclic equivalence

Consider directed graph $G$ with finite number of nodes $V = \{1, ..., n\}$ and with edges set $W$. Define this graph by a contiguity matrix $||d_{ij}||_{i,j=1}^n$, with
Each cluster nodes set. Each node of the sets a⪰ characterizing a partial order relation is a cycle containing these nodes [4]. Divide the set V into classes of cyclic equivalence (clusters) and denote [V] the set of these clusters. On the set [V] define the following binary relation: [v₁ ⪰ v₂], [v₁], [v₂] ∈ [V], if there is a way from a node of the class [v₁] to a node of the class [v₂]. It is obvious that this binary relation is reflexive, transitive and antisymmetric and so is a relation of a partial order [5, § 4].

**Sequential algorithm of clusters and partial order construction.** Following [6] put that on the step 1 there is a single node 1, creating a cluster [1], and a clusters set K = {[1]}. Introduce a matrix a = ||a([p],[q])||ₜₜ=[p],[q]∈K, characterizing a partial order relation ⪰: a([p],[q]) = 1, if [p] ⪰ [q], else a([p],[q]) = 0. On the step 1 this matrix is defined by the equality a([1],[1]) = 1.

Assume that on the step t − 1 the clusters set K defines a dividing of the nodes set {1,...,t − 1} into non intersected sub sets and we have the matrix a. Each cluster [k] ∈ K is indexed by maximal number k of its nodes in the set V. On the step t define following sets

\[ P = \{k ∈ \{1,...,t − 1\} : d_{ik} = 1\}, \quad Q = \{k ∈ \{1,...,t − 1\} : d_{kt} = 1\}. \]

Each node of the sets P, Q contains in some cluster. Denote [P], [Q] sets of these clusters and define

\[ K_P = \bigcup_{[p] ∈ [P]} \{[k] ∈ K : a([p],[k]) = 1\}, \quad K_Q = \bigcup_{[q] ∈ [Q]} \{[k] ∈ K : a([k],[q]) = 1\}, \]

\[ A = K_P \cap K_Q, \quad A_1 = K_P \setminus A, \quad A_2 = K_Q \setminus A, \quad B = K \setminus (A \cup A_1 \cup A_2). \]

New node t and nodes from the set A create new cluster [t] and the set of clusters K is transformed as follows:

\[ [t] := \{t\} ∪ A, \quad K := (K \setminus A) ∪ \{[t]\}. \]

A recalculation of the partial order matrix a on renewed set of clusters K is following:

\[ a([t],[i]) := 1, \quad [i] ∈ A_1 \cup \{[t]\}, \quad a([i],[j]) := 1, \quad [i] ∈ A_2, \quad [j] ∈ A_1 \cup \{[t]\}, \]

\[ a([i],[j]) := 0, \quad [i] ∈ A_1, \quad [j] ∈ A_2 ∪ \{[t]\} \cup B, \quad a([i],[j]) := 0, \quad [j] ∈ A_2, \quad [i] ∈ B \cup \{[t]\}, \]

\[ a([t],[i]) = a([i],[t]) := 0, \quad [i] ∈ B. \]

All other meanings of the matrix a coincide with previous ones defined on the step t − 1. So the matrix a on the step t has the following cell structure (see Table 1). It was shown that this algorithm is much more fast than calculation of max-min product of contiguity matrices for graphs with large number of nodes (few thousands).
<table>
<thead>
<tr>
<th>matrix $a$</th>
<th>clusters of set $A_1$</th>
<th>cluster $[t]$ meanings on step $t-1$</th>
<th>0</th>
<th>clusters of set $A_2$</th>
<th>clusters of set $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>clusters of set $A_1$</td>
<td>meanings on step $t-1$</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cluster $[t]$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clusters of set $A_2$</td>
<td>1</td>
<td>meanings on step $t-1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| clusters of set $B$ | meanings on step $t-1$ | 0 | meanings on step $t-1$

Table 1. Transformation of partial order matrix $a$ on step $t$.

Rectangular sub matrices in Table 1 denoted by 1 (denoted by 0) consist of unit elements 1 (consist of zero elements).

**Algorithm of factor graph construction.**
Construct from the graph $G$ with extracted clusters $[i]$ and with partial order $\succeq$ matrix $a$ a factor-graph in which nodes are clusters in $G$. Between clusters $[i]$, $[j]$, $[i] \succeq [j]$ we introduce directed edge if and only if there are nodes $k \in [i]$, $t \in [j]$, connected in the graph $G$ by edge $(k, t) \in W$. It is obvious that such factor-graph $[G]$ with the nodes set $[V] = \{[i] : i \in V\}$ and the edges set $[W]$ is acyclic.

## 2 Ranking of nodes in acyclic directed graph

Construct an analogy of the Floyd-Worshall algorithm to calculate a matrix of maximal lengths for ways between nodes of acyclic directed graph $T$. Denote $D^k = ||d^k_{i,j}||_{i,j=1}^N$, $k = 1, \ldots, N$ in which $d^k_{i,j}$ is maximal length of ways between the nodes $i$, $j$ passing through intermediary nodes $1, \ldots, k$ if such ways exist. But if such ways do not exist then $d^k_{i,j} = \infty$. It is not difficult to prove the following statement.

**Theorem 2.1** Matrices $D^k = ||d^k_{i,j}||_{i,j=1}^N$, $k = 1, \ldots, N$ satisfy recurrent relations

$$d^k_{i,j} = \max(d^k_{i,j}, d^k_{i,k} + d^k_{k,j}) \text{ if } \max(d^k_{i,j}, d^k_{i,k} + d^k_{k,j}) < \infty, \quad (1)$$

$$\text{else } d^k_{i,j} = \min(d^k_{i,j}, d^k_{i,k} + d^k_{k,j}). \quad (2)$$

The matrix $D^N = ||d^N_{i,j}||_{i,j=1}^N$ define maximal lengths of ways between the graph $T$ nodes if such ways exist. In a case of an absence of such ways corresponding matrix elements equal the infinity.
Factorization of directed graph describing protein network

Assume that in the graph $T$ we define input nodes in which directed edges do not enter. Add the graph by a fictive node 0 and directed edges from 0 to input nodes. Using Theorem 2.1 it is possible to calculate maximal ways lengths from the node 0 to the nodes $i = 1, \ldots, N$ : $l_i = d^N_{0,i}$. In such a way it is possible to define ranks $l_i$ of nodes $i$ in acyclic graph so that edges in the graph $T$ direct only from nodes with smaller ranks to nodes with larger ranks [8].

3 Hierarchical classification and definition of centres in cluster

Consider a cluster - directed graph $H$ with the set $S$ of cyclically equivalent nodes. Denote $R = ||r_{i,j}||_{i,j \in S}$ the matrix of minimal lengths of ways between cluster nodes. Such matrix may be defined using the Floid-Worshall algorithm.

Definition of cluster centres.
For any node $i$ calculate $M_i = \max_{j \in S} r_{i,j}$ and determine $M = \min_i M_i$. Nodes $k : M_k = M$ are defined as centres of the cluster. This definition is analogous to concepts of a mean or a median for a distribution function of a random variable.

Hierarchical classification of cluster nodes.
Transform the matrix $R$ into the matrix $F = ||f_{i,j}||$, $f_{i,j} = r_{i,j} + r_{j,j}$ of minimal lengths of cycles containing nodes $i, j$. Further fix some critical level $\alpha$ and calculate symmetric zero-one matrix $L^\alpha = ||l^\alpha_{i,j}||$, where $l^\alpha_{i,j} = 1$, if $f_{i,j} \leq \alpha$, else $l^\alpha_{i,j} = 0$. Construct non directed graph $H^\alpha$ with a contiguity matrix $L^\alpha$. In the graph $H^\alpha$ define connectivity components which create classes at level $\alpha$. Sequentially increasing $\alpha$, we obtain a merger of these classes in more large ones [9].

4 Example

Factorize a protein network using for a study of the key players required for connecting ABA signaling and ABA-mediated drought and thermotolerance (see Figure 1). In this network we define one node clusters: input, output and intermediary and two multi node intermediary clusters (encircled by blue and red colour curves).
Output proteins DREB2C, ABA receptors PYLs, which are the most important for thermostability of plants by their biochemical characteristics. Multi node cluster encircled by red colour curve does not have edges connected with these proteins.

The most interesting is multi node intermediary cluster encircled by blue colour curve. It has edges connected with output nodes DREB2C, ABA receptors PYLs. In this multi node cluster there are the following centres ABF4, ABF2/AREB1, PP2CA protein phosphatases, which are the most important for further biochemical tests. This statement is confirmed by biochemists also.

Hierarchical classification shows that centres ABF4, ABF2/AREB1, PP2CA protein phosphatases, together with ABF1, DREB2A, SNRK2.2, ABI5, ABF3, KEG create the largest class in the hierarchical classification of the cluster encircled by the blue colour curve.

5 Conclusion

Numerical experiment and discussion with biochemists showed that it is not enough to construct single, may be convenient algorithm to analyse sufficiently complicated biological network. It is necessary to use a complex of algorithms for an obtaining results and for a checking them from different viewpoints. In this paper we used as algorithms of a directed graph factorization so algorithms of clusters analysis by means of their hierarchical classification and a definition of their centres.
Factorization of directed graph describing protein network

References


Received: July 7, 2017; Published: July 20, 2017