2018

# Associative Learning in Biochemical Networks

Yasmin S. Sepulveda
*Portland State University*

# Associative Learning in Biochemical Networks

Yasmin S. Sepulveda

**Prepared for Christof Teuscher**

**Table of Contents**

**Abstract**

Emerging evidence suggests that biochemical networks can be modeled by exploiting their ability to learn through associative learning. This type of learning in biomolecular structures gives it a the advantage to be able to be computationally model, and condition. Associative learning in biochemical networks is a developing area of study that once understood, can further develop diagnostic applications, and be used as tools for data analysis. Although it is a open ended project the motive of this research was to find the the best method of association learning being used in current work. After reading current work three associative learning methods were presented, such models were then analyzed to see how they implemented the associative learning methods . After understanding the models, they will be evaluated by both their challenges, and highlights. Due to time constraints trails of such associative learning methods will not be performed, but the conclusion will be based off the readings and a outline of the intended evaluation of models will be included.

**Introduction**

A goal within in computational modeling of biochemical structures is not to enhance their realism particularly, but model behaviors that can lead to scientific advancement. Chemical processes have been found to display, or have the ability to learn through association learning. This learning can be modeled computationally which allows more efficient, and rapid analysis of biochemical structures, and processes that would otherwise be both monetary, and time expensive through wet lab procedures. Associative learning has become interpretable through

association learning algorithms that are presented as methods for interpreting biochemical network process behavior by quantifying the association between things such as molecules, or chemical reactions. Such quantifications allow the program to perform natural biochemical process actions numerically in a computational manner that is unnatural for such processes.

**1.1 Associative Learning and its Behavior.**

The principle of associative learning is that our experiences and our ideas reinforce each other because they are mentally linked to each other. This demonstrates that our brain recalls information by key words meaning it groups the information as a associative memory, and calls out the associative memory to access the stored information. The associative learning theory states that behavior can be both learned and modified by the relationship of a stimulus and a response, and such learning may also be forgotten. Whether something has been learned or forgotten can be measured by the response it generates, the lack of a response demonstrates it has been forgotten while the learning is shown in the response.

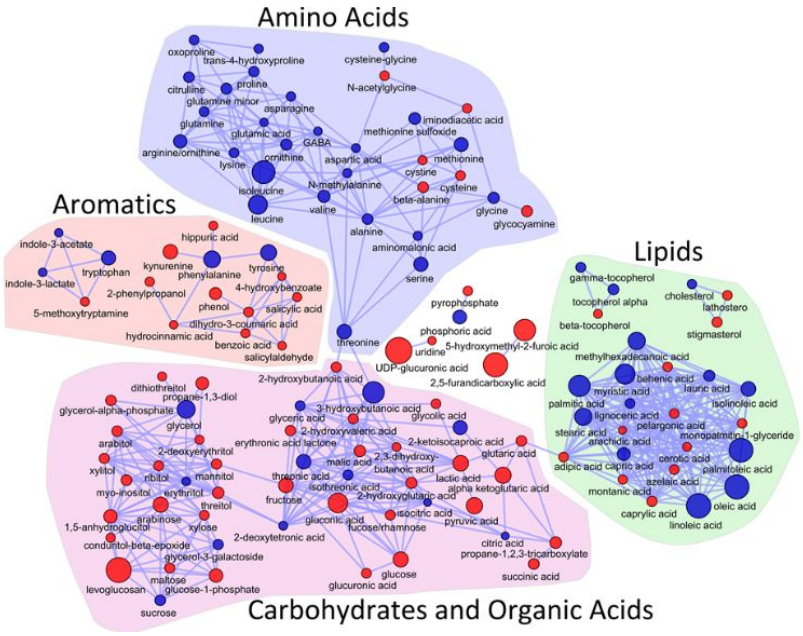
To be more specific, associative learning is a learning process that works by repeatedly introducing paired stimuli to your subject that causes it to have a response. Once the linkage from the stimuli to the memory is created the response will be generated by the first stimulus alone. This can be seen in Pavlov's classical conditioning experiment which involved a dog, a tuning fork, and food. He repeatedly presented the paired stimuli which was the ringing of the tuning fork along with the food; we can label the sound of the tuning fork as the first stimulus and the food as the second stimulus. The dogs love for food caused it to salivate which was the response to the stimuli. Once the dog created the memory of getting fed every time the tuning

fork rang, the response was then induced by the bell ringing alone. The association of the tuning fork with the food resulted in learning a behaviour, hence Associative learning.

### 1.2 What Are Biochemical Networks?

Biochemical networks display how molecules interact with a abstract graph.

**Figure 1**




The abstract graphs use nodes to represent molecules and the connections displayed represent a physical or functional interaction between two molecules. This type of modeling helps make predictions about cellular functions based off of their interactions with other molecules that would not be able to be obtained from studying a individual cell.

### 1.3 Associative Learning in Biochemical Networks.

The ability to learn through associative learning has been expressed in different biochemical networks. Such networks include the immune system, the immune system demonstrates that it has a associative memory that can be drawn out by a stimulus. “After the immune system fights off an infection, a fraction of the antibody-producing B-cells turns into memory cells, that allow the immune system to rapidly respond to future invasions by a given antigen.”[9] In this example the antigen represents the stimulus or experience that is mentally linked to the memory, which is the memory cell that evolved from the B-cell. The network shows that it learned to recognize the antigen because it executes a fighting response when the antigen enters its environment.

**1.4 Associative Learning and Biochemical Network Modeling.**

The ability to quantify the associations allows computational interpretation of biochemical networks, or models. Exploiting the biochemical processes with in biological networks that express associative learning can further evolve areas such as diagnostics, pharmaceuticals and data analysis for scientific advancements. The algorithms are used to mimic the steps within processes (put example of substrate production, show how each step is translated with algorithm, and how it shows the flow, hence: the model of how it works**).** Within today's work there are three approaches that are used for modeling biochemical networks, Hebbian approach, Hopfield approach, and the Bayesian approach.

Hebbian learning is method based on the mechanics of biological systems and is also known also as correlation rule based learning. Hebbian learning follows the claim that vast majority of learning is established from both the increase and decrease of likelihood from neuron to neuron activation (neuron meaning nerve cell).

Hopfield networks are “content-addressable (associative memory) systems with binary threshold nodes.” [S1]. This means that the memory (node) is activated by a binary input (binary input being one or zero) that is meets threshold activation requirement, which is where hebbian manner is followed to learn the binary patterns. Hopfield networks are known for their recurrent network, meaning the nodes connect within each other instead of connecting to the next layer, which makes continuous networks useful for optimization problems.

### 1.5 Association Rules

When using Association rule generating algorithms they will generate all the rules they were designed to generate, in order to select meaningful associations from the set of all possible association rules constraints are set in place.  Such constraints are used as a measurement , the most well known association rule constraints are set to satisfy a specified minimum support or confidence, that is user- specified and varies depending on the user’s objective. They are implemented within code to specify what is of value to the algorithm and what results to display. If a association value does not meet the minimum support or confidence, it is dropped and not considered. The constraints can be defined as follows.

The support  measures how often a item, and itemset appears within a dataset. A mathematical representation is shown where the support of **X** is in respect to **T** (all elements within dataset of choice) is the proportion of the item within the item set.

**Figure 2**

$$\operatorname{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

A great example of support is presented in "Introduction to arules – A computational environment for mining association rules and frequent item sets" by Michael Hahsler where "The itemset **X= {beer,diapers}**, has a support of **⅛ = 0.2** since it occurs in 20% of all transactions (1 out of 5 transactions). The argument of **supp()** is a set of preconditions, and thus becomes more restrictive as it grows (instead of more inclusive)."

Next we have the confidence, which is the value of how often a rule is found to be true within element sets, and can be defined as,

**Figure 3**

$$\operatorname{conf}(X \Rightarrow Y) = \operatorname{supp}(X \cup Y) / \operatorname{supp}(X)$$

Where the confidence value of a association rule **( X → Y)** is measured with respect to **T** which would be the sets of elements. It then evaluates the proportion of itemsets that include both **X** and **Y.**

Other constraints that are used is lift, and conviction. There are also other constraints that are considered, such as lift, and conviction.

Lift is the ratio of the support expected if X and Y were independent.

If the association rule is found to have a lift of 1, it means that no rule can be drawn because the antecedent and consequent are independent of each other. If the lift is greater than one it shows

that that the antecedent and consequent are dependent of each other, and if the lift is less than one, it means that the presence of  antecedent and consequent presence have a negative effect on each other, which can be referred to as the elements being substitute to each other.

On the other hand, conviction can be seen as the frequency of how often a rule makes a wrong prediction, which would be the ratio of the amount of times that **X** occurs without **Y.**

**Figure 4**

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$$

## 2.1 Associative Learning Approaches

**Hebbian Approach**

Molecular circuits for associative learning in single-celled organisms" which was published in the Journal of the Royal Society Interface propocess a gene regulatory network capable of associative learning between a set of chemical signals represented as transcription factors in a Hebbian manner within individual cells. They developed a mathematical model that  that was supported by the simulation results that showed it had  learned a response.  The mathematical model of the gene regulatory network shows that it was optimized by implementing a Hebbian approach. The equations that govern the circuit, and the Generalized Hebbian Algorithm can be seen in figure 1. 1

**Figure 5**

| **Generalized Hebbian Algorithm** |
|---|

$$\Delta w_{ij} = \eta \left( y_i x_j - y_i \sum_{k=1}^{i} w_{kj} y_k \right)$$

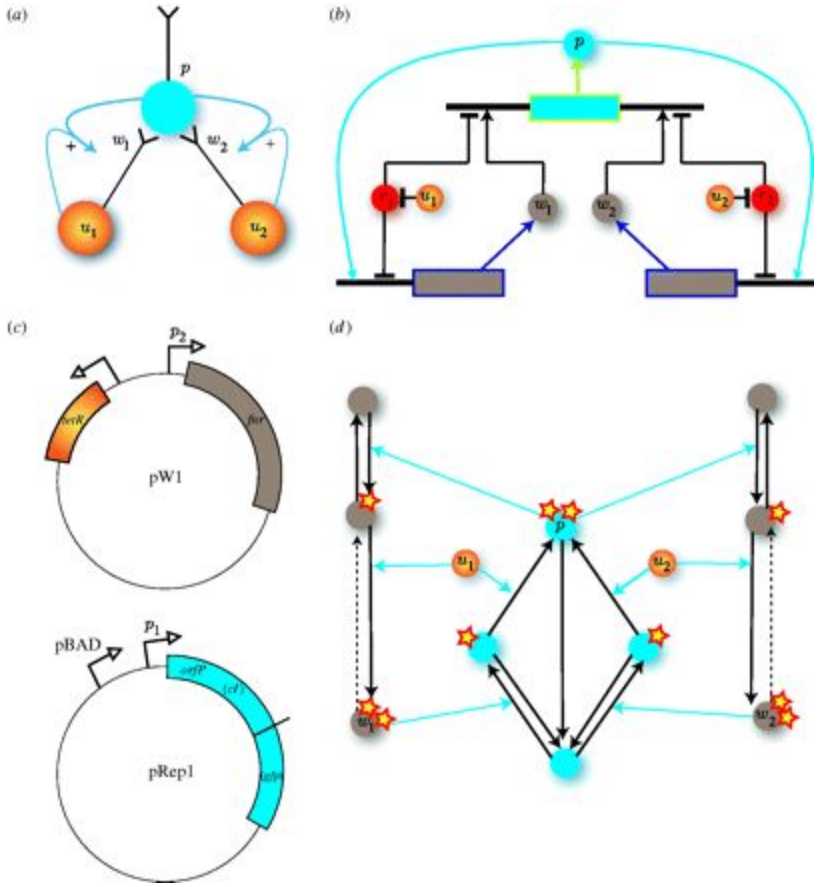
**Equations That Govern The Circuit**

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \sum_{j=1}^{N} \left[ v_p \left( \frac{w_j^4}{K_w^4 + w_j^4} \right) \left( 1 - \frac{r_j^2}{K_r^2 + r_j^2} \right) \right] - \delta_p p, \tag{(2.1)}$$

$$\frac{\mathrm{d}w_j}{\mathrm{d}t} = v_w \left( \frac{p^2}{K_p^2 + p^2} \right) \left( 1 - \frac{r_j^2}{K_r^2 + r_j^2} \right) - \delta_w w_j + \varepsilon_1. \tag{(2.2)}$$

## Hebbian Learning Implementation

## Figure 6



"The neural network implementation of Hebbian learning for two inputs $u1$ and $u2$. The orange circles represent pre-synaptic neurons that project onto a single post-synaptic neuron (blue). The simultaneous firing of the input neurons causes the synaptic weights $w1$ and $w2$ to increase, reinforcing their association. The blue curved lines show how this Hebbian positive feedback works, e.g. the weight $w1$ increases as a product of the output firing rate $p$ and the input firing rate $u1$." [2]

| | |
|---|---|
| Orange Circles | Inputs u1 and u2. |
| Pre-synaptic Neuron | Axon terminal of nerve cell. |
| Post-synaptic Neuron | Cell body or dendrite of nerve cell. |
| Synaptic Weights | Strength of the amplitude in between nodes. |

Although the hebbian approach cannot account the input-output characteristic for the overall system, they implemented its rule based learning for local performance. [2]Its local performance means that it changes the weight according to immediate activations surrounding the weight, this is beneficial because it eliminates the problem of high interdependence that would occur if it considered the rest of the weights. Its limitation to immediate activations allows more clarity of to which weights are of greater importance.

**Bayesian Approach**

Other work such as "Evolution of Associative Learning in Chemical Networks" [10] demonstrates that chemical networks can take on certain area of associative learning tasks with only one core chemical reaction. They use Bayesian methods to analyze the parts of the network that are in charge of learning (they tested out a linear combination of chemical concentrations.)

and the bayesian analysis showed that it had “memory traces” of the chemical network. They are implying that there is minimal reason to conclude that unsuitable phenotypic variation can obstruct associative learning from growing in single, or a mixture of cell networks.

**Bayesian Implementation**

They use probability distribution to modify the responses according to the evidence that can be arranged systematically through code. This is known as Bayesian inference where the initial observations are used as a stepping stone because the initial observation becomes the prior probability that influences the next observation. The Bayesian method is then repeatedly applied to increase the evidence in a rational manner in order to model the environment.[10]

**Hopfield**

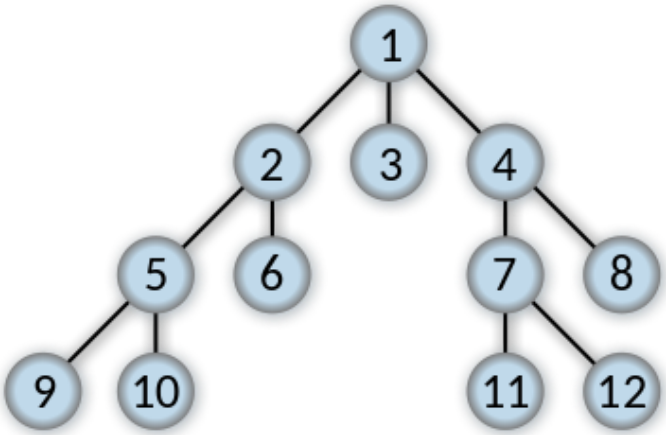
Within Hopfield method work [6] of transforming circuits into DNA strand displacement cascades implements hopfield associative memory. Within this work they transform a artificial neural network model into displacement cascades that pertain to DNA by using DNA gate architecture that allows a multilayer digital circuit. They find the approach to be versatile allowing it to be implemented into Hopfield associative memory. After conditioning, it remembers four single-stranded DNA patterns and recognizes incomplete patterns. They suggest that DNA strand displacement cascades could be used in autonomous chemical systems because they have the ability of recognizing patterns during molecular events, and also it can interpret the environment and use it to make its decisions

**3.1 Results- Addressing Cell Classification By Implementing Associative Rule Algorithm.**

Two data sets that were presented in a proof of concept study in 1999 by Golub et al. were used to implement our own proof of concept for the A priori algorithm. One data set that contained 38 samples of measurements corresponding to Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), and the second included 34 independent test samples. The objective of using this dataset was to generate the most frequent occurring pair of gene expressions, one of which was the cancer gene expression. By successfully generating such pairs we would obtain proof of concept that the A priori algorithm can be implemented within gene expression data and can generate meaningful items sets that can be used for cell classification.

In order to generate the gene expression pairs the apriori algorithm was chosen due to its breadth first search (BFS), this type of search has prefered over depth first search (DFS) because BFS searches in the order the elements expand instead of composing the longest patterns first before backtracking. This benefited us because we were only looking for a small set that consisted of two gene expressions, so its characteristic that made it more efficient would resulted as a fault to us because it's longer process is more time expensive.

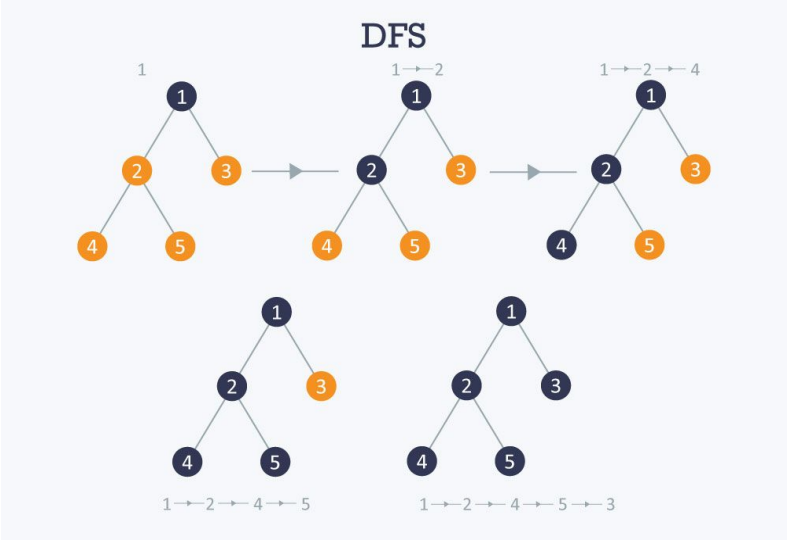
**Figure 7**




Source: https://en.wikipedia.org/wiki/Breadth-first_search

It illustrates how it searches for the pairs according to how they expand by using numbers to follow that illustrate the order.

**Figure 8**




Source; https://www.hackerearth.com/practice/algorithms/graphs/depth-first-search/tutorial/

It illustrates how it searches for the sets by finding the longest pattern first, using numbers to

follow that illustrate the order.After the algorithm is then implemented in python (a computer language) it generates the itemsets and their support, where the support represents how often it was found within the data set together.

| support | itemsets | support | itemsets | Difference |
|---|---|---|---|---|
| 1 | [4051, 'AML'] | 0.1538462 | [4051, 'ALL'] | 0.846154 |
| 0 | [2353, 'AML'] | 0.8461538 | [2353, 'ALL'] | 0.846154 |
| 0 | [4327, 'AML'] | 0.8461538 | [4327, 'ALL'] | 0.846154 |
| 0.818182 | [3319, 'AML'] | 0 | [3319, 'ALL'] | 0.818182 |
| 0.909091 | [6200, 'AML'] | 0.1153846 | [6200, 'ALL'] | 0.793706 |
| 0 | [5771, 'AML'] | 0.7692308 | [5771, 'ALL'] | 0.769231 |
| 0.818182 | [5898, 'AML'] | 0.0769231 | [5898, 'ALL'] | 0.741259 |
| 1 | [6796, 'AML'] | 0.2692308 | [6796, 'ALL'] | 0.730769 |
| 0.272727 | [1927, 'AML'] | 1 | [1927, 'ALL'] | 0.727273 |
| 0.727273 | [2401, 'AML'] | 0 | [2401, 'ALL'] | 0.727273 |
| 0.727273 | [4846, 'AML'] | 0 | [4846, 'ALL'] | 0.727273 |
| 0.818182 | [1248, 'AML'] | 0.1153846 | [1248, 'ALL'] | 0.702797 |
| 0.818182 | [3257, 'AML'] | 0.1153846 | [3257, 'ALL'] | 0.702797 |
| 1 | [2120, 'AML'] | 0.3076923 | [2120, 'ALL'] | 0.692308 |
| 1 | [6802, 'AML'] | 0.3076923 | [6802, 'ALL'] | 0.692308 |

AML stands for the expression that expresses Acute Myeloid Leukemia, and ALL stands for Acute Lymphoblastic Leukemia. These cancer expressions are paired along with a gene expression that was being expressed within the same time and space, these gene expressions are represented as integers which give their location within the first dataset. To check if it had generated meaningful sets,  we looked up the functions of the occompaning gene to measure their relevance.

| ACUTE MYELOID LEUKEMIA | | |
|---|---|---|
| Accompanying Gene | Function | Relevant? |
| 6200 = Interleukin-8 Precursor | Growth Factor for Certain Cancers | yes |

| ACUTE LYMPHOBLASTIC LEUKEMIA | | |
|---|---|---|
| Accompanying Gene | Function | Relevant? |
| C-MYB | Encodes MYB Proto-oncogene Protein | yes |

The algorithm generated a set which included Interleukin- 8 precursor which is a known growth factor for certain cancers along with Acute Myeloid Leukemia. For Acute Lymphoblastic Leukemia the algorithm generated a set where the accompanying gene was C-myb. Both of the accompanying genes have functions that are related to the development of cancer, which indicates that using associative learning rule algorithms can be useful to identify core

**3.2 Gene Expression data for Bayesian Learning**

From the methods that are presented within associative learning the Bayesian approach fits biochemical networks the most fitting based off of the nature and resources available for such models, along with the less time consuming process involved in these networks. Bayesian statistics makes a valuable framework available for use with statistical data analysis and conceptualising how models can update their environment. The main concept of Bayesian statistics is that assumptions can be made off of the probability distribution, each statistical observation gives rise to the activation of further observations. Such ability allows lower time expenses because it can make its next assumptions on its own given the information fed. Bayesian performance considers things to that of what a rational observer would consider, by regarding its environment which is the information provided to it, that would be the gene

expression data. It then would adjusts its responses to the evidence which could be seen as the association sets of gene expression data .[10] The assumption (belief) is dependent on the task it is performing, meaning "most generous attribution of belief given the task''. [10] Meaning it finds the best fitting by averaging the correlation between the "fitted logistic regression model and the analytic posteriors" This was translated as efficiency within the work, because it allowed it make high fitting assumptions even when noise was present. Within Bayesian method work it was also found that "The information required to perform the noisy clocked task is relatively easy to accumulate in a detectable form: for random networks, the mean model/posterior correlation is fairly high (0.82)." [10] This is important due to high noise in biochemical networks and insufficient data available for such networks.

**Conclusion**

The results that associative learning rule algorithms can generate relevant, shows that with further analysis that it can be useful to identify core gene expressions that, has potential to be further developed into a cell classification program that classifies genes by using the core underlying gene expressions that are present in cells.

**Bibliography**

[1] A. P. Mills, B. Yurke, and P. M. Platzman, "Article for analog vector algebra computation," *Biosystems*, vol. 52, no. 1–3, pp. 175–180, Oct. 1999.

[2] C. T. Fernando *et al.*, "Molecular circuits for associative learning in single-celled organisms," *J. R. Soc. Interface*, vol. 6, no. 34, pp. 463–469, May 2009.

[3] D. Blount, P. Banda, C. Teuscher, and D. Stefanovic, "Feedforward Chemical Neural Network: An In Silico Chemical System That Learns xor," *Artificial Life*, vol. 23, no. 3, pp. 295–317, Aug. 2017.

[4] K. MacVittie, J. Halámek, V. Privman, and E. Katz, "A bioinspired associative memory system based on enzymatic cascades," *Chemical Communications*, vol. 49, no. 62, pp. 6962–6964, 2013.

[5] J. Macia, B. Vidiella, and R. V. Solé, "Synthetic associative learning in engineered multicellular consortia," *J R Soc Interface*, vol. 14, no. 129, Apr. 2017.

[6] L. Qian, E. Winfree, and J. Bruck, "Neural network computation with DNA strand displacement cascades," *Nature*, vol. 475, no. 7356, pp. 368–372, Jul. 2011.

[7] M. Tikhonov and W. Bialek, "Complexity of generic biochemical circuits: topology versus strength of interactions," *Phys. Biol.*, vol. 13, no. 6, p. 066012, Dec. 2016.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze

Expression Data," *Journal of Computational Biology*, vol. 7, no. 3–4, pp. 601–620, Aug. 2000.

[9] N. Gandhi, G. Ashkenasy, and E. Tannenbaum, "Associative learning in biochemical networks," *Journal of Theoretical Biology*, vol. 249, no. 1, pp. 58–66, Nov. 2007.

[10] S. McGregor, V. Vasas, P. Husbands, and C. Fernando, "Evolution of Associative Learning in Chemical Networks," *PLOS Computational Biology*, vol. 8, no. 11, p. e1002739, Nov. 2012.

**Figures**

Figure 1- *http://dgrapov.github.io/MetaMapR/gallery.html*

Figure 2- https://en.wikipedia.org/wiki/Association_rule_learning

Figure 3- https://en.wikipedia.org/wiki/Association_rule_learning

Figure 4- https://en.wikipedia.org/wiki/Association_rule_learning

Figure 5- https://en.wikipedia.org/wiki/Association_rule_learning

Figure 6- [2]

Figure 7- https://en.wikipedia.org/wiki/Breadth-first_search

Figure 8- https://www.hackerearth.com/practice/algorithms/graphs/depth-first-search/tutorial/