

Ben-Gurion University of the Negev

Faculty of Engineering Science

Department of Software and Information Systems Engineering

**Thesis Proposal**

**Toward machine learning framework for metabolic pathway classification based on metabolite correlation networks.**

**November 2016**

**Submitted by:**

**Kostya Kilimnik**

**Supervised by:**

**Dr. Rami Puzis**

Table of Contents

1. Background, scope and purpose ....................................................................................... 3

2. Introduction .............................................................................................................................. 5

3. Scientific background and related work ........................................................................ 7

3.1. Graph embedding ................................................................................... 7

3.1.1. Related work – Graph embedding ............................................................................. 10

3.2. Subgraphs and local structures ............................................................................................................................... 16

3.2.1. Motifs ................................................................................................. 16

3.2.2. Graphlets ............................................................................................................ 18

3.3. Graph features ........................................................................................... 20

3.3.1. Global features....................................................................................................... 25

3.3.2. Local features ................................................................................................................... 28

3.3.3. Other features………………………………………………………… ......................................... 30

3.4. Graph learning with neural networks .............................................................................. 33

3.5. Graph generation from vectorial data ……………………………………………………

3.5.1 Correlation networks ………………………………………………………………………….

3.5.1.1 Metabolic Correlation Networks ……………………………………………………….

3.5.2 Other types of network generation ………………………………………………………

3.6. Learning Metabolites and Metabolic Pathways

4. Research objectives and expected significance ........................................................ 35

4.1. Research Objectives .................................................................................................................. 35

5. Detailed description of the proposed research ........................................................ 37

5.1. Research Hypotheses ................................................................................................................ 37

5.2. Dataset Overview ....................................................................................................................... 38

5.3. Research Method ........................................................................................................................ 41

5.4. Evaluation Methods and Measures ..................................................................................... 49

5.5. Research Plan ............................................................................................................................... 51

5.6. Preliminary Results ................................................................................................................... 53

6. References ............................................................................................................................... 55

**1. Background, scope and purpose**

The research described in this work is being done at the intersection of several disciplines, such as network science, bioinformatics and statistical learning. The aim of the study is to explore and develop new methods for sub-graph classification. Specifically, the project seeks to formalize the analytical pipeline, proposed by David Toubiana, and implement the framework for metabolic pathway classification and prediction. The ultimate solution will have three main stages: (1) generation of metabolic correlation network based on metabolic profiles relational data input (2) topological feature extraction for complex sub-graph structures (3) learning statistical model for sub-graph binary classification.

Beyond the implementation of this framework, we plan to apply it to several available metabolic profile datasets of different species (e.g. tomatoes, corn and others), analyze the topology of its correlation networks, and make predictions for previously unmapped pathways based on obtained models. Despite the fact that current research experiments will be held in the domain of metabolomics, we will attempt to build generic framework suitable for use in other fields of study (e.g. malware detection via function correlation networks).

The result of this project will be able to serve as a supporting tool for bioinformatics researchers. Combination of human expertise with top predictions from statistical model holds great potential, as it can guide biological research and help to make smarter decisions in design of high cost and time consuming lab experiments.

High level solution requirements:

* Each of mentioned three stages in the framework should be implemented as an extensible module, providing convenient infrastructure for future research.
* Multiple configuration parameters of the solution should be controlled by the user and have the defaults in order to combine usability and flexibility.
* Multiple machine learning techniques should be attempted to achieve best accuracy in cross-validation.
* The framework should serve as a tool for reproducible research in any field of application.

There is an abundance of available metabolic profile datasets that can be used during this study. The only constraint with all this diversity is lack of uniform layout and multiple annotations for the same metabolites and pathways given in different datasets. We will attempt to overcome such limitation by requiring minimal rules from input datasets. In addition, the solution will strive to address following gaps in knowledge:

* There are no methods for subgraph embedding that can take into account the context of its enclosing network.
* There are no methods to generate global network features that are not based on expert knowledge.
* The problem of subgraph classification within the context of its enclosing network has not been addressed in the literature.
* Relationships between biochemically distant pathways are difficult to extrapolate from metabolite networks

The detailed method and motivation will be covered in next section.

**2. Introduction**

Metabolic pathways are genetically driven series of chemical reactions that produce or degrade substrates (other metabolites). As more complete genome is sequenced and more metabolic profile datasets are available for different species, the task of analyzing the resulting data becomes increasingly important. Experimental elucidation of biochemical processes taking place in the cell is very laborious, time consuming and costly endeavor. Computational approaches for metabolic pathway prediction and discovery may serve as decision support systems to facilitate and optimize the design of the experiment. The effectiveness and accuracy of existent automated analysis techniques is still arguable in this field, but constant progress and innovation in network science and machine learning algorithms push the community to seek for better advanced solutions. The main goal of our study is to create framework that will be able to predict probability of given metabolic pathway to be found in the given organism. Our framework takes as input an annotated metabolic profile data of any available cultivar. The second input is a metabolic pathway collection like PlantCyc, MetaCyc and etc. Based on these inputs framework will build statistical model capable to classify pathways into relevant and irrelevant to the cultivar of interest. Our method will exploit metabolite correlation networks (MCN) constructed by pairwise correlating the profiles of metabolites. Vertices in the network represent the molecular components (metabolites), and the links between them, the correlation coefficients. An MCN captures the coordinated behavior of metabolites, potentially contextualizing metabolic pathways into the network. Usually, metabolic pathways are presented as genome-scale hypernetworks where vertices (metabolites) are connected by hyperlinks (biochemical reactions). In our study pathway representation is simplified: metabolites participating in a pathway induce an undirected weighted subgraph in the MCN. In order to not being limited to network mining methods we use and develop sub-graph embedding techniques. Such embedding allows transition from graph to multi-dimensional vector space where multitude of relational machine learning algorithms can be applied. Research of sub-graph classification within the context of its enclosing network is expected to become another contribution of this study. This aspect of research goes beyond specific domain of metabolomics. Many complex structures exist in a larger context, for example: text documents and the semantic web; communities or groups of interest and Facebook; software and shared libraries; metabolic pathways and metabolic networks; and many more. All these structures can be defined as subgraphs within the context of a larger network. We are not aware of prior works that tackle the problem of sub-graph classification by statistical learning methods and we aim at building the framework suitable for extensions and future research in this field.

The present work is divided into two parts, namely, engineering efforts to build modular framework for end-to-end solution and then the research part for the development of novel approaches to subgraph embedding.

The rest of this work is structured as follows:

**3. Scientific background and related works**

In this research our goal is to propose method for augmented learning. The new algorithm will get vectorial dataset with n instances and m features as an input, and will produce vectorial dataset with same n instances and k features at the output. The method can be considered as a type of feature extraction process, whereby given dataset is transferred to graph space, and specific node attributes are then extracted as new features for further learning. At the end of the current work we aim to show supremacy of suggested approach for classification tasks on several datasets from different domains. We will show that new graph-based features grasp hidden relationships between instances and thus improve accuracy of off-the-shelf classification algorithms. Particularly such approach will stand out for semi-supervised learning, where deficiency in labeled instances can be filled in by relational information from a big number of unlabeled instances. Besides, we have strong indications that our method will be more robust for missing values problems. During the research we will strive to explore the best method and measure for graph construction from any vectorial dataset and to reveal the most efficient graph-based features for the task.

The most common assumption in statistical learning field is of data represented by points in high-dimensional space. For any specific domain task, we can usually collect some low-level features (e.g. words for text and pixels for image classification) and apply standard tools for vector representation. While sufficient for many purposes and development of elegant, effective algorithms, in many cases such abstraction masks rich intrinsic structure of underlying data, preventing findings of more complex patterns and solving more general problems. In this research we build a framework to exploit hidden relational structure of instances by constructing similarity based graphs and extracting its nodes neighborhood attributes. Having in hand newly learned graph-based features we apply standard machine learning methods to improve classification accuracy and reveal new, previously unknown patterns. We plan to extract all possible node structural attributes from constructed graph and choose small set of most informative features that can effectively separate objects of different classes. More than that, we aim at identifying nodes with well-defined topological characteristics, also called roles. Such role exploration can be of great importance in some domains like protein or metabolite data, where it can shed light on protein functions or metabolite pathways. Particularly, we apply our methods on Italian sparkling wine metabolites dataset.

One of the most powerful points of this approach is its application to semi-supervised learning tasks. In the presence of small number of labeled and a vast number of unlabeled instances, we generate graph using similarities of all of them, thus, embedding important information into learning process.

Not all datasets were born equal. Mining some of them is relatively an easy task, given modern toolbox of statistical learning algorithms and computational power. However, majority of high dimensional datasets collected today are complex and contain hidden structure patterns.

**“Drowning in information but starved for knowledge”**

**Problem statement**: We have vectorial dataset with some number of instances and features. We aim at classifying or clustering our dataset into some known number of classes. The most intuitive way to start dealing with this problem is to impute missing values if exist, and apply wide range of classification/clustering algorithms on the data after some specific to algorithm preprocessing routine. In some cases it can work out very well, but in many cases we are not happy with the results. There could be multiple reasons for poor results, for example, high dimensionality of dataset, small number of independent features, or disability of given data representation to express intrinsic structure of data set. One can tackle the challenge in different ways. This work proposes the framework for graph-based learning process that includes the following steps:

* Construction of similarity graph from vectorial dataset
* Local feature extraction from constructed graph
* Semi-supervised learning from augmented feature space

**Background and motivation for graph-based data representation**

Generally there are two major data representations suited for data mining task of classification or clustering: vectorial and graph based data representation. Frequently “classic” vector space is referred as encapsulation of the extrinsic or global structure of the collected data. Graphs tend to represent intrinsic or local structure of data set. There is no sharp line between extrinsic and intrinsic structures of same data set. Very common example to demonstrate the difference between these 2 structures is to imagine a walk in the big building with multiple halls, rooms and floors. One can walk up and down the floors, enter various rooms and at last understand the topology of the building and neighborhood of all halls and rooms. But looking out of the window would not help to realize how this building looks from outside or where it is located with respect to the external world.

Most of datasets naturally are collected as relational data modelling each row as an instance or observation, and each column as a feature or variable. But, with a rise of social networks and web mining fields, more and more data is represented in graph data structures, where vertices are objects and edges encode agreed type of relationships between them. Yet, it is not the only reason. Last decade growing interest in graph based object representation evoked due to the unique ability to represent entity properties and binary relations under different parts of that entity at the same time [1]. Such relationships can be of spatial, temporal or conceptual nature; they can be easily added, removed, or weighted. The trigger for a new wave of interest in this area was released by big buzz around studies on common graph properties such as “small worlds” and scale free networks [2] [3]. That said, graphs are still not the common data structure in pattern recognition field due to the complexity of the algorithms, as any even basic mathematic operation is not trivial when applied to graphs. The most fundamental and intrinsic problem in graph theory is of graph isomorphism [4]. A report on found solution with quasi polynomial time [5] caused a flurry of excitement among the scientific community and foreshadows new breakthroughs in the field of graph theory and network science. Meanwhile each one of both data representations has its own advantages and drawbacks. Representing data by feature vectors offers mathematical wealth of available operations, rather low computational complexity of algorithms and its variety. The same time using vectorial space push us to predefined set of features and lack of binary relations between different parts of a pattern. Probably in some specific cases, where patterns are simple and base on statistical distribution of known variables, such limitations are not severe. However, there are a number of application domains with strongly structured data, and where disregarding of topological information would cause loss of knowledge. These drawbacks can be solved by graph based representations that in its turn are limited by algorithm complexity. The very natural need to overcome limitations of discussed data approaches pushed scientific community to strive to combine the best from two worlds and to find hybrid solutions.

Next section will contain introduction to few related fields that, combined together, form basis for the proposed research.

1. Similarity graph generation from relational dataset
2. Graph-based clustering
3. Local graph features (graphlets/motifs/spectra)

**Related works**

**Metabolic profiles analysis**

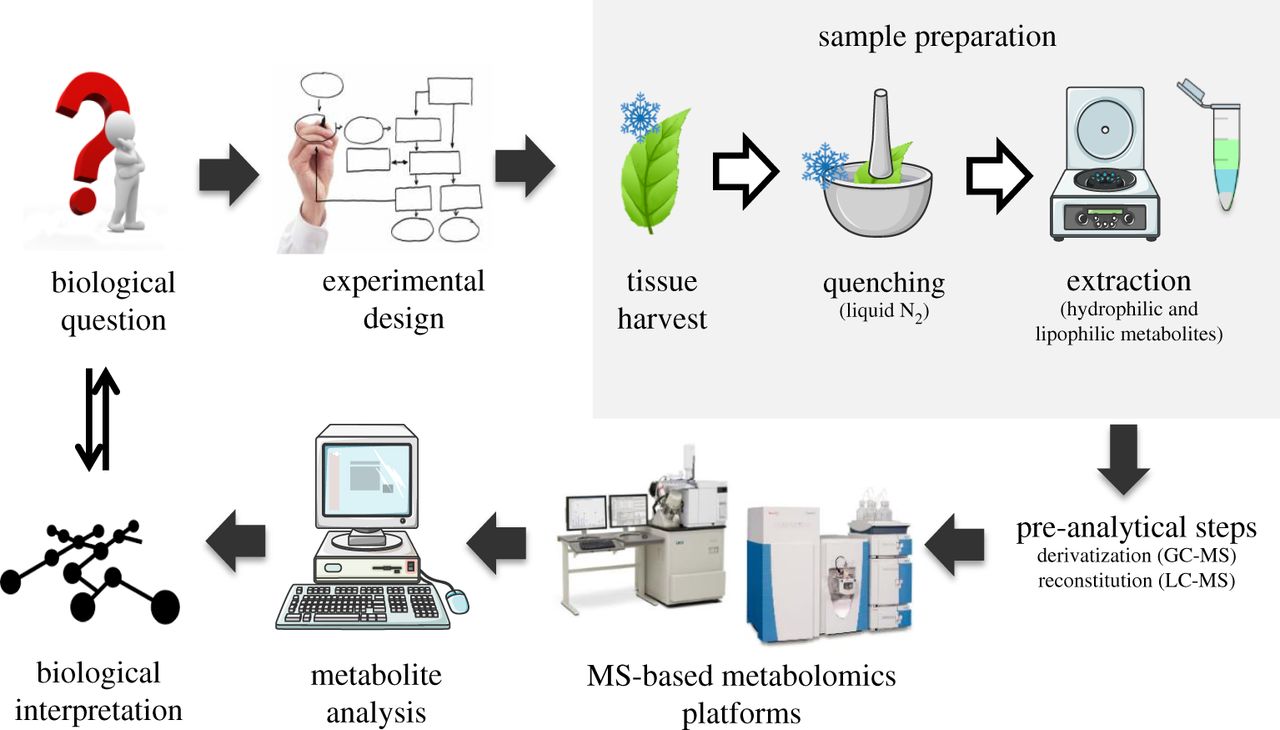
FromJorge, Tiago F., Ana T. Mata, and Carla António. "Mass spectrometry as a quantitative tool in plant metabolomics." *Phil. Trans. R. Soc. A* 374.2079 (2016): 20150370.

Metabolomics is a research field used to acquire comprehensive information on the composition of a metabolite pool to provide a functional screen of the cellular state. Studies of the plant metabolome include the analysis of a wide range of chemical species with very diverse physico-chemical properties, and therefore powerful analytical tools are required for the separation, characterization and quantification of this vast compound diversity present in plant matrices.

“Quantitative plant metabolomics is a tool that helps to improve our understanding of plant biochemistry and metabolism by delivering the accurate measurement, prior to statistical and bioinformatics analysis, of the concentrations of known metabolites that occur in different levels in plant samples [[**1**](http://rsta.royalsocietypublishing.org/content/374/2079/20150370#ref-1),[**2**](http://rsta.royalsocietypublishing.org/content/374/2079/20150370#ref-2)]. The two most commonly used analytical technologies driving quantitative plant metabolomics studies are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. However, due to its high sensitivity relative to NMR, MS is by far the technology of choice in most plant metabolomics studies, and, when coupled to powerful chromatographic techniques (e.g. liquid chromatography–mass spectrometry (LC-MS), capillary electrophoresis–mass spectrometry (CE-MS) and gas chromatography–mass spectrometry (GC-MS)), allows the separation and characterization of the extremely high compound diversity present in the plant metabolome [[**3**](http://rsta.royalsocietypublishing.org/content/374/2079/20150370#ref-3)–[**5**](http://rsta.royalsocietypublishing.org/content/374/2079/20150370#ref-5)]. Nonetheless, NMR methods are still increasingly being applied as a particular approach for structure elucidation in plant metabolomics studies [[**6**](http://rsta.royalsocietypublishing.org/content/374/2079/20150370#ref-6)]. Indeed, both approaches provide complementary as well as supplementary information on the concentrations of metabolites, their ranges and changes in complex plant matrices, and, when combined with new experimental and computational methods, enable routine quantitative analysis of hundreds of plant metabolites[[**7**](http://rsta.royalsocietypublishing.org/content/374/2079/20150370#ref-7)–[**10**](http://rsta.royalsocietypublishing.org/content/374/2079/20150370#ref-10)].” See figure 1.

GC-MS metabolite profiling approaches have the advantage over LC-MS and CE-MS of a relatively broad coverage of compound classes, and interest in applying it will continue to grow in the field of plant metabolite responses to various genetic and/or environmental perturbations (abiotic/biotic stress factors).

In [35] authors demonstrate usefulness of metabolic profiling obtained from gas chromatography–mass spectrometry technologies. Besides they show how metabolic profiling in conjunction with data mining tools such as hierarchical cluster analysis and principle component analysis can contribute to comprehensive characterization of a plant genotype.

****

**Similarity graph construction from vectorial data**

When set of data points is modelled as a graph, it’s natural that vertices represent data points and edge weights represent its pairwise similarity. The common name for such graphs in literature is similarity graphs (rarely affinity graphs). The choice of similarity measure is very important and depends on specific domain. After the choice, some verification needs to be done to check that closely related data points get high value of similarity measure. There is a number of approaches to connect neighboring objects together: ε- neighboring, k- neighboring and fully connected graph [20].

In case of ε-neighboring approach two objects x and y are connected by an edge if they are lying in an ε radius environment (d(x,y) < ε, where d yields the ‘distance’ of the objects x and

y, and ε is a small real number). Weights are rarely applied in this case as scale of distance between connected vertices is bounded by ε. Applying the k-neighboring approach, two objects are connected to each other if one of them is in among the k-nearest neighbors of the other, where k is the number of the neighbors to be taken into account. By definition resulting graph is a directed graph. The author mentioned two options to make it undirected: just ignore direction or connect two vertices only if each of them is in k-neighborhood of another. First option will produce so called knn graph, the latter option – mutual knn graph. The edges of the graph can be weighted several ways. In simplest case, Euclidean distance of the objects can be assigned to the edge connecting them together. Of course, there are other possibilities as well, for example the number of shared neighbors can also characterize the strength of the connectivity. In some domains like biochemistry and computational biology, networks are constructed based on Pearson correlation pair wised values [7] [8]. To ensure robustness of the similarity graph, first the corresponding p-value of Pearson product-moment correlation is determined and afterwards, matching connectivity threshold for Pearson coefficient is chosen by evaluating four different network properties. It is important to point out that all above methods are sensitive to normalization scheme applied on dataset – different normalizations results in different structured graphs. Unfortunately, there is no one clear research that compares and concludes on recommended and universal normalization for similarity graph construction.

Given raising complexity of collected datasets and its high dimensionality, not all available features can equally contribute to similarity graph-based representation. Moreover, in many cases usage of l2-norm metric as a similarity measure is not justified, as data does not possess the Euclidean behavior. Thus, recently more and more research is devoted to search for more robust methods for affinity graph construction. Few of such efforts were conducted in direction of adapting to the local data structures [23-24], particularly focusing on learning adaptive scaling factor σ for the Gaussian kernel . On the side note, these methods are still not immune to noisy and irrelevant features. Another work [25] proposed algorithm for forming tight neighborhoods by selecting maximal cliques. This method claimed to construct graphs with fewer false affinity edges. To address the same problem, kNN based graph generation method appeared [26], where consensus information from multiple kNNs is used to discard noisy edges and again detect tight neighborhoods. All mentioned methods still use all features to build the graph. Few more algorithms to filter out all irrelevant features were developed, all are based on random forest method [27-29]. Recently generalized model for affinity inference method was proposed [21] and it also exploits the clustering random forests method. This approach eliminates need in Gaussian kernel as it naturally possess local neighborhood.

Additional work, which deals with protein function prediction, deserves attention for its original approach to similarity graph modelling. All dataset is divided into subsets of samples according to labelled classes. Instead of single graph, graph pyramid is constructed for each subset. The highest level in pyramid contains similarity graph with lowest density (only strongest relationships are preserved) modelling global features only. In contrast the lowest level in pyramid has the highest density graph (even very weak relationships) and models local features.

Then such hierarchal graph structure is exploited to classify unlabeled protein to the right pyramid (class).

In real world problems, collected data can have a big number of observations. Calculating weighted graph with all pair wised similarity measure values can be intractable. In such case some sort of data reduction is required. The simplest approach is to perform stratified sampling from data, but that can lead to loss of some vital data needed to construct adequate structural data representation, which reflects real intrinsic properties of the dataset. Vector Quantization methods can choose data representatives called codebook vectors (fingerprints or cluster centres), which in their turn will be used for graph similarity construction. Widely used vector quantization algorithm include: k-means clustering, neural gas [32], growing neural gas [33], and topology representing networks [34].

*Overall, pathway reconstruction is a laborious, long-lasting procedure, requiring a priori knowledge of the stoichiometric balances between compounds and thermodynamic information of the pathway’s reactome, as well as its cellular compartmentalization. Existing methods neglect the crosstalk and concerted regulation between biochemically distant pathways. Genome based methods do not take into account endogenous (e.g. developmental) and exogenous (external) factors that influence metabolic ties.*

Complementary to the constrained-based approach, metabolite networks, constituted on high-throughput data metabolite profiles, provide an attractive method to study the coordinated behaviour of metabolites without the need for a priori knowledge. Profiles of single metabolites, quantified in different biological samples, are pairwise correlated. The correlations are transformed into network form, where vertices represent the metabolites, and links the correlation coefficients between them [105], forming a correlation network (CN). The resulting network illustrates a holistic view of metabolite relationships, contextualising the metabolic pathways into the network, and reflecting the actual state of coordinated behaviour at the time of sampling. Metabolite CNs are often reconstructed based on the exploitation of the natural variability of mapping populations [106, 107] or collections of different varieties or cultivars, as they provide large sample sizes stabilizing the correlation and reducing the error rate. Structural network properties can be used to interpret metabolite CNs and even to propose hypotheses. For instance, a network property analysis approach has been successfully performed to identify loci regulating branched-chain amino acids in tomato seeds [108]. *Although, CNs can be studied with a variety of graph theory tools (see above), most current studies suffice themselves employing CN analysis to describe the global structure and relationships of metabolite data. To the best of our knowledge, structural analysis of metabolite CN was never employed for identifying metabolic pathways.*

**Research Objectives**

Our research focuses on subgraph classification in correlation networks. Our main domain of research is computational biology or, more specifically, metabolic pathways that can be easily modelled as subgraphs in correlation network which, in its turn, is built above any relational dataset of metabolite profiles. The main objective of current research is to develop generic and pluggable data analysis framework for subgraph exploration and classification which can be used in multiple domains and various collected datasets that fit our analytical model. We divide the research goals and its significance into 2 pivotal groups.

1. Framework development and algorithm generalization

* Design and implement basic extensible framework based on analytical pipeline developed by D.Toubiana. It should be self-contained and include the whole chain of data analysis: data sources preprocessing, correlation network generation, subgraph feature extraction and, finally, subgraph classification model training.
* Explore and analyze the whole process of subgraph feature extraction. Propose and implement new methods for feature extraction of complex structures like subgraphs
* Compare subsets of significant features for classification among different datasets and conclude on specificity or generality of useful feature groups
* Enhance ML methods and algorithms used for classification in order to improve accuracy and robustness of the predictions
* Conduct research of influence of negative instances sampling method on quality and robustness of classification algorithm based on cross validation evaluation
* Make framework applicable for any domain of research (not only biological data) where datasets can be stored in a form that fits our data model
* Distribute the framework as a package for open-source community.

1. Framework application on various datasets of metabolic profiles

* Reproduce results of previous research on tomatoes species dataset using our framework
* Apply the framework and train models on other available metabolic profile datasets, e.g. corns, soya, cannabis and etc.
* Compare and analyze results provided by different trained models on same metabolic pathway candidates or randomly sampled subgraphs
* Examine metabolism behavior of same type of species under normal and stressed environment conditions in context of correlation network topology. Check what graph-based features differ significantly.
* Conduct proof of concept experiments for simple search strategy of new metabolic pathways in correlation network

# Detailed description of the proposed research

# Research hypotheses

The main hypothesis of this research is that metabolic pathway existence in specific type of cells is predictable based only on knowledge of metabolite profiles data extracted over multiple sample population of that cell. We aim to explore the hidden knowledge in pairwise correlations of single metabolite profiles and develop analytical easy-to-use tool for metabolic pathway prediction. Such tool will be capable to provide probability of pathway occurring in cell of interest even if that pathway was never experimentally found in that type of cells.

The research hypotheses are:

* **H1:** Accuracy and general positive results achieved on tomato seeds dataset in previous research are reproducible on other species metabolic profile data by usage of developed framework.
* **H2:** Subgraph feature extraction and selection methods developed in this research will improve levels of classification accuracy and AUC achieved in previous research.
* **H3:** Machine learning classification algorithms tuned for our framework needs will increase accuracy and other evaluation metrics compared to previous research results produced by D.Toubiana.
* **H4:** Combining random sets of metabolites and real pathways as negative pathway instances in classification methods will improve classification results and make them more robust.
* **H5:** Model trained with developed framework on one dataset will produce significant classification results when applied to other dataset correlation networks.
* **H6:** *Formulate hypothesis related to analysis of pathway role in dataset with normal conditions compared to dataset collected for cell in stressed conditions*

**5.2 Dataset overview**

The metabolite data set presented is based on field-grown ILs from two seasons (2001 and 2003). The field trials were conducted at Western Galilee Experimental Station in Akko, Israel. Plants were grown in a completely randomized design with one plant per m2. Seedlings were grown in

greenhouses for 35–40 d and then transferred to the field. The field was irrigated with 320 m3 of water per 1,000 m2 of field area throughout the season.

**Metabolite measurement by gas chromatography/mass spectrometry.**

Relative metabolite content was determined essentially as described [35] with modifications specific to tomato. IL mapping. To map the metabolites, a two-way ANOVA was used to partition

metabolic variation into genotype, environment and genotype \_ environment

interaction effects. This method has been commonly applied to transcriptional

analysis and shows excellent robustness. A metabolic effect in a specific IL was

designated significant when the control and the IL had at least four replications

in each year, and the genotype factor for the combined 2-year analysis was

significant (a ¼ 0.05) but the interaction factor (a ¼ 0.01) was not significant

(or if the IL was significantly different (a ¼ 0.05) from M82 in each year

separately (t-test) and in the same direction). Genotypes missing in one year or

with less than four replications were not subjected to ANOVA.

Network analysis. Correlation between all trait (metabolite plus morphological)

pairs was tested by using IL mean values (n ¼ 76 lines). Because the 83

traits yield 3,403 pairs, we choose strict significance levels (0.0001). The 302

pairs that resulted were considered as a network in which a vertex corresponds

to a trait and a link between two vertices corresponds to significant correlations

between these two traits.

**Correlation Networks**

(From Toubiana 2016) Correlation-based network analysis (CNA), on the other hand, provides a method to illustrate the relationship between molecular components without prior knowledge of the underlying chemistry. The relational ties established between different cellular components via CNA can represent coordinated changes of abundances in response to a given genetic or environmental perturbation (Toubiana et al., 2013). Furthermore, the topology of correlation networks can be analyzed with well-defined network properties from graph theory and communities can be identified with community detecting algorithms

(from A correlation network approach to metabolic data analysis

for tomato fruits 2007) Metabolite correlations are believed to provide a

‘fingerprint’ of the underlying biophysical system

(Steuer et al. 2003b; Steuer et al. 2003a; Weckwerth

et al. 2004; Morgenthal et al. 2006). Elucidating the

origin of metabolite correlations will give us insight

into biochemical processes and their regulation

(Camacho 2005; Steuer 2006).

**References :**

[1] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. Int. Journal of Pattern Recognition and Artificial Intelligence, 18(3):265-298, 2004.

[2] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of ‘small world’ networks." Nature 393.6684 (1998): 440-442.

[3] Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." science 286.5439 (1999): 509-512.

[4] Köbler, Johannes, Uwe Schöning, and Jacobo Torán. The Graph Isomorphism Problem: Its Structural Complexity. Boston: Birkhäuser, 1993. Print.

[5] Babai, László. "Graph isomorphism in quasipolynomial time." arXiv preprint arXiv:1512.03547 (2015).

[6] Sandhan, Tushar, et al. "Graph pyramids for protein function prediction."BMC medical genomics 8.2 (2015): 1.

[7] Carlin, Silvia, et al. "Regional features of northern Italian sparkling wines, identified using solid-phase micro extraction and comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry." Food chemistry 208 (2016): 68-80.

[8] Toubiana, David, et al. "Network analysis: tackling complex data to study plant metabolism." Trends in biotechnology 31.1 (2013): 29-36.

[9] Xia, Tian, et al. "On defining affinity graph for spectral clustering through ranking on manifolds." Neurocomputing 72.13 (2009): 3203-3211.

[10] Yang, Shuang-Hong, et al. "Variational graph embedding for globally and locally consistent feature extraction." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2009.

[11] Huang, Pu, et al. "Feature extraction using graph discriminant embedding."Image and Signal Processing (CISP), 2013 6th International Congress on. Vol. 1. IEEE, 2013.

[12] Bunke, Horst, and Kaspar Riesen. "Improving vector space embedding of graphs through feature selection algorithms." Pattern Recognition 44.9 (2011): 1928-1940.

[13] Bunke, Horst, and Kaspar Riesen. "Graph classification based on dissimilarity space embedding." Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer Berlin Heidelberg, 2008.

[14] Gago-Alonso, Andrés, Alfredo Muñoz-Briseño, and Niusvel Acosta-Mendoza. "A new proposal for graph classification using frequent geometric subgraphs." Data & Knowledge Engineering 87 (2013): 243-257.

[15] Gibert, Jaume, Ernest Valveny, and Horst Bunke. "Graph embedding in vector spaces by node attribute statistics." Pattern Recognition 45.9 (2012): 3072-3083.

[16] N. Prˇzulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scalefree

or geometric? Bioinformatics, 20(18):3508–3515, 2004.

[17] Yaveroğlu, Ömer Nebil, et al. "Revealing the hidden language of complex networks." Scientific reports 4 (2014).

[18] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. In NIPS, pages 849–856. MIT; 1998, 2002.

[19] Schaeffer, Satu Elisa. "Graph clustering." Computer science review 1.1 (2007): 27-64.

[20] Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.

[21] Zhu, Xiatian, Chen Change Loy, and Shaogang Gong. "Constructing robust affinity graphs for spectral clustering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[22] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.

[23] J. Wang, S.-F. Chang, X. Zhou, and S. Wong. Active microscopic cellular image annotation by superposable graph transduction with imbalanced labels. In CVPR, 2008.

[24] ] L. Zelnik-manor and P. Perona. Self-tuning spectral clustering. In NIPS, 2004.

[25] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. TPAMI, 29(1):167–172, 2007.

[26] V. Premachandran and R. Kakarala. Consensus of k-nns for robust neighborhood selection on graph-based manifolds. In CVPR, 2013.

[27] Shi, Tao, and Steve Horvath. "Unsupervised learning with random forest predictors." *Journal of Computational and Graphical Statistics* (2012).

[28] Criminisi, Antonio, Jamie Shotton, and Ender Konukoglu. "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning." *Foundations and Trends® in Computer Graphics and Vision* 7.2–3 (2012): 81-227. [29] X. Zhu, C. C. Loy, and S. Gong. Video synopsis by heterogeneous multi-source correlation.

[30] K. Riesen and H. Bunke. Graph classification based on vector space embedding. Int. Journal of Pattern Recognition and Artificial Intelligence, 23(6):1053–1081, 2009

[31] Martinetz, T.M., Shulten, K.J.: Topology representing networks. Neural Netw. 7(3), 507–522 (1994)

[32] Martinetz, T.M., Shulten, K.J.: A neural-gas network learns topologies. In Kohonen, T., Mäkisara,

K., Simula, O., Kangas, J. (eds.) Artificial Neural Networks, pp. 397–402, Elsevier Science Publishers B.V, North-Holland (1991)

[33] Fritzke, B.: A growing neural gas network learns topologies. Adv. Neural Inf. Proces. Syst. 7, 625–632 (1995)

[34] Toubiana, David, et al. "Correlation-Based Network Analysis of Metabolite and Enzyme Profiles Reveals a Role of Citrate Biosynthesis in Modulating N and C Metabolism in Zea mays." *Frontiers in Plant Science* 7 (2016).

[35] Roessner, Ute, et al. "Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems." *The Plant Cell* 13.1 (2001): 11-29.

[36] Schauer N, Semel Y, Roessner U, Gur A, Balbo I, et al. (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for

tomato improvement. Nat Biotechnol 24: 447–454.

[37] Schauer N, Semel Y, Balbo I, Steinfath M, Repsilber D, et al. (2008) Mode of inheritance of primary metabolic traits in tomato. Plant Cell 20: 509–523.

**Research objectives and expected significance**

The main motivation for our study is to find solution for identifying sets of vertices whose interrelationships match a given class of subgraphs. In this work we focus on domain of metabolomics where given class of subgraphs represents known metabolic pathways in investigated organism. As previously discussed, the secondary goal of this research is to create framework for metabolite pathways classification and prediction based on developed sub-graph embedding techniques and machine learning algorithms. Suggested analytical pipeline implemented in the modular framework will combine and exploit advantages of two representations of collected data: structural (graph-based) and vectorial (feature based). Until recently, metabolic pathways prediction was not studied by means of network analysis and by pathway modelling as a subgraph in correlation metabolic network. Lately, such method was suggested and proved to successfully predict unknown metabolic pathway in tomato pericarp. However, systematic study on subgraph embedding for metabolic data was never held. This work attempts to explore different approaches to subgraph embedding and utilize feature selection methods attached to classification algorithms in order to find generic efficient configuration for metabolomics domain of application. Proposed framework will allow user to customize and configure the system based on his preferences including preprocessing parameters, correlation threshold, sub-graph interpretation and classification algorithm. Our top priority is to come up with good defaults, find out and explore regularities in different configurations of the system.

**Research Objectives**

Despite the fact that our main application domain is metabolomics, we aim at developing universal methods potentially used in multiple domains and various collected datasets that fit our analytical model. Thus, we divide the research goals and its significance into two groups:

1. Framework development and algorithm generalization

* Design and implement basic extensible framework based on analytical pipeline developed by D.Toubiana. It should be self-contained and include the whole chain of data analysis: data sources preprocessing, correlation network generation, subgraph feature extraction and, finally, subgraph classification model training.
* Explore and analyze the whole process of subgraph feature extraction. Propose and implement new methods for feature extraction of complex structures like subgraphs
* Compare subsets of significant features for classification among different datasets and conclude on specificity or generality of useful feature groups
* Enhance ML methods and algorithms used for classification in order to improve accuracy and robustness of the predictions
* Conduct research of influence of negative instances sampling method on quality and robustness of classification algorithm based on cross validation evaluation
* Make framework applicable for any domain of research (not only biological data) where datasets can be stored in a form that fits our data model
* Distribute the framework as a package for open-source community.

1. Framework application on various datasets of metabolic profiles

* Reproduce results of previous research on tomatoes pericarp dataset using our framework
* Apply the framework and train models on other available metabolic profile datasets, e.g. corns, soya, cannabis and etc. Identify and suggest existence of new previously unmapped metabolic pathways
* Compare and analyze results provided by different trained models on same metabolic pathway candidates or randomly sampled subgraphs
* Examine metabolism behavior of same type of species under normal and stressed environment conditions in context of correlation network topology. Check what graph-based features differ significantly.
* Conduct proof of concept experiments for simple search strategy of new metabolic pathways in correlation network

**Expected Significance**

The main expected outcome is an automated framework for prediction and analysis of components in complex networks. Subgraph embedding is a unique approach to facilitate study of complex network structures within a larger context. It will extend the paradigm of graph embedding and define features that characterize substructures in context of the whole network in vectorial space.

Until now most of expensive and time consuming lab work in metabolomics field of study is based on human expertise. The proposed framework will serve as a decision tool for metabolic pathways researchers to guide their lab research and make smarter decisions on experiment design. As a result of applying framework on several different datasets of species, we will create list of previously unknown metabolic pathways with high potential to be found in specific organism. Comparing different framework configurations on several metabolic profile datasets will provide systematic view of what configuration achieves best classification accuracy in cross validation terms.

**Detailed description of the proposed research**

**Research Hypotheses**

The main hypothesis of this research is that metabolic pathway existence in specific type of cells can be inferred from metabolite profiles data extracted over multiple samples population of that cell. This can be achieved by our developed modular and configurable framework. We wish to explore capabilities and parameters of this framework focusing on sub-graph embedding aspects. The following hypotheses are composed in order to guide research direction and priorities.

The research hypotheses are:

* **H1:** Accuracy and general positive results achieved on tomato pericarp dataset in previous research are reproducible on other species metabolic profile data by usage of developed framework.
* **H2:** Including of features groups calculated on sub-graphs in context of whole network significantly increases classification results in cross validation
* **H3:** Including of aggregated graphlet based features significantly increases classification results in cross validation
* **H4:** Normalization of relevant features by partial pathway length (number of pathway metabolite members found in whole graph) improves InfoGain of those features
* **H5:** Including all sub-graph features of 1 hop extended sub-graph (pathway members with its neighbors) contribute to classification results in cross validation
* **H6:** Including all sub-graph features of 2 hope extended sub-graph has no significant influence on classification results.
* **H7:** Classification results are sensitive to choice of statistical learning method
* **H8:** Combining random sets of metabolites and real pathways as negative pathway instances in classification methods improves classification results and make them more robust.
* **H9:** Model trained with developed framework on one dataset will produce significant classification results when applied to other dataset correlation networks.

Group subgraph extensions and subgraph features. Features of subgraph in context of whole network and features of sub-graph as is

Subgraph extensions

**Dataset overview**

Our main metabolite dataset was obtained from introgression population of tomato fruit pericarp from three seasons (2001 - season I , 2003 – season II, 2004 – season III). Values in dataset are measured levels of more than 70 primary metabolites in three populations of tomato resulting from an interspecific cross between cultivated variety (S. lycopersicum) and its distant wild relative (Solanum pennellii). Relative metabolite contents were determined using an established GC-MS protocol allowing the quantification, in methanol extracts, of sugars, sugar alcohols, organic and amino acids, and the vitamins. Additionally, the procedure was optimized for tomato fruit and mass-spectral libraries were utilized for peak identification.

This data was used multiple times in previous researches (36,37). Besides, same data was used by D.Toubiana for his proof of concept research on metabolic pathway prediction.

The data in all three datasets was being normalized by average M82 values and log function. Missing values are imputed by PCA organizing map algorithm commonly accepted in metabolomics domain, but another options can be tested.

Additional datasets:

We will use metabolic profiles from other species to test our implemented analytical method for generality and determine whether accuracy levels in cross validation vary a lot from data to data.

Currently we possess data for following plants in two environmental conditions of water/drought:

* Watermelon
* Pepper
* Maze
* Sunflower
* Melon
* Eggplant

Number of samples in these datasets is much less than in tomato case (12-20 samples), but is enough to determine correlation levels between metabolites and build correlation network.

|  |  |  |  |
| --- | --- | --- | --- |
| **Season** | **2001** | **2003** | **2004** |
| **Number of genotypes(IL)** | 76 | 73 | 69 |
| **Samples for each IL** | ~18 | ~6 | ~6 |
| **Number of parent (M82) samples** | 307 | 90 | 217 |
| **Detected metabolites** | 75 | 75 | 78 |
| **Missing values %** |  |  |  |
| **Number of unique metabolites in 3 seasons** | 90 | | |
| **Number of shared metabolites in 3 seasons** | 50 | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Amino Acids** | **Organic Acids** | **Other** | **Sugar** | **Sugar alcohol** | **Phosphate** |
| Alanine | 1.4\_Lactone\_  GALACTORINATE | 2\_oxoglutarate | Fructose | Erythritol | Fructose\_6\_P |
| Arginine | Benzoate | Adenosine\_5\_  monophosphate | Fucose | Glycerol | Glucose\_6\_P |
| Asparagine | Citramalte | Alpha\_Tocopherol | Galactose | Inositol | Glycerate\_3\_P |
| Aspartate | Citrate | Calystegine\_A3 | Glucose | Maltitol | Glycerol\_1\_P |
| Beta\_Alanine | Dehydroascorbate | Calystegine\_B2 | Maltose | Xylitol | Phosphate |
| Cysteine | Fumarate | Galactinol | Mannose |  |  |
| GABA | Galacturonate | Guanine | Ribose |  |  |
| glutomate | Gluconate | Nicotin\_amid | Sucrose |  |  |
| glutomine | Glycerate | Nicotinate\_2 | Trehalose |  |  |
| Glycine | Isocitrate | Phenylamine | Xylose |  |  |
| Histidine | L\_Ascorbate | Putrescine |  |  |  |
| Homoserine | Malate | Pyruvate |  |  |  |
| Isoleucine | Octadecanoate | Suqalene |  |  |  |
| leucine | Quinqate | Uracil |  |  |  |
| Lysine | Salicylate | Urea |  |  |  |
| Methionine | Succinate |  |  |  |  |
| Phenylalanine | Threnoate |  |  |  |  |
| Proline |  |  |  |  |  |
| Serine |  |  |  |  |  |
| Threonine |  |  |  |  |  |
| Tryptophan |  |  |  |  |  |
| Tyrosine |  |  |  |  |  |
| Valine |  |  |  |  |  |

**Research Methods**

In this section the planned research methodology will be presented. Besides, all framework building blocks will be described together with its potential configurable parameters. The learning process will consist of feature generation for metabolic pathways modelled as sub-graphs and model construction for binary classification. Our final goal is inference of any given pathway as existent or non-existent in the specie of interest.

The system method proposed is comprised of following main stages:

* Framework design and initial implementation of all its modules
* Iterative process of experiments on tomato fruit data in order to check research hypotheses
* Tuning and refining of the framework based on experimental results
* Experiments on other datasets

A schematic view of proposed framework is shown in Fig.

**Detailed description of framework modules**

**Preprocessing**

The purpose of this module is to process raw metabolic profile data and prepare it for correlation network generation. Input relational data is assumed to be in comma separated values format (csv) and contain unique mass intensity values for each annotated compound (metabolite). It is read as numeric matrix where rows represent genotype samples (in case of tomato dataset - introgression lines) and columns represent metabolites detected. Input matrix undergo the following data wrangling process:

* Row based averaging of values for samples of same genotype
* Empty rows/columns elimination
* Row values are divided by chosen control line average
* Decimal logarithm is applied on all intensity values to better resemble a normal distribution
* Missing values imputation

The reason for normalization by control line lays in randomized batch design usually applied for growth of plant samples. For example, in case of tomato dataset, to grow 76 genotypes with 6 samples for each one quite a big field was used. Within the field environment conditions can change such as temperature, sun light, different minerals. To ascertain test for genotype differences only, control line genotype (M82) was grown near each plant. Taking average of all control line samples and dividing all other genotypes intensities by its value statistically removes variability due to environmental conditions. Not all datasets require such normalization and not all of them include control line values, thus it will be optional by configuration parameter.

Another configuration parameter can control imputation algorithm. By default we use nipals PCA method commonly used in bioinformatics.

**Metabolic correlation network generation**

As mentioned in related works section, networks (graphs) are widely used in modern biology to represent and analyze systematic interplay of biologic components. The advantage of correlation network use in field of metabolomics is the fact, that CNs can be composed without any a priori information. All we need is metabolic profile data and few preprocessing steps. Metabolic correlation networks (MCN) are constructed by pairwise correlating the profiles of molecular components of living organisms, where vertices in the network represent the molecular components (metabolites), and the links between them represent the correlation coefficients. A MCN captures the coordinated behaviour of metabolites, potentially contextualising metabolic pathways into the network. Some metabolites that participate in the cycle are highly correlated in most cases because all of them are required for the process to take place. In some cases the correlations between metabolites participating in the same pathways are impaired by other pathways that involve the same metabolites. Metabolite correlations constitute large number of weak clues for existence of metabolic pathways. In our study we keep using commonly accepted Pearson correlation to construct MCN. First, Pearson correlation coefficient matrix (r values) and p-value matrix are generated. Afterwards, optimal thresholds for r and p values are found in order to filter out not robust and weak correlations. The following describes process to set these thresholds.

* Filter out all correlations with p-value less than chosen significance level (0.05)
* Generate multiple graphs based on different combinations of r and p thresholds ( 50 values grid)
* Calculate number of edges for each network and its standard deviation for each value of r threshold. By that we want to determine robustness level of graph at each level of correlation threshold.
* Find minimum threshold for which different levels of significance threshold don’t affect graph density. Condition: std value < Sensititivity \* (number of edges in fully connected graph)
* To account for multiple hypothesis testing we find threshold that guarantees a pre-defined false discovery rate [ reference]
* Generate weighted undirected graph with edges filtered by chosen thresholds. Weight of edge is a correlation value of vertices it connects.

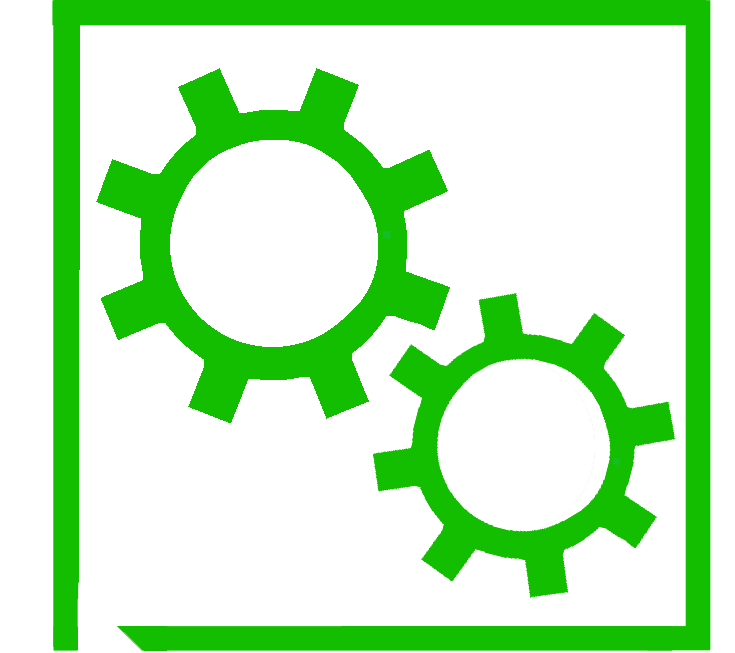
**Sub-graph mapping**

Any node in MCN is a metabolite detected in the genotype. Thus metabolic pathways, that can be viewed as a chain of metabolites and enzymes, might be modelled as sub-graphs of MCN. For training of our classification model we need balanced set of positive and negative instances. Positive instance will be sub-graph that encodes partial or full metabolic pathway known to be present in tested organism ( tomato in our main dataset ). There are few online repositories that contain information on all known pathways for different plants (reference for tomatoCyc, plantCyc). So we just need to scan the repository and query it for pathways that contain at least two compounds from our network. Negative instance will be pathway from another plant that has not been yet discovered in tested specie or any random subset of nodes within length range of known pathways. Combination pattern of these two types of negative samples is point of interest in our research. There are few challenges in mapping task of partial or full pathways from public repository to our annotated network. Names of pathway member compounds not always match metabolite names in our dataset, even though they are the same one. Solution for this challenge might use and maintain list of synonyms or manual annotation for small part of nodes.

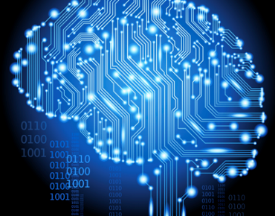
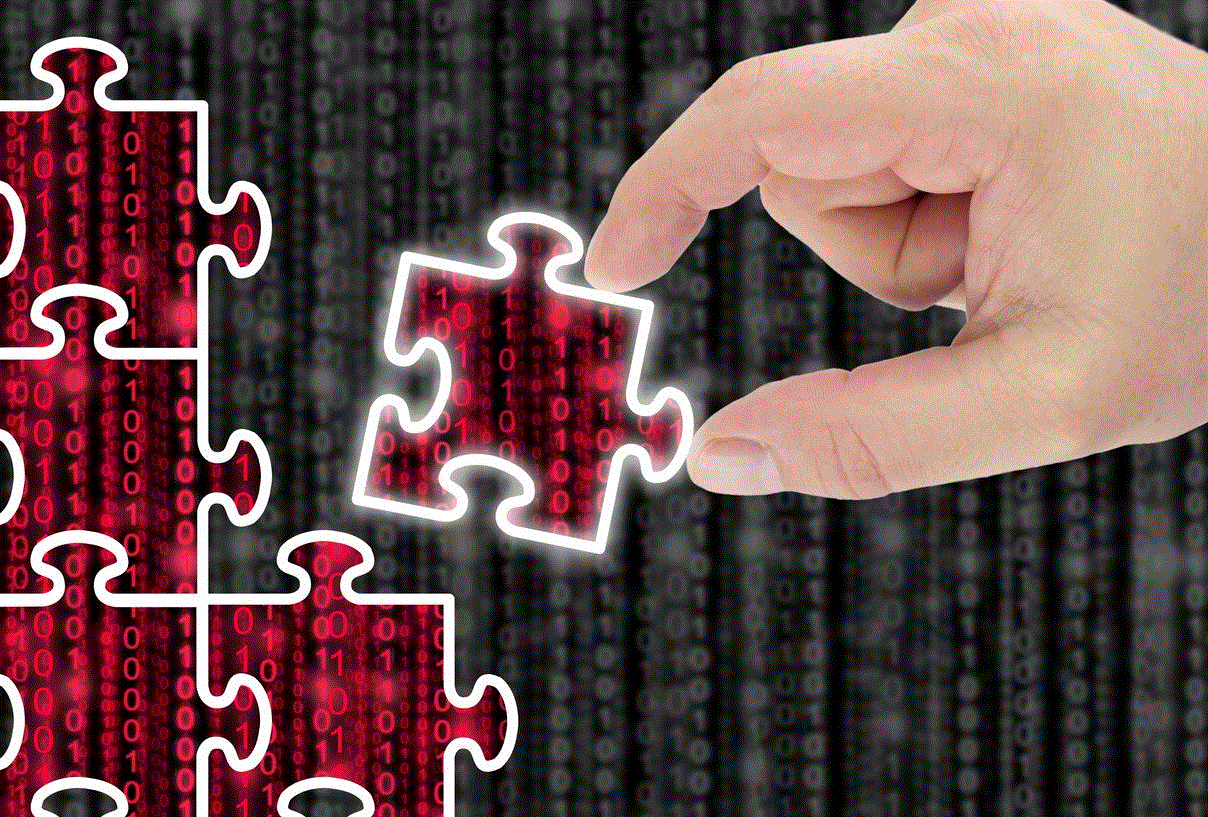
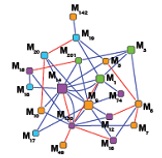
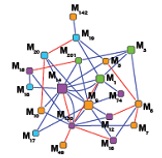
Metabolites

Samples

Metabolic Profile



Preprocessing



**Model**



MCN generation

Sub-graph embedding

ML algorithm

Sub-graph mapping

Pathway repository