# ASAPP: Architectural Similarity-based Automated Pathway Prediction System and its Application in Host-Pathogen Interactions

Rishika Sen, Somnath Tagore, and Rajat K. De, *Senior Member, IEEE*

**Abstract**—The significance of metabolic pathway prediction is to envision the viable unknown transformations that can occur provided the appropriate enzymes are present. It can facilitate the prediction of the consequences of host-pathogen interactions. In this article, we have proposed a new algorithm ASAPP (Architectural Similarity-based Automated Pathway Prediction) to predict metabolic pathways based on the structural similarity among the metabolites. ASAPP takes two-dimensional structure and molecular weight of metabolites as input, and generates a list of probable transformations without the knowledge of any externally established reactions, with an accuracy of 85.09%. ASAPP has also been applied to predict the outcome of pathogen liberated toxins on the carbohydrate and lipid pathways of the hosts. We have analyzed the disruption of host pathways in the presence of toxins, and have found that some metabolites in Glycolysis and the TCA cycle have a high chance of being the breakpoints in the pathway. The tool is available at http://asapp.droppages.com/.

**Index Terms**—Pathogen informatics, Toxins, Perturbation, Metabolic pathway, Similarity, Chemoinformatics, Chemical structure.

## 1 Introduction

Pathogens are infectious agents that disrupt the proper functioning of the host and cause diseases. One of the modus operandi by which pathogens ambush the host is via protein secretion, using the mechanism of secretion systems [1]. These secretion systems discharge protein(s), called effectors, into the body of the hosts which have the capability to distort the usual metabolic pathways leading to the occurrence of unfamiliar transformations. Other than effectors, small molecules called toxins, secreted by pathogens into the host, cause diseases on contact with or absorption by body tissues interacting with biological macromolecules. This results in perturbation of the host system [2]. The significance of pathway prediction is to comprehend the possible undisclosed transformation(s) (reaction(s)) that can materialize provided the appropriate enzymes are available. Our algorithm is an attempt towards achieving this goal.

Multiple attempts have already been made for pathway prediction. *In silico* prediction of pathway came into existence when Karp *et al.* developed the PathoLogic tool [3], followed by the PathMiner [4], Pathway-Hunter [5], Oh *et al.* [6], PathPred [7], and UM-PPS [8] predicting xenobiotic biodegradation pathways, and Rahnuma [9]. The mechanism behind the PathoLogic algorithm was hard-coded, with complicated interactions among various rules, making the algorithm difficult to maintain and extend. Following PathoLogic Tool, McShan *et al.* developed PathMiner [4], a heuristic-based path inferring algorithm. SMILES representation of chemical compound was used to represent metabolites in PathMiner [4] and Pathway-Hunter [5]. However, SMILES representation lacks a standard methodology to generate the representation. Canonical SMILES attempted to alleviate this issue, but there could be some variance in canonical SMILES depending on what tool was used to create them. For each canonical SMILES string of length $n$, there are $[n \times (n+1)]/2$ different sequence of atoms [10]. Different representation of SMILES of the same metabolite leads to different similarity scores between two metabolites.

Similarly, PathMiner [4] uses Manhattan distance between the SMILES sequences of all the metabolite pairs to determine the similarity between the two, thus predicting transformations among the metabolites. However, this method predicts a linear pathway without considering the possibility that branching in the pathway may exist. Likewise, InChI format based software may generate different InChI strings for the same molecule, depending on the choice of a multitude of options [11]. It also lacks the ability to represent polymers. Pathway Hunter tool aims to find the minimum pathway between two metabolites. Soon after PathMiner, specialized tools like PathPred [7] and UM-PPS [8] attempted to predict only the xenobiotic biodegradation pathway. In reality, the metabolic pathways are not restricted to xenobiotic pathways. In fact, xenobiotic pathways make up for only 12% of the metabolic pathways (there are 181 pathways listed in KEGG, among which 21 are xenobiotic). Oh *et al.* and PathPred [7] used RDM (R: Reaction center; D: Difference atom; M: Matched atom) patterns for pathway prediction. In xenobiotic pathways, 80% of the RDM patterns corresponding to each of the

- *R Sen and R K De are with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: rishikasen_r@isical.ac.in, st3179@cumc.columbia.edu, rajat@isical.ac.in*
- *S. Tagore is with Califano Laboratory of Systems Biology, Columbia University Medical Center, Herbert Irving Cancer Research Center, New York, USA.*

*Manuscript received April 19, 2005; revised August 26, 2015.*

transformations in a pathway was unique [6], and could be used to uniquely identify a transformation pair. Thus, the rule for transformation of one metabolite to another was more certain for the xenobiotic metabolite, provided the RDM patterns were taken into consideration. Similarly, UM-PPS [8] [12] was solely applicable for the prediction of bio-degradation pathways. It has a predefined set of transformation rules which needs to be manually updated in order to upgrade the algorithm. Another pathway prediction system, known as Rahnuma [9], used the existing experimentally verified reactions to create a pathway. It has consciously overlooked a set of metabolites and assumed an upper threshold value for the length of the pathways, above which the pathways were not taken into account.

In this article, we have designed a novel generalized algorithm, called Architectural Similarity-based Automated Pathway Prediction (ASAPP) which is used to predict pathways based on the structural resemblance of the metabolites. It has been seen that in a considerable number of pathways, there is structural similarity among the primary metabolites. ASAPP is a versatile algorithm which considers two-dimensional structure (atoms and bonds as well as molecular weight) of the metabolites, as inputs to build an array of probable transformations independently. It does not depend on any externally established reactions. Moreover, ASAPP has an accuracy of 85.09% when tested on 41 predefined pathways (Supplementary Information Table S1-S3). We have applied the algorithm in the domain of host-pathogen interactions to analyze the effect of toxins on the metabolic pathways of the host. The tool ASAPP has been made available at http://asapp.droppages.com/.

## 2 METHOD

In this section, we describe the proposed methodology for automated pathway reconstruction. A pool of metabolites has been considered as input in the form of atoms and bonds as well as molecular weight. The output is a list of probable transformations in the form of compound pairs, indicating that the transformation between these two compounds are highly probable. We have extracted structural information of the metabolites from the KEGG database [13]. KEGG has been considered as the primary database due its versatility, routine updation and robustness. Consider for example, a pathway given in Figure 1. It is the Oxidative phase of the pentose phosphate pathway, where Glucose 6P (C01172[1]) is the initial metabolite and Ribulose 5P (C00199) is the final metabolite. The arrows indicate the transformation of metabolites via the reactions[2]. For example, the metabolites D-Glucono-1,5-lactone 6-phosphate (C01236) and 6-Phospho-D-gluconate (C00345) are transformable via the reaction R02035. Using the present Architectural Similarity-based Automated Pathway Prediction (ASAPP) algorithm, we have computed the chance of these transformations of one metabolite to another, depending on the two dimensional structural similarity between the metabolites.

1. Each metabolite in KEGG is identified by its unique ID of the format C*****.

2. Each reaction in KEGG is identified by its unique ID of the format R*****.

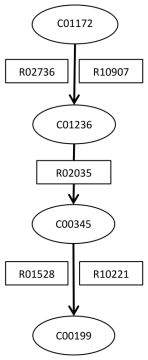
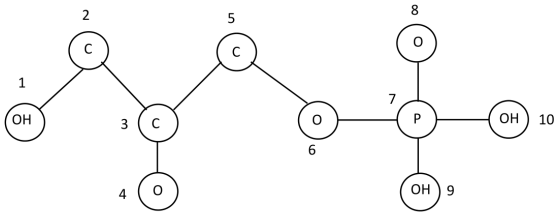



Fig. 1. The oxidative phase of the Pentose Phosphate Pathway. The ovals contain the metabolite IDs and the rectangles stand for reactions. For example, metabolite beta-D-Glucose 6-phosphate (C01172) gets transformed into D-Glucono-1,5-lactone 6-phosphate (C01236) via the reactions R02736 and R10907 (as given in KEGG).




| Edges | 3 sized segment | Atom format | 5 sized segment | Atom format | 7 sized segment | Atom format |
|---|---|---|---|---|---|---|
| **1-2** | 1-2-3 | OCC | 1-2-3-5-6 | OCCCO | 1-2-3-5-6-7-8 | OCCCOPO |
| **2-3** | 2-3-4 | CCO | 2-3-5-6-7 | CCCOP | 1-2-3-5-6-7-10 | OCCCOPO |
| **3-4** | 2-3-5 | CCC | 3-5-6-7-8 | CCOPO | 1-2-3-5-6-7-9 | OCCCOPO |
| **3-5** | 3-5-6 | CCO | 3-5-6-7-10 | CCOPO | | |
| **5-6** | 5-6-7 | COP | 3-5-6-7-9 | CCOPO | | |
| **6-7** | 6-7-8 | OPO | | | | |
| **7-8** | 6-7-9 | OPO | | | | |
| **7-9** | 6-7-10 | OPO | | | | |
| **7-10** | 8-7-9 | OPO | | | | |
| | 8-7-10 | OPO | | | | |
| | 9-7-10 | OPO | | | | |

Fig. 2. Two dimensional structure of the metabolite Glycerone phosphate (C00111) has been laid out as given in KEGG KCF (XML format) files, where each atom has been numbered. Segments of length three, five and seven have been constructed, and their constituent atoms have been shown. The edges represent the bonds between the atoms.

### 2.1 Algorithm

The algorithm ASAPP has been designed to predict a pathway involving possible reactions among metabolites, based on two-dimensional structural similarities between a pair of metabolites. Each metabolite has been perceived as a undirected graph containing bonds and atoms as shown in Figure 2. The symbols used in the algorithm and their meaning have been summarized in Table 1. The modulated flow of ASAPP has been depicted in Figure 3.

#### *2.1.1 Reading metabolite information from KEGG*

Atoms, bonds among the atoms and molecular weights of the metabolites has been automatically extracted from KEGG by the algorithm. The algorithm reads the metabolite names as input and maps a name to a KEGG ID. For each

TABLE 1
Description of symbols used in ASAPP

| Symbol | Description |
|---|---|
| $n\ (\in \mathbb{N})$ | Number of metabolites |
| $m_i\ (\in \mathbb{N})$ | Number of atoms in $i^{th}$ metabolite |
| $\alpha_{ik}$ | $k^{th}$ atom of $i^{th}$ metabolite |
| $M_n$ | Set of $n$ input metabolite names obtained from KEGG |
| $T$ | Set of metabolite pairs |
| $X_p^{(3)},\ X_q^{(5)},\ X_r^{(7)}$ | Sets of atoms involved in $p^{th}$, $q^{th}$ and $r^{th}$ segments of length three, five and seven, i.e., each segment consisting of three, five and seven atoms respectively |
| $x_p^{(3)},\ x_q^{(5)},\ x_r^{(7)}$ | Sequence of atoms in the $p^{th}$, $q^{th}$ and $r^{th}$ segments of length three, five and seven, i.e., each segment consisting of three, five and seven atoms respectively |
| $A_i^{(3)},\ A_i^{(5)},\ A_i^{(7)}$ | Sets of segments of length three, five and seven, generated from $i$th metabolite |
| $\delta_i\ (\in \mathbb{N})$ | Number of bonds in $i^{th}$ metabolite |
| $\lambda_i\ (\in \mathbb{N})$ | Number of atoms in $i^{th}$ metabolite |
| $\epsilon_{ij}^{(3)},\ \epsilon_{ij}^{(5)},\ \epsilon_{ij}^{(7)}\ (\in \mathbb{N})$ | The number of common three, five and seven-atom segments, respectively, between $i^{th}$ and $j^{th}$ metabolites |
| $\beta_{ij}^{(3)},\ \beta_{ij}^{(5)},\ \beta_{ij}^{(7)}\ (\in \mathbb{R})$ | Standardized score of the number of common three, five and seven-atom segments, respectively, between $i^{th}$ and $j^{th}$ metabolites |
| $\zeta_i\ (\in \mathbb{R})$ | Molecular weight of $i^{th}$ metabolite |
| $\xi_{ij}\ (\in \mathbb{R})$ | Standardized difference in molecular weight between $i^{th}$ and $j^{th}$ metabolites |
| $\omega_{ij}\ (\in \mathbb{R})$ | Final score depicting the similarity between $i^{th}$ and $j^{th}$ metabolites |
| $\mathscr{C}_i^{(1)},\ \mathscr{C}_i^{(2)},\ \mathscr{C}_i^{(3)}$ | Sets of metabolites/compounds having highest, second highest and third highest similarity score values, respectively, with $i^{th}$ metabolite |

metabolite, the corresponding two dimensional structure, in the form of atoms and bonds, has been obtained online from the KEGG KCF (XML format) files, along with its molecular weight. Using this information, the process of segmentation of metabolite has been carried out. We are considering the connections between the atoms in a metabolite as edges.

#### 2.1.2 Segmentation of the metabolites

After accumulation of information, the next stage is segmentation. In a reaction, product metabolite have been formed by integrating multiple segments of two or more reactants. Segments are continuous linear sequence of connected atoms, such that an $n$-atom sequence has $n-1$ bonds. Three, five or seven-atom segments have been considered for representing a metabolite. Some metabolites are so small that a five or seven-atom segment cannot be used represent the metabolite in totality, while they can form segments of size three. For larger metabolites, the seven-atom segments are able to represent the structural similarity in a better way than the three or five-atom segments. Two structurally dissimilar molecules may have common three-atom segments, but the chance of having five-atom or seven-atom segment is comparatively less. In Section 2 of Appendix, it has been mathematically proved that a metabolite can be broken down into multiple 3-atom segments. Joining these three-atom segments will lead to the formation of the original atom. On the other hand, for five and seven-atom segments, one or more atoms may not find its place in any of the segments formed. Hence their amalgamation would not lead to the original 2D structure of the metabolite.

Let us consider $p^{th}$ three-atom segment $x_p^{(3)} = \alpha_{ik-1}\alpha_{ik}\alpha_{ik+1}$, $q^{th}$ five-atom segment $x_q^{(5)} = \alpha_{ik-2}\alpha_{ik-1}\alpha_{ik}\alpha_{ik+1}\alpha_{ik+2}$ and $r^{th}$ seven-atom segment $x_r^{(7)} = \alpha_{ik-3}\alpha_{ik-2}\alpha_{ik-1}\alpha_{ik}\alpha_{ik+1}\alpha_{ik+2}\alpha_{ik+3}$ of $i^{th}$ metabolite. Thus,

$$X_p^{(3)} = \{\alpha_{ik-1}, \alpha_{ik}, \alpha_{ik+1}\}; \tag{1}$$

$$X_q^{(5)} = \{\alpha_{ik-2}, \alpha_{ik-1}, \alpha_{ik}, \alpha_{ik+1}, \alpha_{ik+2}\}; \tag{2}$$

and

$$X_r^{(7)} = \{\alpha_{ik-3}, \alpha_{ik-2}, \alpha_{ik-1}, \alpha_{ik}, \alpha_{ik+1}, \alpha_{ik+2}, \alpha_{ik+3}\}; \tag{3}$$

where $1 \leqslant i \leqslant n$, and $\alpha_{ik}$ is not a terminal atom; $p, q, r \in \mathbb{N}$; $p, q$ and $r = 1, 2, ...$, such that

$$A_i^{(3)} = \{x_p^{(3)} | p \in \mathbb{N}\} \tag{4}$$

$$A_i^{(5)} = \{x_q^{(5)} | q \in \mathbb{N}\} \tag{5}$$

$$A_i^{(7)} = \{x_r^{(7)} | r \in \mathbb{N}\} \tag{6}$$

Two dimensional structure of a metabolite can be depicted in the form of these segments. The segments can be combined to form a larger segment of any length. Initially two bonds with one common atom have been combined to form a three-atom segment. For example, bonds $\alpha_{ik-1}\alpha_{ik}$ and $\alpha_{ik}\alpha_{ik+1}$ have been combined together to form segment $\alpha_{ik-1}\alpha_{ik}\alpha_{ik+1}$, where $\alpha_{ik}$ is the common atom between the bond atoms. Subsequently, two three-atom segments having only one common terminal atom have been concatenated to form a five-atom segment. Likewise, a five-atom segment has been concatenated with a three-atom segment to form a seven-atom segment. For example, consider a certain three-atom segment $\alpha_{ik_1-1}\alpha_{ik_1}\alpha_{ik_1+1}$ and a certain five-atom segment $\alpha_{ik_2-2}\alpha_{ik_2-1}\alpha_{ik_2}\alpha_{ik_2+1}\alpha_{ik_2+2}$. If $k_1 - 1 = k_2 - 2$ or $k_1 - 1 = k_2 + 2$ or $k_1 + 1 = k_2 - 2$ or $k_1 + 1 = k_2 + 2$, these two segments can be concatenated to form a seven-atom segment.

The segments are formed following the rule such that all segments, except the first one should contribute to the addition of only one new atom. Consider a set $F$ containing all the atoms $\alpha_1, \alpha_2, ...\alpha_{m_i}$ of $i^{th}$ metabolite whose segments need to be formed. Let $x_p^{(3)} = \alpha_{ik-1}\alpha_{ik}\alpha_{ik+1}$ be the first continuous segment of length three. Initially, $A_i^{(3)} = \phi$. Since $x_p^{(3)}$ is the first segment formed, $A_i^{(3)}$ is modified as $A_i^{(3)} = A_i^{(3)} \cup \{x_p^{(3)}\}$. The atoms in $X_p^{(3)}$ are

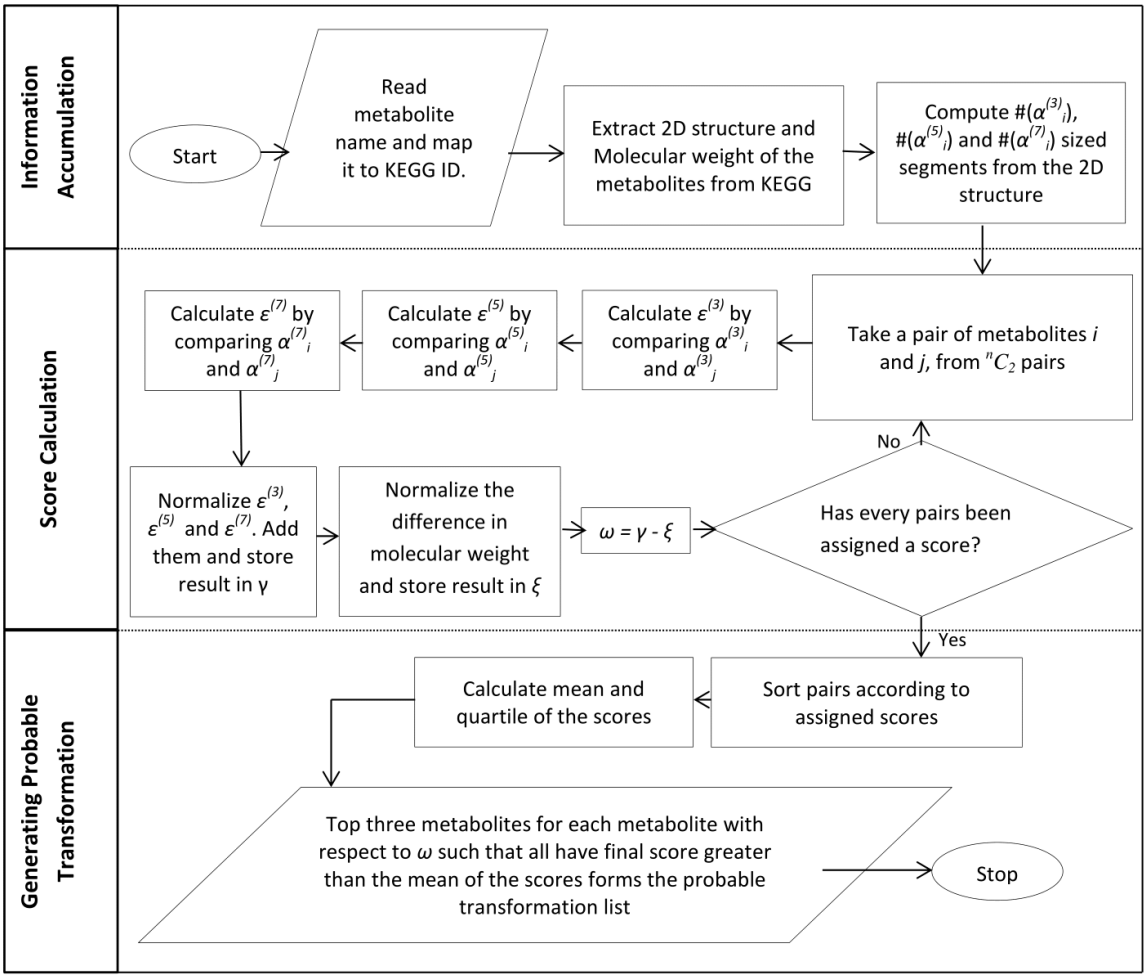



Fig. 3. Flowchart of ASAPP

now removed from $F$. Hence, $F = F - X_p^{(3)}$. The second segment of length three has been formed in a way such that any one of the terminal atoms must be present in $F$ while the other two atoms must not be present in $F$. Let the new segment formed be $x_p^{(3)} = \alpha_{ik}\alpha_{ik+1}\alpha_{ik+2}$. Previously, the atoms $\alpha_{ik-1}, \alpha_{ik}$ and $\alpha_{ik+1}$ were removed from $F$. Comparing the previous and the new segment formed, $\alpha_{ik}$ and $\alpha_{ik+1}$ are common atoms. These two atoms were already removed from $F$. The terminal atom $\alpha_{ik+2}$ is present in $F$. This atom, which is common in the new segment $X_p^{(3)}$ and $F$, has been removed from $F$. Thus, $A_i^{(3)} = A_i^{(3)} \cup x_p^{(3)}$. Hence, for the segments, except the first one, we have

$$F = \begin{cases} F - (F \cap X_p^{(3)}), & \text{if } |F \cap X_p^{(3)}| = 1; \\ F, & \text{otherwise.} \end{cases} \quad (7)$$

Segments have been formed until $F$ becomes empty, and only those segments have been retained, which have led to the removal of only one atom from $F$. Formation of the segments of size five and seven is a tweak of the above rule, such that, $F$ may not be empty even after all possible unique segments are formed.

Consider Figure 2 for an example. There are $m = 10$ atoms in the metabolite, such that, $F = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}\}$, where $\alpha_1$=OH, $\alpha_2$=C,... and so on. We aim at forming three-atom segments initially. The first segment $x_p^{(3)} = \alpha_1 - \alpha_2 - \alpha_3$ is formed such that $X_p^{(3)} = \{\alpha_1, \alpha_2, \alpha_3\}$. Initially, $A_i^{(3)} = \phi$. Since it is the first segment, $A_i^{(3)} = A_i^{(3)} \cup \{x_p^{(3)}\}$. The atoms in the segments are removed from $F$. New $F$ becomes $F = \{\alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}\}$. Let the next segment formed be $x_p^{(3)} = \alpha_2 - \alpha_3 - \alpha_5$ such that $X_p^{(3)} = \{\alpha_2, \alpha_3, \alpha_5\}$. According to the rule, $\alpha_5$ is the only atom that is common in both $X_p^{(3)}$ and $F$, hence $\alpha_5$ is removed from $F$, leading to $F = \{\alpha_4, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}\}$ and $A_i^{(3)} = A_i^{(3)} \cup \{x_p^{(3)}\}$. Repeating the previous operation, the next segment formed is $x_p^{(3)} = \alpha_4 - \alpha_3 - \alpha_5$ such that $X_p^{(3)} = \{\alpha_3, \alpha_4, \alpha_5\}$. According to the rule, $\alpha_4$ is the only atom that is common in $X_p^{(3)}$ and $F$, hence $\alpha_4$ is removed from $F$, leading to $F = \{\alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}\}$ and $X_p^{(3)}$ is retained. Suppose the next segment formed is $\alpha_2 - \alpha_3 - \alpha_4$ such that $X_p^{(3)} = \{\alpha_2, \alpha_3, \alpha_4\}$. According to the rule, since $|F \cap X_p^{(3)}| \neq 1$, no deduction is performed in this step and $X_p^{(3)}$ is discarded. In this particular example, five-atom segments can be formed so that $F$ becomes empty at the end. During the formation of seven-atom segment, the atom

$\alpha_4$ remains in $F$ even after all the seven-atom segments have been formed. No seven-atom continuous segment containing the atom $\alpha_4$ can be formed. The next step is to find the similarity between pairs of metabolites in terms of common segment count.

#### 2.1.3 Computing similarity between a pair of metabolites

Following the process of segmentation, the next step is to quantify the similarity between a pair of $i^{th}$ and $j^{th}$ metabolites. Considering a pair of metabolites, the number of common three-atom, five-atom and seven-atom segments between these are counted as follows:

$$\epsilon_{ij}^{(l)} = |A_i^{(l)} \cap A_j^{(l)}|, l = 3, 5, 7 \tag{8}$$

The score $\beta_{ij}^{(l)}$, corresponding to $\epsilon_{ij}^{(l)}$ $(l = 3, 5, 7)$, has been obtained by standardizing the number of common segments as

$$\beta_{ij}^{(l)} = \frac{\epsilon_{ij}^{(l)}}{|A_i^{(l)} \cup A_j^{(l)}| + \delta_i + \delta_j + m_i + m_j} \tag{9}$$

Molecules of metabolites are of varying sizes. Hence, the count of common segments requires standardization. The scores have been standardized based on the complete structure of each metabolite. The similarity between two metabolites depends primarily on four factors:

1) Number of three-atom common segments between two metabolites.
2) Number of five-atom common segments between two metabolites.
3) Number of seven-atom common segments between two metabolites.
4) Difference in molecular weight of two metabolites.

The similarity score between a pair of metabolites has been found to increase with the number of common three-atom, five-atom and seven-atom segments. Higher the number of matched segments, higher is the structural similarity between a pair of metabolites. Factor 4 above has been found to have an inverse association with the similarity score. For most of the metabolites, it has been noticed that closer the two dimensional structures of two metabolites, lower is the difference in the molecular weights. The standardized difference in molecular weights has been considered as a contributing factor for computing the similarity scores, and is defined as:

$$\xi_{ij} = \frac{abs(\zeta_i - \zeta_j)}{\delta_i + \delta_j} \tag{10}$$

Thus, the final score for each metabolite pair is

$$\omega_{ij} = \beta_{ij}^{(3)} + \beta_{ij}^{(5)} + \beta_{ij}^{(7)} - \xi_{ij} \tag{11}$$

#### 2.1.4 Probable transformations

The metabolite pairs have been sorted in descending order of their final scores $\omega_{ij}$, from highly probable to highly improbable transformation pairs. Mean, quartile and triplets have been used as the threshold values to isolate the probable transformations from the improbable ones. Using mean, the set of probable transformations are:

$$feasible\_pair = \{(\mathscr{C}_i, \mathscr{C}_j) | \omega_{ij} > \frac{\sum_{(i,j)=(1,1)}^{(n,n)} \omega_{ij}}{\binom{n}{2}}, i \neq j\} \tag{12}$$

The third quartile of the scores has been computed as another threshold value. For each metabolite, three metabolites (other than the metabolite under consideration) have been filtered on the basis of similarity scores which has the maximum resemblance with the metabolite under consideration. The similarity score for $i^{th}$ metabolite with the rest of the metabolites in the list have been sorted as:

$$\omega_{ij_1} \leq \omega_{ij_2} \leq \omega_{ij_3} \ldots \leq \omega_{ij_{n-1}}, i \neq j \tag{13}$$

Three metabolites having the highest similarity values with the $i^{th}$ metabolite, are extracted as follows:

$$\mathscr{C}_i^{(1)} = \{\mathscr{C}_j | score(\mathscr{C}_j) = \omega_{ij_{n-1}}, i \neq j\} \tag{14}$$

$$\mathscr{C}_i^{(2)} = \{\mathscr{C}_j | score(\mathscr{C}_j) = \omega_{ij_{n-2}}, i \neq j\} \tag{15}$$

$$\mathscr{C}_i^{(3)} = \{\mathscr{C}_j | score(\mathscr{C}_j) = \omega_{ij_{n-3}}, i \neq j\} \tag{16}$$

Here $\mathscr{C}_i^{(1)}$ stands for the metabolite with maximum similarity to the $i^{th}$ metabolite, $\mathscr{C}_i^{(2)}$ designates the metabolite with next best similarity with respect to $\mathscr{C}_i^{(1)}$, and $\mathscr{C}_i^{(3)}$ denotes the metabolite with the next to next best similarity with respect to $\mathscr{C}_i^{(1)}$. Due to the better performance of triplet method, the final list of transformations for $i^{th}$ metabolite is $\mathscr{C}_i^{(1)}$, $\mathscr{C}_i^{(2)}$, and $\mathscr{C}_i^{(3)}$ respectively.

**Algorithm 1** ASAPP

Procedure $ASAPP$
$n \leftarrow nom$
Perform initialization
**while** $i \leq n$
  Compute all possible unique 3,5 and 7-atom segments and store
  them in $A_i^{(l)}$ where $l = 3, 5, 7$.
**while** $i \leq n$
  $j \leftarrow$ i + 1
  **while** $j \leq n$
    Compute the number of common sized segments in $A_i^{(l)}$
    and $A_j^{(l)}$ and store the value in $\epsilon_{ij}^{(l)}$.
Standardize the common segment count $\epsilon_{ij}^{(l)}$ as $\beta_{ij}^{(l)}$.
Compute the segment score $\gamma_{ij}$ by summing $\beta_{ij}^{(l)}$.
Calculate the effect of molecular weight $\xi_{ij}$ $(\frac{abs(\zeta_i - \zeta_j)}{\delta_i + \delta_j})$
Generate the final score $\omega_{ij}$ $\left(\gamma_{ij} - \xi_{ij}\right)$
Sort $\omega_{ij}$ in descending order. Find the mean value of $\omega$
**while** $i \leq n$
  Prune $\omega$ for 3 metabolite with maximum similarity to $i$th metabolite
  Discard metabolites having $\omega_{ij}$ greater than the mean of $\omega$.
Output probable transformations

As a precautionary measure to ensure that unnecessary transformations are not reported, we have used the combined mean and the triplet parameters to generate the probable list of transformations. After the top three metabolites have been obtained based on the final similarity score $\omega_{ij}$, these metabolites are filtered using the mean value $\left(\sum_{\substack{i,j=1 \\ i\neq j}}^{n} \omega_{ij}/\binom{n}{2}\right)$. The metabolites which have scores greater than the mean value, are taken into consideration and the rest are discarded.

The overall complexity of ASAPP is $O(n^2x^2)$, where $x$ is the maximum number of atoms in a metabolite and $n$ is the total number of metabolites. The detailed complexity estimation has been given in Appendix Section 1. The mathematical validation of ASAPP has been given in Appendix Section 2. Algorithm 1 describes the step-wise computation of ASAPP.

## 3 Results

In this section, we shall describe how the algorithm has been applied to predict possible transformations in multiple crucial carbohydrates, lipid/fat and amino acid metabolic pathways. We have compared our results with the already established sets of transformations in KEGG.

### 3.1 Performance Comparison

ASAPP has been applied on 41 pathways (Table S1-S3 in Supplementary Information) involving 782 metabolites and 17556 transformation pairs as enlisted in KEGG. In order to analyze the performance of ASAPP, we have considered the transformations not enlisted in KEGG as not occurring at all. Such a consideration may not be correct since the presence of appropriate (yet unknown) enzymes may lead to the occurrence of such transformations. The summary of the performance measures has been depicted in Table 2.

A detailed description of the performance of three groups (carbohydrate, lipid/fat and amino acid) have been given in the Supplementary Information Figures S1-S3. Among the carbohydrate metabolic pathways, amino sugar and nucleotide sugar metabolism pathway has obtained the highest accuracy of 95.22%. Similarly, among the lipid pathways, primary bile biosynthesis ASAPP has achieved the highest accuracy of 93.98%. Finally, among the amino acid metabolism pathway, tryptophan metabolism has obtained the highest accuracy of 93.07%.

Considering all the pathways, a trade-off has been noticed among the accuracy, sensitivity and specificity (Table 2). When using the mean value of scores as a threshold, a high sensitivity but a low accuracy and specificity have been noticed, while on the other hand, the triplet method, a high accuracy and specificity, and a low sensitivity have been found. The quartile method has an average performance. Considering the three performance measures, we have chosen triplet method for prediction since it has given better performance in terms of accuracy and specificity, and has generated the least number of false positives.

Figure 4 shows the flow of synthesis and degradation of ketone bodies pathway formation using ASAPP involving 6 metabolites. The algorithm starts with one single compound. The initial metabolite considered here is $a$. High scores obtained by $a$ is with the metabolites $b$ ($\omega_{a,b}$=0.310) and $c$ ($\omega_{a,c}$=0.301). Hence, we have obtained two new metabolites from $a$. Considering the newly obtained metabolite b, high score obtained is with $a$ ($\omega_{a,b}$=0.310) and $c$ ($\omega_{b,c}$=0.2888). Since $a$ already exists in the pathway, the transition from $b$ to $c$ is added. Considering the newly obtained metabolite $c$, the high scores obtained are $b$ ($\omega_{b,c}$=0.2888), $a$ ($\omega_{a,c}$=0.301), and $d$ ($\omega_{c,d}$=0.0960). Since $a$ and $b$ are already in the pathway, $d$ is added to the existing pathway and a transition is made from $c$ to $d$. With metabolite $d$, the high score obtained is with $c$ ($\omega_{c,d}$=0.2960), $e$ ($\omega_{d,e}$=0.1851) and $f$ ($\omega_{d,f}$=0.2326). Metabolite $c$ is already in the pathway, $e$ and $f$ are now added. Apart from the above mentioned pathway, the formation of six other pathways (alpha linoleic acid metabolism, linoleic acid metabolism, glycolysis pathway, TCA cycle, alanine aspartite and glutamate metabolism, and valine, leucine and isoleucine biosynthesis) have been depicted in the Supplementary (Figure S1-S6).

For a particular pathway, if the scores of most of the transformations are close to each other, then it can be concluded that the pathway constitutes structurally similar metabolites. Considering the alpha linoelic acid metabolism pathway (Figure S1) under the group of amino acid metabolism, it has been seen that apart from the transition between the molecule no. 24 (Traumatic acid) and 25 ((9Z,15Z)-(13S)-12,13-Epoxyoctadeca-9,11,15-trienoic acid), other transformations are associated with similar score among themselves, ranging from 0.265 to 0.294 (short interval) indicating that the compounds involved in this pathway are structurally similar to each other.

### 3.2 Application of ASAPP in the field of host-pathogen interactions

Toxins are substances created by plants and animals that are poisonous to humans. These toxins, once in the body of the host, intervene with the normal functioning of the metabolism of the host [14] (Supplementary Information Section 3). Pathogen liberated toxins have been seen to have a spectrum of upshots on their hosts. The transformation mechanism of natural toxins need to be studied in details as these help in proper drug designing ( [15]).

The two dimensional structural similarity of the toxins with the metabolites are of significance and needs to be examined. Consider a simple pathway consisting of the transformations A $\rightarrow$ B, B $\rightarrow$ C, and C $\rightarrow$ D. Consider a toxin X having high similarity with the metabolite B. Occurrence of unknown reaction may block the transformation of B $\rightarrow$ C. The other metabolites which react with B to produce C may as well, due to structural similarity and in the presence of appropriate enzyme, react with X to produce a different metabolite which is not C. Besides, if B is structurally similar to X, B can transform to X in the presence of appropriate

TABLE 2
Performance comparison of various thresholding methods used in ASAPP

| Performance measures | Mean | Quartile | Triplet |
|---|---|---|---|
| Accuracy | 45.95 | 74.45 | 84.20 |
| Sensitivity | 79.80 | 49.14 | 29.00 |
| Specificity | 43.14 | 75.77 | 86.13 |

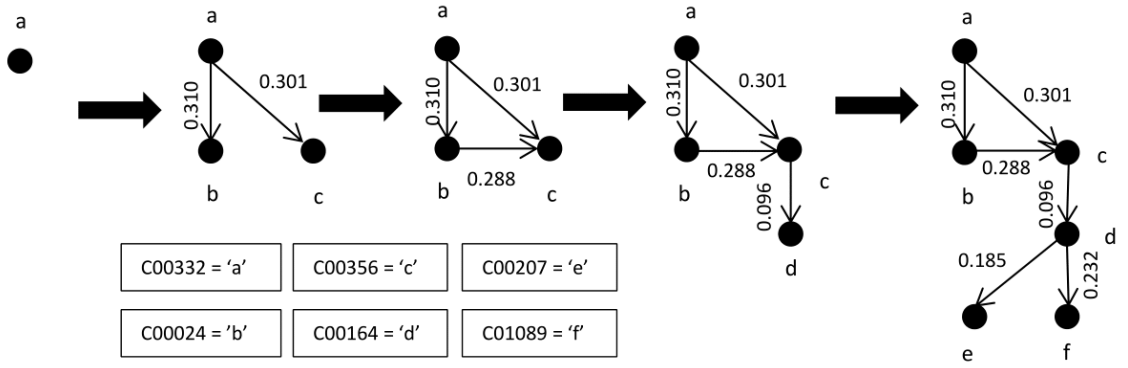



Fig. 4. Step-by-step formation of the synthesis and degradation of ketone bodies pathway using ASAPP. $a$ (Acetoacetyl-CoA), $b$ (Acetyl-CoA), $c$ (Hydroxymethylglutaryl-CoA), $d$ (Acetoacetate), $e$ (Acetone) and $f$ ((R)-3-Hydroxybutanoate) are the compounds whose corresponding KEGG IDs are given. In each time step, one compound, whose transformations have not been considered previously and which is a recent addition to the pathway, is considered for finding the transformations related to that compound.

enzyme and other metabolites. As soon as X is produced, the other metabolites, A, C, and D have a chance of reacting with X in the presence of the appropriate enzymes and thus breaking the pathway. The summary of the probable toxin transformations to/from metabolites from KEGG have been documented in Supplementary Information Table S4.

### 3.3 Prediction of possible pathway breaks due to the presence of toxins

We have executed ASAPP on the metabolites involved in the Glycolysis and the TCA cycle. We have considered 52 toxins from KEGG, one toxin from each of the categories of toxin. None of these toxins have any reported set of reactions in KEGG. For each of these toxins, we have predicted the consequence of its presence in the glycolysis (Figure 5 a) and the TCA cycle (Figure 5 b).

Considering Glycolysis pathway metabolite acetyl-coa (C00024), thiamin diphosphate (C00068) and s-acetyldihydrolipoyllysine (C16255) have the maximum chance of being the breakpoints of the pathway as depicted in Figure 5 a and Table 3. For example, toxin Anisatin (C09294) has high structural similarity with the metabolite beta-D-Fructose 1,6-bisphosphate (C05378). In presence of this toxin and appropriate enzyme, the metabolites that reacted with beta-D-Fructose 1,6-bisphosphate (C05378) to form D-Glyceraldehyde 3-phosphate (C00118) or Glycerone phosphate (C00111) may react with the toxin to produce unknown compounds in such a way that pathways can redirect from its usual course. Among the rest of the metabolites, 1,3-bisphospho-d-glycerate (C00236), 2,3-bisphospho-d-glycerate (C01159), pyruvate (C00022), l-lactate (C00186), acetate (C00033), acetaldehyde (C00084), and ethanol (C00469) have been observed to have the least chance of being the breakpoints, *i.e.*, the pathway has a high chance of not getting perturbed at these points.

TABLE 3
Toxins having structural similarity with the metabolites of Glycolysis

| KEGG Compound | KEGG Toxin |
|---|---|
| Thiamin diphosphate (C00068) | Brucine (C09084), Echimidine (C10299), Cylindrospermopsin (C19999), Gonyautoxin 1 (C16855), Philanthotoxin (C20052), Arenobufagin (C20035) |
| Acetyl-CoA (C00024) | alpha-Chaconine (C10796), Nodularin (C15713), Okadaic acid (C01945), Brevetoxin A (C16839), Azaspiracid (C16907) |
| S-acetyldihydrolipoyllysine (C16255) | alpha-Chaconine (C10796),Nodularin (C15713), Okadaic acid (C01945), Azaspiracid (C16907), Cephalostatin 1 (C20060) |

Considering the TCA cycle, the possible breakpoint metabolites are acetyl-coa (C00024), S-acetyldihydrolipoyllysine (C16255), succinyl-coa (C00091) and s-succinyldihydrolipoyllysine (C16254) as depicted in Figure 5 b and Table 4. Among the rest of the metabolites, pyruvate (C00022) and fumarate (C00122) have the least possibility of being the breakpoints. Further analysis leads to finding that the toxin azaspiracid (C16907) has the maximum likelihood to affect the Glycolysis and the TCA cycle as azaspiracid (C16907) has a high structural similarity with s-succinyldihydrolipoyllysine (C16254). Closely following these two toxins are the toxins nodularin (C15713) and okadaic acid (C01945), which too have a high chance of disrupting the pathways. A detailed result of the presence of toxin in the pathways have been documented

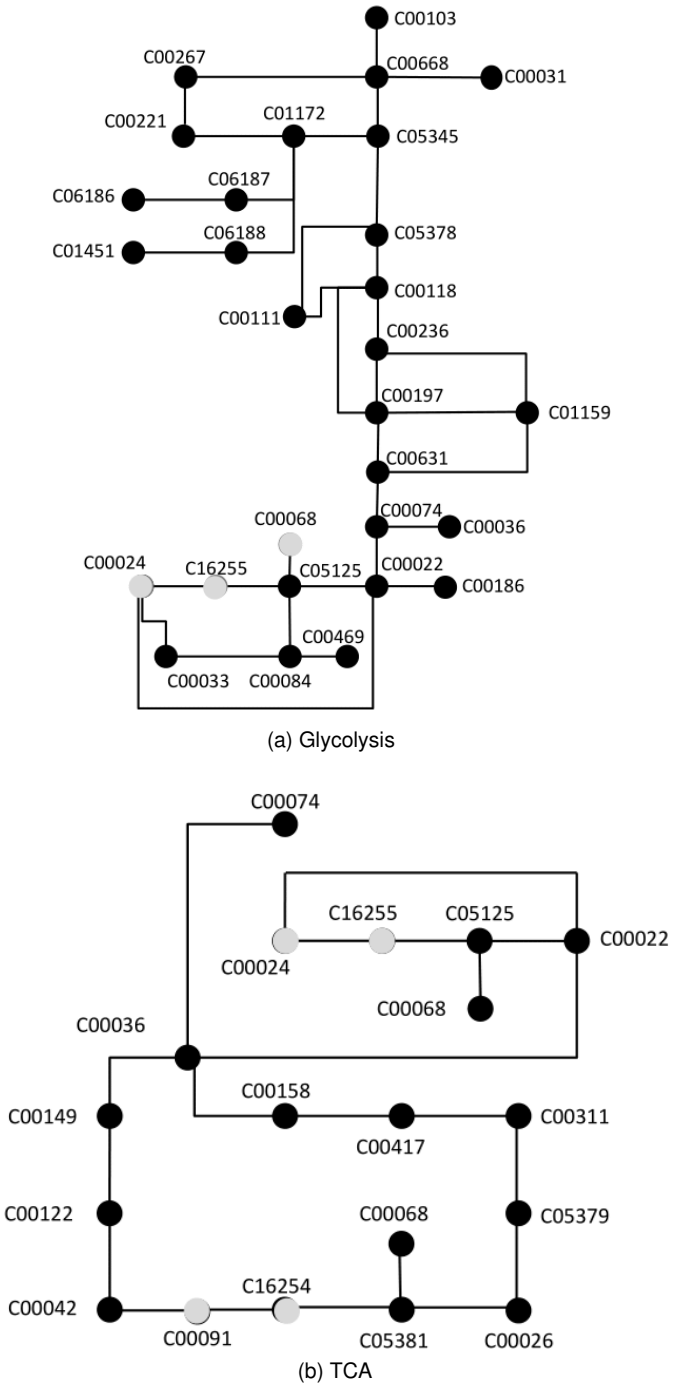



Fig. 5. The pathway models depicting the transformations within the (a) Glycolysis and (b) TCA pathway. The gray dots represent the breakpoints in the pathway. The black dots signify other metabolites which have a lower probability of being the breakpoints in the pathway.

in the Supplementary Information Table S4.

### 3.4 Analysis of ASAPP with respect to other tools

There exist several pathway prediction tools, viz., PathoLogic [3], PathMiner [4], Pathway hunter [5], Um-PPS [8], and Rahnuma [9]. However, ASAPP has a different aim compared to these methods. Besides, it has been noticed that apart from PathPred, none of the other tools are publicly available. Although PathPred is available, there is a fundamental difference between the functionality of PathPred and ASAPP. PathPred takes in one metabolite as input and finds all the pathways involving that metabolite from KEGG database. PathPred does not predict a new pathway. ASAPP, on the other hand, takes a group of metabolites as inputs and predict possible pathways involving them. Moreover, unlike ASAPP, the functionality of PathPred is limited to only xenobiotic pathways. Unavailability of the prediction tools and the limited functionality of PathPred make ASAPP more significant and the need for its availability crucial. A summary of the analysis of ASAPP with respect to the other tools has been given in Table 5.

TABLE 4
Toxins having structural similarity with the metabolites in the TCA cycle

| KEGG Compound | KEGG Toxin |
|---|---|
| Acetyl-CoA (C00024) | alpha-Chaconine (C10796), Nodularin (C15713), Okadaic acid (C01945), Brevetoxin A (C16839), Azaspiracid (C16907) |
| S-acetyldihydrolipoyllysine (C16255) | alpha-Chaconine (C10796), Nodularin (C15713), Okadaic acid (C01945), Azaspiracid (C16907), Cephalostatin 1 (C20060) |
| S-succinyldihydrolipoyllysine (C16254) | alpha-Chaconine (C10796), Pectenotoxin 1 (C16871), Brevetoxin A (C16839), Azaspiracid (C16907), Cephalostatin 1 (C20060) |

Prediction of host-pathogen interactions has been done at the population level [16], gene-level [17] and protein-level [18] [19] [20] [21]. At the population level, the statistics of the population of pathogen species interacting with host species are taken into consideration to predict novel interactions between a new pathogen and a host species [16]. At the gene-level, the pair of genes, one from the host and the other from the pathogen, is predicted to be interacting [17]. Host-pathogen interactions at the protein level are well studied. Host proteins, which interact with pathogen proteins, are predicted [19]. However, none has been done on the basis of metabolites and disruption of pathways. A summary of the analysis of ASAPP in the domain of host-pathogen interactions has been given in Table 6. ASAPP is one of a kind tool using which one can predict the probable pathway breaks in the host due to toxins from pathogens.

## 4 Conclusion

We have developed a novel algorithm ASAPP (Architectural Similarity-based Automated Pathway Prediction), which predicts biochemical transformations from the 2D structure of metabolites. We have predicted the chance of a transformation of one metabolite to another, depending on the two dimensional structural similarity among the metabolites and the difference in their molecular weights. Depending on these factors, we have given a score to each transformation and applied various thresholding policies to determine the final list of probable transformations. Unlike other similar tools for pathway prediction, ASAPP has been made publicly available at http://asapp.droppages.com/.

TABLE 5
Comparative analysis of ASAPP with some existing tools

| Tool name | Aim | Input | Output | Application domain | Web Availability |
|---|---|---|---|---|---|
| ASAPP | Predict possible pathway (linear/non-linear) among them | List of metabolites | Pathway with all the given metabolites predicted | All metabolic pathways | Available |
| PathoLogic [3] | Creating pathway genome database (PGDB) file | Annotated genome of an organism | PGDB file | Xenobiotic pathways | Unavailable |
| PathMiner [4] | Find linear path between these two compounds from KEGG | Initial metabolite, final metabolite in SMILES format | Linear pathway | Xenobiotic pathways | Unavailable |
| Pathway hunter [5] | Find shortest path between two metabolites using KEGG pathway information | Two metabolites in SMILES format | Shortest linear pathway | Xenobiotic pathways | Unavailable |
| PathPred [7] | Predict all pathways in which that metabolite is present from KEGG | One metabolite | Set of pathways | Xenobiotic pathways | Available |
| Um-PPS [8] | Recognize functional group in metabolite and apply group to group transformation as enlisted in UM-BBD database | One metabolite, draw the metabolite on MarvinView Java applet | Predict all pathways in which that metabolite is present | Xenobiotic pathways | Unavailable |
| Rahnuma [9] | Predict pathways using the metabolites from KEGG | KEGG pathways, metabolites | Pathways in which the metabolites occur | Bio-degradation pathways | Unavailable |

TABLE 6
Analysis of prediction systems in the domain of host-pathogen interactions

| Tool Description | Level | Aim |
|---|---|---|
| ASAPP | Metabolite | Predict possible pathway breaks due to toxins produced by pathogens |
| Dallas et. al. [16] | Population | Predict connections between host species and pathogen species on a population level |
| Reid et. al. [17] | Gene | Predict genes involved in host-pathogen interactions |
| Alguwaizani et. al. [18] | Protein | Predict unknown PPI |
| Mariano et. al. [19] | Protein | Predict unknown PPI |
| Nourani et. al. [20] | Protein | Predict unknown PPI |
| Dyer et. al. [21] | Protein | Predict unknown PPI |

By *in silico* analysis, we have shown how the presence of toxin in the host body may adversely affect its metabolic pathways. Here, we have predicted the outcome of 52 such toxins on the Glycolysis pathway and the TCA cycle. The effect of toxins on other pathways still needs to be explored. The field of host-pathogen interactions is emerging as a crucial area of infectious disease research in the post-genomic era. It is a budding research field where new discoveries are getting announced almost each day throughout the globe. The discovery of the dynamics of pathway perturbation during host-pathogen interactions will aptly facilitate further development in the field of discovering new drugs and new therapies for different diseases. Likewise, pathway perturbation is a crucial aspect of pathogen infection. Hence, further study in this field is needed in future.

Appendices are given in separate file titled Appendix.pdf.

## REFERENCES


[1] R. Sen, L. Nayak, and R. De, "A review on host–pathogen interactions: classification and prediction," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 35, no. 10, pp. 1581–1599, 2016.

[2] E. Oswald, J.-P. Nougayrède, F. Taieb, and M. Sugai, "Bacterial toxins that modulate host cell-cycle progression," *Current opinion in microbiology*, vol. 8, no. 1, pp. 83–91, 2005.

[3] P. D. Karp, S. Paley, and P. Romero, "The pathway tools software," *Bioinformatics*, vol. 18, no. suppl 1, pp. S225–S232, 2002.

[4] D. C. McShan, S. Rao, and I. Shah, "PathMiner: predicting metabolic pathways by heuristic search," *Bioinformatics*, vol. 19, no. 13, pp. 1692–1698, 2003.

[5] S. A. Rahman, P. Advani, R. Schunk, R. Schrader, and D. Schomburg, "Metabolic pathway analysis web service (pathway hunter tool at cubic)," *Bioinformatics*, vol. 21, no. 7, pp. 1189–1193, 2005.

[6] M. Oh, T. Yamada, M. Hattori, S. Goto, and M. Kanehisa, "Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways," *Journal of Chemical Information and Modeling*, vol. 47, no. 4, pp. 1702–1712, 2007.

[7] Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto, and M. Kanehisa, "PathPred: an enzyme-catalyzed metabolic pathway prediction server," *Nucleic Acids Research*, p. gkq318, 2010.

[8] L. B. Ellis, J. Gao, K. Fenner, and L. P. Wackett, "The university of minnesota pathway prediction system: predicting metabolic logic," *Nucleic Acids Research*, vol. 36, no. suppl 2, pp. W427–W432, 2008.

[9] A. Mithani, G. M. Preston, and J. Hein, "Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison," *Bioinformatics*, vol. 25, no. 14, pp. 1831–1832, 2009.

[10] S. Tagore and R. K. De, "SAGPAR: Structural grammar-based automated pathway reconstruction," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 4, no. 2, pp. 116–127, 2012.

[11] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "Inchi, the iupac international chemical identifier," *Journal of Cheminformatics*, vol. 7, no. 1, p. 1, 2015.

[12] J. Gao, L. B. Ellis, and L. P. Wackett, "The university of minnesota biocatalysis/biodegradation database: improving public access," *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D488–D491, 2010.

[13] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[14] H. S. Hussein and J. M. Brasel, "Toxicity, metabolism, and impact of mycotoxins on humans and animals," *Toxicology*, vol. 167, no. 2, pp. 101–134, 2001.

[15] X. Yang, J. Tong, L. Guo, Z. Qian, Q. Chen, R. Qi, and Y. Qiu,

"Bundling potent natural toxin cantharidin within platinum (iv) prodrugs for liposome drug delivery and effective malignant neuroblastoma treatment," *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 13, no. 1, pp. 287–296, 2017.
[16] T. Dallas, A. W. Park, and J. M. Drake, "Predicting cryptic links in host-parasite networks," *PLoS computational biology*, vol. 13, no. 5, p. e1005557, 2017.
[17] A. J. Reid and M. Berriman, "Genes involved in host–parasite interactions can be revealed by their correlated expression," *Nucleic acids research*, vol. 41, no. 3, pp. 1508–1518, 2012.
[18] S. Alguwaizani, B. Park, X. Zhou, D.-S. Huang, and K. Han, "Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids," *Journal of Healthcare Engineering*, vol. 2018, 2018.
[19] R. Mariano and S. Wuchty, "Structure-based prediction of host–pathogen protein interactions," *Current opinion in structural biology*, vol. 44, pp. 119–124, 2017.
[20] E. Nourani, F. Khunjush, and S. Durmuş, "Computational approaches for prediction of pathogen-host protein-protein interactions," *Frontiers in microbiology*, vol. 6, p. 94, 2015.
[21] M. D. Dyer, T. Murali, and B. W. Sobral, "Computational prediction of host-pathogen protein–protein interactions," *Bioinformatics*, vol. 23, no. 13, pp. i159–i166, 2007.

**Rishika Sen** received her BSc and MSc degrees in Computer Science from University of Calcutta, India, in 2012 and 2014 respectively. She is currently working towards the Doctorate Degree at Machine Intelligence Unit in Indian Statistical Institute, Kolkata, India. Her current research interest includes bioinformatics,computational biology and machine learning.

**Somnath Tagore** has received his BSc degree from University of Calcutta in 2003, MSc degree from Manipal University in 2005, MTech degree from West Bengal University of Technology in 2007 and PhD(Engg) engineering from ISI (Jadavpur Univ) in 2014. He was a Post-doctoral fellow at Cancer Genomics and BioComputing Lab, The Faculty of Medicine, Bar-Ilan University, Safed, Israel. Currently, he is working as a Research Scientist at Califano Laboratory of Systems Biology, Columbia University Medical Center, Herbert Irving Cancer Research Center, New York, USA. His current research interests include Systems Biology, Network Medicine, Infectious disease modeling, Cancer metabolomics, Data mining, In-silico metabolic engineering, Algorithms, Graphs and Optimization.

**Rajat K. De** Rajat K. De is a Professor working at Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. He obtained his Ph.D. degree from the same Institute in 2000. He was a Distinguished Postdoctoral Fellow at the Whitaker Biomedical Engineering Institute, Johns Hopkins University, USA, during 2002-2003. During the last 15 years, Professor De has been working in the area of bioinformatics and in silico systems biology. Recently, he has started working on Big Data Analytics and Deep Learning in the domain of bioinformatics, systems biology and healthcare. Professor De visited the Department of Medicine, University of California, San Diego, in 2017 and 2018, with a Fulbright-Nehru Academic and Professional Excellence Fellowship. He has published about 90 research papers in international journals, conference proceedings and in edited books, and co-edited three books.