Interdiscip Sci Comput Life Sci (2012) 4: 116–127
DOI: 10.1007/s12539-012-0119-8

SAGPAR: Structural Grammar-based Automated Pathway Reconstruction

Somnath TAGORE[1], Rajat K. DE[2*]
[1](Department of Biotechnology and Bioinformatics, Dr DY Patil University, Navi Mumbai 400614, India)
[2](Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India)

Received 5 January 2011 / Revised 5 July 2011 / Accepted 3 September 2011

Abstract: *In-silico* metabolic engineering is a very useful branch of systems biology for modeling, analysis and prediction of various outcomes of metabolic pathways. It can also be used for detecting interactions and dynamics within a network. Various protocols have been proposed for modeling a pathway. But most of these protocols have various disadvantages and shortcomings with respect to automated pathway modeling and analysis. In the present article, we have proposed a novel algorithm for automated pathway reconstruction. We have also made a comparative study of our algorithm with other standard protocols and discussed its advantages over others. We present StructurAl Grammar-based automated PAthway Reconstruction (SAGPAR), a fast and robust algorithm that generates any metabolic pathway using some given structural representations of metabolites. Users can model any pathway based on some pre-required features that are asked as an input by the algorithm. The algorithm also takes into considerations various thermodynamic thresholds and structural properties while modeling a pathway. The given algorithm has been tested on the standard pathway datasets of 25 pathways of *Mycoplasma pneumoniae M129* and 24 pathways of *Homo sapiens*. The dataset is taken from KEGG and PubChem Compound data repositories. SAGPAR performs much better than some already present metabolic pathway analysis tools like Copasi, PHT, Gepasi, Jarnac and Path-A.
Key words: boolean, connectivities, graph, lavenshtein distance, perturbation, similarity, SMILES, topological index.

1 Introduction

Metabolic pathways consist of a series of bio-chemical reactions that occur within a cell, catalyzed by enzymes resulting in the formation of a metabolic product. A suitable model of a metabolic pathway is developed to understand and visualize networks, analyze various enzymes involved within the pathways, study the gene expression levels, and to analyze the change in product output with respect to initial reactants (Klipp, 2005). A well-developed model can also be used to predict the outcomes of various alterations made to the cells and can identify intracellular targets for drugs and for genetic engineering (Palsson, 2006).

The concept of synthetic modeling can be traced back to early 1980s when a system was assumed as a black box with an input, and a certain output (Periwal *et al.*, 2006). Thus, properties of various systems can be studied using a model or a schema. Furthermore, a system is a collection of interrelated objects that in turn consist of some elemental unit upon which observations can be made. Thus, systems are anything humans wish to discuss and models are tools that facilitate the discussion (Kitano, 2001). One of the branches of synthetic modeling is metabolic pathway modeling, where hypothetical models are proposed for predicting the outcome of a certain input and studying its effectiveness over time (Westerhoff *et al.*, 2005).

Only a certain amount of progress has been done for automated pathway reconstruction. Automated reconstruction is an *ab-initio* approach where initial input for modeling a pathway is either raw or incomplete (in case of diseased pathways) (Periwal *et al.*, 2006; Palsson, 2006). Also, given a set of metabolites it is quite cumbersome to predict their relations. This is because biological moieties are extremely complex and may have more than one relation. Thus, one of the preliminary steps for automated pathway reconstruction is finding the criterion for relating two moieties so that they can be linked together (Klipp, 2005).

We have used the concepts of path mining for linking these biological moieties and thereby reconstruct-

*Corresponding author.
E-mail: rajat@isical.ac.in

ing the pathway. Path mining is the application of graphs in metabolic engineering for building hypothetical models and for analyzing topological parameters in metabolome networks. Here, mining can be based on structural grammars, keys, patterns and indexing. We have taken the concept of grammars and keys for automated pathway reconstruction (Oltvai *et al.*, 2004). We assume that given a set of metabolite, the probability that one would be converted to another, thus giving rise to a reaction link, is higher if they are structurally similar. Now, there should be some standard nomenclature for predicting the structural similarity among two biological moieties. In this work, we use the concept of SMILES string representation for representing the moieties (Navarro, 2001). Here, we propose an algorithm called structural grammar-based automated pathway reconstruction (SAGPAR), that can automatically reconstruct the complete pathway with reaction links among the set of metabolites, given the metabolites.

We have taken the SMILES representation for metabolites as input to our algorithm. The next step for pathway reconstruction is predicting reaction links among this given set of inputs. For this we have used the concept of topological indices, used mainly in pharmacokinetics analysis (Balaban *et al.*, 2007). These indices predict the similarity and dissimilarity among moieties, given some initial inputs. But, another important concept that we should keep in mind while handling biological moieties is that these are naturally occurring and tend to behave differently while in natural environment (Boncher, 1983). Also, whatever model we propose may not be valid when actually imposed. Thus, while using these topological indices we also need to take into account some physiological measures and some pattern based measures for modeling pathways.

We have studied the effectiveness of our algorithm over both large and small size networks. Thus, the organisms whose information has been used are *Mycoplasma pneumoniae M129* and *Homo sapiens*. *Mycoplasma pneumoniae M129* has been selected as the metabolite datasets of its pathways are easily available in standard repositories like Kyoto Encyclopedia for Genes and Genomes (KEGG) (Kanehisa *et al.*, 2004). The entire genome of this bacterium has been sequenced. The modeling is done on 25 pathways of *M. pneumoniae M129*. Similarly, *H. sapiens* has been selected to test our method over large datasets. The modeling has been done on 24 pathways of *H. sapiens*. The metabolites involved in the pathways were collected from Kyoto Encyclopedia for Genes and Genomes (KEGG) (Kanehisa *et al.*, 2004) and the SMILES string of the metabolites were gathered from PubChem Compound database (Shivakumar *et al.*, 1995; Assenov *et al.*, 2008) respectively.

2 StructurAl Grammar-based automated PAthway Reconstruction (SAGPAR)

Here we develop an algorithm for construction of a metabolic pathway from the set of metabolites, using some structural features, symmetry and thermodynamic properties of the metabolites present in the pathway. We consider structures of biomolecules in terms of graphs and compute various similarity-based measures that directly compare the topology and properties of the graphs. All biomolecules (in our case, in the metabolic pathways of *M. pneumonia M129* and *H. sapiens*) can be represented in terms of certain standard notations and formulae. We have considered the structural formula-based representations of these biomolecules, viz., SMILES, *i.e.*, Simplified Molecular Input Line Entry System (Navarro, 2001). We start with the idea that any two given biomolecules can be structurally compared on the basis of this SMILES string notation (Xue *et al.*, 2003). For example, D-glucose has the SMILES, C(C1C(C(C(C(O1)O)O)O)O)O, whereas ethanol has the SMILES, CCO. This representation is quite compact compared to most other methods of representing structure. They may be of two types, Canonical, where chirality and hydrogen contributions are not considered, and Isomeric, where chirality as well as hydrogen is considered (Xue *et al.*, 2003). Here we have considered canonical type. That is why no hydrogen atom is considered in the above SMILES.

As SMILES are 1D representation of a 3D biomolecule, so whenever we take them into consideration, we neglect some of their properties. These include linkages, interaction properties, to name a few. Thus, in order to remove these drawbacks, SAGPAR converts the SMILES strings to certain specific patterns to store information about linkages in terms of bonds in the biomolecule or metabolite (discussed in Section 'Structural grammar representation'). After converting the SMILES to patterns, a comparison needs to be done between each metabolite pair for finding out similarity between them. The comparison schema we have taken is using a similarity score (discussed in Section 'Score Formulation'). One of the stringencies of using this score is that it checks the presence or absence of a binary digit (1 or 0) in the considered metabolite pair for comparison.

Thus, the patterns generated by SAGPAR are further converted into its binary counterpart first and a 'similarity score' is generated by comparing the binary strings of both the metabolites (discussed in Section 'Structural grammar representation', 'Score formulation'). Another important factor that we have kept in our mind is the method of correlating this 'similarity

score' with the actual biological properties of metabolites. We have performed various tasks like taking into consideration the 'weight contribution' (discussed in Section 'Linking pools and non-pools'), 'branching patterns' (discussed in Section 'Other factors'), structural features' (discussed in Section 'Other factors') and combining them with 'similarity score' for generating the 'final score' (discussed in Section 'Other factors') between the compared metabolite pairs.

2.1 Structural grammar representation

Bio-molecules can be represented *in silico* using structural grammars, the concept that we have used in this study (Klipp, 2005). They represent biomolecules using certain binary arrays. For a particular biomolecule, various patterns can be generated, which can be further converted into their corresponding binary arrays (*i.e.* having 1 or 0). These binary arrays can be further used for analyzing the bio-molecules. The various patterns for a particular molecule's structural grammar are generated from the molecule itself. For instance, the molecule ethanol (CH_3CH_2OH) having SMILES CCO, generates the patterns such as 'C', 'C', 'O' (1-atom pattern); 'CC', 'CO' (2-atom pattern); 'CCO' (3-atom pattern) (Periwal *et al.*, 2006). One may notice that, in any pattern generation for structural grammars, only adjacent atoms are considered for higher patterns (*i.e.* > 1-atom), whereas, in 1-atom patterns, all the atoms are considered (Palsson, 2006). The number of patterns generated for a SMILES of length n is $[n \times (n+1)]/2$.

After the patterns are generated, they are combined to form a single array, *i.e.* 'CCOCCCOCCO'. The next step is to convert this pattern array into its corresponding binary array or string (Steinbeck *et al.*, 2003). The reason for converting this pattern array to its binary counterpart is for calculating a similarity value or coefficient between a metabolite pair. These coefficients work only on binary counterpart of metabolites. Also, comparing 1's and 0's is easier than comparing a series of characters present in SMILES. The binary string is formed by studying the occurrence of various atoms in a given pathway. It is seen that in any given pathway, the occurrence of C, O, H, P, N and S is more compared to other atoms (De Luca *et al.*, 2000). For converting the pattern array into its corresponding binary string, two approaches may be implemented. One is using a random binary generator and the other is static binary generator. As, the number of significant atoms present in metabolites is limited (*i.e.* 6 in number), we have selected the static binary generator (Westerhoff *et al.*, 2005). Thus, for C we have assigned a binary string, '111111', for O, it is '111110', for H, '111100', for P, '111000', for N, '110000' and for S, the binary string is '100000'. The above pattern array for ethanol (CH_3CH_2OH) contains only two distinct atoms, C and O. Thus, the combined binary string formed for ethanol is '111111 111111 111110 111111 111111 111111 111110 111111 111111 111110' (Klipp, 2005). This representation is known as structural grammar representation in pharmacokinetics (Xue *et al.*, 2003; Ei-Basil, 2008).

2.2 Score formulation

For checking the similarity between metabolites in a pair, a standard schema should be followed. SAGPAR uses the structural grammar representation (as discussed in Section 'Structural grammar representation') for predicting a similarity value or score between two metabolites (Oltvai *et al.*, 2004). Thus, higher the score more is the similarity between the metabolites. The binary strings are calculated for all the metabolites that are taken as input dataset. We now compare each pair of metabolites for finding out their similarity on the basis of these binary strings. If two metabolites are similar to one another in terms of structural similarity, the chance that one would be converted into another is more than the pair of metabolites which are dissimilar in structure (Steinbeck *et al.*, 2003; Whittle *et al.*, 2003; Davies *et al.*, 2006). Structural similarity, identified using 'structural grammars', between two metabolites is an essential criteria for automated identification of pathways. For instance, the chance that L-noradrenaline would be converted to L-adrenaline in tyrosine metabolism of *H. sapiens* is more than that of conversion of L-noradrenaline to 3-methoxy-4-hydroxy-phenyl-ethylene glycol. This is due to the fact that, L-noradrenaline is more similar to L-adrenaline than 3-methoxy-4-hydroxyphenyl-ethylene glycol (Kitano, 2001).

A similarity score is found for each pair of metabolites. This score is calculated by comparing the binary strings of two metabolites and checking their position specific occurrence in the structural grammar representations of both the metabolites. For two structural grammars of metabolites i and j, we count the number of 1's in i but not in j(termed as $sub1$), the number of 0's in j but not in i (termed as $sub2$), the number of 1's in both i and j (termed as $sub3$), and the number of 0's in both i and j (termed as $sub4$) (Davies *et al.*, 2006). On the basis of these four parameters, a score $EI = [(sub3 + sub4)/tot]^{1/2}$, $tot = sub1 + sub2 + sub3 + sub4$ (Tada *et al.*, 2003), is calculated signifying the similarity between two metabolites (Balaban *et al.*, 2007; Tada *et al.*, 2003). Fig. 1 discusses the major steps in SAGPAR.

2.3 Linking pools and non-pools

But, only this comparison score is insufficient for predicting the conversion of one metabolite into another, as for a reaction to occur successfully, certain side-products (*i.e.* pools) need to be considered too along with primary metabolites (*i.e.* non-pools) (Kerber *et al.*, 2007). Thus, pool metabolites are those whose occurrence is more than that of non-pool metabolites, e.g., NADP, ATP whereas occurrence of non-pool metabolites, e.g., D-glucose in glycolysis, are those whose oc-

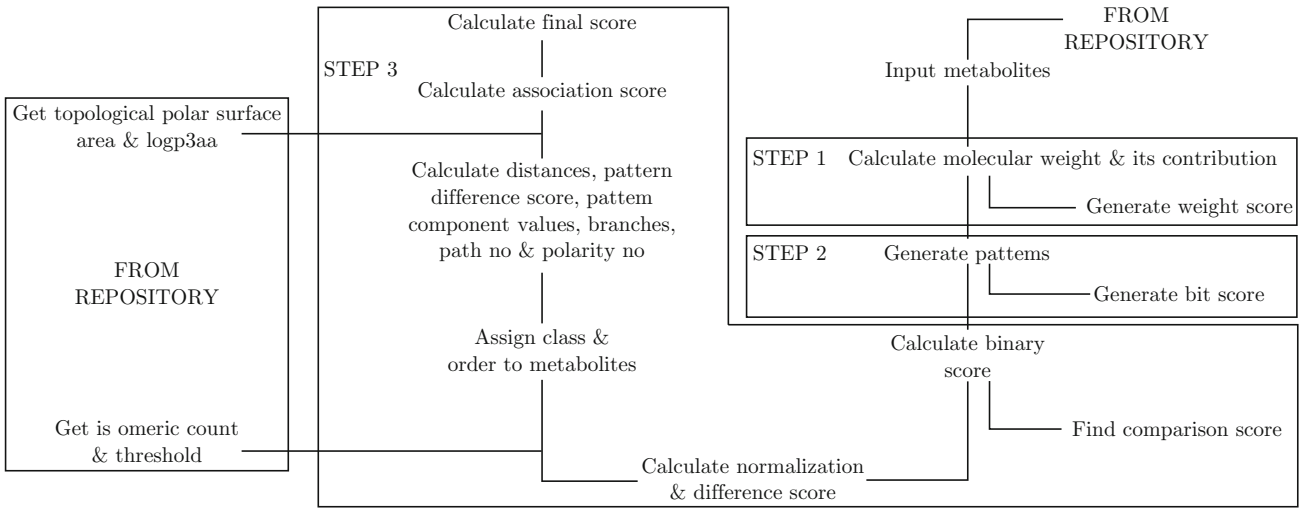

Fig. 1 Steps in SAGPAR

currence is not as high as pools, but contribute significantly in the functioning of the process. This identification is very essential because pools and non-pools significantly contribute to the combined biological activity of the pathways (Boncher, 1983). But a problem that one may face is how to predict, which pools are acting together with the metabolites being compared.

SAGPAR successfully comes up with a solution to this too. Pool metabolites are high in number in any given pathway dataset, but they are highly repetitive also, *i.e.*, there may be NADP, ATP, H_2O, but each of these pools may occur more than once, as they may be present in more than one reaction (Kerber *et al.*, 2007). SAGPAR first identifies the distinct pools from the dataset. Next, it finds out the molecular weights of all the pool and non-pool metabolites. Now let us consider the given two conversions, $A \rightarrow B + C$ and $A \rightarrow L + M$ (Briem *et al.*, 1996). Here, C and M are pools whereas A, B, and L are non-pools. But, in automated pathway reconstruction, this conversion is not known previously. Using, the previously explained structural grammar concept, only $A \rightarrow B$ and $A \rightarrow L$ can be predicted (Moorthy, 2007; Briem *et al.*, 1996).

After finding the molecular weight of the metabolites, a separate list is parsed, consisting of all the pools present in the dataset. This list is generated by SAGPAR by identifying some key properties of pools like their higher occurrence than non-pools, repetitive in nature, lower molecular weight than non-pools, to name a few. Thus for linking the pools with non-pools we have argued that for a successful conversion to occur, the difference between the weight contributions of all the reactants (pools and non-pools) and that of the products (pools and non-pools) should be minimum, *i.e.*, difference between weight contributions of A and weight contributions of B + weight contribution of C is minimum, whereas, difference between weight contribution of A and weight contribution L + weight contribution of M is minimum respectively (Moorthy, 2007). The 'weight contribution' of an atom in a metabolite is its contribution in the overall metabolite weight. For example, weight contribution of 'O' in 'CCO' is 16/40=0.4=40%.

Similarly, 'weight contribution' of a metabolite in a reaction is its contribution in the overall reaction weight (calculated by adding the molecular weight of all the participating metabolites in the reaction). Thus, in the given equation, *ATP+acetate* $\rightarrow$ *acetylphosphate+ADP*, acetate and acetyl phosphate are non-pools, whereas ATP and ADP are pools. SAGPAR uses this basic property for linking the pools and non-pools. It compares all the pools with each metabolite pair (i, j) and finds whether weight contribution of $i \simeq$ weight contribution of $j+$ weight contribution of each pool, or weight contribution of $i+$ weight contribution of each pool $\simeq$ weight contribution of j (Briem *et al.*, 1996). The pools satisfying this criterion are linked with their respective metabolite pair.

2.4 Other factors

As discussed previously, input data for SAGPAR is collected from KEGG. It may be possible that while collecting information from repositories, there are some repetitive entries in the datasets, which may give rise to erroneous results. This is because SAGPAR needs to differentiate between pools and non-pools from the given metabolite set by itself. One of the properties of pools that SAGPAR considers for differentiating it from non-pools is that they are more repetitive in occurrence than non-pools. There are some entries in KEGG which are non-pools but still are repetitive, viz. guanosine that occurs 5 times in purine metabolism of *H. sapiens*. SAGPAR nullifies this problem using the concept of normalization (West, 1996). It removes the repet-

itive entries of non-pools which make SAGAR easily differentiate between pools and non-pools. We also use this concept for standardizing datasets having multiple properties by dividing them using a common variable for negating its effect on the complete dataset. This allows the underlying characteristics of the datasets to a common measurable scale (Ei-Basil, 2008).

Furthermore, while converting a SMILES string into its corresponding structural grammar, certain anomalies may arise. This is due to the fact that conversion from SMILES (representing structural properties of a metabolite) to binary string format (1-dimensional) may give rise to loss in biological activities (Pogliani *et al.*, 2008). To overcome this problem, SAGPAR combines the total effect of structural grammars and pools, and finds out how far apart the metabolites in each pair are. It acts in coordination with the comparison score measures and is essential to preserve the biological properties of our datasets from repeated transformations in different scales (Diestel, 2005). In D-glucose, having a 3-dimensional structure is converted to its SMILES 'C(C1(C(C(C(C(O)O)O)O)O)O)O' string causing loss in biological activity due to its linear structure. SAGPAR calculates this loss by checking the number of connectivities in smiles with the number of patterns generated from it. For example, in 'C(C(O)O)O', the loss is the number of branches divided by number of patterns generated (Δ_i in Sections 'Algorithm', 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*') which is $2/15 = 0.0133$ (SD_i in Sections 'Algorithm', 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*').

SAGPAR also considers a problem faced while dealing with isomers that have the same structure but different functions. It takes into consideration two fundamental properties of bio-molecules called isomeric thresholds and isomeric count, whose values are taken from PubChem Compound (http://www.ncbi.nlm.nih.gov/pccompound) data repositories. Isomeric thresholds store the measures of already defined properties based on experimental datasets, calculated from 3-D structure of these metabolites, whereas isomeric count keeps an account of the various configuration modes of metabolites in 3-D space as well as their symmetrical measures (Kitano, 2001). For example, isomeric count & isomeric threshold of L-erythro-4-hydroxyglutamate are 168 and 4 respectively.

Before calculating the final score that is used for pathway reconstruction, SAGPAR assigns a class to the metabolite-pair taken into consideration. This classification is necessary for distinguishing various types of metabolites, as a metabolite that is cyclic in nature cannot be compared with another that is un-branched in form. This class is assigned whether the metabolites are linear (L) (e.g., carbon dioxide: OCO), cyclic (C) (e.g., benzene: C1=CC=CC=C1), acyclic (A) (e.g., pentetic acid: C(CN(CC(=O)O)CC(=O)O)N(CCN(CC(=O)O)CC(=O)O)CC(=O)O), branched (B) (e.g., glycogen: C(C1C(C(C(C(O1)OCC2C(C(C(O2)OC3C(OC(C(C3O)O)O)CO)O)O)OC4C(C(C(O4)CO)O)O)O)O)O)O) and unbranched (U) (e.g., hexane: CCCCCC) forms, which is calculated from the SMILES itself. For example, D-glucose is given a 'B' class.

This is followed by giving an order to the metabolites (pools and non-pools both) taken for comparison. This measure is very important as we can keep the count of the pools and non-pools that are being compared. The order is calculated by finding out the number of occurrences of the metabolites (that are being compared) in the dataset (Moorthy, 2007). It is noticed that the order of pools are always higher than that of non-pools. For instance, in a given data set, NADPH is given order 15 whereas glucose is given order 2, based upon their occurrence.

Another effective consideration by SAGPAR is by calculating the Lavenshtein distance between the metabolite pair (Diestel, 2005). This is calculated for the class of metabolites that are already assigned previously. It is the minimum number of operations needed to transform the structural grammar of one metabolite into the other, where an operation is an insertion, deletion, or substitution of a single binary digit. For example, converting a structural grammar of one molecule '110010100110' into another molecule '100010000111' requires 3 operations, as the second grammar differs from first in 2nd, 7th and 12th positions ($L_{i,j}$ in Sections 'Algorithm', 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*'). But it may also be possible that a non-pool may be repeated and is compared, while generating the comparison score (Briem *et al.*, 1996). In this case, no further operations are performed. Also, since the calculations performed by SAGPAR is based on the patterns and the binary strings generated from the SMILES for each metabolite, we also consider the number of patterns generated for each metabolite and the corresponding significant atoms in the patterns (6 in our case). For example, SMILES of 'C(C(O)O)O' & 'C(O)O' generate total number of 15 & 6 patterns respectively, giving rise to a pattern difference score of $(15 - 6) \times 6 = 54$($Kpat_{i,j}$ in Sections 'Algorithm', 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*'). Furthermore, to link the effect of class, order, lavenshtein distance & pattern difference for a given metabolite pair, SAGPAR calculates a pattern component value ($Ncom_{i,j}$ in Sections 'Algorithm', 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*').

Moreover, it is also possible that while comparing two metabolites, they may be of different classes, e.g., one may be linear chain, and another may be branched (Rouvray, 1986). In this case, two important mea-

sures are calculated by SAGPAR, viz., *path number* and *polarity number*. The *path number* is the sum of the number of steps needed to traverse to the central atom from both the sides of more compact the metabolite, smaller is the *path number*. For example, in case of 'C(C(O)O)O', for traversing to the innermost atom, it takes 2 + 2 = 4 steps from both sides (w_i in Sections 'Algorithm', 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*'). The *polarity number* of a metabolite is the number of pairs of atoms separated by three bonds. For example, in case of 'C(O)O', *polarity number* is 0 ($p3cc_i$ in Sections 'Algorithm', 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*'). Both these measures are important in predicting the difference in orientation of two different classes of metabolites (Balaban *et al.*, 2007). A final score is generated by combining all these previously calculated values for reconstructing a pathway. The various symbols used and the algorithm developed are explained below.

2.5 Various symbols

(1) M_i = SMILES string of i^{th} metabolite

(2) P_i = pattern array for i^{th} metabolite

(3) BS_i = combined binary string of i^{th} metabolite

(4) sub1 = number of 1's in BS_i but not in BS_j

(5) sub2 = number of 1's in BS_j but not in BS_i

(6) sub3 = number of 1's in both BS_i and BS_j (Whittle *et al.*, 2003)

(7) sub4 = number of 0's in both BS_i and BS_j (Whittle *et al.*, 2003)

(8) tot = sub1 + sub2 + sub3 + sub4

(9) S_I = (2 × sub1 × sub2) / (sub1 + sub2) (Davies *et al.*, 2006)

(10) T_I = sub3 / (sub1 + sub2 + sub3) (Davies *et al.*, 2006; Balaban *et al.*, 2007)

(11) W_i = molecular weight of metabolite i

(12) P_k = set of pool metabolites involved in the k^{th} metabolite pair

(13) W_{TOTAL} = molecular weight of all the metabolites in the pair including associated pool metabolites

(14) $WS_{i,j}$ = (W_i + W_j) / W_{TOTAL}

(15) $CS_{i,j} = [(sub3 + sub4)/tot]^{1/2}$ (Balaban *et al.*, 2007)

(16) $NS_{i,j}$ = (2 × number of common bits in BS_i & BS_j) / total bits in BS_i & BS_j

(17) SD_i = (number of connectivities in SMILES) / number of patterns produced

(18) IC_i = Isomeric count for i^{th} metabolite

(19) IT_i = Isomeric threshold for i^{th} metabolite

(20) $IS_{i,j}$ = Isomeric scoring function = $NS_{i,j} \times SD_i \times SD_j \times IC_i \times IC_j \times IT_i \times IT_j$

(21) C_i = class of i^{th} metabolite

(22) g_i = order of i^{th} metabolite

(23) $L_{i,j}$ = Lavenshtein distance between i^{th} and j^{th} metabolites

(24) $Kpat_{i,j} = 6 \times (|BS_i| - |BS_j|)$

(25) $Ncom_{i,j} = 1/((g_i + g_j)(L_{ij} \times Kpat_{ij})$ (Rouvray, 1986)

(26) Δ_i = number of branches in BS_i.

(27) w_i = path number

(28) $p3cc_i$ = polarity number

(29) a_i = topological polar surface area for i^{th} metabolite

(30) b_i = Logp3-AA for i^{th} metabolite

(31) $AC_i = a_i\Delta_i w_i + b_i\Delta_i p3cc_i$

(32) $Cac_{i,j} = AC_i + AC_j$, if BS_i & BS_j belong to different classes

(33) FSi,j = *final score* = ($WSi,j \times CSi,j \times ISi,j$) × $Ncomi,j \times Caci,j$

(34) T-value = threshold value

Here, IC_i, IT_i, a_i and b_i-values are taken from PubChem compound database (http://www.ncbi.nlm.nih.gov/pccompound). Also, T is selected according to the specificity required by SAGPAR while generating the links between the metabolites. It may change for different pathways.

2.6 Algorithm

Input: A set of n metabolites M = {M_1, M_2, M_3, ..., M_n}.

Output: For each pair of metabolites M_i & M_j, $FS_{i,j}$ is generated.

Steps:

(1) Calculate weights, W_i's for all the metabolites in M.

(a) Pool metabolites are identified.

(b) Associate nonpool metabolites with randomly selected pools.

(c) Calculate $WS_{i,j}$ and identify the highest $WS_{i,j}$.

(2) For each metabolite, M_i in M, do,

(a) Generate patterns, P_i corresponding to i-edge, i=0 to 'm', where 'm' is the length of smiles for M_i.

(b) For each pattern, convert to the corresponding bit value, and combine the bit values of all patterns for each metabolite, denote by BS_i.

(3) For a pair of i^{th} & j^{th} metabolites, do,

(a) Calculate sub1, sub2, sub3, sub4 and thereby, $CS_{i,j}$.

(b) Calculate $NS_{i,j}$, SD_i & SD_j.

(c) Get IC_i, IC_j and IT_i, IT_j from PubChem Compound and thereby $IS_{i,j}$.

(d) Assign C_i, C_j and g_i, g_j to i^{th} & j^{th} metabolites.

(e) Calculate $L_{i,j}$, $Kpat_{i,j}$, $Ncom_{i,j}$, Δ_i, w_i and $p3cc_i$.

(f) Get a_i and b_i from PubChem Compound.

(g) Calculate AC_i and $Cac_{i,j}$.

(h) Calculate $FS_{i,j}$.

(i) Keeping M_i fixed, identify $FS'_{i,j}s > T$-value, for a pair M_i & M_j, and generate links between M_i & M_j.

The final score *FS* takes weight assigning score, similarity measure, comparison score, stereochemical thresholds and thermodynamic parameters into account for modeling a pathway. After generating the *FS*-values for all metabolite pairs, the reaction links

need to be established. This is achieved by identifying all the *FS*-values keeping a particular metabolite fixed and changing its counterpart while generating the pair. SAGPAR uses a threshold value, T-value, for identifying more than one reaction links for a metabolite based upon the criteria that *FS*-values must be greater than T-value. The specificity of SAGPAR depends on the proper selection of T-value. It ranges from 0 to 1. It is also recommended that T-value should be kept optimum to refine correct path regeneration.

It has already been discussed previously that if two metabolites are similar to one another in terms of structural similarity, the chance that one would convert to another would be more than that pair of metabolite which are dissimilar in structure (Steinbeck *et al.*, 2003; Whittle *et al.*, 2003; Davies *et al.*, 2006). SAGPAR effectively uses *FS*-values to calculate this similarity between the metabolites. It also considers side products by linking the metabolites (non-pools) with the pools. Another important consideration by SAGPAR is to predict the actual reaction link when there are isomeric compounds. The pathway regeneration is carried out from the initial metabolites that are given to SAGPAR in the form of SMILES. The effectiveness of SAGPAR along with its performance comparison over other algorithms is discussed in the Section 'Performance comparison'.

2.7 Selection of *T*-value

In metabolic pathways, there are many metabolites that participate in more than one reaction. Selecting a proper T-value is an important criterion for detecting such metabolites. Before selecting a suitable T-value for pathway regeneration, SAGPAR assigns a minimum and maximum for it. The minimum is the lowest *FS*-value and maximum is the highest *FS*-value for a pathway dataset without considering the T-value. We have found out the specific range of the T-values for all the pathways by testing SAGPAR on the wide range of T-values between minimum and maximum. For instance, the minimum and maximum for arginine, proline metabolism in *M. pneumonia M129* are 0.325786 and 0.712987 respectively. Now, for $T = 0.33$, the scoring pairs $> T$ are 1-4, 1-2, 1-5, 2-3, 2-6, 2-8, 2-9, 3-4, 3-5, 3-6, 3-9, 4-5, 4-6, 4-7, 4-10, 5-6, 5-9, 5-10, 7-10, 7-12, 7-13, 9-10, 9-11, 9-12, 10-11, 11-13 and 12-13 (details in Sections 'Pathway prediction of arginine, proline metabolism in *M. pneumoniae M129*' and Section C in Supplementary Materials). As, the T-value is lowered, multiple links among metabolites increase, and as T-value is increased, multiple links among metabolites decrease, but may result in decreasing the specificity of SAGPAR as many correct links are neglected due to high T-value. Thus, an optimum T-value needs to be selected for proper prediction.

The T-values have been selected based upon the already generated *FS*-values, *i.e.* the lowest *FS*-value is taken as the lower limit of T-value whereas the highest *FS*-values are taken as the upper limit of T-value. Thus, for any given metabolic pathway, the range of T-values is 'lowest *FS*-value' - 'highest *FS*-value'. For example, the lowest *FS*-value and highest *FS*-value for arginine, proline metabolism in *M. pneumoniae M129* are 0.325786 and 0.712987 respectively. Thus, the range of T-values arginine, proline metabolism in *M. pneumoniae M129* is $0.325786 - 0.712987$. But, a problem with this assumption is that the range is too large and it is difficult to choose a proper T-value measure. For this purpose, we have formulated a strategy for predicting the exact T-value for any given metabolic pathway (Fig. 2).

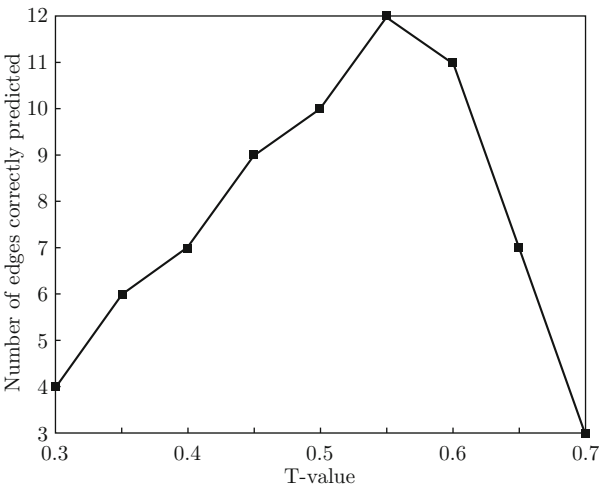

Fig. 2 T-value selection in case of arginine, proline metabolism in *M. pneumoniae M129*

First we constructed a graph with T-values in X-axis and number of correct edges predicted in Y-axis (Fig. 2). The set of T-values were generated with an interval of 0.05 and corresponding number of edges correctly predicted were plotted. Thus, the most probable and best result was obtained at T-value = 0.55, where the highest number of correct edges were predicted, *i.e.* 12. Due to space constraints, we are only including the graph of T-value calculation for arginine, proline metabolism in *M. pneumoniae M129* here, whereas the range of T-values for other metabolic pathways is presented in the Tables 5-6 (in Supplementary Materials). However, we have not included such plots for the other pathways, in order to restrict the sizes of both the paper and the Supplementary Materials. We have analyzed 49 pathways in pathways of *M. pneumoniae M129* and *H. sapiens* for determining the range of T-value while testing our algorithm. The range of T-value in the pathways of *M. pneumonia M129* and *H. sapiens* are present in Tables 5-6 (in Supplementary Materials).

3 Results

This section gives an insight into the testing results of SAGPAR on 25 metabolic pathways of *M. pneumoniae M129* and 24 metabolic pathways of *H. sapiens*. Subsection 3.1 highlights the implementation of SAGPAR on arginine, proline metabolism in *M. pneumoniae M129*, followed by the test results on other 24 metabolic pathways in *M. pneumoniae M129* (subsection 3.2). The pathway reconstruction results of SAGPAR carried out on 24 metabolic pathways of *H. sapiens* is focused in subsection 3.3. The last and final subsection 3.4 deals with the comparison of SAGPAR with some other standard pathway analysis and reconstruction tools, namely, Jarnac (Sauro, 2000), Pathway Hunter Tool (PHT) (Rahman *et al.*, 2005), Gepasi (Mendes, 1993), Copasi (Hoops *et al.*, 2006) and Path-A (Pireddu *et al.*, 2006) respectively.

3.1 Pathway prediction of arginine, proline metabolism in *M. pneumonia M129*

Here we show the step-by-step execution of the proposed SAGPAR algorithm on arginine & proline metabolism in *M. pneumoniae M129*. We have started with all the 13 non-pool metabolites and 10 pool metabolites present in the pathway. Non-pool metabolites form the set M and that for pool metabolites is N. The non-pool metabolites are L-erythro-4-hydroxyglutamate (1), N-carbamoyl-sarcosine (2), carbamoyl-P (3), L-citrulline (4), L-arginine (5), L-arginyl-tRNA (Arg) (6), L-ornithine (7), trans-4-hydroxy-L-proline (8), L-glutamate (9), L-glutamate-semialdehyde (10), L-1-pyrroline-3-hydroxy-5-carboxylate (11), L-prolyl-tRNA (12) and L-proline (13) respectively (Section A in Supplementary Materials).

Similarly, the pool metabolites, N, taken into consideration are NADPH (p1), O (p2), H (p3), ATP (p4), NH_3 (p5), CO_2 (p6), ADP (p7), $NADP^+$ (p8), NO_2 (p9) and H_2O (p10) respectively (Section B in Supplementary Materials). The number i in brackets represent i^{th} non-pool metabolites and p's are pool metabolites. W_i, $\forall_i$ are calculated from SMILES strings in Step (1).

For example, $W_1 = 121$, is calculated by counting the number of atoms in SMILES. Ten pool metabolites are identified from KEGG in Step (1a).

(1) Step (1b): For each metabolite pair (i, j) in M, two pool metabolites are randomly selected from N, and one of them is associated with i^{th} and the other with j^{th} non-pool metabolite. Now, $WS_{i,j}$ is computed. Furthermore, the best scoring pool-nonpool pair is taken into consideration based on highest $WS_{i,j}$. For non-pool metabolite pair (1, 2), $WS_{1,2} = 205/221 = 0.9276018$, where pool metabolite p2 is associated with 1 and no pool metabolite is associated with 2. We have selected 12 best possible pool-nonpool combinations. Higher the *WS*-value, better is the score. Thus, from a total of 78 metabolite pairs, $12 \times 78 = 936$ pool-nonpool pairs are taken into consideration.

(2) Step (2a)- (2b): For each metabolite, patterns P_i and their corresponding binary string BS_i are generated. Some patterns and binary strings of various metabolites are as follows. P_1: C, C, C, O, O, N, C, C, O, O, O ...; BS_1: 111111111111111 ...; P_2: C, N, C, C, O, O, C, O, N, CN ...; BS_2: 111111110000111 ...; P_3: C, O, N, O, P, O, O, O, CO ...; BS_3: 111111111110110 ...; P_4: C, C, C, C, O, O, N, C, N ...; BS_4: 111111111111111 ...; P_5: C, C, C, C, O, O, N, C ...; BS_5: 111111111111111 ...; P_6: C, C, C, C, O, O, N, C ...; BS_6: 111111111111111 ...; P_7: C, C, C, C, O, O, N, C, N ...; BS_7: 111111111111111 ...; P_8: C, C, C, C, O, O, N, C, C ...; BS_8: 111111111111111 ...; P_9: C, C, C, O, O, C, C, O, O ...; BS_9: 11111111111111111...; P_{10}: C, C, C, O, O, C, C, C, O, O ...; BS_{10}: 11111111111111111 ...; P_{11}: C, C, C, N, C, C, O, O, O, CC...; BS_{11}: 111111111111111111 ...; P_{12}: C, C, C, N, C, C, C, O, O, CC ...; BS_{12}: 111111111111111111 ...; P_{13}: C, C, C, N, C, C, O, O, CC, CC ...; BS_{13}: 111111111111111111 ...

(3) Step (3a): For metabolite pair (1, 2), sub1=869, sub2=869, sub3=826, sub4=64, whereas for metabolite pair (1, 3), sub1=751, sub2=148, sub3=473, sub4=70. Similarly, these are calculated for all other metabolite pairs. For (1, 2), $CS_{1,2}$=0.6217762, whereas, for (1, 3), $CS_{1,3}$=0.613645. *CS*-values are calculated for all other metabolite pairs.

(4) Step (3b): For (1, 2), $NS_{1,2} = 2 \times (826+64)/15100 = 0.1178806$, for (1, 3), $NS_{1,3} = 0.083861$ *etc.* Similarly, $SD_1 = 5 \,/\, (11 \times 12)/2 = 0.0757575$, $SD_2 = 0.06666666$ *etc.*

(5) Step (3c): $IC_1 = 168$, $IC_2 = 134$, $IC_3 = 135$ etc. Also, $IT_1 = 04$, $IT_2 = 02$, $IT_3 = 02$ *etc.* Similarly, $IS_{1,2} = 0.117886 \times 0.0757575 \times 0.066666 \times 168 \times 134 \times 4 \times 2 = 107.22015$, $IS_{1,3} = 128.07836$ *etc.*

(6) Step (3d): C_1 = B, C_2 = B, C_3 = B etc. Also, $g_1 = 5$, $g_2 = 2$, $g_3 = 3$ *etc.*

(7) Step (3e): $L_{1,2} = 612$, $L_{1,3} = 591$ *etc.* $Kpat_{1,2} = 126$, $Kpat_{1,3} = 180$ *etc.* $Ncom_{1,2} = 1/(5+2)(612 \times 126) = 0.0000018$, $Ncom_{1,3} = 0.0000011$ *etc.* Δ_1 =05, $\Delta_2 = 03$, $\Delta_3 = 04$ *etc.* $w_1 = 20$, $w_2 = 04$, $w_3 = 00$ *etc.*, whereas $p3cc_1 = 02$, $p3cc_2 = 0$, $p3cc_3 = 0$ *etc.*

(8) Step (3f): $a_1 = 121$, $a_2 = 836$, $a_3 = 110$ *etc.*, whereas b_1 = -4.1, b_2 = -1.3, $b_3 = -2.1$ *etc.*

(9) Step (3g): $AC_1 = 121 \times 5 \times 20 - 1.4 \times 5 \times 2 = 12086$, $AC_2 = 10032$, $AC_3 = 00$ *etc.*, whereas, $C(1)$ = B, $C(2)$ = B, $AC_{1,2}$ = N/A and $C(1)$ = B, $C(3)$ = B, $AC_{1,3}$ = N/A to name a few. $RF2_{1,2} = 0.0000018$, $RF2_{1,3} = 0.0000011$ and $RF2_{1,4} = 0.025$. Here, 'N/A' stands for 'Not Applicable'.

(10) Step (3h): Some of the scoring pairs are $FS_{1,4} = 0.6278126$, $FS_{2,3} = 0.6741456$, $FS_{3,4} = 0.5777381$ *etc.*

(11) Step (3i): T-value is set to 0.55. The scoring pairs that are $> T$-value are $FS_{1,4} = 0.6278126$, $FS_{2,3} =$

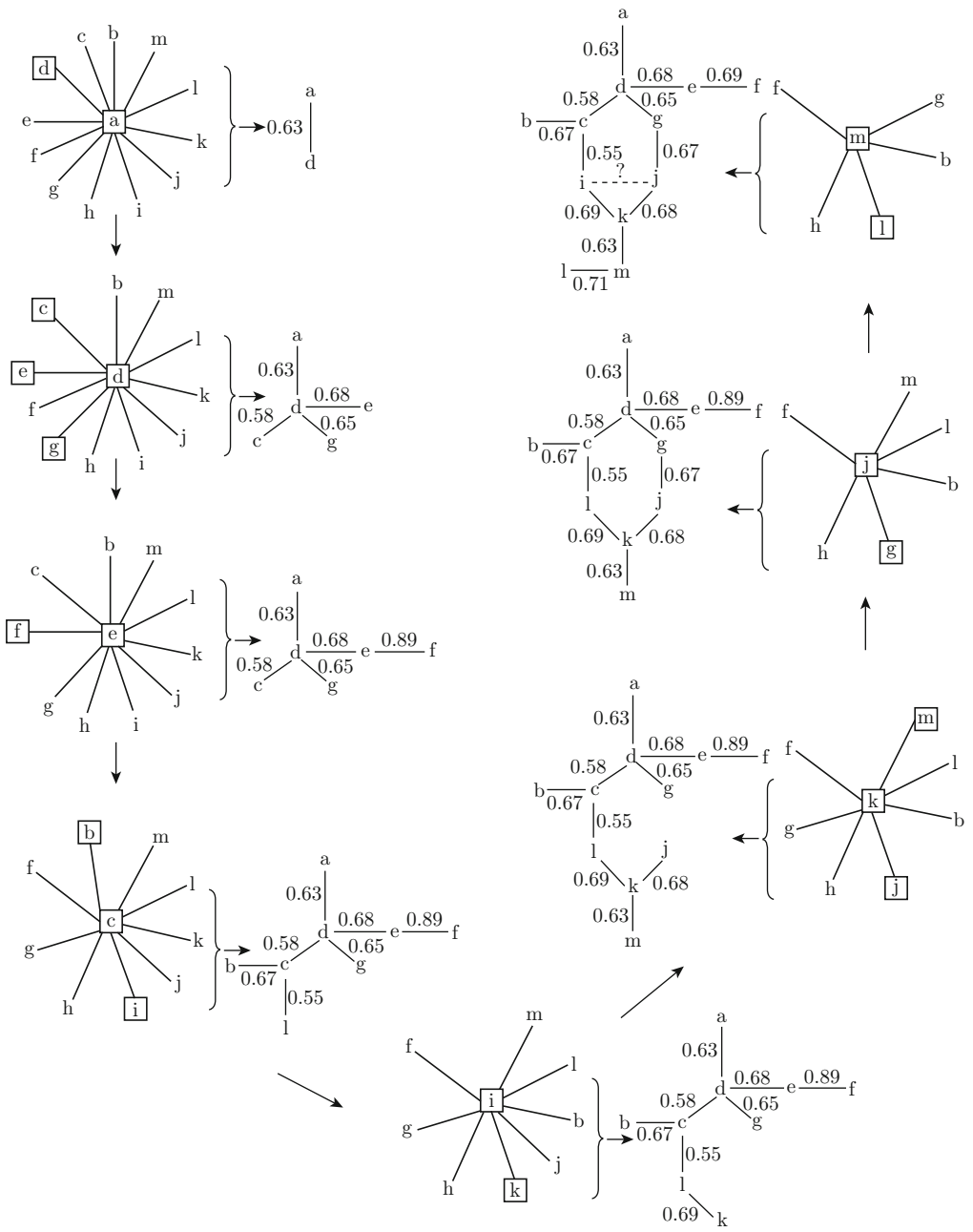

Fig. 3 Step-by-step simulation of arginine, proline metabolism in *M. pneumoniae M129*

0.6741456, $FS_{3,4} = 0.5777381$, $FS_{3,9} = 0.5537142$, $FS_{4,5} = 0.67328$, $FS_{4,7} = 0.648238$, $FS_{5,6} = 0.6931696$, $FS_{7,10} = 0.678974$, $FS_{9,11} = 0.6978823$, $FS_{10,11} = 0.687169$, $FS_{11,13} = 0.634826$ and $FS_{12,13} = 0.712987$.

(12) Step (3j): Links are established between the metabolites as: 1-4, 2-3, 3-4, 3-9, 4-5, 4-7, 5-6, 7-10, 9-11, 10-11, 11-13 and 12-13.

Fig. 3 represents the step-wise reconstruction of arginine, proline metabolism in *M. pneumoniae M129* based upon the *FS*-values > T-value in the above steps. The links generated between the metabolite pairs are also shown simultaneously (Step 3j). In this case, the overall accuracy of SAGPAR is 91.67% (details are present in Section 'Performance comparison').

3.2 Reconstruction of some other pathways in *M. pneumoniae M129*

Some other 24 pathways of *M. pneumoniae M129* are also considered to test SAGPAR's performance over smaller datasets. The pathways include glycolysis (mp1), pentose phosphate (mp2), pentose & glucoronate interconversion (mp3), fructose & mannose metabolism (mp4), galactose metabolism (mp5), ascorbate & aldarate metabolism (mp6), starch & sucrose metabolism (mp7), pyruvate metabolism (mp8), glyc-

ero phospholipid metabolism (mp9), purine metabolism (mp10), pyrimidine metabolism (mp11), glutamate metabolism (mp12), alanine & aspartate metabolism (mp13), glycine & serine & threonine metabolism (mp14), methionine metabolism (mp15), valine & leucine & isoleucine degradation (mp16), valine & leucine & isoleucine biosynthesis (mp17), arginine & proline metabolism (mp18), phenylalanine, tyrosine & tryptophan biosynthesis (mp19), selenoamino acid (mp20), glutathione metabolism (mp21), thiamine metabolism (mp22), riboflavin metabolism (mp23), pantothenate & CoA biosynthesis (mp24) and folate biosynthesis (mp25).

The number of wrong and unrecognized edges (Table 2 in Supplementary Materials) in carbohydrate pathways range from 0 to 2; in lipids it is 1; this number in nucleotides ranges from 0 to 2; in amino acid 0 to 2; and in co-factor & vitamins 0 to 1 respectively. The maximal unrecognized edge (2) in found in pyruvate, purine, pyrimidine metabolism; and valine, leucine & isoleucine biosynthesis respectively whereas the maximal wrong edge (1) is found in 14 pathways. The prediction accuracy of correct nodes and edges ranges from 80% to 100% respectively. Figs. 28 to 52 (in Supplementary Materials) show the *M. pneumoniae M129* pathways reconstructed using SAGPAR.

The best result (100%) is obtained in case of glutamate metabolism (Fig. 39 in Supplementary Materials), methionine metabolism (Fig. 42 in Supplementary Materials), phenylalanine, tyrosine & tryptophan biosynthesis (Fig. 46 in Supplementary Materials) and riboflavin metabolism (Fig. 50 in Supplementary Materials), whereas for glycine, serine & threonine metabolism (Fig. 41 in Supplementary Materials), SAGPAR gives result around 80%. Similarly, the accuracy of pool prediction (Table 4 in Supplementary Materials) in carbohydrate pathway datasets range from 87.5% to 94.44%; 93.33% in lipid; 93.33% to 95% in nucleotide; 83.33% to 96% in amino acid; 89.47% to 90.48% in co-factors & vitamins; and in xenobiotics varying from 86.67% to 90.91% respectively. The maximal accuracy (96%) is found in valine, leucine & isoleucine degradation, while the lowest accuracy (83.33%) is found in phenylalanine, tyrosine & tryptophan biosynthesis.

3.3 Reconstruction of some pathways in *H. sapiens*

We have considered 24 metabolic pathways of *H. sapiens* to test SAGPAR's effectiveness over large and complex datasets. The pathways include pentose phosphate (hs1), pentose & glucoronate interconversion (hs2), fructose & mannose metabolism (hs3), galactose metabolism (hs4), asorbate & aldarate metabolism (hs5), starch & sucrose metabolism (hs6), pyruvate metabolism (hs7), biosynthesis of steroids (hs8), purine metabolism (hs9), pyrimidine metabolism (hs10), alanine & aspartate metabolism (hs11), lycine & serine & threonine metabolism (hs12), methionine metabolism (hs13), valine & leucine & isoleucine degradation (hs14), valine & leucine & isoleucine biosynthesis (hs15), arginine & proline metabolism (hs16), tyrosine metabolism (hs17), tryptophan metabolism (hs18), phenylalanine & tyrosine & tryptophan biosynthesis (hs19), glutamate metabolism (hs20), lysine biosynthesis (hs21), porphyrin metabolism (hs22), drug metabolism - cytochrome P450 (hs23) and metabolism of xenobiotics (hs24) by cytochrome P450. The number of wrong and unrecognized edges (Table 1 in Supplementary Materials) in carbohydrate pathway datasets range from 0 to 2; in lipid range from 0 to 2; in nucleotide range from 0 to 3; in amino acid range from 0 to 2; in co-factors & vitamins is 1; and in xenobiotics varies from 0 to 1 respectively. The maximal unrecognized edges (3) and wrong edges (2) are found in purine metabolism. Furthermore, the prediction accuracy of correct nodes and edges range from 90% to 100% respectively.

Figs. 4 to 27 (in Supplementary Materials) show the *H. sapiens* pathways reconstructed using SAGPAR. The best result (100%) is obtained in case of asorbate & aldarate metabolism (Fig. 8 in Supplementary Materials), alanine & aspartate metabolism (Fig. 14 in Supplementary Materials), methionine metabolism (Fig. 16 in Supplementary Materials), phenylalanine, tyrosine & tryptophan biosynthesis (Fig. 23 in Supplementary Materials) and lysine biosynthesis (Fig. 24 in Supplementary Materials), whereas fructose & mannose metabolism (Fig. 6 in Supplementary Materials) gives result around 90.91%. Similarly, the accuracy of pool prediction (Table 3 in Supplementary Materials) in carbohydrate pathway datasets range from 96.88% to 100%; in lipid range from 97% to 98%; in nucleotide range from 97.78% to 98.18%; in amino acid range from 93.33% to 98.89%; in co-factors & vitamins is 98.86%; and in xenobiotics varies from 96.63% to 97.78% respectively. The maximal accuracy (100%) is found in galactose metabolism, while the lowest accuracy (93.33%) is found in valine, leucine & isoleucine biosynthesis.

3.4 Performance comparison

Furthermore, we have compared the performance of SAGPAR with other pathway analyzers, viz., Jarnac (Sauro, 2000), Pathway Hunter Tool (PHT) (Rahman *et al.*, 2005), Gepasi (Mendes, 1993), Copasi (Hoops *et al.*, 2006) and Path-A (Pireddu *et al.*, 2006). The pathways tools that are taken for comparison have their own underlying algorithms for working. The comparison SAGPAR with these algorithms cannot be done as they are completely different in methodology and course of action. Thus, a comparison has been made on the basis of their ability to detect correct linkages among metabolites while regenerating and forming reaction links among metabolites. The parameters used for comparison are '%Correct connectivity prediction

or %ccp', '%Pool prediction or %pp' and '%Accuracy or %acc'. '%ccp' corresponds to the correct reaction links predicted by the tools; '%pp' corresponds to correct pool metabolite prediction; whereas '%acc' corresponds to the total accuracy of the tool calculated from the correct nodes as well as calculating the number of unrecognized edges as predicted by the tools (Tables 1 and 2 in Supplementary Materials).

In *H. sapiens* (Table 8 in Supplementary Materials), %cpp of SAGPAR ranges from 95.23% to 100%; compared to Jarnac (70% to 100%), PHT (82.45% to 88.88%), Gepasi (71.22% to 77.77%), Copasi (70.56% to 78.88%) and Path-A (62.44% to 66.66%). %pp of SAGPAR ranges from 93.10% to 100%; compared to Jarnac (72.12% to 77.77%), PHT (82.45% to 88.88%), Gepasi (70.00% to 100%), Copasi (70.34% to 83.45%) and Path-A (61.23% to 66.66%). %acc of SAGPAR ranges from 90.90% to 100%; compared to Jarnac (70.70% to 100%), PHT (80.34% to 88.34%), Gepasi (70.23% to 77.77%), Copasi (70.70% to 87.86%) and Path-A (61.12% to 66.66%). In terms of total accuracy, SAGPAR is followed by PHT, Jarnac, Copasi, Gepasi and Path-A respectively (*H. sapiens*). Similarly, in *M. pneumoniae M129* (Table 7 in Supplementary Materials), %cpp of SAGPAR ranges from 90.90% to 100%; compared to Jarnac (70.76% to 77.76%), PHT (80.45% to 88.89%), Gepasi (70.23% to 77.77%), Copasi (70.34% to 77.77%) and Path-A (60.45% to 66.67%). %pp of SAGPAR ranges from 91.67% to 95.45%; compared to Jarnac (70.46% to 76.77%), PHT (80.58% to 88.88%), Gepasi (75.65% to 77.77%), Copasi (70.46% to 77.78%) and Path-A (60.34% to 66.54%). %acc of SAGPAR ranges from 84.62% to 100%; compared to Jarnac (70.43% to 77.77%), PHT (80.34% to 88.88%), Gepasi (70.77% to 79.97%), Copasi (70.43% to 100%) and Path-A (60.45% to 66.66%). Thus, in terms of total accuracy, SAGPAR is followed by Copasi, PHT, Gepasi, Jarnac and Path-A (*M. pneumoniae M129*).

4 Conclusions

We have presented here a novel algorithm, called structural grammar-based automated pathway reconstruction (SAGPAR), which is able to generate a pathway from some given structural formulae. This methodology eliminates the idea of predefined patterns by generating dynamic patterns, as these are generated from the molecule itself. The algorithm is also able to distinguish between 'pool' and 'non-pool' metabolites, and is able to correctly assign connectivities. The algorithm also considers various stereochemical parameters like isomerism, chirality along with thermodynamic thresholds. The coefficients of similarity give a very good measure in terms of similarity between two metabolites.

Structural Grammars are easily generated and can be used to generate the parent structure. Regarding computational feasibility, for each SMILES string of length n, there are $[n\times(n+1)]/2$ patterns. The number of patterns generated depends on the size of the binary string. The quality of the final similarity score may differ based upon the choice of similarity coefficients. It can also be noted that metabolic pathways are highly complex and the kind of metabolites participating in various biochemical reactions is even more complex. The issue of multiple reactants as well as products participating in the same reaction needs to be addressed in more detail.

We intend to implement SAGPAR for studying these complex reactions, thereby increasing its efficiency. Moreover, some of the major areas where this algorithm can be implemented are reconstructing pathways for drug metabolism, diseases and analyzing nodal points in pathways that can be used for identifying leads in drug designing. Furthermore, our algorithm can be used for identifying missing links in incomplete pathways.

We tested SAGPAR on 49 metabolic pathways in *M. pneumoniae M129* and *H. sapiens*, showing pathway reconstruction average accuracy of 91.0924% in *M. pneumoniae M129* and 96.63125% in *H. sapiens*. It is also able to detect multiple reaction links for metabolites, based upon a pre-assigned threshold value. Similarly, the connectivity prediction average accuracy of SAGPAR is 97.33375% in *M. pneumoniae M129* and 96.6944% in *H. sapiens*. Overall, SAGPAR performs much better than some already present metabolic pathway tools like Copasi, PHT, Gepasi, Jarnac and Path-A.

Electronic Supplementary Material

Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s12539-012-0119-8 and is accessible for authorized users.

References

[1] Assenov, Y., Schelhorn, S.E., Lengauer, T., Albrecht, M., Ramrez, F. 2008. Computing topological parameters of biological networks. Bioinformatics 24 (Suppl 2), 282–284.

[2] Balaban, A.T., Devillers, J. 2007. Topological Indices and Related Descriptors in QSAR and QSPR. CRC Press, Florida.

[3] Boncher, D. 1983. Information Theoretic Indices for Characterization of Chemical Structures. Research Studies Press, Hertfordshire.

[4] Briem, H., Kuntz, I.D. 1996. Molecular similarity based on DOCK generated fingerprints. J Med Chem 39 (Suppl 17), 3401–3408.

[5] Davies, J.W., Glick, M., Deng, Z., Nettles, J.H., Bender, A., Jenkins, J.L. 2006. Bayes affinity fingerprints' improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget

drugs a feasible concept? J Chem In Model 46, 2445–2456.

[6] De Luca, V., Romeo, J.T., Ibrahim, R., Varin, L. 2000. Evolution of Metabolic Pathways (Recent Advances in Phytochemistry). Pergamon, Oxford.

[7] Diestel, R. 2005. Graph Theory. Springer, Heidelberg.

[8] EI-Basil, S. 2008. Combinatorial properties of graphs and groups of physicochemical interest. Comb Chem High Throughput Screen 11 (Suppl 9), 707–722.

[9] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., Kummer, U. 2006. COPASI – a COmplex PAthway SImulator. Bioinformatics 22 (Suppl 24), 3067–3074.

[10] Kanehisa, M., Kawashima, S., Okuno, Y., Hattori, M., Goto, S. 2004. The kegg resource for deciphering the genome. Nucleic Acids Res 32, D277–D280.

[11] Kerber, A., Laue, R., Meringer, M., Rucker, C. 2007. Molecules in silico - a gradescription of chemical reactions. J Chem Inf Model 47(Suppl 3), 805–817.

[12] Kitano, H. 2001. Foundations of Systems Biology. MIT Press, Cambridge.

[13] Klipp, E. 2005. Systems Biology in Practice: Concepts, Implementation And Application. John Wiley & Sons Inc., New York.

[14] Mendes, P. 1993. GEPASI: A software package for modelling the dynamics, steady states and control of biochemical and other systems. Comput Appl Biosci 9 (Suppl 5), 63–71.

[15] Moorthy, K. 2007. Fundamentals of Biochemical Calculations. CRC Press, Florida.

[16] Navarro, G. 2001. A guided tour to approximate string matching. ACM Computing Surveys 33 (Suppl 1), 3188.

[17] Oltvai, Z.N., Barabasi, A.L. 2004. Network biology: Understanding the cell's functional organization. Nat Rev Genet 5, 101–113.

[18] Palsson, B. 2006. Systems Biology: Properties of Reconstructed Networks. Cambridge University Press, Cambridge.

[19] Periwal, S., Szallasi, Z., Stelling, J., Alon, V. 2006. Systems Modeling in Cellular Biology. MIT Press, Cambridge.

[20] Pireddu, L., Szafron, D., Lu, P., Greiner, R. 2006. The Path-A metabolic pathway prediction web server. Nucleic Acids Res 34, W714–W719.

[21] Pogliani, L., de Julian Ortiz, J.V., Galvez, J., Garcia-Domenech, R. 2008. Some trends in chemical graph theory. Chem Rev 108 (Suppl 3), 1127–1169.

[22] Rahman, S.A., Advani, P., Schunk, R., Schrader, R., Schomburg, D. 2005. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics 21, 1189–1193.

[23] Rouvray, D.H. 1986. Mathematics and Computational Concepts in Chemistry. Horwood Publishers, Chichester.

[24] Sauro, H.M. 2000. Jarnac: A system for interactive metabolic analysis. Stellenbosch University Press, Stellenbosch.

[25] Shivakumar, N., Narendran, B., Agarwal, P., Srreran, C. 1995. The concord algorithm for synchronization of networked multimedia streams. In: 2^{nd} IEEE International Conference on Multimedia Computing and System'95 (ICMCS'95), Washington DC, 31–40.

[26] Steinbeck, C., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E., Han, Y. 2003. The chemistry development kit (cdk): An open-source java library for chemo and bioinformatics. J Chem Inf Comput Sci 43, 493–500.

[27] Tada, M., Shijima, H., Nakamura, M. 2003. Smiles-type free radical rearrangement of aromatic sulfonates and sulfonamides: Syntheses of arylethanols and arylethylamines. Org Biomol Chem 1 (Suppl 14), 2499–2505.

[28] West, D. 1996. Introduction to Graph Theory. Prentice Hall, New Jersey.

[29] Westerhoff, H., Alberghina, L. 2005. Systems Biology: Definitions and Perspectives. Springer, New York.

[30] Whittle, M., Klaffke, W., van Noort, P., Willett, P. 2003. Evaluation of similarity measures for searching the dictionary of natural products database. J Chem Inf Comput Sci 43, 449–457.

[31] Xue, L., Stahura, F.L., Bajorath, J., Godden, J.W. 2003. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. J Chem Inf Comput Sci 43, 1151–1157.