Internet Electronic Journal of
Molecular Design

June 2004, Volume 3, Number 6, Pages 350–360

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday
Part 12

Guest Editor: Douglas J. Klein

An Information–Theoretical Measure of Similarity and a Topological Shape and Size Descriptor for Molecular Similarity Analysis

Chandan Raychaudhury[1,2] and Indira Ghosh[3,4]

[1] Indian Institute of Chemical Biology, 4 Raja S. C. Mullick Road, Jadavpur, Kolkata 700032, India
[2] Present address: Accelrys K.K., Elite House, #16, 3rd Cross, 13th H Main, Doopanahalli, HAL 2nd Stage, Bangalore 560008, India
[3] AstraZeneca R&D India, Bellary Road, Hebbal, Bangalore, 560024, India
[4] Present address : Institute of Bioinformatics & Biotechnology, University of Pune, Ganeshkhind, Pune 411007, India

Received: July 30, 2003; Revised: November 24, 2003; Accepted: February 27, 2003; Published: June 30, 2004

Citation of the article:

C. Raychaudhury and I. Ghosh, An Information–Theoretical Measure of Similarity and a Topological Shape and Size Descriptor for Molecular Similarity Analysis, *Internet Electron. J. Mol. Des.* **2004**, *3*, 350–360, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

An Information–Theoretical Measure of Similarity and a Topological Shape and Size Descriptor for Molecular Similarity Analysis[#]

Chandan Raychaudhury[1,2] and Indira Ghosh[3,4,]*

[1] Indian Institute of Chemical Biology, 4 Raja S. C. Mullick Road, Jadavpur, Kolkata 700032, India
[2] Present address: Accelrys K.K., Elite House, #16, 3rd Cross, 13th H Main, Doopanahalli, HAL 2nd Stage, Bangalore 560008, India
[3] AstraZeneca R&D India, Bellary Road, Hebbal, Bangalore, 560024, India
[4] Present address : Institute of Bioinformatics & Biotechnology, University of Pune, Ganeshkhind, Pune 411007, India

Received: July 30, 2003; Revised: November 24, 2003; Accepted: February 27, 2003; Published: June 30, 2004

Internet Electron. J. Mol. Des. **2004**, *3* (6), 350–360

Abstract

Motivation. Molecular similarity studies play an important role in today's drug discovery research for finding new lead compounds with the expectation that similar molecules would exhibit similar biological activities. Such approaches are of special importance when one tries to find lead compounds from the databases of thousands to millions of compounds. Although a number of similarity measures are available in the literature, it appears that a similarity measure that takes care of the variety/diversity in structural components, such as substructures, of chemical compounds might find useful applications in this regard and Shannon's measure of information content of a discrete system may be useful in formulating such a similarity index. However, for doing similarity analyses using molecular descriptors, availability of suitable molecular descriptors becomes an essential requirement too. Graph–theoretical descriptors have been found to be very useful in this purpose since they are fast to compute and take care of important structural aspects. Also, there is a special interest for the descriptors reflecting shape and size aspects of molecules since they are related to the fitting/docking of small molecules in the macromolecular receptor sites. We intend to address these aspects in the present communication.

Method. In this paper, we have proposed a new information–theoretical measure of similarity, INFSIM, based on Shannon's measure of information content of a discrete system. In our study, we have also used a topological shape and size index, TSS, defined for small molecules. These indices have been used to carry out molecular similarity analysis for qualitative discrimination (active/inactive) of eleven beta–lactams, taken from the literature, with respect to the anti–bacterial activity of penicillin G. These studies have been carried out using the software AZMOLTOP.

Results. It appears from the present study that the molecules under consideration may be effectively classified using INFSIM and TSS indices since the similarity values seem to be reflected in the experimentally determined activities of the compounds. A comparative study with Tanimoto similarity indicates that the proposed method has performed relatively better similarity assessment for the studied compounds.

Conclusions. The results indicate that the similarity index INFSIM may find useful application in classifying

[#] Dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday.

* Correspondence author; phone: 91–20–5690195; E–mail: indira@bioinfo.ernet.in.

compounds of a database qualitatively according to their activities on the basis of the structural features encoded by TSS. Since TSS is believed to translate the topological shape and size of chemical compounds effectively, it may find applications in database screening for the identification of new lead compounds in drug discovery.

Keywords. Molecular similarity; topological shape and size; information content; graph–theoretical index; database; beta–lactam.

1 INTRODUCTION

Molecular similarity study [1,2] has become one the methods of special interest in drug discovery research toward finding new lead compounds, and in a bigger scale, screening potential therapeutic candidates from large databases before going for experimental work. Presumably, this should help save appreciable amount of time and money involved in drug discovery programs. However, in doing similarity analyses using quantitative descriptors of molecular structures / substructures, two factors play crucial role in doing such studies, the similarity index used and the descriptor(s) considered. While molecular descriptors reflect structural characteristics of the compounds, similarity indices help identify similar compounds from databases with respect to a given molecule (hereafter referred to as query molecule) on the basis of the molecular descriptor values. Although a number of similarity indices exist in the literature [3] with their strength and limitations, the need for doing more work in this direction is very much felt. In particular, methods that can take care of diversities / varieties in the elements of a system, such as Shannon's measure of the information content of a (discrete) system [4], may find useful application in formulating widely acceptable similarity index. So far as molecular descriptors are concerned, descriptors derived from graph–theoretical models of chemical compounds [5] are believed to be useful in molecular similarity studies [6] considering present day's requirement for fast identification of similar compounds from large databases to facilitate drug discovery research.

In this paper, we have proposed an information–theoretical measure of similarity, INFSIM, using Shannon's formula for measuring information content of a discrete system [4]. Since information content of a system reflects the amount of variety [7] present among the elements of the system, we feel that such a measure, if defined suitably, could help get useful measure of similarity. We have considered the atoms in a molecule to form a system where the atoms are the elements of the system. Precisely, we have used Shannon's information theoretical measure of Redundancy of a system [4] to derive the similarity measure. Regarding molecular descriptor, we have used a topological shape and size index (TSS), a TopoPhysical Molecular Descriptor (TPMD), for our study. The index is computed from the vertex weighted graph models of chemical compounds. Such an index for peptides was found to be effective in identifying allele–specific T cell epitopes [8]. Moreover, since shape and size are considered to be two important factors governing fitting / docking of small molecules in the macromolecular receptor cavities for getting desired biological activity, an index taking care of these aspects is believed to produce significant results in identifying similar molecules exhibiting similar biological activities. To investigate the usefulness of the similarity measure INFSIM and the molecular descriptor TSS in qualitative discrimination/ranking of compounds from molecular similarity analysis with respect to a given biological activity, they have been used in our study to classify qualitatively eleven beta–lactams, taken from the literature, as actives and inactives with respect to the antibacterial activity of penicillin G [9].

The computations of INFSIM and TSS indices have been carried out using the new software AZMOLTOP [10]. Furthermore, we have carried out a comparative study with one of the widely used similarity measures, the Tanimoto similarity [11], and the proposed method has been found to produced relatively better assessment for the studied compounds.

2 MATERIALS AND METHODS

In this paper, an information–theoretical similarity measure, INFSIM, and a topological shape and size index for small molecules, TSS, have been used for molecular similarity analysis. A concept, Flexibility Percentage (FP), for identifying similar vertices of two molecular graphs has also been introduced. All these measures have been incorporated in newly developed software AZMOLTOP [10] that has been used for computing the indices for the present study. In this section, the methods of computing the indices have been described. Moreover, SYBYL ver 6.9 [11] has been used to compute Tanimoto similarity values for the compounds under consideration.

2.1 TSS Index

Let *G* be the graph model of a chemical compound penicillin G. Now, with respect to a vertex v in *G*, one can find shortest paths connecting other vertices of *G* with v. It is evident that there may be more than one shortest path between v and some other vertex of *G*, if *G* contains cycle(s). However, one of them may be sorted out for serving a purpose such as for computing a graph–theoretical molecular descriptor.

Also, in the process of identifying such shortest paths, one may get branched vertices in *G* and more importantly, may not need to pass through some of the edges of *G*. If all the shortest paths between v and all other vertices of *G* were identified, one would get a shortest path tree in *G* with respect to v connected to some pendant vertices. Now, for the present purpose, in computing TSS index, we consider the paths going out of *v* and the branching of the paths. If there be, say, two paths going out of *v*, then to take care of this branching factor, a probability value ½ is assigned to each edge emerging from v (it is assumed that the paths are starting from *v* and going to other vertices). If again, say, three paths go out of one of the adjacent vertices of *v*, then a probability value 1/6 (= ½ × 1/3) is assigned to each edge emerging from the adjacent vertex.

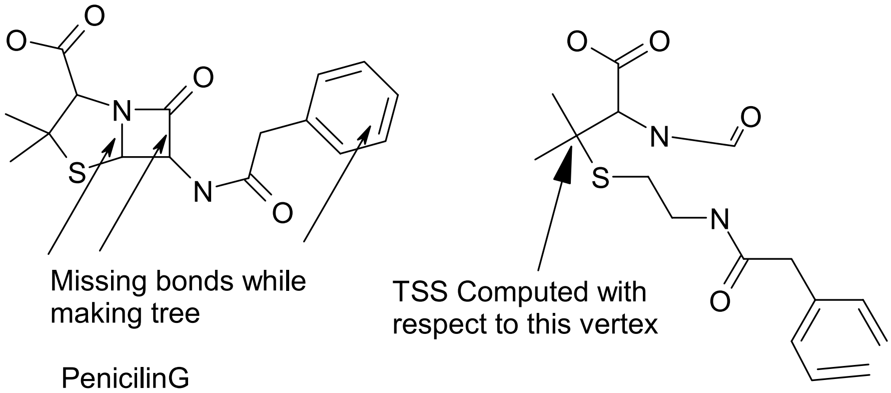

Missing bonds while making tree
PenicilinG
TSS Computed with respect to this vertex
O
N
S

Figure 1. Molecular structure and graph of penicillin G are displayed (a) indicating (by arrows) the unconsidered bonds while making tree and (b) the actual tree with the indication of the vertex for which TSS has been computed in the illustration.

This probability assignment for branching may be continued until one reaches the pendant vertices. In this process, one would get the shortest path tree in *G* with respect to *v* and this tree would touch all the vertices of *G* (although would not pass through all the edges in a cycle). Evidently; if shortest path trees with respect to all the vertices were traced out, all those trees would touch each and every vertex of *G*. Thus, an index computed on the basis of such a shortest path tree in *G* from *v* would reflect some topological shape of *G* with respect to *v*. Now, since we are interested in incorporating both topological shape as well as the size of a molecule in one descriptor, we also consider atomic weight as the weight of a vertex, to take into account the size effect. Thus, the weight of v would represent the atomic weight of the corresponding atom in the molecule depicted by *G*. Furthermore, to take care of the saturated and unsaturated bonds in a molecule, '1' would be added to the weight of each vertex for each hydrogen attached to the heavy atom represented by that vertex. Taking all these into consideration, the index TSS may be computed in the following way:

Let a path *P* in *G* be connecting *v* with a pendant vertex u in *G*. If *p* is the probability value of the edge connected to *u*, then we compute the path value P^{u}_{p} of the path *P* as:

$$P^{u}_{p} = p(w_v + w_1 + w_2 + \ldots..) \quad (1)$$

where w_i, $i = 1, 2, \ldots..$, are the weights of the vertices in *P* other than that of *v* which is w_v. Thus, if there are *h* such paths having path values P^{u1}_{p1}, P^{u2}_{p2}, P^{uh}_{ph} connecting *h* pendant vertices to *v*, then the index TSS for v may be computed from Eq. (2):

$$\mathrm{TSS}\,(v) = \sum_{j=1}^{h} P^{uj}_{pj} \quad (2)$$

The method may be illustrated as follows taking one vertex (indicated by arrow in Fig.1) in the molecular graph of penicillin G as an example.

The paths which are connecting the given vertex, say, *v* for which TSS has to be computed, with the pendant vertices of the tree, a subgraph of the molecular graph of penicillin G, obtained with respect to the given vertex v and the corresponding path values (PV) are:

I. C–CH3 (2 paths);

PV = 1/4(12 + 15) = 6.75 (for each path)

II. C–CH–C=O;

PV = 1/16(12 + 13 + 12 + 16) = 3.3125

III. C–CH–C–OH;

PV = 1/16(12 + 13 + 12 + 17) = 3.375

IV. C–CH–N(H)–C(H)=O;

PV = 1/8(12 + 13 + 15 + 13 + 16) = 8.625

V. C–S–C(H)–C(H)–NH–C=O;

PV = 1/8(12 + 32 +14 + 14 + 15 + 12 + 16) = 14.375

VI. C–S–C(H)–C(H)–NH–C–CH2–C–CH=C(H2);

PV = 1/16(12+32+14+14+15+12+14+12+13+14) = 9.5

VII. C–S–C(H)–C(H)–NH–C–CH2–C–CH–CH–C(H2);

PV = 1/16(12+32+14+14+15+12+14+12+13+13+14) = 10.3125

Therefore, TSS(v) = (2 x 6.75)+3.3125+3.375+8.625+14.375+9.5+10.3125 = 63.00

It may be noted again that the TSS values are computed for the non–hydrogen atoms only and the paths of the trees, used for computing TSS values, also contain only non–hydrogen atoms. However, to take into account hydrogen too, as mentioned in the method, `H's have been attached to the heavy atoms in the path descriptions and '1' has been added to the atomic weights of the heavy atoms for each hydrogen attached to them. Furthermore, since in identifying the tree subgraphs, some of the bonds (paths) are taken out being redundant, '(H)' has been attached to each of the heavy atoms where such edges do not appear in the shortest path tree. It may also be noted that, since we are using graph–theoretical methods for computing TSS index, stereochemical features in the molecules have not been considered. Moreover, for the O–C=O (carbonyl) group, present in all the compounds under consideration, one oxygen has been connected by a double bond and the other by a single bond with the carbon atom.

2.2 Similarity measure INFSIM

Shannon's measure of information content of a system [4] may be used to get a measure of similarity between two objects *e.g.*, two molecules. To formulate such an index, we proceed in the following manner:

Let there be n_1 vertices in a molecular graph G_1 and n_2 vertices in a molecular graph G_2. If m_1 (< or = n_1) vertices of G_1 and m_2 (< or = n_2) vertices of G_2 take part in the similarity determination in such a way that one or more vertices out of m_1 are found to be similar with one or more vertices out of m_2, then we say that $(m_1 + m_2)$ vertices out of the $(n_1 + n_2)$ vertices in G_1 and G_2, taken together, form a disjoint partitioned class and the remaining $[(n_1 + n_2) – (m_1 + m_2)]$ vertices form that many number of single element disjoint classes. Thus, a measure of information content, InfCon, may be obtained with respect to this partition using Shannon's formula [4]. We propose to use this measure to have a measure of similarity between two molecules and the similarity index, INFSIM, may be obtained as follows:

$$\mathrm{InfCon} = [\{(m_1 + m_2)/(n_1 + n_2)\} \log_2 \{(n_1 + n_2)/(m_1 + m_2)\}] + [(n_1 + n_2) – (m_1 + m_2)][1/(n_1 + n_2) \log_2 (n_1 + n_2)] \tag{3}$$

Subsequently, a measure of Redundancy, RInfCon, may be obtained from (4):

$$\mathrm{RInfCon} = 1 – [\mathrm{InfCon} / \log_2 (n_1 + n_2)] \tag{4}$$

We propose to use this measure of Redundancy as a measure of similarity, INFSIM (5), which may be obtained by substituting InfCon from (3) into (4):

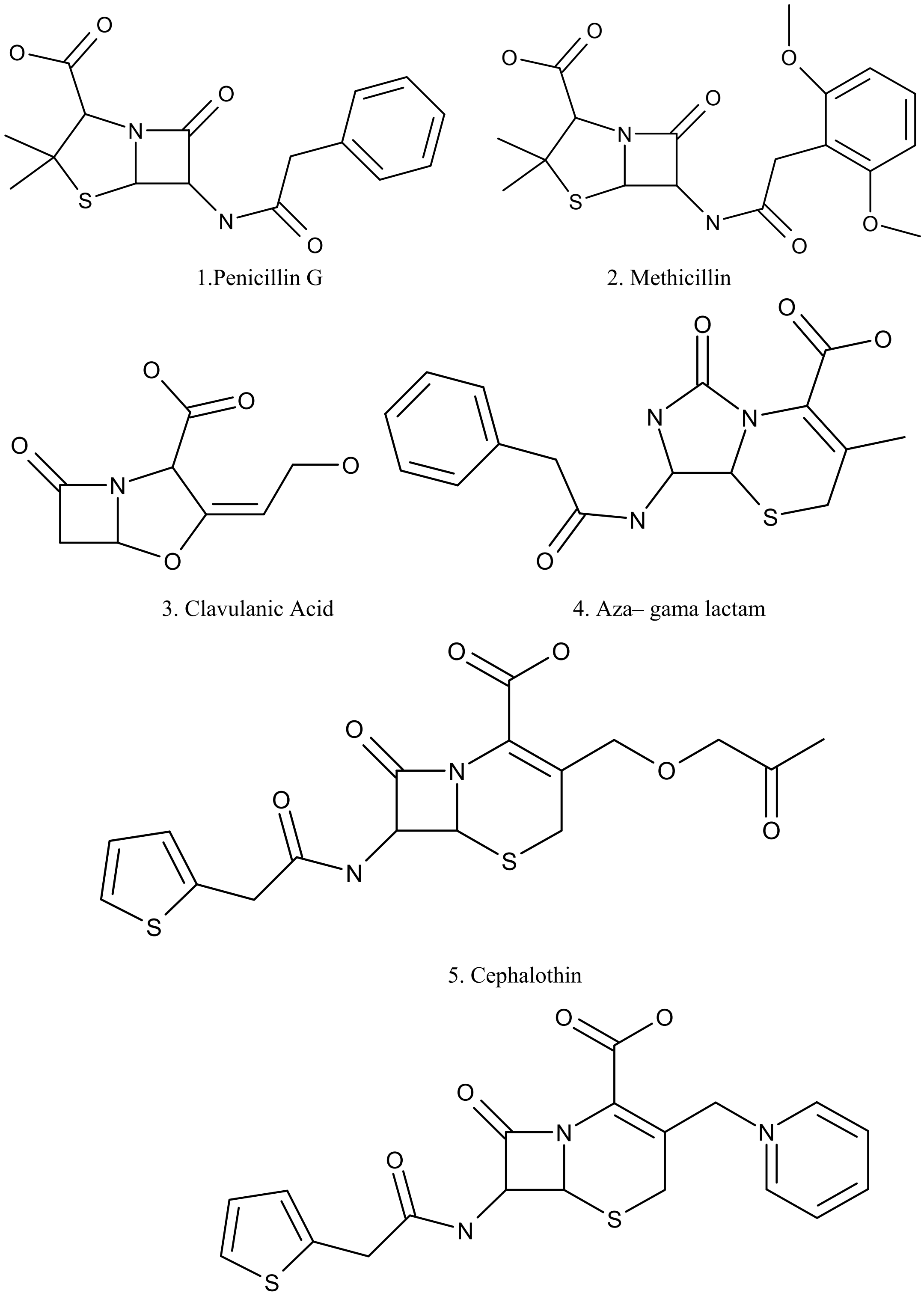
$$\mathrm{INFSIM} = 1 – [\{((m_1 + m_2) / (n_1 + n_2)) \log_2 ((n_1 + n_2) / (m_1 + m_2)) + ((n_1 + n_2) – (m_1 + m_2)) (1/(n_1 + n_2)) \log_2 (n_1 + n_2))\} / \log_2 (n_1 + n_2)] = [(m_1 + m_2) / (n_1 + n_2)] – [(m_1 + m_2) \log_2 (n_1 + n_2) / (n_1 + n_2) \log_2 (n_1 + n_2)] + [(m_1 + m_2) \log_2 (m_1 + m_2) / (n_1 + n_2) \log_2 (n_1 + n_2)] = (m_1 + m_2) \log_2 (m_1 + m_2) / (n_1 + n_2) \log_2 (n_1 + n_2) \tag{5}$$

It may be noted that INFSIM value will lie between 0 and 1 (both inclusive) where 0 would signify no similarity and 1 would correspond maximum similarity. As more and more vertices from the two molecular graphs become similar, $(m_1 + m_2)$ would increase giving higher INFSIM *i.e.*, similarity values. Clearly, if each vertex of one molecular graph becomes similar to one or more vertices of the other molecular graph, then $(m_1 + m_2)$ becomes equal to $(n_1 + n_2)$ and INFSIM would be one. As a particular case, if two molecular graphs are identical in all respect then one would get maximum similarity, l, which is quite in agreement with the intuitive notion of similarity.

In the present study, we have used TSS values of the vertices of the molecular graphs for calculating similarities between two compounds from their INFSIM values.

2.3 Flexibility Percentage (FP)

The TSS indices for the vertices of one molecular graph may not be exactly equal to those of another one although they may be very close. Such close TSS values indicate the closeness of the corresponding substructures (shortest path trees emerging from the vertices in the respective molecular graphs). Thus, it seems reasonable to consider those substructures as similar where TSS value for one falls within a range of values obtained from a given percentage of tolerance on that of the other.



1.Penicillin G

2. Methicillin

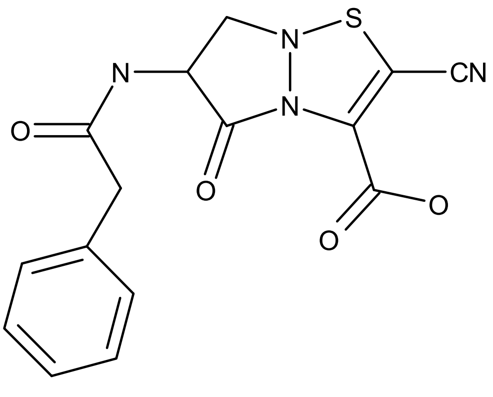
3. Clavulanic Acid

4. Aza– gama lactam

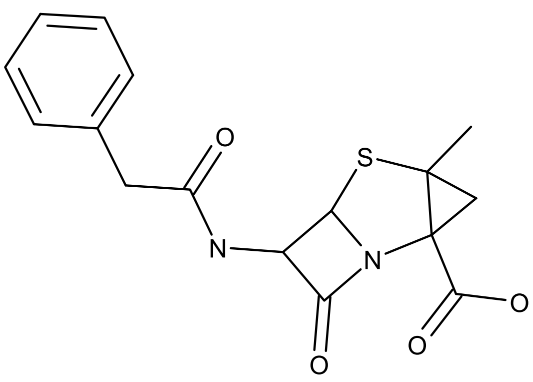
5. Cephalothin

6. Cephaloridine

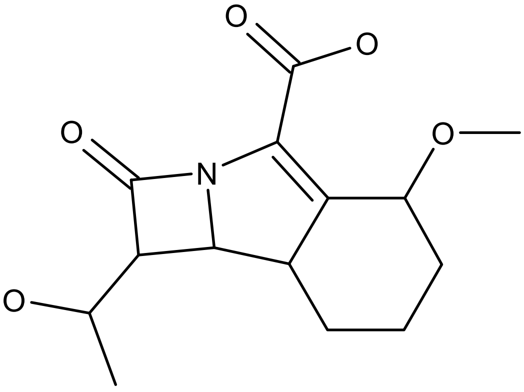
Figure 2. Schematic diagram of penicillin G and eleven other classical and non–classical lactams.



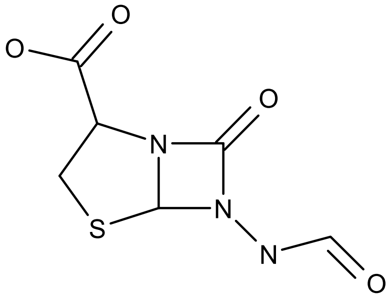
7. LY193375



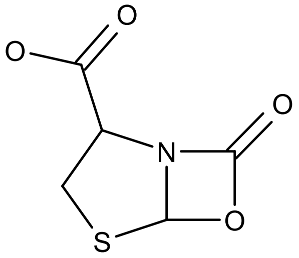
8. Cyclopropylpenam



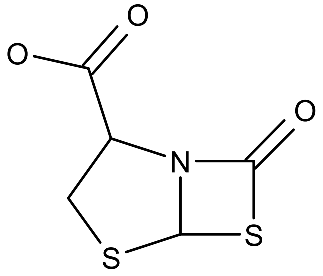
9. Sanfenitrem



10. Aza beta lactam



11. Oxo beta lactam



12. Thio beta lactam

Figure 2. (Continued).

We have, therefore, introduced a new factor, named Flexibility Percentage (FP), such that if FP is given, say, 5, then for a TSS value x of a vertex u in one molecule would be considered to be similar to another vertex v in another molecule if the value of v is in the range $\{x - (5\% \text{ of } x)\}$ and $\{x + (5\% \text{ of } x)\}$. Clearly, low FP values would judge similarity more strictly compared to the higher ones. It is also obvious that for topologically (connectivity wise) identical molecules, one would always get 100% similarity between the two molecules for any FP value.

3 RESULTS AND DISCUSSION

In the present study, we have used a newly developed information–theoretical similarity index INFSIM and a topological shape and size descriptor, TSS, for small molecule to investigate the usefulness of such measures in classifying active and inactive compounds from molecular similarity analysis. This is important in that such a method can be useful in identifying/screening potential drug candidates from the databases of chemical compounds. Here, we have carried out our studies with penicillin G and eleven other classical and non–classical lactams, (Figure 2) a group of compounds of special interest for research in developing antibacterial drugs. The data have been taken from Coll *et al.* [9]. In order to investigate the usefulness of the proposed method we have carried out the present study to see whether the present method can help discriminate those compounds, which are known to have antibacterial activity like penicillin G from those, which do not have that.

The finding of this study has been given in Table 1 where the INFSIM values for FP = 3%, 5%, 8% and 10% as well as the Tanimoto similarity values for the eleven compounds are shown. The activity column of the table shows the experimentally found activities (active/inactive) of the compounds [9]. It is clear from the data that INFSIM values of the active molecules are mostly higher than those of the inactive molecules. In fact, if the cut–off similarity value is set to 50% then only the actives would be screened out for FP =3% (the program AZMOLTOP has this facility such that one can set *x*% as cut–off and thus the molecules which are *x*% or more similar to the query molecule would be shown). This seems to be an encouraging result so far as the efficacy of INFSIM in distinguishing actives and inactive is concerned.

Table 1. INFSIM similarities (expressed in %) for different Flexibility Percentage (FP) and the Tanimoto similarity values of eleven beta–lactam derivatives with respect to penicillin G (query molecule, Q). INFSIM in % at FP= 3%, 5%, 8%, 10% respectively and Tanimoto coefficient

No	Compound name	Activity [a]	FP = 3%	5%	8%	10%	Tanimoto
1	Penicillin G	Q					
2	Methicillin	I	34.77	34.77	38.99	52.17	82.15
3	Clavulanic acid	I	38.94	44.84	57.09	63.40	33.24
4	Aza–gama–lactam	I	42.15	46.81	71.29	83.03	52.21
5	Cephalothin	A	58.16	58.16	67.65	72.48	56.58
6	Cephaloridine	A	55.31	59.79	68.93	83.03	45.17
7	LY193375	A	51.56	66.26	66.26	66.26	39.30
8	Cyclopropylpenam	A	73.25	73.25	78.49	94.54	80.98
9	Sanfenitrem	A	52.38	68.57	79.77	79.77	35.64
10	Aza–beta–lactam	N	44.84	57.09	63.40	69.83	44.20
11	Oxo–beta–lactam	N	24.87	37.00	43.39	49.97	42.33
12	Thio–beta–lactam	N	30.82	43.39	49.97	49.97	41.67

[a] Activity = anti–bacterial activity, Q = Query molecule, I = inactive, A = active, N = not (clearly) known

Looking into the data in more details, it is found that the similarity values of actives and inactives are more distinctly classified for lower FP values (3% and 5%). This is, again, in agreement with what one would expect since due to considering larger ranges of index values for searching similar vertices, INFSIM for higher FP values may accommodate some relatively less similar vertices (ie, vertices having relatively larger differences in their index values) in one class producing lesser discrimination between actives and inactives.

It is also interesting to note that INFSIM values of Cyclopropylpenam is very high for all the FP values and the experimental studies indicate that this compound belongs to the same class with penicillin G [9]. Furthermore, even though Methicillin, an inactive compound, has some structural similarity with penicillin G besides only the substituents in the phenyl ring, the program has successfully identified the structural differences producing lower similarity values. In finding similarity of another inactive compound, aza–gama–lactam, with penicillin G, the program has produced some high similarity for higher FP values although for lower FP values it is not so, indicating strict similarity consideration would place them as different groups of compounds. So far the other inactive compound Clavulanic acid is concerned, it is structurally quite different from penicillin G and has been rightly shown by the program as a molecule of low similarity, particularly for lower FP values. On the basis of these findings, one may perhaps infer that compounds numbers 9–11 (table 1) might show poor or low antibacterial activity as they have low structural similarities with penicillin G. This is also in agreement with the inference drawn by Coll et. al [9].

In addition to these, the Tanimoto similarity values of the compounds (Table 1) have also been computed to investigate the relative efficacy of the proposed method . The notable finding here is that Methicillin has been picked up as highly similar (82.15%) to penicillin G by Tanimoto measure and therefore may be regarded as an active compound which goes against the experimental finding. Again, two active compounds, Sanfenitrem and LY193375, have 35.64% and 39.30% Tanimoto similarities respectively with penicillin G which might lead one to consider them as inactive compounds but not so if one uses INFSIM values (>50% for FP = 3). Therefore, for the studied compounds, the proposed method seems to have an edge over Tanimoto similarity.

4 CONCLUSIONS

All these analyses indicate that the proposed method may find useful applications in identifying similar molecules. In appears from the present study that TSS, the TopoPhysical Molecular Descriptor (TPMD), is capable of identifying the structural differences effectively. It also seems reasonable to say that an index like TSS which is computed from vertex index values, takes care of the structural and chemical aspects in great details. Furthermore, an index, taking care of some kind of shape and size aspects of molecules, can find useful application in doing molecular similarity studies related to drug discovery research.

So far as the information–theoretical similarity measure, INFSIM, is concerned, it has been able to produce similarities that appear to help classify (active / inactive) the studied compounds with significant accuracy. It may also be noted that although we have used TSS indices of the vertices in the present study for computing INFSIM, one can always use any other vertex/substructural index for computing INFSIM. Since Shannon's measure of information takes care of the variety/diversity of the elements in a system, INFSIM may be found to be a measure of interest for molecular similarity studies. Moreover, since the proposed method has produced notably better similarity results than what has been obtained from the Tanimoto measure for some of the studied compounds, discussed earlier, it is tempting to believe that the method may produce significant results in virtual screening of compounds from databases and in molecular similarity studies in general.

Acknowledgment

The authors are thankful to AstraZeneca Research Foundation India, Bangalore, India, for the financial support in developing the software AZMOLTOP. The authors are extremely thankful to Dr. Anand Kumar, Dr. Chhabinath Mandal and Mr. Partha Bag for their contribution in developing AZMOLTOP.

5 REFERENCES

[1] P. Willett, *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth, 1987.

[2] M Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Wiley & Sons Inc, New York, 1990.

[3] J. D. Holliday, C–Y. Hu, and P. Willett, Grouping of Coefficients for the Calculation of Inter–Molecular Similarity and Dissimilarity using 2D Fragment Bit–Strings, *Combinatorial Chemistry and High Throughput Screening* **2002**, 5, 155–166.

[4] C. Shannon and W. Weaver, *Mathematical Theory of Communication*, University of Illinois Press, Urbana, Illinois, 1949.

[5] N. Trinajstić, *Chemical Graph Theory*, *Vol.2*, *Chapter4*, CRC Press, Boca Raton, FL, 1983.

[6] S. C. Basak, B. D. Gute and G. D. Grunwald, Characterization of Molecular Similarity of Chemicals using Topological Invariants; in: Advances in Molecular Similarity, Eds. R. Carbo–Dorca and P. G. Mezey, JAI Press, Stanford, Connecticut, Vol 2, 1998, pp 171–185.

[7] W. Ashby, An Introduction to Cybernetics, Wiley–Interscience, New York, 1956.

[8] C. Raychaudhury, A. Banerjee, P. Bag and S. Roy, Topological Shape and Size of Peptides: Identification of Potential Allele Specific Helper T Cell Antigenic Sites, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 248–254.

[9] M Coll, J. Frau, B. Vilanova, J. Donoso and F. Munoz, Electrostatic and Structural Similarity of Classical and Non–Classical Lactam Compounds, Journal of Computer–Aided Molecular Design **2001**, *15*, 819–833.

[10] The program AZMOLTOP was developed at Indian Institute of Chemical Biology (IICB), Kolkata, India, for AstraZeneca Research Foundation India (AZRFI), Bangalore, India, and is an exclusive property of AZRFI. For queries regarding AZMOLTOP, please contact the communicating author, Dr Indira Ghosh.

[11] Tanimoto similarity values were computed using SYBYL software, ver 6.9, from Tripos, St. Louis, MO.