Springer

Dear Author,

Here are the proofs of your article.

- You can submit your corrections **online**, via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For fax submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/ corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI].
If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information go to: http://www.link.springer.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

Please note: Images will appear in color online but will be printed in black and white.

ArticleTitle	Investigation of similarity and diversity threshold networks generated from diversity-oriented and focused chemical libraries
Article Sub-Title	
Article CopyRight	Springer International Publishing Switzerland (This will be the copyright line in the final PDF)
Journal Name	Journal of Mathematical Chemistry

Corresponding Author	Family Name	**Sukumar**
	Particle	
	Given Name	**N.**
	Suffix	
	Division	Department of Chemistry and Center for Informatics
	Organization	Shiv Nadar University
	Address	Gautam Budh Nagar, 201314, Dadri, UP, India
	Email	n.sukumar@snu.edu.in
	ORCID	http://orcid.org/0000-0002-2724-9944
Author	Family Name	**Prabhu**
	Particle	
	Given Name	**Ganesh**
	Suffix	
	Division	Department of Chemistry
	Organization	Shiv Nadar University
	Address	Gautam Budh Nagar, 201314, Dadri, UP, India
	Email	gp771@snu.edu.in
	ORCID	
Author	Family Name	**Bhattacharya**
	Particle	
	Given Name	**Sudeepto**
	Suffix	
	Division	Department of Mathematics and Center for Informatics
	Organization	Shiv Nadar University
	Address	Gautam Budh Nagar, 201314, Dadri, UP, India
	Email	sudeepto.bhattacharya@snu.edu.in
	ORCID	
Author	Family Name	**Krein**
	Particle	
	Given Name	**Michael P.**
	Suffix	
	Division	
	Organization	Lockheed-Martin Advanced Technology Laboratories

	Address	3 Executive Campus, Suite 600, 08002, Cherry Hill, NJ, USA
	Email	michael.krein@lmco.com
	ORCID	

Schedule	Received	8 October 2015
	Revised	
	Accepted	20 June 2016

Abstract

Topological properties of chemical library networks, such as the average clustering coefficient, average path length, and existence of hubs, can serve as indicators to describe the inherent complexities of chemical libraries. We have used Diversity-Oriented Synthesis (DOS) and Focussed Libraries to investigate the appearance of scale-free properties and absence of small-world behavior in chemical libraries. DOS aims to elicit structural complexity in small compounds with respect to skeleton, functional groups, appendages and stereochemistry. Complexity here indicates incorporation of sp^3 carbons, hydrogen bond acceptors and donors in the molecule. Biological studies have shown how structural complexity enhances the interaction of molecules with complex biological macromolecules. In contrast, Focussed Libraries concentrate on specific scaffolds against a specific biological target. We have quantified the diversity in several DOS and Focussed Libraries based on properties of similarity and dissimilarity threshold networks formed from them. Similarity and dissimilarity networks were generated from diverse chemical libraries at various Tanimoto similarity coefficients (t_c) using FP2 and MACCS fingerprints. The dissimilarity networks at very low t_c threshold led to the absence of small-world behaviors, as evidenced by low average clustering coefficient and high average path length in comparison to Erdös–Renyi networks. Dissimilarity networks exhibit scale free topology as evidenced by a power law degree distribution. The similarity networks at high t_c threshold have shown high clustering coefficients and low average path lengths, without the appearance of hubs. Combining dissimilarity and similarity threshold graphs revealed assortative and dissortative behaviors in the DOS libraries, leading to the conclusion that the vertices of the dissimilarity communities are more likely to share similarity edges, but it is quite unlikely for the vertices in a similarity community to share dissimilarity edges. We propose a simple and convenient diversity quantification tool, QuaLDI (Quantitative Library Diversity Index) to quantify the diversity in DOS and Focussed libraries. We anticipate that these topological properties can be used as descriptors to quantify the diversity in chemical libraries before proceeding for synthesis.

Keywords (separated by '-') Dissimilarity - Similarity - Diversity - Small-world - Chemical space networks

Footnote Information

J Math Chem
DOI 10.1007/s10910-016-0657-0

ORIGINAL PAPER

Investigation of similarity and diversity threshold networks generated from diversity-oriented and focused chemical libraries

Ganesh Prabhu[1] · Sudeepto Bhattacharya[2] · Michael P. Krein[3] · N. Sukumar[4]

Received: 8 October 2015 / Accepted: 20 June 2016

Abstract Topological properties of chemical library networks, such as the average clustering coefficient, average path length, and existence of hubs, can serve as indicators to describe the inherent complexities of chemical libraries. We have used Diversity-Oriented Synthesis (DOS) and Focussed Libraries to investigate the appearance of scale-free properties and absence of small-world behavior in chemical libraries. DOS aims to elicit structural complexity in small compounds with respect to skeleton, functional groups, appendages and stereochemistry. Complexity here indicates incorporation of sp^3 carbons, hydrogen bond acceptors and donors in the molecule. Biological studies have shown how structural complexity enhances the interaction of molecules with complex biological macromolecules. In contrast, Focussed Libraries concentrate on specific scaffolds against a specific biological target. We have quantified

✉ N. Sukumar
n.sukumar@snu.edu.in

Ganesh Prabhu
gp771@snu.edu.in

Sudeepto Bhattacharya
sudeepto.bhattacharya@snu.edu.in

Michael P. Krein
michael.krein@lmco.com

1 Department of Chemistry, Shiv Nadar University, Gautam Budh Nagar, Dadri, UP 201314, India

2 Department of Mathematics and Center for Informatics, Shiv Nadar University, Gautam Budh Nagar, Dadri, UP 201314, India

3 Lockheed-Martin Advanced Technology Laboratories, 3 Executive Campus, Suite 600, Cherry Hill, NJ 08002, USA

4 Department of Chemistry and Center for Informatics, Shiv Nadar University, Gautam Budh Nagar, Dadri, UP 201314, India

the diversity in several DOS and Focussed Libraries based on properties of similarity and dissimilarity threshold networks formed from them. Similarity and dissimilarity networks were generated from diverse chemical libraries at various Tanimoto similarity coefficients (t_c) using FP2 and MACCS fingerprints. The dissimilarity networks at very low t_c threshold led to the absence of small-world behaviors, as evidenced by low average clustering coefficient and high average path length in comparison to Erdös–Renyi networks. Dissimilarity networks exhibit scale free topology as evidenced by a power law degree distribution. The similarity networks at high t_c threshold have shown high clustering coefficients and low average path lengths, without the appearance of hubs. Combining dissimilarity and similarity threshold graphs revealed assortative and dissortative behaviors in the DOS libraries, leading to the conclusion that the vertices of the dissimilarity communities are more likely to share similarity edges, but it is quite unlikely for the vertices in a similarity community to share dissimilarity edges. We propose a simple and convenient diversity quantification tool, QuaLDI (Quantitative Library Diversity Index) to quantify the diversity in DOS and Focussed libraries. We anticipate that these topological properties can be used as descriptors to quantify the diversity in chemical libraries before proceeding for synthesis.

Keywords Dissimilarity · Similarity · Diversity · Small-world · Chemical space networks

1 Introduction

With the increasing popularity of automated screening technologies and the availability of cheap data storage and powerful computers, compound collections have grown into large molecular libraries, often containing millions of chemical substances, and representing valuable intellectual property. However, the potential combinations of just one hundred atoms create a chemical space far exceeding the total number of particles in the universe. All the molecules known to chemists since the dawn of alchemy represent an infinitesimal subspace of this vast chemical space.

There are as many ways to assess molecular similarity as there are distinct molecular properties. Representations using molecular descriptors or fingerprints are often employed to quantify similarities between molecules. Molecular descriptors are constitutional, topological, and geometrical or quantum chemical features of molecules that quantify the relationship between the molecular structure and molecular properties. Threshold networks have been constructed using descriptors representing physicochemical properties [1], such as molecular weight, partition coefficients, and constitution (e.g. the number of sp^2 hybridized atoms), and choosing similarity or dissimilarity threshold values. All pairs of molecules with similarity greater than or equal to the threshold produces a similarity network; pairs of molecules with similarity less than or equal to the threshold produces the corresponding dissimilarity network. Molecular fingerprints are the representations of the molecular structures encoded as bit strings. The bit patterns are characteristic of a given molecule. The fingerprints [2–7] for the molecules computed via Open Babel [5]. In networks char-

Journal: **10910-JOMC** Article No.: **0657** TYPESET DISK LE CP Disp.:**2016/6/24** Pages: **26** Layout: **Small-X**

acterising chemical libraries, molecules in the libraries are treated as vertices (or nodes) and the relationships between pairs of molecules form the edges of the network. Similarity networks constructed on large compound collections using different sets of descriptors have revealed some common features [8–10], such as the small-world property and scale-free degree distributions. The idea of the small-world was inspired from Milgram's *Six Degrees of Separation* [11] and popularized by Watts and Strogatz [12] among physicists and biologists. It refers to communities with highly connected vertices in the network. Scale-free networks, where the probability that a node has k links decays as power-law $p(k) \sim k^{-\alpha}$ (α is the exponent and usually lies between 2 and 3) are often characterized by a small number of highly connected vertices (hubs). A scale-free network's degree distribution is a straight line on a log–log plot. Many real world networks with complex topology have been reported to follow scale-free distributions with dissortative (high degree vertices connecting low degree vertices) degree mixing, such as the world wide web [13], internet [14], protein-protein interaction networks [15], and with assortative (high degree vertices connecting high degree vertices) mixing, such as networks of film actors [12] and business people [16]. There are reports of networks showing both small-world and scale-free behavior [17]. This raises the question of how commonplace and important are small-world and scale-free properties within classes of chemical libraries.

Chemists construct molecular libraries for a variety of reasons, using different synthetic strategies. We adopt focused library design strategies when the objective is to look at molecules that are chemically similar to known drug leads. In other situations, it is more useful to cast the net wide with the hope of discovering new types of molecules. One popular strategy to create large molecular libraries is combinatorial chemistry [18], where various combinations of functional groups are attached at different substitution points to a molecular scaffold, but it has been argued that most combinatorial approaches fail to deliver truly novel compounds [19]. Willet [20], Agrafiotis et al. [21], and Wintner et al. [22], have proposed different algorithms to quantify the diversity in a chemical library.

Quantitative structure activity relationships (QSAR [23]) are an effective tool in drug design used to predict the biological activities of molecules based on their structural similarities. However, factors such as solubility, permeability, polymorphism, cytotoxicity, mutation and drug resistance represent major challenges encountered by chemists, biologists and pharmacists, forcing researchers to broaden the spectrum of new chemical entities. Broadening the chemical spectrum requires a chemically diverse set of compounds obtained from synthetically feasible number of steps targeting various regions of biological space. The quest for diverse compounds is supported by the Diversity-oriented synthesis (DOS) [24] strategy. DOS helps to synthesize molecular libraries possessing structural complexity as well as skeletal and stereochemical diversity. For the present study, we are using DOS (expressing diversity) [25–28], and focussed (expressing similarity) libraries [29] to quantify the diversity through network theory.

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

2 Dataset, pair-wise similarity and dissimilarity measure

We represented the DOS libraries [25–28], and Focussed Libraries [29] in SDF format, converting molecules into fingerprints using Open Babel [5]. We used the simple and popular Open Babel fingerprints such as FP2 [17] (a path based fingerprint characterized by 7-atom linear chain fragments that correspond to 1024 bits) and MACCS [30] fingerprints (that uses SMARTS patterns to describe the molecular sub structure or subgraph).

In order to compare keys we used the Tanimoto similarity distance, t_c, a pair-wise measure represented by the equation

$$t_c = \frac{A \cap B}{(A \cup B) - (A \cap B)}, \tag{2.1}$$

where $A \cap B$ is the number of common bits in the structural fingerprints of compounds 'A' and 'B' and $A \cup B$ is the sum of the numbers of bits in the fingerprints of compounds 'A' and 'B'.

Networks based on Tanimoto coefficient cut-offs of the structural fingerprints were generated using Eq. 2.11 and the igraph package [6,31]. Dissimilarity networks at thresholds $t_c \leq 0.5, 0.4, 0.3, 0.2, 0.22$ and similarity networks at thresholds $t_c \geq 0.8, 0.9, 0.95, 0.98, 0.99$ and 0.995 were generated. The similarity and dissimilarity networks were compared at equivalent density of edges to maintain consistency in the properties of the threshold networks.

2.1 Pattern of labelling threshold network

The scheme used for labelling the threshold networks is described in Table 1. The threshold networks were compared with their corresponding Erdös–Renyi random networks (ERNs) at equivalent edge densities. For example, the ERN corresponding to the threshold network $DOS118\_F\_t_c \leq 0.22$ is generated from the equivalent number of molecules (N = 118) by connecting them randomly with a probability of connection ($p = 0.0016$) nearly equivalent to the edge density of the threshold network as mentioned in Table 2.

3 Network properties

A graph [11] is an algebraic object represented by an ordered triple comprising a non-empty set (V, E, Ψ_G), where $vertices$, $V = \{v_i | i = 1, 2, 3, \ldots, n\}$, $edges$, $E = \{e_i | i = 1, 2, 3, \ldots, m\}$, such that $V \cap E = \varnothing$ and an incident function is defined by, $\Psi_G : E \rightarrow [V]^2; e \mapsto \Psi_G(e) = \left\{v_i, v_j\right\}$. A network is a dynamical object defined by the four-tuple, $G = \left(V_t, E_t, \Psi_{N_t}, J_t\right)$, where t is a time parameter, simulated or real; J_t is an algorithm for defining the behavior of vertices and edges of the network with time. Our study involves the static libraries (non-dynamic chemical libraries) in conjugation with algorithm J; nevertheless, we refer to chemical similarity and dissimilarity graphs as networks throughout this paper. Here every $v_i \in V$ is represented as a *vertex*

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

Table 1 Labelling scheme used for the threshold networks

Network label	Chemical library/Network	No. of compounds./ Size of the library (n)	Fingerprint F= FP2 M = MACCS	$t_c \leq$ threshold for dissimilarity networks	$t_c \geq$ threshold for similarity networks	*p*
DOS118_F_$t_c \leq 0.22$	DOS	118	F	$t_c \leq 0.22$	NA	NA
DOS118_F_$t_c \geq 0.95$	DOS	118	F	NA	$t_c \geq 0.95$	NA
DOS41_F_$t_c \leq 0.36$	DOS	41	F	$t_c \leq 0.36$	NA	NA
DOS41_F_$t_c \geq 0.95$	DOS	41	F	NA	$t_c \geq 0.95$	NA
DOS32_M_$t_c \leq 0.2$	DOS	32	M	$t_c \leq 0.2$	NA	NA
DOS32_M_$t_c \geq 0.8$	DOS	32	M	NA	$t_c \geq 0.8$	NA
FL41_M_$t_c \leq 0.3$	FL	41	M	$t_c \leq 0.3$	NA	NA
FL41_M_$t_c \geq 0.98$	FL	41	M	$t_c \leq 0.22$	$t_c \geq 0.98$	NA
ERN (41, 0.022)	ERN	41	NA	NA	NA	0.022

The entry in the table follows a unique network-labelling pattern. The threshold networks discussed in this paper follow the same pattern mentioned in this table. The hyphen '_' used here is to separate the significant characters in the label. *DOS* Diversity oriented synthesis, *FL* Focussed library, *ERN* Erdös–Renyi random network, F = *FP2* Fingerprint 2 (open babel fingerprint), M = *MACCS* Molecular access system, t_c Tanimoto similarity coefficient, *p* probability of wiring, *NA* Not applicable

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

Table 2 Network properties of DOS libraries (N = 118, 41, 32) and Focussed library (FL, N = 41) at various dissimilarity and similarity thresholds using FP2 fingerprints

Dissimilarity networks-FP2	$C(G)$	$L(G)$	$C(G) > C_E(G)$	$C(G) \gg C_E(G)$	$L(G) < L_E(G)$	Average degree	No. of edges	$D(G)$
DOS118_F_$t_c \leq 0.22$	0	1.64	No	No	No	1.6	11	0.0016
ERN (118, 0.0016)	0	1.125				0.12	7	0.001
DOS41_F_$t_c \leq 0.36$	0	1.9	No	No	Yes	4.14	29	0.017
ERN (41, 0.017)	0	1.7				0.73	15	0.018
DOS32_F_$t_c \leq 0.4$	0.05	2.13	No	No	Yes	4.5	54	0.11
ERN (32, 0.11)	0.1	2.7				3.06	49	0.1
FL41_F_$t_c \leq 0.5$	0	1.75	No	No	Yes	2.75	11	0.013
ERN (41, 0.013)	0.16	2.4				0.7	14	0.017
Similarity networks								
DOS118_F_$t_c \geq 0.95$	0.75	1.03	Yes	Yes	Yes	1.07	28	0.004
ERN (118, 0.004)	0	1.35				0.37	22	0.003
DOS41_F_$t_c \geq 0.95$	1	1	Yes	Yes	Yes	1.43	7	0.012
ERN (41, 0.012)	0	1.25				0.6	12	0.015
DOS32_F_$t_c \geq 0.8$	0.81	1.33	Yes	Yes	Yes	2.24	28	0.06
ERN (32, 0.06)	0.08	3.4				2.12	34	0.07
FL41_F_$t_c \geq 0.995$	1	1	Yes	Yes	Yes	4.0	18	0.022
ERN (41, 0.022)	0	1.4				0.7	14	0.04

The abbreviation scheme for similarity and dissimilarity networks follow Table 1. The table describes feature of dissimilarity threshold networks showing absence of small-world behavior, whereas similarity threshold networks show small-world behavior. The dissimilarity networks abbreviated as DOS118_F_$t_c \leq 0.22$, DOS41_F_$t_c \leq 0.36$ and DOS32_F_$t_c \leq 0.4$ refers to networks generated from DOS library comprising 118, 41 and 32 compounds using Open Babel fingerprint FP2 at Tanimoto similarity coefficient, $t_c \leq 0.22$, 0.36 and 0.4. The dissimilarity network abbreviated as FL41_F_$t_c \leq 0.5$, refers to a network generated from a Focussed library (FL) comprising 41 compounds using Open Babel fingerprint FP2 (F) at Tanimoto similarity coefficient, $t_c \leq 0.5$. $C(G)$ *and* $C_E(G)$ are the average clustering coefficient of threshold and Erdös–Renyi network. $L(G)$ *and* $L_E(G)$ are the average path lengths of threshold and Erdös–Renyi network. $D(G)$ = Network density

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

(or node) and the discretized similarity measure forms the connection or *edge* between adjacent vertices. The total number of vertices in the network represents the *order* of the network, $(V) = |V| = n$ and the total number of connections or edges between the vertices represents the *size* of the network, $(E) = |E| = m$. The total possible number of connections or edges in the network is given by $|E|_{max} = \frac{n((n-1)}{2}$. The network density, $D(G) = \frac{number\ of\ edges}{total\ number\ of\ edges} = \frac{number\ of\ edges}{total\ number\ of\ edges} = \frac{m}{n(n-1)/2}$.The structure of the network (V, E) can be represented as an adjacency matrix (n, n),

$$\boldsymbol{A} = (\boldsymbol{a_{ij}}) = \begin{pmatrix} \boldsymbol{a_{1i}} & \cdots & \boldsymbol{a_{1j}} & \ldots & \boldsymbol{a_{1n}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{a_{2i}} & \cdots & \boldsymbol{a_{2j}} & \ldots & \boldsymbol{a_{2n}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{a_{ni}} & \ldots & \boldsymbol{a_{nj}} & \ldots & \boldsymbol{a_{nn}} \end{pmatrix}$$, where the entries a_{ij} = if i and j are adjacent; $a_{ij} = 0$ if i and j are not adjacent. Network properties act as functions of the thresholds [32]. The current study focuses on properties such as the vertex degree (k), the degree distribution $p(k)$, the average clustering coefficient $C(G)$, average path length $L(G)$, degree assortativity (r) and modularity (Q), defined below. The network properties of these chemical libraries were compared with the corresponding Erdös–Renyi random networks at comparable edge density. Hubs in a network are the vertices with maximum degree, $H = \max\{k(v_i)\}$ in a local neighborhood.

3.1 Small-world property

The Clustering Coefficient (C_{i_G}) [12] of a vertex in a graph/network is defined as the actual number of triangles (t_r) which pass through the vertex 'i' divided by the total number of possible triangles of vertex 'i'.

$$C_{i_G} = t_{r_i} / (n(n-2)), \tag{3.1}$$

A clique C_q is a maximal complete subgraph in a graph, *i.e.* a subgraph in which every pair of vertices is connected [33]. For example, the hexagonal array of vertices with similarity edges (colored blue) in Fig. 3c, wherein every node pair is connected, represents the maximal complete subgraph or clique in a network. Detection and analysis of the cliques in the network reveals a detailed view of the community structure (highly connected vertices with similar or dissimilar feature) within it. The neighborhood S_N of a vertex i in a network G is the set of all its adjacent vertices j. The neighborhood of i in G is given by $\Gamma(i) = \{j \in V : i, j \in E\}$.

3.1.1 Average clustering coefficient $C(G)$

Average Clustering Coefficient $C(G)$ of a network (G) is the clustering coefficient C_{i_G} of the node 'i' averaged over all 'n' vertices of the network G.

$$C(G) = \sum_{i=1}^{n} \frac{C_{i_G}}{n}, \tag{3.2}$$

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

The dissimilarity networks generated from FP2 and MACCS fingerprints exhibit $C(G) \cong C_E(G)$. Fig. 3a shows mostly second-order clustering characterised by a minimal ring size of four in similarity networks with $C(G) \gg C_E(G)$, as shown in Tables 2 and 3.

3.1.2 Average path length $L(G)$

Average Path Length $L(G)$ [12] is the shortest path $d_{i,j}$ connecting a pair of vertices, averaged over all pairs of vertices 'n_p' in the network G (Eq. 3.3).

$$L(G) = \sum_i \sum_j \frac{d_{i,j}}{n_p} \tag{3.3}$$

A community of highly connected vertices with very high average clustering coefficient $C(G)$ and relatively short average path length $L(G)$ in a network is known as a small-world. The existence of hubs in a network acts as an indicator of the presence of the small-world property.

This property plays a significant role in describing the existence or absence of the small-world behavior in the network with low or high $L(G)$ in comparison with $L_E(G)$, the average path length of ERN measured at nearly equivalent network density.

The dissimilarity networks generated from FP2 and MACCS fingerprints exhibit $L(G) \cong L_E(G)$, while similarity networks show very high $L(G) < L_E(G)$ as seen from Tables 2 and 3.

3.1.3 Small-world metric

The existence of the small-world property in a network can be characterised by the following metrics:

(a) $C(G) \gg C_E(G)$, where $C(G)$ = Average Clustering Coefficient of a network, G and $C_E(G)$ = Average Clustering Coefficient of the Erdös–Renyi random network constructed from the same vertices at nearly equivalent edge density [13,34],
(b) $L(G) < L_E(G)$, where L_G = Average Path Length [12] of a network, G and $L_E(G)$ = Average Path Length of the corresponding Erdös–Renyi random network constructed from the same vertices at nearly equivalent edge density,
(c) $L(G) \propto LogN$, where $L(G)$ = Average Path Length of a network, G should be proportional to $LogN$. The third metric refers to the growing network (dynamic) but can be ignored since the study involves only analysis of static networks (N = constant).

If the network fails to satisfy any of the above metrics, it lacks small-world character.

The dissimilarity networks mentioned in the Tables 2 and 3 fail to satisfy the metrics (a) and (b), thereby establishing absence of the small-world property.

The networks at dissimilarity thresholds $t_c \leq 0.3$–0.7 show properties, $C(G) \cong C_E(G)$ *and* $L(G) \cong L_E(G)$ resembling the corresponding Erdös–Renyi random network constructed from the same vertices (N = 118) at nearly equivalent edge density

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

Table 3 Network properties of DOS libraries (N=118, 32, 41) and Focussed library (FL, N=41) at various dissimilarity and similarity thresholds using MACCS(M) fingerprints

Dissimilarity networks-MACCS	$C(G)$	$L(G)$	$C(G) > C_E(G)$	$C(G) \gg C_E(G)$	$L(G) < L_E(G)$	Average degree	No. of edges	$D(G)$
DOS118_M_$t_c \leq 0.2$	0	1.81	No	No	Yes	2	28	0.004
ERN (118, 0.0045)	0	2.11				0.52	31	0.0045
DOS41_M_$t_c \leq 0.2$	0	1.66	No	No	No	1.66	5	0.0061
ERN (41, 0.0061)	0	1.4				0.3	6	0.0097
DOS32_M_$t_c \leq 0.2$	0	1.86	No	No	No	1.66	5	0.01
ERN (32, 0.01)	0	1.25				0.19	3	0.006
FL41_M_$t_c \leq 0.3$	0	1.92	No	No	Yes	3.64	31	0.038
ERN (41, 0.038)	0	2.63				1.27	26	0.032
Similarity networks-MACCS								
DOS118_M_$t_c \geq 0.9$	0.67	1.1	Yes	Yes	Yes	1.6	28	0.004
ERN (118, 0.004)	0	2.11				0.52	31	0.0045
DOS41_M_$t_c \geq 0.8$	0.5	1.37	Yes	Yes	Yes	1.21	14	0.02
ERN (41, 0.02)	0	3.3				0.97	20	0.024
DOS32_M_$t_c \geq 0.8$	0.75	1.16	Yes	Yes	Yes	1.54	10	0.02
ERN (32, 0.02)	0	1.38				0.56	9	0.018
FL41_M_$t_c \geq 0.98$	1	1	Yes	Yes	Yes	4.9	22	0.026
ERN (41, 0.026)	0	1.77				0.78	16	0.019

The table describes features of dissimilarity threshold networks (generated using MACCS fingerprints) showing the absence of small-world behavior, whereas similarity threshold networks exhibit the small-world behavior. The dissimilarity networks abbreviated as DOS118_F_$t_c \leq 0.22$, DOS41_F_$t_c \leq 0.36$ and DOS32_F_$t_c \leq 0.4$ refer to the networks generated from DOS library comprising 118, 41 and 32 compounds using Open Babel fingerprint FP2 at Tanimoto similarity coefficient, $t_c \leq 0.22$, 0.36 and 0.4. The dissimilarity network abbreviated as FL41_F_$t_c \leq 0.5$, refers to a network generated from Focussed library (FL) comprising 41 compounds using Open Babel fingerprint FP2 (F) at Tanimoto similarity coefficient, $t_c \leq 0.5$. $C(G)$ *and* $C_E(G)$ are the average clustering coefficient of threshold and Erdös–Renyi network. $L(G)$ *and* $L_E(G)$ are the average path lengths of threshold and Erdös–Renyi network. $D(G)$ = Network density

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

(Eg., ERN (N = 118, D(G) = 0.18); this reflects pseudo random behavior of the network, as mentioned in Table 4.

3.1.4 Degree assortativity [35,36]

The nature of community structures in the threshold networks depends on the assortative and dissortative mixing of degrees of vertices in the network. Degree assortativity is the correlation coefficient between the degrees of connected vertices, given by Eq. 3.4:

$$r = \frac{\sum_{1 \le i, j \le n} \left(A_{ij} - \frac{k_i k_j}{2m} \right) k_i k_j}{\sum_{1 \le i, j \le n} \left(k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) k_i k_j}, \tag{3.4}$$

where an element of the adjacency matrix of the network,
$A_{ij} = \begin{cases} 1 \; if \; nodes \; i \; and \; j \; are \; connected, \\ 0 \; otherwise \end{cases},$
$k_i k_j$ are the degrees of node i and j, respectively; δ_{ij} = Kronecker delta function, $\begin{cases} 0 \; i = j \\ 1 \; i \neq j \end{cases}$; n and m are the order (total number of vertices in the network) and the size (total number of edges in the network) of the network [32]. The degree assortativity provides information about the vertices of high degree connecting vertices of high degree and the low degree vertices connecting low degree vertices; its value can be positive or negative. Negative values represent degree dissortativity, characterised by vertices of high degree connecting low degree vertices.
Assortative and Dissortative degree mixing in DOS libraries The nature of community structures in the threshold networks depends on the assortative and dissortative mixing of degrees of vertices in the network. Table 5 illustrates the assortative and dissortative mixing of degrees of vertices in threshold networks. The dissimilarity networks show the dissortative (negative degree assortativity) behavior characterised by high degree vertices connected to low degree vertices, as evidenced by the star subgraphs shown in Fig. 3a. The high negative assortativity associated with absence of cliquishness represents high dissortative mixing of vertices in the networks, thus reflecting dissimilarity or diversity in the library, as illustrated in Table 5. On the contrary, similarity networks show high degree assortativity accompanied by cliquishness featuring high degree vertices linked to other high degree vertices, thereby leading to small-world structures in the network (Fig. 3b).

3.1.5 Modularity

Modularity [37] of a network quantifies the community structures in the network, separating the vertices into groups in such a way that there exist enough edges between the vertices within a group but very few edges between the groups. A number of algorithms have been recently proposed to find the community structures in networks, such as the Edge betweenness divisive algorithm proposed by Girvan and Newman [38,39], and the walktrap algorithm proposed by Pons and Latapy [40].We used Newman's fast

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

Table 4 Dissimilarity threshold networks of DOS library (118 molecules) using FP2 and MACCS fingerprints showing pseudo random behavior at $t_c \leq 0.3–0.7$

Dissimilarity network-FP2	$C(G)$	$L(G)$	$C(G) > C_E(G)$	$C(G) \cong C_E(G)$	$L(G) \cong L_E(G)$	Average degree	No. of edges	$D(G)$
DOS118_F_$t_c \leq 0.3$	0.09	2.18	Yes	Yes	Yes	8.83	521	0.074
ERN (118, 0.074)	0.07	2.46				8.3	488	0.07
DOS118_F_$t_c \leq 0.4$	0.36	1.65	Yes	Yes	Yes	40.4	2383	0.34
ERN (118, 0.34)	0.33	1.67				38.5	2272	0.33
DOS118_F_$t_c \leq 0.5$	0.67	1.35	Yes	Yes	Yes	76	4479	0.64
ERN (118, 0.64)	0.64	1.36				74.77	4412	0.64
DOS118_F_$t_c \leq 0.6$	0.84	1.16	Yes	Yes	Yes	97.66	5762	0.83
ERN (118, 0.83)	0.84	1.16				98.1	5786	0.84
DOS118_F_$t_c \leq 0.7$	0.93	1.06	Yes	Yes	Yes	109.4	6453	0.93
ERN (118, 0.93)	0.93	1.07				108.75	6416	0.93
Dissimilarity Networks-MACCS								
DOS118_M_$t_c \leq 0.3$	0.15	1.7	No	No	Yes	21.15	1248	0.18
ERN (118, 0.18)	0.17	1.84				20.64	1218	0.18
DOS118_M_$t_c \leq 0.4$	0.6	1.38	Yes	No	Yes	69.2	4083	0.59
ERN (118, 0.59)	0.58	1.41				68.54	4044	0.59
DOS118_M_$t_c \leq 0.5$	0.8	1.2	Yes	No	Yes	94.5	5574	0.80
ERN (118, 0.8)	0.8	1.2				93.66	5526	0.8
DOS118_M_$t_c \leq 0.6$	0.92	1.1	Yes	No	Yes	108.23	6386	0.92
ERN (118, 0.92)	0.92	1.07				108.3	6390	0.92
DOS118_M_$t_c \leq 0.7$	0.97	1.03	Yes	No	Yes	113.3	6683	0.96
ERN (118, 0.96)	0.96	1.04				112.45	6635	0.96

The networks at dissimilarity threshold $t_c \leq 0.3–0.7$ using FP2(F) and MACCS(M) fingerprints showing properties ($C(G) \cong C_E(G)$ *and* $L(G) \cong L_E(G)$) resembling those of the corresponding Erdös–Renyi random network constructed from the same vertices (N = 118) at nearly equivalent edge density, which is characteristic of pseudo random behavior. $C(G)$ *and* $C_E(G)$ are the average clustering coefficient of threshold and Erdös–Renyi network. $L(G)$ *and* $L_E(G)$ are the average path lengths of threshold and Erdös–Renyi network. *D(G)* Network density

Journal: **10910-JOMC** Article No.: **0657** TYPESET DISK LE CP Disp.:**2016/6/24** Pages: **26** Layout: **Small-X**

Table 5 Degree assortativity and Modularity of various threshold networks from DOS libraries (N = 118, 41, 32) and Focussed library (FL, N = 41) generated using FP2(F) and MACCS(M) fingerprints and the corresponding Erdös–Renyi random network (ERN) constructed from the same vertices at nearly equivalent edge density

Dissimilarity networks-FP2	r	Q and Q_E	$D(G)$	Dissimilarity network-MACCS	r	Q and Q_E	$D(G)$
DOS118_F_$t_c \leq 0.22$	−0.158	0.64	0.0016	DOS118_M_$t_c \leq 0.2$	−0.57	0.42	0.004
ERN (118, 0.0016)	−0.166	0.81	0.001	ERN (118, 0.0045)	0.2	0.93	0.0045
DOS41_F_$t_c \leq 0.36$	−0.53	0.15	0.017	DOS41_M_$t_c \leq 0.2$	−1	0	0.0061
ERN (41, 0.017)	−0.41	0.7	0.018	ERN (41, 0.0061)	−0.61	0.61	0.0097
DOS32_F_$t_c \leq 0.4$	−0.58	0.18	0.11	DOS32_M_$t_c \leq 0.2$	−0.74	0.22	0.01
ERN (32, 0.11)	−0.23	0.4	0.1	ERN (32, 0.01)	−0.5	0.44	0.006
FL41_F_$t_c \leq 0.5$	−0.668	0.062	0.013	FL41_M_$t_c \leq 0.3$	−0.72	0.087	0.038
ERN (41, 0.013)	−0.095	0.58	0.017	ERN (41, 0.038)	−0.03	0.66	0.032
Similarity networks-FP2				*Similarity networks-MACCS*			
DOS118_F_$t_c \geq 0.95$	0.70	0.95	0.004	DOS118_M_$t_c \geq 0.9$	0.8	0.85	0.004
ERN (118, 0.004)	0.2	0.88	0.003	ERN (118, 4-E03)	0.2	0.88	0.0045
DOS41_F_$t_c \geq 0.95$	1	0.73	0.012	DOS41_M_$t_c \geq 0.8$	0.69	0.82	0.02
ERN (41, 0.012)	0.3	0.73	0.015	ERN (41, 0.02)	0.14	0.74	0.024
DOS32_F_$t_c \geq 0.8$	0.59	0.77	0.06	DOS32_M_$t_c \geq 0.8$	0.35	0.72	0.02
ERN (32, 0.06)	−0.09	0.46	0.07	ERN (32, 0.02)	−0.53	0.64	0.018
FL41_F_$t_c \geq 0.995$	1	0.28	0.022	FL41_M_$t_c \geq 0.98$	1	0.2	0.026
ERN (41, 0.022)	0.144	0.74	0.04	ERN (41, 0.026)	−0.19	0.73	0.019

The dissimilarity threshold networks show negative degree assortativity (dissortativity), describing networks with high degree vertices connecting to low degree vertices. Dissimilarity networks demonstrate modularity $Q < Q_E$. The similarity threshold networks show positive degree assortativity, describing networks with high degree vertices linked to high degree vertices. Their modularity $Q \cong Q_E$ except for DOS41_F_$t_c \geq 0.995$ and DOS41_M_$t_c \geq 0.98$. r = degree assortativity, Q and QE = Modularity of threshold and Erdös–Renyi network

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

greedy algorithm [2] to find the communities in dissimilarity and similarity networks. Modularity is described by Eq. 3.5:

$$Q = \frac{1}{2m} \sum_{1 \le i, j \le n} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(i, j) \tag{3.5}$$

where $\delta(i, j) \begin{cases} 1 & \text{if } i \text{ and } j \text{ belong to the same community} \\ 0 & \text{otherwise,} \end{cases}$, the network with more community structure reflects high modularity. Modularity ranges between [−1 and 1). Dissimilarity networks show low modularity values as compared to similarity networks at comparable edge density, as mentioned in Table 5.

3.2 Scale-free dissimilarity networks

Scale-free networks, where the probability that a node has k links decays as a power-law

$$p(k)\, \alpha\, k^{-\alpha}, \tag{3.6}$$

are often characterized by a small number of highly connected vertices (hubs). A scale-free network's degree distribution is linear on a log–log plot. While there have been studies of large real world complex networks (including chemical libraries) which exhibit scale-free topology based on similarity measures, there have been no studies using dissimilarity measures.

3.2.1 Power law fit [41,42]

Scale free networks and complex systems may not always follow power law degree distributions (Eq. 3.6). In a scale free network, the probability $p(k)$ that a node has degree ׳k׳ decays exponentially, where α is the exponent of the fitted power-law distribution. The minimum value of k is k_{min}, above which the theoretical degree distribution starts fitting the data plot; k_{min} can be estimated by Kolmogorov and Smirnov's (KS) [43] test. The KS statistic D estimates the maximum separation between the data and the fitted cumulative distribution function (CDF),

$$D = max_{k \ge k_{min}} \left| S(k) - P(k) \right|, \tag{3.7}$$

where $S(k)$ and $P(k)$ are the CDFs of the data and the fitted model, respectively. The appropriate k_{min} is the value which minimizes D.

For the given value of k_{min} the scaling parameter is estimated by using a maximum likelihood estimator (MLE) [42] optimising maximum log-likelihood, defined as

$$\hat{\alpha} \cong 1 + n \left[\sum_{i=1}^{n} \log \left(\frac{k_i}{k_{min} - 0.5} \right) \right]^{-1}, \tag{3.8}$$

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

Table 6 Power law fitting

Dissimilarity networks-FP2 or MACCS	α	k_{min}	Log likelihood	KS. stat	KS.p
DOS118_F_$t_c \leq 0.3$	2.44	8	−137.21	0.066	0.99
DOS118_M_$t_c \leq 0.2$	2.29	1	−32.93	0.064	0.99
DOS41_F_$t_c \leq 0.4$	4.87	9	−11.12	0.16	0.99
DOS41_M_$t_c \leq 0.3$	3.3	10	−24.23	0.127	0.99
DOS32_F_$t_c \leq 0.4$	2.42	3	−37.55	0.11	0.98
DOS32_M_$t_c \leq 0.2$	2.13	1	−7.01	0.07	1
FL41_F_$t_c \leq 0.5$	3.19	3	−5.22	0.11	1.0
FL41_M_$t_c \leq 0.3$	3.18	4	−14.31	0.15	0.99

Power law fits of dissimilarity networks using igraph package in R, which performs a test to determine whether a power law distribution is plausible or not. Dissimilarity network-FP2: Dissimilarity threshold networks generated from FP2 fingerprints. Dissimilarity network-MACCS: Dissimilarity threshold networks generated from MACCS fingerprints. α = Numeric scalar, the exponent of the fitted power-law distribution. k_{min} = Numeric scalar, the minimum value from which the power-law distribution was fitted. In other words, only values of k larger than k_{min} were used from the input vector. Log likelihood = Numeric scalar, the log-likelihood of the fitted parameters. KS.stat = Numeric scalar, the test statistic of a Kolmogorov–Smirnov test that compares the fitted distribution with the input vector. Smaller scores denote better fit. KS.p = Numeric scalar, the *p* value of the Kolmogorov–Smirnov test. Small *p* values (less than 0.05) indicate that the test rejects the null hypothesis. The dissimilarity graphs such as DOS118_F_$t_c \leq 0.3$, DOS32_F_$t_c \leq 0.4$, DOS118_M_$t_c \leq 0.2$ and DOS32_M_$t_c \leq 0.2$ exhibit power-law like behavior with α value between 2 and 3

where $\hat{\alpha}$ is the KS statistic derived from data, k_i, $i = 1, \ldots, n$, are the observed values of k such that $k_i \geq k_{min}$. The results obtained using Eq. 3.8 are listed in Table 6.

3.2.2 Degree distribution p(k)

Degree distribution $p(k)$ is the probability that a fraction of the vertices in the network has k links. However, a change in the similarity/dissimilarity threshold leads to variation in the edge density, which in turn changes the degree distribution. Degree distributions of dissimilarity threshold networks have demonstrated power law (scale free behavior) as shown in Fig. 1a–e. However, other distributions such as exponential, lognormal and Poisson may also fit the data, as demonstrated in the CDF v/s Degree plots (Fig. 2a–d) in Sect. 3.23. Further, at very high similarity thresholds, the resulting networks did not follow any conventional degree distribution, irrespective of the type of fingerprints used.

3.2.3 Cumulative distribution function (CDF)

The dissimilarity networks were found to fit various degree distributions at different values of k_{min}. Fig. 2a–d demonstrates that the CDF may be fit by power law (at the tail end), lognormal and exponential distributions. The trends show that dissimilarity networks demonstrate best fit to power law and lognormal distribution as compared to other statistics.

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

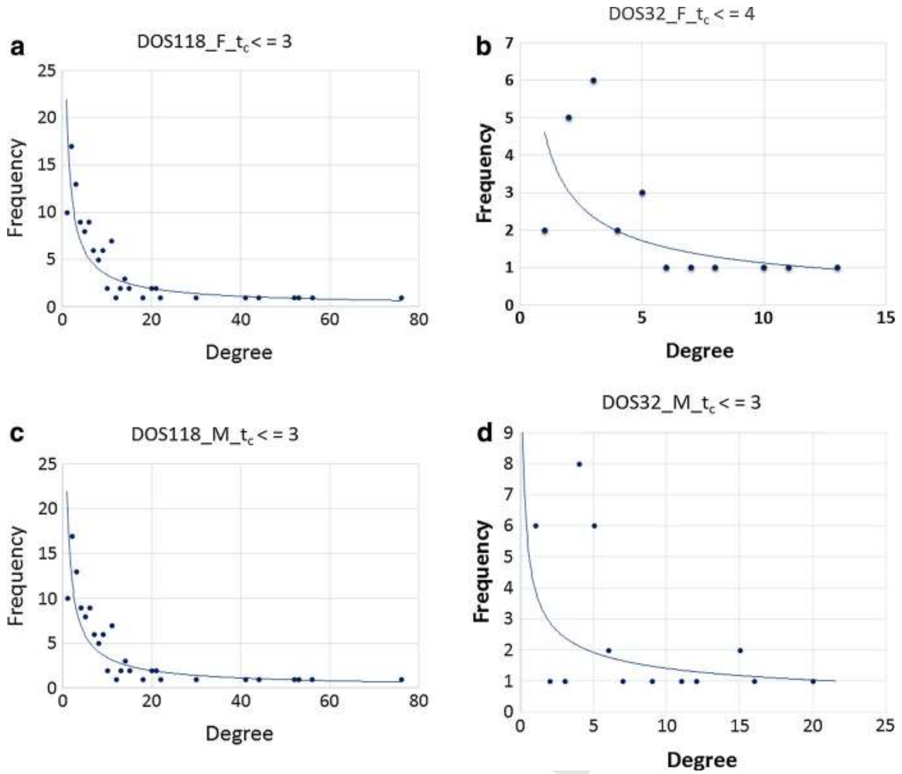

Fig. 1 Degree distributions of various dissimilarity networks. The DOS libraries (N = 118, 32) and Focussed Library (N = 41) at FP2 and MACCS threshold values $t_c \leq 0.3 - 0.4$ display power law (scale free behavior) and exponential (1a, c, e) distribution. The Focussed library (N = 41) at FP2 and MACCS threshold value $t_c \leq 0.3 - 0.5$ display skewed lognormal distributions (1b and d)

Fitting a power law to a continuous or discrete data is done by testing the null hypothesis. The *p* value generated quantifies the plausibility of the hypothesis. To test whether they follow a power law, we used a null model (H_0) as well as an Alternative model (H_1). Null hypothesis, H_0 = data follows power law distribution. Alternative hypothesis, H_1 = data does not follow power law distribution. If the *p* value is > 0.1 then one cannot reject the null hypothesis. If the *p* value is < 0.1 then one has to reject the null hypothesis. To assess the scale-free ehaviour in the network, the best way is to fit the data to a power law.

From Table 6, it is evident that dissimilarity graphs such as DOS118_F_$t_c \leq 0.3$, DOS32_F_$t_c \leq 0.4$, DOS118_M_$t_c \leq 0.2$ and DOS32_M_$t_c \leq 0.2$ demonstrate power-law like ehaviour with the exponent α lying between 2 and 3. However, these networks also fit a log-normal distribution, as shown in Fig. 2a–d.

Journal: **10910-JOMC** Article No.: **0657** TYPESET DISK LE CP Disp.:**2016/6/24** Pages: **26** Layout: **Small-X**

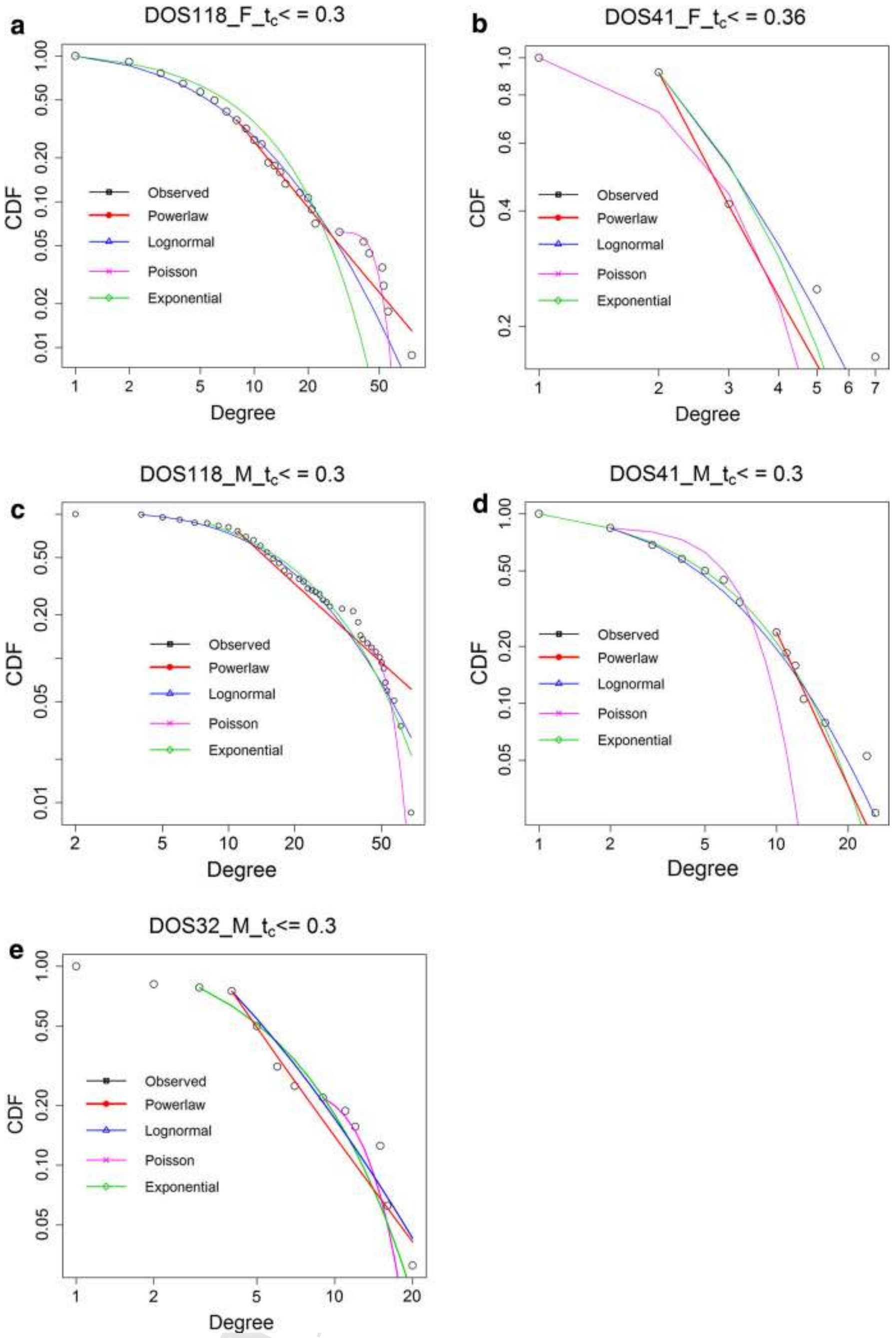

Fig. 2 Fits of the cumulative distribution function (CDF) v/s degree of dissimilarity networks. The CDF of DOS libraries (N = 118, 41) at FP2 and MACCS threshold values $t_c \leq 0.3 - 0.36$ fit a power law at the tail end of the distribution besides fitting to lognormal and exponential distributions

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

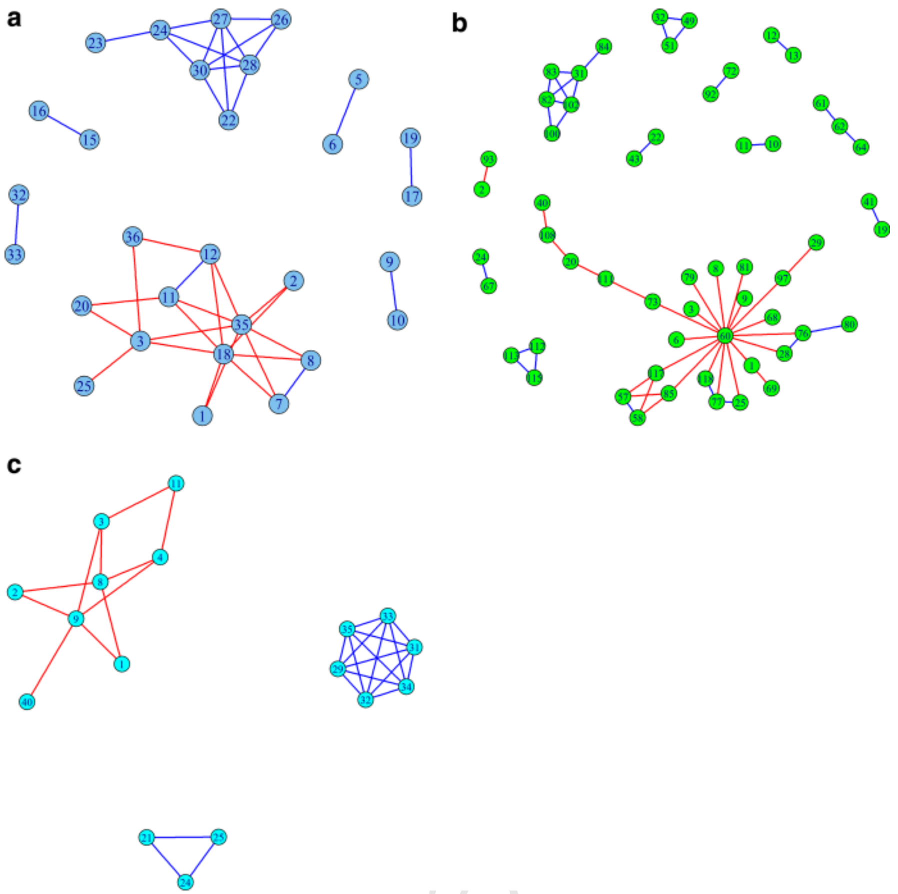

Fig. 3 Visual representations of similarity and dissimilarity threshold networks. The Fruchterman Reingold layout (force field directed layout) used to visualize the undirected threshold networks. *Red* and *blue* edges represent dissimilarity and similarity edges, respectively. Such combined representations of dissimilarity and similarity networks can be used to conveniently visualize the characteristics of a chemical library at a glance. The Fig. 3a–c show that the combined dissimilarity and similarity networks, DOS41_F_$t_c \leq 0.36 + t_c \geq 0.9$ (DOS Library with 41 molecules, FP2 fingerprints with threshold $t_c \leq 0.36$ and $t_c \geq 0.9$), DOS118_M_$t_c \leq 2 + t_c \geq 9$ (DOS Library with 118 molecules, MACCS fingerprints with threshold $t_c \leq 0.2$ and $t_c \geq 0.9$) and FL41_F_$t_c \leq 5 + t_c \geq 995$ (Focussed library with 41 molecules, FP2 fingerprints with threshold $t_c \leq 0.5$ and $t_c \geq 0.995$) exhibit either homophily or diversity. The combined networks demonstrate more likely sharing of similarity edges between the vertices in dissimilarity subnetworks and less likely sharing of dissimilarity edges between the vertices in the similarity subnetworks

4 Quantitative library diversity index (QuaLDI)

Exploring hitherto unexplored regions of chemical space is critically important for identifying novel diverse chemical structures with drug-like properties. Quantifying the diversity in a chemical library is important for optimising structural diversity along with other features. This helps in choosing optimal reagents and substrates for generating highly diverse compounds that can be further explored in drug design.

Journal: **10910-JOMC** Article No.: **0657** TYPESET DISK LE CP Disp.:**2016/6/24** Pages: **26** Layout: **Small-X**

Table 7 Dissimilarity and Similarity threshold networks of DOS library (118 molecules) generated using FP2 fingerprints

Dissimilarity network-FP2	$C(G)$	$L(G)$	$C(G) > C_E(G)$	$C(G) \gg C_E(G)$	$L(G) < L_E(G)$	d_G	(r)	(Q)	$D(G)$	QuaLDI %
DOS118_F_$t_c \leq 0.23$	0	1.64	No	No	No	3.0	−0.63	0.62	0.004	61
ERN (118, 0.004)	0	1.35				5.0	0.2	0.9	0.004	54.5
Similarity network										
DOS118_$t_c \geq 0.95$	0.75	1.03	Yes	Yes	Yes	2.0	0.70	0.95	0.004	35.5
ERN (118, 0.004)	0	1.35				5.0	0.2	0.9	0.004	54.5

The network properties and quantified diversity indices of dissimilarity network DOS118_F_$t_c \leq 0.22$, similarity network DOS118_F_$t_c \geq 0$. The labelling pattern for the threshold networks follows the scheme described in Table 1. The dissimilarity network, DOS118_F_$t_c \leq 0.22$ and similarity network, DOS118_F_$t_c \geq 95$ are studied at equivalent network density and compared with the corresponding Erdös–Renyi network. For the dissimilarity network $C(G) < C_E(G)$, $L(G) > L_E(G)$ and QuaLDI 61 %, in contrast to the small-world characteristics of the similarity network, where $C(G) \gg C_E(G)$, $L(G) < L_E(G)$ and QuaLDI 35.5 %. $C(G)$ *and* $C_E(G)$ are the average clustering coefficient of threshold and Erdös–Renyi network. $L(G)$ *and* $L_E(G)$ are the average path length of threshold and Erdös–Renyi network. r = degree assortativity; Q = modularity, d_G = diameter of network, D(G) = Network density; QuaLDI = Quantitative Library Diversity Index measured in percentage (%)

Table 8 Network properties of DOS library (41 molecules) at various similarity and dissimilarity thresholds using FP2 fingerprints

Dissimilarity network	$C(G)$	$L(G)$	$C(G) > C_E(G)$	$C(G) \gg C_E(G)$	$L(G) < L_E(G)$	d_G	r	Q	$D(G)$	QuaLDI %
DOS41_F_$t_c \le 0.36$	0	1.95	No	No	Yes	3.0	−0.63	0.19	0.023	64
ERN (41, 0.023)	0	3.31				9.0	−0.2	0.73	0.022	60
Similarity network										
DOS41_F_$t_c \ge 0.9$	0.75	1.31	Yes	Yes	Yes	3.0	1.0	0.73	0.024	24
ERN (41, 0.024)	0.0	3.41				9.0	−0.3	0.76	0.026	60

The network properties and quantified diversity indices of dissimilarity network, DOS41_F_$t_c \le 0.36$, similarity network, DOS41_F_$t_c \ge 0.9$. The labelling pattern for the threshold networks follows the scheme described in Table 1. The dissimilarity network, DOS41_F_$t_c \le 0.36$ and similarity network, DOS41_F_$t_c \ge 0.9$ are studied at equivalent network density and compared with the corresponding Erdös–Renyi network. The dissimilarity network shows $C(G) = C_E(G) = 0$, $L(G) < L_E(G)$ and QuaLDI 64 %, in contrast to the small-world characteristics of the similarity network, where $C(G) \gg C_E(G)$, $L(G) < L_E(G)$ and QuaLDI 24 %. $C(G)$ *and* $C_E(G)$ are the average clustering coefficient of threshold and Erdös–Renyi network. $L(G)$ *and* $L_E(G)$ are the average path length of threshold and Erdös–Renyi network. r = degree assortativity; Q = modularity, d_G = diameter of network, D(G) = Network density; QuaLDI = Quantitative Library Diversity Index measured in percentage (%)

Journal: **10910-JOMC** Article No.: **0657** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2016/6/24** Pages: **26** Layout: **Small-X**

Table 9 Network properties of DOS library (32 molecules) at similarity and dissimilarity thresholds using FP2 fingerprints

Dissimilarity network	$C(G)$	$L(G)$	$C(G) > C_E(G)$	$C(G) \gg C_E(G)$	$L(G) < L_E(G)$	d_G	r	Q	$D(G)$	QuaLDI %
DOS32_F_$t_c \leq 0.39$	0.015	2.25	No	No	Yes	4.0	−.53	0.26	0.08	64
ERN (32, 0.08)	0.1	3.4				7.0	−0.1	0.53	0.08	55
Similarity network										
DOS32_F_$t_c \geq 0.72$	0.73	1.82	Yes	Yes	Yes	5.0	0.43	0.77	0.08	33
ERN (32, 0.08)	0.1	3.4				7.0	−0.1	0.53	0.08	55

The network properties and quantified diversity indices of dissimilarity network, DOS32_F_$t_c \leq 0.39$ and similarity network, DOS32_F_$t_c \geq 0.72$. The labelling pattern for the threshold networks follows the scheme described in Table 1. The dissimilarity network, DOS32_F_$t_c \leq 0.39$ and similarity network, DOS32_F_$t_c \geq 0.72$ are studied at equivalent network density and compared with the corresponding Erdös–Renyi network. For the dissimilarity network $C(G) < C_E(G)$, $L(G) > L_E(G)$ and QuaLDI 64 %, in contrast to the small-world characteristics of the similarity network, where $C(G) \gg C_E(G)$, $L(G) < L_E(G)$ and QuaLDI 33 %. $C(G)$ *and* $C_E(G)$ are the average clustering coefficient of threshold and Erdös–Renyi network. $L(G)$ *and* $L_E(G)$ are the average path length of threshold and Erdös–Renyi network. r = degree assortativity; Q = modularity, d_G = diameter of network, D(G) = Network density; QuaLDI = Quantitative Library Diversity Index measured in percentage (%)

Table 10 Network properties of Focussed library (41 molecules) at similarity and dissimilarity thresholds using FP2 fingerprints

Dissimilarity network	$C(G)$	$L(G)$	$C(G) > C_E(G)$	$C(G) \gg C_E(G)$	$L(G) < L_E(G)$	d_G	r	Q	$D(G)$	QuaLDI %
FL41_F_$t_c \leq 0.51$	0	1.75	No	No	Yes	3.0	−0.54	0.012	0.026	67
ERN (41, 0.026)	0	2.4				4.0	−0.2	0.73	0.02	53
Similarity network										
FL41_F_$t_c \geq 0.995$	1	1	Yes	Yes	Yes	1.0	1.0	0.28	0.022	9
ERN (41, 0.022)	0	1.4				5.0	−0.007	0.8	0.028	60

The network properties and quantified diversity indices of dissimilarity network, FL41_F_$t_c \leq 0.5$ and similarity network, FL41_F_$t_c \geq 0.995$. The labelling pattern for the threshold networks follows the scheme described in Table 1. For the dissimilarity network $(G) = C_E(G) = 0, L(G) > L_E(G)$ and QuaLDI = 67 %, in contrast to the small-world characteristics of the similarity network, where $C(G) \gg C_E(G), L(G) < L_E(G)$ and QuaLDI = 9 %. $C(G)$ *and* $C_E(G)$ are the average clustering coefficient of threshold and Erdös–Renyi network. $L(G)$ *and* $L_E(G)$ are the average path length of threshold and Erdös–Renyi network. r = degree assortativity; Q = modularity, d_G = diameter of network, D(G) = Network density; QuaLDI = Quantitative Library Diversity Index measured in percentage (%)

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

Over the past two decades, there have been many studies exploring various diversity measures for a chemical library, but none using network topology to quantify diversity [20–22].

In the present study, we propose a simple, convenient and novel Quantitative Library Diversity Index to quantify the diversity of a chemical library based on network topology:

$$QuaLDI\% = \left(1 - \frac{\sum_{\omega=1}^{n} \omega}{\lambda}\right) * 100 \tag{4.1}$$

where ω is a scaled network topological property; λ is the total number of properties used for quantification. For the present study, have used the four properties: average clustering Coefficient $C(G)$, average path length $L(G)$, degree assortativity r and modularity Q. The values of these properties lie in the range: $C(G)$ between (0,1); $L(G)$ between (0, d_G), where d_G is diameter of the network G; r between $(-1,1)$ and Q between $(-1,1)$. To equalize the contributions of the different network properties to the index, each property ω is scaled between 0 and 1 using Eq. 4.2:

$$Scale_0^1\omega = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \tag{4.2}$$

where x_i is the value of i^{th} property: $i = \{1, 2, 3, \ldots, n\}$; x_{min} is the minimum of all values of x; and x_{max} is the maximum of all values of x.

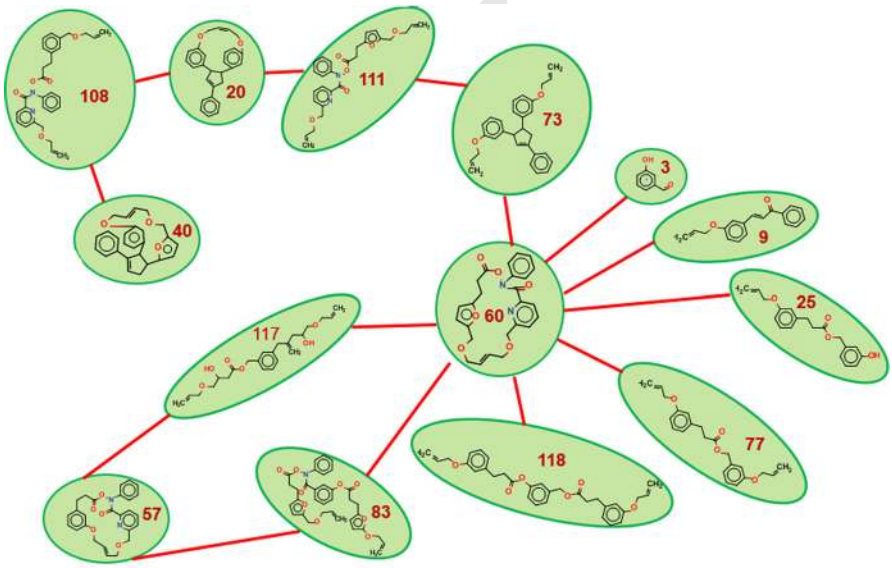

Fig. 4 Structures of the molecules belonging to the dissimilarity network **DOS118_M_t$_c$ $\leq$ 0.2** constructed from a DOS library. The dissortative nature of dissimilarity network DOS118_M_t$_c$ $\leq$ 0.2 with structurally diverse compounds. Molecule 60 (hub) is structurally very dissimilar to the rest of the library, being connected by dissimilarity edges to most of the other compounds

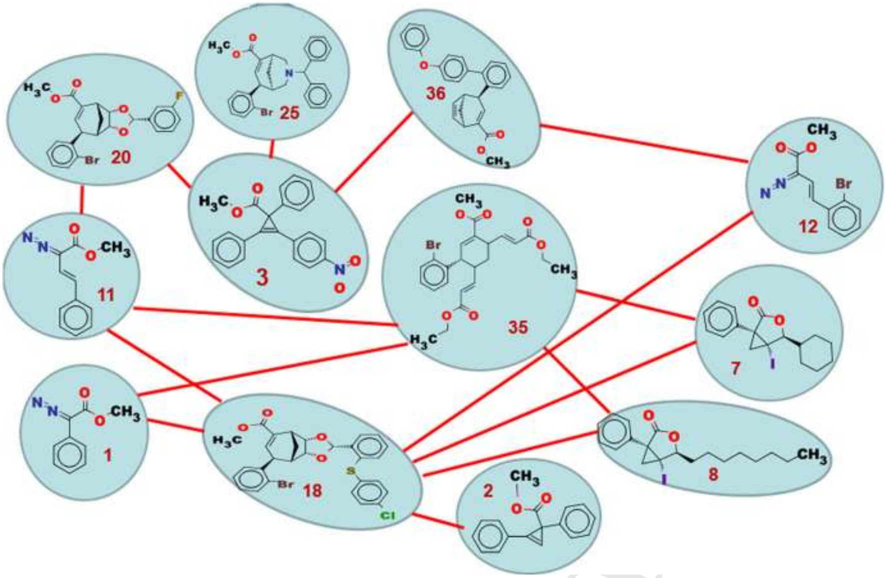

Fig. 5 Structures of the molecules belonging to the dissimilarity network $\mathbf{DOS41\_F\_t_c \leq 0.36}$ constructed from a DOS library. The dissortative nature of dissimilarity network DOS41_F_$t_c \leq 0.36$ with structurally diverse compounds. Molecules 3, 18 and 36 (hubs) are structurally very dissimilar to the rest of the library, being connected through dissimilarity edges to many other molecules, but they do not share any edges between them, as their similarity coefficients tc are just above 0.36

Equation 4.1 was employed for the quantification of library diversity, and the results are reported in the Tables 7, 8, 9 and 10. The dissimilarity sets of compounds are shown in Figs. 4, 5 and 6.

5 Network visualisation

The undirected threshold networks are visualized in the 'Fruchterman Reingold' force field directed layout. The dissimilarity networks show the absence of cliques in the community, which lead to the absence of the small-world property, as illustrated in 2
Figs. 3a–c and 4, 5, 6. The molecule 60 (hub) in Fig. 4, and molecules 3, 18 and 36 (hubs) in Fig. 5, are structurally dissimilar to the rest of their respective libraries. The maximally diverse set of compounds in the focussed library (Fig. 6) seems to be less diverse in comparison to the dissimilarity subsets of the DOS libraries.

The vertices in the dissimilarity network display dissortative hubs characterised by star graphs. The dissimilarity network also show second order clustering characterised by minimal ring size of four. However, the high similarity networks show high clustering coefficient and low average path length within the islands/small-world communities in comparison to the corresponding Erdös–Renyi random networks at equivalent edge density, as in Fig. 3a–b. As previously discussed, a chemical library is a combination of threshold networks exhibiting properties that reflect homophily or diversity in the subnetworks of the chemical library. Fig. 3a–b illustrates the common

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

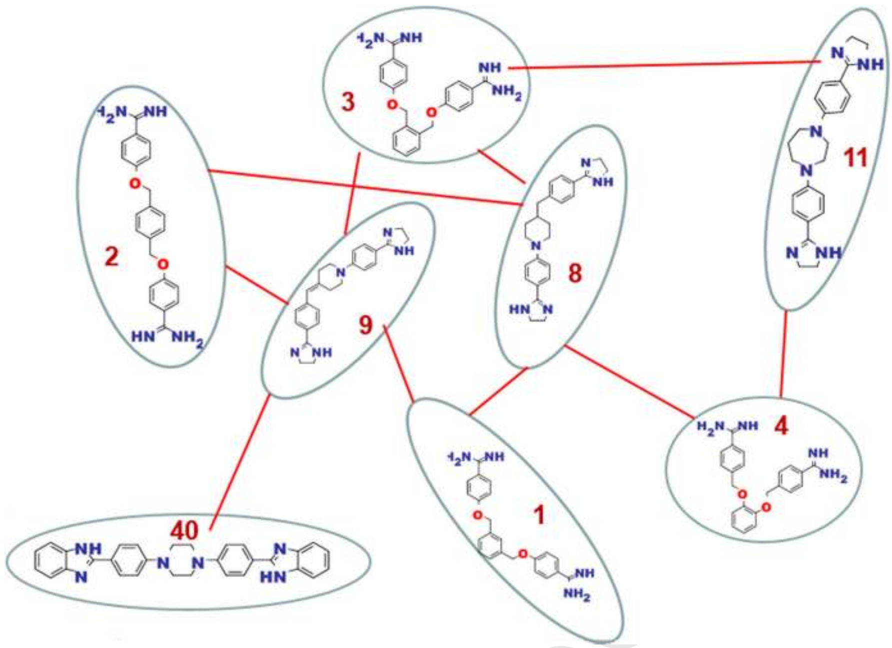

Fig. 6 Structures of the molecules belonging to the dissimilarity network **FL41_F_t$_\mathbf{c}$ $\leq$ 0.5** constructed from a Focused library. The dissortative nature of dissimilarity network FL41_F_t$_c$ $\leq$ 0.5 identifying the maximally diverse set of compounds within the library, but these are less diverse in comparison to the dissimilarity subsets of the DOS libraries

motif of sharing of similarity edges between the vertices in dissimilarity networks and the rarer sharing of dissimilarity edges between the vertices in the similarity network communities.

6 Conclusion

In the present research, we studied the design and properties of various dissimilarity and similarity threshold networks generated from DOS and focussed libraries using FP2 and MACCS fingerprints. The dissimilarity networks show the absence of small-world behavior, as evidenced by very low average clustering coefficients and high average path lengths in comparison to the Erdös–Renyi networks. The dissimilarity networks exhibit scale-free topology compatible with power-law, exponential and log-normal distributions. Both similarity and dissimilarity networks show the presence of hubs. The hubs in dissimilarity networks reveal dissortative behavior, whereas the hubs in similarity networks show assortative behavior. The dissimilarity networks display pseudo random network behavior, while the similarity networks demonstrate small-world behavior. High average clustering coefficient, assortativity and high modularity of the network are hallmarks of a high similarity threshold network of a chemical library. Low average clustering coefficient, dissortativity and low modularity ($Q < Q_E$) of the network are the signatures of a high dissimilarity threshold network

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

of a chemical library. In dissimilarity networks, the mixing of degrees of vertices is more dissortative, in contrast to the assortative behavior of similarity networks.

Quantifying the diversity in a virtual chemical library prior to synthesis is highly desirable when building a diverse library of molecules for screening, for which we propose a diversity measure QuaLDI based on network properties.The diversity of small DOS and Focussed libraries were assessed and quantified, based on the properties of similarity and dissimilarity threshold graphs of the chemical libraries. As illustrated in the present study, QuaLDI may be used to quantify the diversity in small chemical libraries (∼30 to 120 compounds). Employing network measures to quantify diversity in a chemical library provides a systematic and unbiased way to prune or grow a library such that each molecule adds maximum information content to a structure-activity relationship. Similar network measures have been used in feature selection [44] to systematically drop inter-correlated descriptors in an unbiased manner while retaining maximum information content in the remaining ones. Our future goal is to construct predictive models for diversity in chemical libraries using network topological properties as descriptors along with other molecular descriptors.

References

1. O. Raevsky, Mini-Rev. Med. Chem. **4**, 1041 (2004)
2. M.E.J. Newman, Phys. Rev. E **69**, 066133 (2004)
3. J. Hert, P. Willett, D.J. Wilton, J. Chem. Inf. Comput. Sci. **44**, 1177 (2004)
4. BioSolveIT, SciTegic (2007) Pipeline Pilot. (Accelrys Software, San Diego, CA) . Version 6.1.5
5. N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, J. Cheminf. **3**, 33 (2011)
6. RStudio Team, RStudio: Integrated Development for R (RStudio, Inc., Boston, MA, 2015), http://www.rstudion.com/
7. D. Rogers, M. Hahn, J. Chem. Inf. Model. **50**, 742 (2010)
8. M.P. Krein, N. Sukumar, J. Phys. Chem. A **115**, 12905 (2011)
9. N. Sukumar, S. Das, M. Krein, R. Godawat, I. Vitol, S. Garde, K.P. Bennett, C.M. Breneman, Computational approaches, in *Cheminformatics and Bioinformatics*, ed. by R. Guha, A. Bender (Wiley, Hoboken, 2011), pp. 107–143
10. R.W. Benz, S.J. Swamidass, P. Baldi, J. Chem. Inf. Model. **48**, 1138 (2008)
11. T.G. Lewis, *Network Science: Theory and Applications* (Wiley, Hoboken, 2009)
12. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
13. A.L. Barabasi, R. Albert, Science **286**, 509 (1999)
14. C. Qian, C. Hyunseok, R. Govindan, S. Jamin, IEEE Infocom. **2**, 608 (2002)
15. H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Nature **411**, 41 (2001)
16. G.F. Davis, M. Yoo, W.E. Baker, Strateg. Organ. **1**, 301 (2003)
17. J.I. Perotti, F.A. Tamarit, S.A. Cannas, Phys. A **371**, 71 (2006)
18. T. Kodadek, Curr. Opin. Chem. Biol. **14**, 713 (2010)
19. J.Y. Ortholand, A. Ganesan, Curr. Opin. Chem. Biol. **8**, 271 (2004)
20. P. Willett, Inform. Res. **2**, 3 (1996)
21. D.K. Agrafiotis, V.S. Lobanov, J. Chem. Inf. Model. **39**, 51 (1999)
22. E.A. Wintner, C.C. Moallemi, J. Med. Chem. **43**, 1993 (2000)
23. H.M. Patel, M.N. Noolvi, P. Sharma, V. Jaiswal, S. Bansal, S. Lohan, S.S. Kumar, V. Abbot, S. Dhiman, V. Bhardwaj, Med. Chem. Res. **23**, 4991–5007 (2014)
24. S.L. Schreiber, Science **287**, 1964 (2000)
25. A. Grossmann, S. Bartlett, M. Janecek, J.T. Hodgkinson, D.R. Spring, Angew. Chem. **53**, 13093 (2014)
26. B. M. Ibbeson, L. Laraia, E. Alza, O. C. CJ, Y. S. Tan, H. M. Davies, G. McKenzie, A. R. Venkitaraman, D. R. Spring, Nature Comm. **5**, 3155 (2014) 3

Journal: 10910-JOMC Article No.: 0657 TYPESET DISK LE CP Disp.:2016/6/24 Pages: 26 Layout: Small-X

27. V.S. Damerla, C. Tulluri, R. Gundla, L. Naviri, U. Adepally, P.S. Iyer, Y.L. Murthy, N. Prabhakar, S. Sen, Chem. Asian J. **7**, 2351 (2012)
28. R. Mamidala, V.S. Babu Damerla, R. Gundla, M.T. Chary, Y.L.N. Murthy, S. Sen, RSC Adv. **4**, 10619 (2014)
29. M. Cruz-Monteagudo, F. Borges, M. Perez Gonzalez, M.N. Cordeiro, Bioorg. Med. Chem. **15**, 5322 (2007)
30. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, J. Chem. Inf. Comput. Sci. **42**, 1273 (2002)
31. G. Csardi, T. Nepusz, InterJournal, Complex Systems, 1695 (2006) 4
32. M. Zwierzyna, M. Vogt, G.M. Maggiora, J. Bajorath, J. Comput.-Aided Molec. Des. **29**, 113 (2015)
33. R.D. Luce, A.D. Perry, Psychometrika **14**, 95 (1949)
34. P. Erdös, A. Renyi, Pub. Math. **6**, 290 (1959)
35. M.E.J. Newman, Phys. Rev. E **67**, 02616 (2003)
36. M.E.J. Newman, Phys. Rev. Lett. **89**, 208701 (2002)
37. A. Clauset, M.E.J. Newman, C. Moore, Phys. Rev. E **70**, 066111 (2004)
38. M. Girvan, M.E. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002)
39. M.E.J. Newman, Phys. Rev. E **74**, 036104 (2006)
40. P. Pons, M. Latapy, J. Graph Algorithms Applic. **10**, 191 (2006)
41. M. E. J. Newman, Contemp. Phys. 46 (2005)
42. A. Clauset, C.R. Shalizi, M.E.J. Newman, SIAM Rev. **51**(4), 661 (2009)
43. D.L. Evans, J.H. Drew, L.M. Leemis, Comm. Stat. - Simul. Comput. **37**, 1396 (2008)
44. K. Wu, N. Sukumar, N. Lanzillo, C. Wang, R. Ramprasad, R. Ma, A. F. Baldwin, G. Sotzing, C. M. Breneman, J. Polymer Sci. (communicated)

Journal: 10910
Article: 657

Springer
the language of science

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	Please confirm if the corresponding author is correctly identified. Amend if necessary.	
2.	As per the information provided by the publisher, Fig. 3 will be black and white in print; hence, please confirm whether we can add "colour figure online" to the caption.	
3.	Please check and confirm the author names for the reference [26].	
4.	Please update the references [31], [41] and [44].	