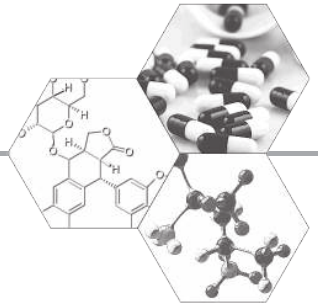
REVIEW



Graphs and networks in chemical and biological informatics: past, present and future

Chemical and biological network analysis has recently garnered intense interest from the perspective of drug design and discovery. While graph theoretic concepts have a long history in chemistry – predating quantum mechanics – and graphical measures of chemical structures date back to the 1970s, it is only recently with the advent of public repositories of information and availability of high-throughput assays and computational resources that network analysis of large-scale chemical networks, such as protein–protein interaction networks, has become possible. Drug design and discovery are undergoing a paradigm shift, from the notion of 'one target, one drug' to a much more nuanced view that relies on multiple sources of information: genomic, proteomic, metabolomic and so on. This holistic view of drug design is an incredibly daunting undertaking still very much in its infancy. Here, we focus on current developments in graph- and network-centric approaches in chemical and biological informatics, with particular reference to applications in the fields of SAR modeling and drug design. Key insights from the past suggest a path forward via visualization and fusion of multiple sources of chemical network data.

N Sukumar*[1,2]†
& Michael P Krein[3]†

[1]Department of Chemistry, Shiv Nadar University, Chithera, Dadri, 203207, India
[2]Department of Chemistry & Chemical Biology, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[3]Lockheed Martin Advanced Technology Laboratories, 3 Executive Campus, Suite 600, Cherry Hill, NJ 08002 USA
*Author for correspondence:
E-mail: n.sukumar@snu.edu.in

†Authors contributed equally

The arc of scientific thought and practice for the past several centuries has followed a definite reductionist paradigm. With the breakthroughs in molecular biology during the latter half of the 20th century, biological sciences have joined this trend, with functions attributed to proteins, amino acid sequences and motifs. However, each gene typically encodes several distinct proteins, whose functions can vary depending upon the stage of development of the organism, the location within the body and other environmental factors. The effect of a drug on a living organism involves complex interactions at vastly different spatial and temporal scales, from the molecular and cellular levels to those at the level of the tissue, organ and the organism itself. A fundamental understanding of complex biological systems is, thus, one of the grand challenges in science of the 21st century. Recent decades have seen the advent of robotic high-throughput bioassays and an increasing availability of whole genome sequences. This capability and the realization that there is no guarantee of emergence of higher level functions from a study of individual molecular-level components have led to a transition from the study of functions of individual genes, proteins and molecular SARs, to a systems-level view of biology and biochemistry. **Systems biology** is the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions. In this view, a structure–function relationship is no longer associated with a specific ligand or biological target, but with the flow of information through this integrated cellular network. Protein–protein interaction networks may be associated with specific disease effects and drug responses, since many diseases are caused by disruption of normal protein interactions, disruption of protein–DNA interactions, or formation of new undesirable protein interactions.

Analogous to biological structure–function relationships, SAR in cheminformatics are commonly envisioned in a high-dimensional space of numerical descriptors of molecular structure (commonly referred to as chemistry space). A similarity network in chemistry space consists of a pair-wise similarity relationship (an edge or connection) between individual molecules (forming the nodes of the network). An example is the network-like similarity graph **(Figure 1)** introduced by Bajorath and co-workers [1]. Local neighborhoods in this space correspond to regions of structural similarity. In reaction networks, the connections between molecules represent some measure of the reaction connecting them. For instance, a metabolic network can be constructed for bacteria, archaea and eukaryotes, with the nodes representing metabolites and the edges representing biochemical reactions for which one metabolite is a substrate and the other a product [2]. Such networks have a modular structure with most of the nodes connected only to other nodes

10.4155/FMC.12.128 © 2012 Future Science Ltd

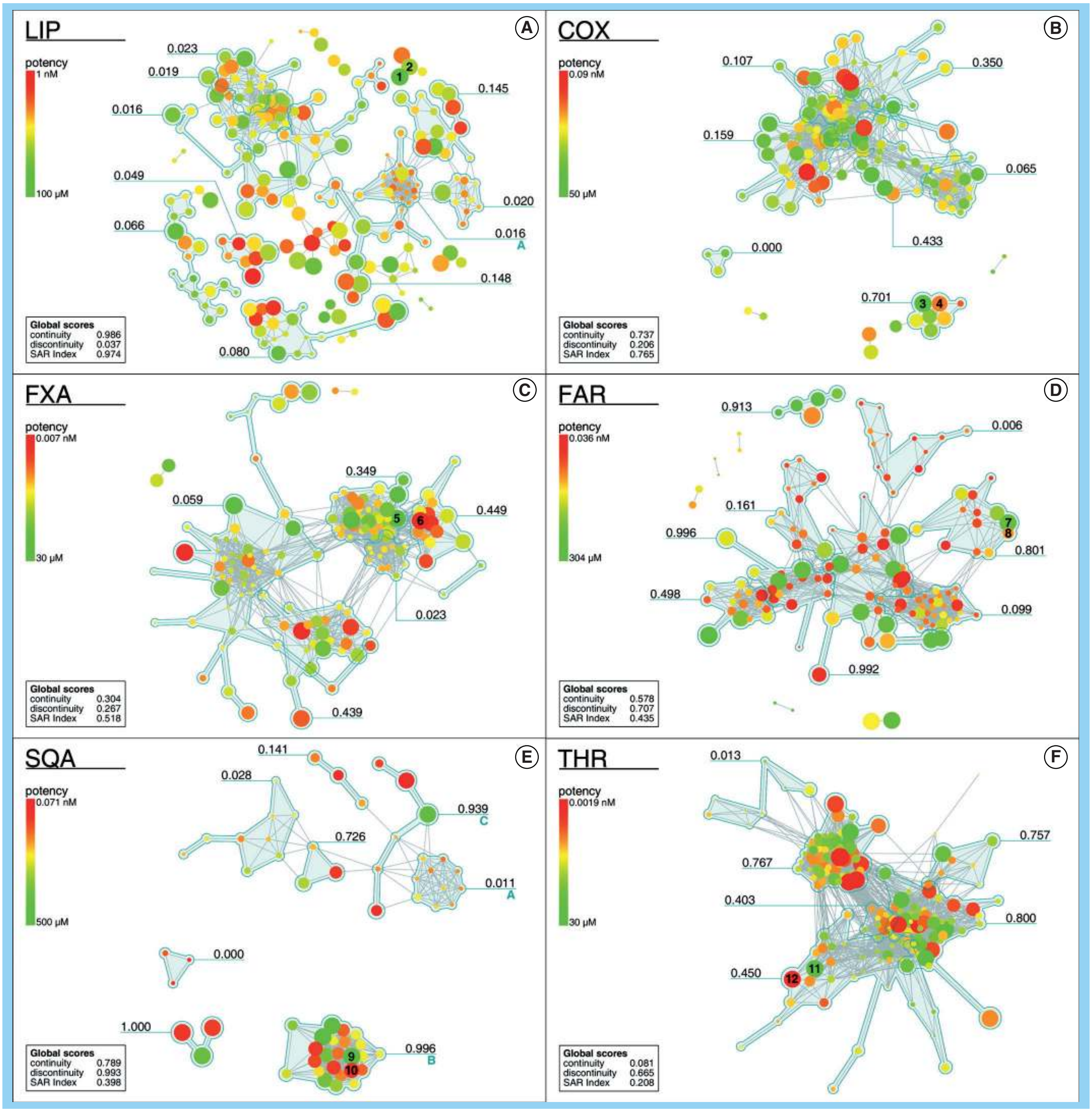


Figure 1. Similarity graphs for six classes of enzyme inhibitors. (A) lipoxygenase, **(B)** cyclooxygenase-2, **(C)** coagulation factor Xa, **(D)** protein farnesyltransferase, **(E)** squalene synthase and **(F)** thrombin. Nodes represent molecules, with edges between them if the pairwise MACCS Tanimoto similarity is greater than 0.65. Nodes are color-coded according to potency using a continuous spectrum from green (lowest potency) to red (highest potency) and scaled according to their local compound discontinuity scores. COX: Cyclooxygenase-2; FAR: Protein farnesyltransferase; FXA: Coagulation factor Xa; LIP: Lipoxygenase; SQA: Squalene synthase; THR: Thrombin.
Reproduced with permission from [1].

within their respective modules. Metabolites participating in only a few reactions, but connecting different modules, are found to be more strongly conserved than hubs whose links are mostly within a single module. Similarly, network topological analyses strive to assign meaning to global and local descriptions of network structure.

Global network topology & its visual representation

Network representations of biology and chemistry allow for succinct quantitation and visualization of relationships between data, and are, thus, well-rooted in scientific history. In biology, the branching Tree of Life can be traced back to Jean-Baptiste Lamarck's 1809 publication *Philosophiezoologiqueou Exposition des Considerations Relatives l'histoire Naturelle Desanimaux* and internationally popularized in Charles Darwin's 1859 work *On the Origin of Species*. Similarly, in chemistry, graph representations of molecular structure date back to the connection diagrams in Kekule's 1857 descriptions of the tetravalence of carbon (*Über die s. g. Gepaarten Verbindungen und die Theorie der Mehratomigen Radicale*) and his now famous 1865 work that described the structure of benzene, *Sur Laconstitution des Substances Aromatiques*. Collapsing connected atoms into single nodes generates what are known as **reduced graphs (Figure 2)** [3]. Many different graph reduction schemes have been developed for similarity searching [4], the objective being to enable compounds sharing the same activity but belonging to different 2D scaffolds to be perceived as similar. Such reduced graphs function as topological pharmacophores, complementing traditional molecular descriptors and offering the potential for scaffold hopping. Reduced graphs have been used in applications ranging from chemical patent searches to identification of SARs and clustering of high-throughput screening data [5]. The reduced graph representation enables heterogeneous compounds, such as those found in high-throughput screening data, to be captured in a single representation with the resulting query encoding SARs in a readily interpretable form [6].

Graph theory permeates modern computational chemistry and biology, where connection tables – lists of atoms and lists of the bonds that connect them – commonly represent structures. Adjacency or distance matrices – matrix representations of graph structure – are produced by systematic comparison of biological or chemical entities, where every element in a data set is compared with every other element via a similarity assessment metric. A common global measure of a network is the **degree distribution**. The **degree of a node** is the number of links between it and other nodes; the degree distribution P(k) is the probability that a specified node has exactly k links. **Scale-free networks** are characterized by a power-law tail in the degree distribution: the probability that a node has k links follows $P(k) \sim k^{-\gamma}$ (seen as a straight line on a log–log plot). Krein and Sukumar investigated the network topology and scaling relationships of several chemistry spaces, which showed qualitatively similar behavior, following power law degree distributions, indicating the small-world nature of the corresponding networks [7]. The small-world behavior of chemistry spaces has also been noted by other authors [8,9].

The properties of a scale-free network are often determined by a relatively small number of highly connected nodes (hubs). Disabling even a substantial number of nodes in a scale-free network does not lead to fragmentation of the network. Thus, such networks are characterized by topological robustness; they are robust against accidental failures because random failures affect mostly the many nodes of low degree and do not disrupt the network's overall integrity. However, this reliance on hubs in a scale-free network carries a cost, in that it implies vulnerability to targeted attack against a few key hubs. In chemistry space, hubs are represented by molecules with high leverage in a SAR. Such molecules are important for maintaining the diversity of a chemical library and for ensuring good predictive performance of QSAR models across a wide **domain of applicability**, that is, across different molecular scaffolds.

Genes and gene products constitute a complex network of interactions. In such biological networks, nodes may represent genes, gene products, drugs, proteins, phenotypes, or metabolites, and the edges may represent interactions or co-occurrence of phenotypes. The small-world topology of protein residue networks has been convincingly demonstrated [6–8] – here, the amino acid residues are the nodes of the network and two residues are connected if they are closer than a certain distance cutoff. Analysis of the networks of DNA-binding proteins [10] – where the edges between amino acid residues are determined by the strengths of the non-covalent interactions between them revealed a strong correlation between the positions of the residues interacting with DNA and highly connected hubs(these are amino acid residues making connections with a large number of other residues). Protein residue networks have been found to exhibit the universal

Key Terms

Systems biology: Study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions.

Reduced graphs: Summary representations of chemical structures generated by collapsing groups of connected atoms into single nodes while preserving the topology of the original structures.

Degree distribution: Probability that a specified node has exactly a given number of links.

Degree of a node: Number of links between it and other nodes.

Scale-free networks: Characterized by a power-law degree distribution: the probability that a node has k links follows $P(k) \sim k^{-\gamma}$ (seen as a straight line on a log–log plot). The properties of a scale-free network are often determined by a relatively small number of highly connected nodes (hubs); the average path length of such a network follows a log log N distribution (where N is the number of nodes).

Domain of applicability: Physicochemical, structural or biological space that has been used to train the model and within which the model may be reliably applied to make predictions for new molecules.

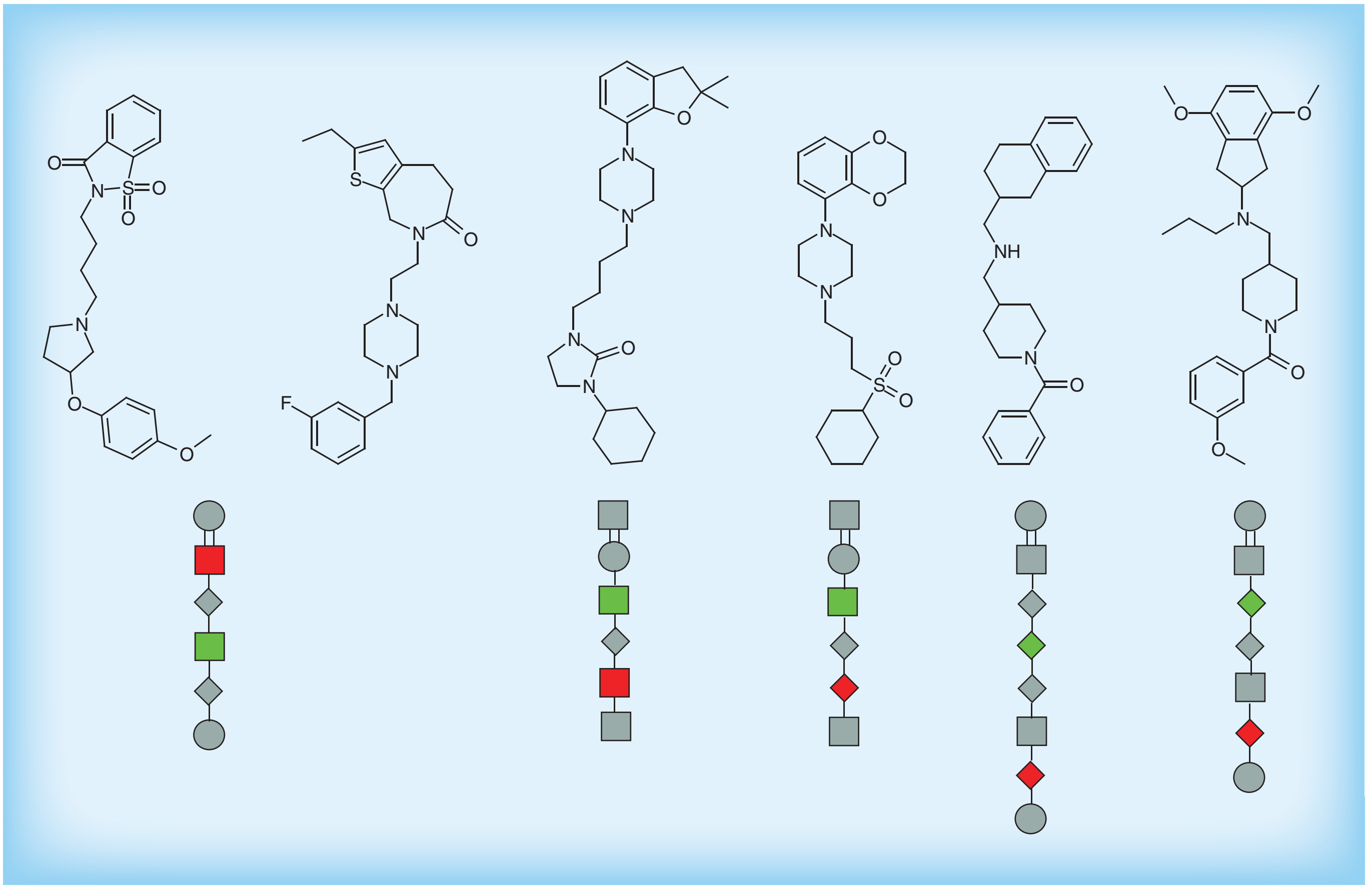


Figure 2. Series of $5HT_{1A}$ agonists and their reduced graph representations. The first two molecules are both represented by the same reduced graph. The reduced graphs of the second pair of molecules differ in the substitution of an aliphatic acceptor node for an acyclic acceptor node. The reduced graphs of the last pair differ by the insertion/deletion of a linker node.
See Table 1 in [6] for the key to graphical representations.
Reproduced with permission from [6].

topological characteristics [11] of modular networks consisting of several interconnected clusters. These clusters are separated by topological voids that represent protein binding sites.

In an excellent review, Fliri *et al.* have shown the limitations inherent in current drug design and discovery: the one-target, one-drug perspective has led to a lengthy and costly discovery route plagued with late-stage failures [12]. These limitations may be mitigated by a better understanding of the cause–effect relationships, the interplay of biology at different length and time scales. Protein–protein interaction network models have been widely explored in their capacity for cause–effect analyses in medicine. In protein-interaction networks [13,14], hubs represent highly connected proteins. Recent years have seen a shift from traditional receptor-specific studies to a cross-receptor view [15,16] of protein–drug interactions. While similar ligands may bind to similar targets, ligands quite frequently have affinity for more than one target [17,18]. Different proteins may have different sequences or folds and yet have similar binding partners. Thus, ligands that would otherwise be considered dissimilar by commonly used ligand-similarity detection algorithms can bind on to the same target [19,20]. This has led to the study of protein–target-based networks that identify related targets by comparing their binding sites [7,20,21], hence estimating the potential for cross-reactivity between the corresponding ligands. This multidimensional view of related targets is further complicated by the notion that our knowledge and representation of individual targets is incomplete, leading to gaps in the structure–activity landscape.

Representation & exploration of activity cliffs

An activity landscape is a graphical representation that integrates similarity and potency

Key Terms

Activity cliffs: Rugged region of chemistry space where pairs of structurally similar molecules have large differences in potency.

Similarity principle: Holds that similar molecules should exhibit similar activities in biological assays.

Structure–activity landscape index: Quantitative measure of the ruggedness of an activity landscape, given by the absolute difference in activities between a pair of molecules divided by their dissimilarity coefficient. Two compounds are connected by a structure–activity landscape index edge if their structure–activity landscape index score exceeds a given threshold value.

relationships between molecules for a specific biological target. **Activity cliffs** are formed by pairs of structurally similar molecules with large differences in potency. One can also define selectivity cliffs between a pair of molecules that have significantly different potencies against one or both targets of a pair. Mechanism cliffs are formed when small structural modifications induce a transition from one molecular mechanism to another in an analogous series.

Conventional QSAR relies on the **similarity principle**, which holds that similar molecules should exhibit similar activities in biological assays. It is now well recognized that very similar molecules may exhibit very different activities in some assays, leading to 'activity cliffs' and deviations from the similarity principle. Such discontinuities or cliffs in the structure-activity landscape are mapped as **structure–activity landscape index** (SALI) networks that highlight abrupt changes in biological activity associated with the steepest cliffs [22]. In SALI graphs, nodes are molecules and edges represent activity cliffs of varying magnitude. SALI is defined by the expression:

$$\mathrm{SALI}_{i,j} = \frac{|A_i - A_j|}{(1 - \mathrm{sim}[i,j])}$$

EQUATION 1

whereA$_i$ and A$_j$ are the activities of the ith and the jth molecules, and sim(i,j) is the similarity coefficient between the two molecules. Two compounds are connected by a SALI edge if their SALI score exceeds a given threshold value. Edges are depicted as arrows directed toward the more potent compound **(FIGURE 3)**. Identifying the locations of activity cliffs in a structure activity landscape through SALI mapping can improve our understanding of where a QSAR model is more or less accurate, especially the ability of the model to correctly predict the relative ordering of activities. A plot of the SALI value versus the normalized similarity threshold is known as the **SALI curve**. While a SALI network graph orders pairs of molecules by activity, the SALI curve tallies how many of these orderings a model is able to predict. The value of the SALI curve at zero similarity threshold measures the ability of the model to capture all of the edges, while the value at a normalized similarity threshold of unity measures its ability to correctly identify the most significant activity cliffs. The integral of the SALI curve is another useful measure of the performance of a model that rank orders molecules by activity.

The presence of activity cliffs in a structure–activity landscape necessitates hands-on data visualization and statistical analysis [23]. Other measures to characterize activity cliffs include **SAR indices** (SARI) [24] and SAR Maps [25]. The SARI is a composite index, a combination of a continuity score and a discontinuity score:

$$\mathrm{SARI} = (\mathrm{score}_{cont} + [1 - \mathrm{score}_{disc}])$$

EQUATION 2

These are, in turn, constructed from normalized sums of local continuity and discontinuity scores:

$$\mathrm{cont}(i) = \frac{\sum_{j<i}\left(\frac{W_{ij}}{1 + \mathrm{sim}[i,j]}\right)}{\Sigma_{(i,j|i>j)} W_{ij}}$$

EQUATION 3

$$W_{ij} = \frac{A_i A_j}{1 + |A_i - A_j|}$$

EQUATION 4

$$\mathrm{disc}(i) = \frac{\sum_{(j|\mathrm{sim}[i,j]>0.65, i \neq j)} (|A_i - A_j|\,\mathrm{sim}[i,j])}{(j \mid \mathrm{sim}[i,j] > 0.65, i \neq j}$$

EQUATION 5

with A$_i$ and A$_j$ being the potencies of i and j. A molecule has a high discontinuity score (near unity) if its potency differs significantly from those of its immediate structural neighbors. Such pairs of structurally similar compounds with significantly different potencies mark activity cliffs. In contrast, SARI locates structurally divergent compounds having similar activity.

Much coverage has been dedicated to the quantification and visualization of activity cliffs [26,27]. In practice, one can use this activity cliff information as a measure of model performance [28]. Several studies have noted the importance of training set selection in determining overall model quality and, thus, these activity cliff measures may help guide model building efforts [29]. Furthermore, this information may be mined to explore just how catastrophic the effect of an activity cliff would be on QSAR model performance [25], that is, the degree to which local predictivity is indicative of the global performance of the model.

Large activity cliffs are not uncommon and have been found to be comparably distributed over different target classes. Namasivayam *et al.* found that in a number of cases, model continuity was preserved near activity cliffs, in

Key Terms

Structure–activity landscape index curve: Plot of the structure–activity landscape index curve value versus the normalized similarity threshold.

SAR indices: Composite index, formed from a combination of a continuity score and a discontinuity score, is another quantitative measure of the topography of an activity landscape.

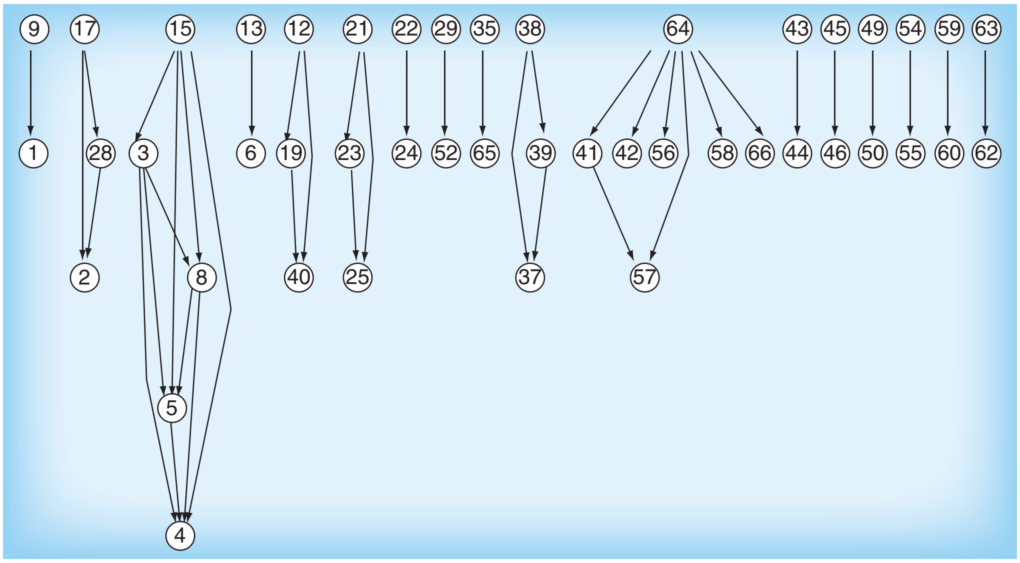


Figure 3. Graph representation of the structure–activity landscape index. An edge occurs between two nodes if the structure–activity landscape index value for that pair is greater than a user-specified cutoff. The arrows point from the molecule with lower activity to that with higher activity.
Reproduced with permission from ref [22].

the sense that local SAR continuity and discontinuity can occur together in a structure–activity neighborhood [30]. Thus, cliffs need not be black holes of model prediction quality. Dimova *et al.* examined activity cliff information over multiple targets [31]. These multitarget activity landscapes identify single-, dual- and triple-target activity cliffs and hierarchical cliff distributions.

This information is only as useful as it is interpretable to the end user; consensus views of activity cliff information [32] seek activity cliffs present over multiple structure representations. Overlaying this activity cliff information over data sets could be useful in comparing structure representations within a data set, leading to local decisions in modeling choices. Mapping the locations of activity cliffs and comparing the global characteristics of SALI sub-networks with those of the underlying chemistry space networks generated from different molecular fingerprint or descriptor representations [7] can guide the modeler in the choice of an optimal representation of the data. A higher local density of SALI edges with a particular representation in a specific region of chemistry space implies a more challenging SAR using that fingerprint. Conversely, Wawer and Bajorath have designed a methodology for large-scale SAR analysis that directly accounts for structural relationships between active compounds, without relying on numerical compound similarity assessment [33].

Bon and Waldmann recently examined hierarchical structural relationships between compound classes and structural similarities in ligand-sensing protein cores [34]. By looking at chemical and biological space in tandem, they demonstrated a methodology, biology-oriented synthesis, that led to the prospective identification of new targets of known biologically active compound classes and to the design of compound libraries.

In a similar manner, Klenner *et al.* have used stochastic proximity embedding to rapidly and automatically identify and visualize areas of interest, or 'activity islands' in chemical space [35]. They successfully applied stochastic proximity embedding and found inhibitors of *Helicobacter pylori* protease HtrA with new molecular scaffolds. This drug-discovery exercise was led by visualization, minimizing experimental effort and costs.

These improvements to data set visualization are critical in the workflow of drug design and discovery, representing a future view of the field.

Future perspective

The success and widespread acceptance of the systems view of chemical and biological spaces hinges on the ability to relate relevant information from multiple sources. Widespread usage of computer-readable structural formats allowed for the initial growth of publicly accessible data repositories, such as PubChem, BindingDB, ChEMBL and Chembank. Efficient mining of these databases' metadata is beginning to make an impact upon scientific practice. Semantic representations of biological data have proven their utility and are now fairly ubiquitous. For instance, Jenssen *et al.* created the PubGene database by automated knowledge extraction from the titles and abstracts of millions of publicly available MEDLINE records [36]. This enabled creation of a gene-to-gene co-citation network of the thousands of named human genes. The associations between genes were annotated by linking genes to terms from the Medical Subject Heading index and from the Gene Ontology database [37,101], the assumption being that if two genes are mentioned in the same MEDLINE record, then there must be an underlying biological relationship between them. The use of prior information extracted from biomedical literature and protein–protein interaction data sets has been found to improve the ability to learn biologically realistic networks from gene expression data [38].

The drug target network developed by Yildirim *et al.* [39] used all known US FDA-approved drugs and their targets from the DrugBank database [40] to construct a bipartite graph of protein–drug interactions, where a drug and a protein were connected to each other if the protein was a known target of the drug. Two network projections were generated from this graph, one whose nodes represented drugs and the connections a shared target protein; and a complementary network whose nodes were proteins, which were connected to each other if they were targeted by a common drug. Although constructed independently of any knowledge of drug classes, the drug subnetwork naturally clustered drugs by major therapeutic classes. Integrating DrugBank with protein–protein interaction data and the Online Mendelian Inheritance in Man™ database showed that drugs were more likely to act within co-expressed modules and were enriched in specific regions of the human disease network. Analysis of experimental drugs showed that the vast majority of new drugs targeted well-known proteins. The analysis also revealed a recent gradual trend of drugs targeting more diversified proteins, with a greater tendency for promiscuity.

In the future, the scope and accessibility of such observations will broaden with the maturation of semantic representations of chemical and biological information, driven by community acceptance of standardized data formats and open access to data [41–43]. The ubiquity of community-designed tools that build on such relationships will be rooted in the economics of cloud computing, which enable public access to large data-mining capabilities. Relatively inexpensive access to tools and data will drive efficiency and allow us to tackle the challenges of designing new drugs and repositioning existing drugs using a more holistic, systems-level approach to biochemistry.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Background

- The field of drug design and discovery is shifting from single-target studies to systems-wide approaches. Computational methods that deal with the resulting information explosion are being developed.

Global network topology & its visual representation

- Network types and network visualization approaches are covered; findings of protein–protein interaction network studies are discussed.

Representation & exploration of activity cliffs

- SARs, their utility and implications of their breakdowns, known as activity cliffs, are discussed. Quantification of a SAR's structural features to assess SAR performance is reviewed.

The future: linking of chemical & biological spaces

- Repositories that utilize experimental metadata to link multiple sources of information and tools that mine these repositories enables the leap from single-target to systems-wide approaches.

References

Papers of special note have been highlighted as:

- ▪ of interest
- ▪▪ of considerable interest

1 Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J. Structure–activity relationship anatomy by network-like similarity graphs and local structure–activity relationship indices. *J. Med. Chem.* 51(19), 6075–6084 (2008).

2 Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. *Nature* 433, 895–900 (2005).

3 Gillet VJ, Downs GM, Ling A *et al.* Computer storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs and their applications in generic chemical structure retrieval. *J. Chem. Inf. Comput. Sci.* 27(3), 126–137 (1987).

4 Birchall K, Gillet VJ. Reduced graphs and their applications in chemoinformatics. In: *Chemoinformatics and Computational Chemical Biology.* Bajorath J (Ed.). Humana Press, NY, USA, 197–212 (2011).

5 Harper G, Bravi GS, Pickett SD, Hussain J, Green DVS. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* 44(6), 2145–2156 (2004).

6 Birchall K, Gillet VJ, Harper G, Pickett SD. Evolving interpretable structure–activity relationships. 1. Reduced graph queries. *J. Chem. Inf. Model.* 48(8), 1543–1557 (2008).

7 Krein MP, Sukumar N. Exploration of the topology of chemical spaces with network measures. *J. Phys. Chem. A* 115(45), 12905–12918 (2011).

8 Benz RW, Swamidass SJ, Baldi P. Discovery of power-laws in chemical space. *J. Chem. Inf. Model.* 48(6), 1138–1151 (2008).

9 Tanaka N, Ohno K, Niimi T, Moritomo A, Mori K, Orita M. Small-world phenomena in chemical library networks: application to fragment-based drug discovery. *J. Chem. Inf. Model.* 49(12), 2677–2686 (2009).

▪ **The realization and implications of small-world network phenominology to drug design are discussed.**

10 Sathyapriya R, Brinda KV, Vesheshwara S. Correlation of side-chain hubs with the functional residues in DNA binding protein structures. *J. Chem. Inf. Model.* 46, 123–129 (2006).

11 Estrada E. Universality in protein residue networks. *Biophys. J.* 98(5), 890–900 (2010).

12 Fliri AF, Loging WT, Volkmann RA. Cause–effect relationships in medicine: a protein network perspective. *Trends Pharmacol. Sci.* 31(11), 547–555 (2010).

▪▪ **Excellent review of work demonstrating the utility of protein–protein interaction network models in cause–effect analyses in medicine.**

13 Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314, 1938–1941 (2006).

14 Sprinzak E, Altuvia Y, Margalit H. Characterization and prediction of protein–protein interactions within and between complexes. *Proc. Natl Acad. Sci. USA* 103, 14718–14723 (2006).

15 Klabunde T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* 152(1), 5–7 (2007).

16 Rognan D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* 152(1), 38–52 (2007).

17 Paolini GV, Shapland RHB, Hoorn WPv, Mason JS, Hopkins AL. Global mapping of pharmacological space. *Nat. Biotechnol.* 24(7), 805–815 (2006).

18 Keiser MJ, Setola V, Irwin JJ *et al.* Predicting new molecular targets for known drugs. *Nature*, 462(7270), 175–181 (2009).

▪ **Chemical similarity measures were employed to examine the efficacy of *in silico* drug repositioning and side effect analysis.**

19 Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE. Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* 5(7), e1000423 (2009).

▪▪ **A systems-wide methodology was employed to reposition drugs to treat multidrug-resitant tuberculosis.**

20 Das S, Kokardekar A, Breneman CM. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* 49(12), 2863–2872 (2009).

21 Das S, Krein MP, Breneman CM. Binding affinity prediction with property-encoded shape distribution signatures. *J. Chem. Inf. Model.* 50(2), 298–308 (2010).

22 Guha R, Van Drie JH. Structure–activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* 48, 646–658 (2008).

23 Cumming JG, Winter J, Poirrette A. Better compounds faster: the development and exploitation of a desktop predictive chemistry toolkit. *Drug Discov. Today* 17(17–18), 923–927 (2012).

24 Bajorath J, Peltason L, Wawer M, Guha R, Lajiness MS, Van Drie JH. Navigating structure–activity landscapes. *Drug Discov. Today* 14(13–14), 698–705 (2009).

25 Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS. SAR Maps: a new visualization technique for medicinal chemists. *J. Med. Chem.* 20, 5926–5937 (2007).

26 Stumpfe D, Bajorath J. Methods for SAR visualization. *RSC Advances* 2(2), 369–378 (2012).

▪ **Excellent review of visualization strategies in a diverse set of SARs.**

27 Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* 55(7), 2932–2942 (2012).

28 Guha R, Van Drie JH. Assessing how well a modeling protocol captures a structure–activity landscape. *J. Chem. Inf. Model.* 48(8), 1716–1728 (2008).

▪ **Local SAR information as determined by structure–activity landscape indicies helps guide modeling decisions.**

29 LeDonne N, Rissolo K, Bulgarelli J, Tini L. Use of structure–activity landscape index curves and curve integrals to evaluate the performance of multiple machine learning prediction models. *J. Cheminformatics* 3(1), 7 (2011).

30 Namasivayam V, Iyer P, Bajorath J. Exploring SAR continuity in the vicinity of activity cliffs. *Chem. Biol. Drug Des.* 79(1), 22–29 (2012).

31 Dimova D, Wawer M, Wassermann AM, Bajorath J. Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *J. Chem. Inf. Model.* 51(2), 258–266 (2011).

32 Medina-Franco JL, Martínez-Mayorga K, Bender A *et al.* Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* 49(2), 477–491 (2009).

33 Wawer M, Bajorath J. Local structural changes, global data views: graphical substructure–activity relationship trailing. *J. Med. Chem.* 54(8), 2944–2951 (2011).

▪▪ **Network views of approved drugs and their protein targets lead to natural partitioning within the space that described structural features such as promiscuity.**

34 Bon RS, Waldmann H. Bioactivity-guided navigation of chemical space. *Acc. Chem. Res.* 43(8), 1103–1114 (2010).

35 Klenner A, Hähnke V, Geppert T *et al.* From virtual screening to bioactive compounds by

visualizing and clustering of chemical space. *Mol. Inf.* 31(1), 21–26 (2012).

36 Jenssen T-K, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28 (2001).

37 Ashburner M, Ball CA, Blake JA *et al.* Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25(1), 25–29 (2000).

38 Quackenbush J. Extracting biology from high-dimensional biological data. *J. Exp. Biol.* 210, 1501–1517 (2007).

39 Yildirim MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug-target network. *Nat. Biotechnol.* 25, 1121 (2007).

40 Wishart DS, Knox C, Guo AC *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36(Database issue), D901–D906 (2008).

41 Murray-Rust P. Semantic science and its communication – a personal view. *J. Cheminformatics* 3(1), 48 (2011).

42 Samwald M, Jentzsch A, Bouton C *et al.* Linked open drug data for pharmaceutical research and development. *J. Cheminformatics* 3(1), 19 (2011).

43 Orchard S, Al-Lazikani B, Bryant S *et al.* Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10(9), 661–669 (2011).

■ Website

101 Gene Ontology Consortium. The Gene Ontology Project (1999–2007). www.geneontology.org