Published in final edited form as:
J Chem Inf Model. 2008 April ; 48(4): 755–765. doi:10.1021/ci8000259.

Quantifying the Relationships among Drug Classes

Jérôme Hert[†], **Michael J. Keiser**[†], **John J. Irwin**[†], **Tudor I. Oprea**[‡], and **Brian K. Shoichet**[*,†]

[†]Department of Pharmaceutical Chemistry, University of California—San Francisco, 1700 4th St., San Francisco, California 94143-2550

[‡]Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, 2703 Frontier NE, Albuquerque, New Mexico 87131

Abstract

The similarity of drug targets is typically measured using sequence or structural information. Here, we consider chemo-centric approaches that measure target similarity on the basis of their ligands, asking how chemoinformatics similarities differ from those derived bioinformatically, how stable the ligand networks are to changes in chemoinformatics metrics, and which network is the most reliable for prediction of pharmacology. We calculated the similarities between hundreds of drug targets and their ligands and mapped the relationship between them in a formal network. Bioinformatics networks were based on the BLAST similarity between sequences, while chemoinformatics networks were based on the ligand-set similarities calculated with either the Similarity Ensemble Approach (SEA) or a method derived from Bayesian statistics. By multiple criteria, bioinformatics and chemoinformatics networks differed substantially, and only occasionally did a high sequence similarity correspond to a high ligand-set similarity. In contrast, the chemoinformatics networks were stable to the method used to calculate the ligand-set similarities and to the chemical representation of the ligands. Also, the chemoinformatics networks were more natural and more organized, by network theory, than their bioinformatics counterparts: ligand-based networks were found to be small-world and broad-scale.

INTRODUCTION

There is much current interest in relating drug targets by the chemical similarities of their ligands,[1–4] using the chemical similarity among ligand sets as a proxy for the pharmacological similarities of the protein targets. The idea exploits the internal similarity of most ligands for a particular target[5] and the observation that similar ligands will have similar protein binding patterns.[6,7] Chemical mapping of pharmacological relationships quantifies these notions, relating targets in a formal network by the similarity of their ligands, where the similarities are the network edges (Figure 1). These networks complement those more familiar from bioinformatics and reveal relationships among targets that would be obscure on the basis of sequence or structural similarities alone. For instance, some drugs acting on μ-opioid receptors have been found to resemble those acting on M3 muscarinic receptors, despite the differences between these receptors, leading to the prediction that the opioid methadone will antagonize M3 muscarinic receptors. Similarly, a drug acting on protein biosynthesis has been predicted

*Corresponding author tel.: +1 415-514-4126, fax: +1 415-514-4260, e-mail: shoichet@cgl.ucsf.edu.

Supporting Information Available: Example of set similarities between the glutamate ionotropic and metabotropic receptor (Table S1). Additional tables of Spearman correlations, percentages of common edges, and percentages of top nearest neighbors for the combinations of methods, databases, and descriptors that were not provided in the main body of the manuscript (Tables S2–S7). Additional plots of the connectivity distributions for the MDDR and the WOMBAT databases using SEA (Figures S1 and S2). This material is available free of charge via the Internet at http://pubs.acs.org.

to bind to adrenergic receptors, even though the ribosome and adrenergic G-protein-coupled receptors (GPCRs) are biologically unrelated. Nevertheless, both predictions have been subsequently confirmed experimentally.[4] Chemoinformatics networks of receptors have been used to predict off-target effects of drugs[3,8] and may be used to predict polypharmacology, side effects, and drug repurposing.

Notwithstanding the precocious successes of this research program, it is sensible to wonder why a chemical organization of targets should reflect biology. Most drugs are synthetic, and even natural products are typically used in contexts unanticipated by their biosynthesis. Thus, there is no historical relationship among drugs other than that conferred by human creativity (or its lack). Conversely, bioinformatics metrics are based on historical divergences among sequences and structures according to molecular evolution. When a particular protein shares a high sequence identity with another, one can be quantifiably certain that the two are related, not only on the basis of the statistics of expected similarity but also on the basis of a profound understanding of the mechanisms of gene replication, sequence divergence, and evolutionary relationship. No such mechanisms or relationships exist among drugs, each of which represents a unique chemical instance that is unrelated to all others except by similarities perceived or desired in the minds of their creators. Creationist relationships, so laughable a canard in biology, hold among drugs and synthetic ligands. Why then should we *expect* to relate biological targets on the basis of the similarity among the sets of drugs and reagents that bind to them? Can we do so reliably, or are the similarities dependent on the peculiarities of the chemical metrics? How quantifiably different are chemoinformatics networks from their bioinformatics analogs, and which sort of network is the most useful and trustworthy for pharmacology?

Here, we quantify the differences between chemoinformatics and bioinformatics similarities for pharmacological targets. To explore the robustness of these differences, we investigate how different chemical similarity metrics and the use of different ligand sets affect the calculated similarities of hundreds of targets. Doing so necessarily involves technical aspects of representing chemical information. We explore seven different ways of representing the ligands and two different methods for comparing the ligand sets (a statistical approach based on the individual similarity between ligands and a model-based approach using Bayesian inference). The details of these approaches will interest specialists in this area. The interest of the general reader will be repaid by the overall features that emerge from these quantitative comparisons. We find there is little similarity between a network of drug targets based on sequences and one based on ligand sets, irrespective of how we represent or compare the ligands. Conversely, the ligand-set relationships are largely preserved for multiple ways of representing chemical information and are robust even to the particular choice of ligands within these sets. Despite the lack of any evolutionary basis for the chemoinformatics networks, they are robust and, by network theory, more natural than are the bioinformatics networks. This apparently baffling result derives from the origins of the targets we compare in pharmacology.

METHODS

Sets of Ligands

An annotated set is one where a function, such as “dihydrofolate reductase inhibitor” or “anticancer agent”, is assigned to the ligands in it. The sets of ligands were created from two databases: (1) the *MDL Drug Data Report* (MDDR) 2006.1 database[9] and (2) the *WOrld of Molecular BioAcTiVity* (WOMBAT) 2006.1 database.[10] After removal of duplicates and molecules that we could not process, a total of 163 963 compounds in 593 annotated sets and 136 068 compounds in 1167 annotated sets for the MDDR and WOMBAT databases, respectively, were available for searching. The MDDR results reported hereafter consider a subset of the MDDR database containing 65 367 compounds organized in 249 sets (with more

than five ligands) that Schuffenhauer et al. could associate with a specific biological target in their ontology.[11]

Similarity Measures

The performance of six different topological fingerprints and one 3D structural fingerprint was evaluated: (1) 2048-bit Daylight,[12] (2) 988-bit Unity,[13] (3) 166-bit MDL Keys,[14] (4) 1024-bit ECFP_4,[15] (5) 1024-bit FCFP_4,[15] (6) 1200-bit CATS,[16,17] and (7) FEPOPS.[18] Daylight fingerprints were generated from an in-house program based on the Daylight toolkit. MDL Keys, ECFP_4, and FCFP_4 descriptors were generated using standard Pipeline Pilot components; ECFP_4 and FCFP_4 representations were subsequently folded to 1024-bit-strings.[19] CATS descriptors were calculated using an in-house Pipeline Pilot protocol; the histogram was not normalized by the number of heavy atoms as reported in the implementation description;[16] instead, the occurrence of each pharmacophore pair was binned into an 8-bit string.[19] The FEPOPS descriptors were generated using a Novartis in-house program. The Tanimoto coefficient was used for all similarity calculations.[20] If two molecules have *a* and *b* bits set to "on" in their bit-strings with *c* bits in common, the Tanimoto similarity is defined to be

$$\mathrm{Tc}=\frac{c}{a+b-c} \tag{1}$$

Relating Ligand Sets through the Similarity Ensemble Approach[4] (SEA)

SEA adapts statistical techniques that BLAST[21–23] uses to calculate expectation values for sequence similarity and applies them to calculate the similarity between sets of ligands. The method has been previously described and so will only be summarized here. SEA calculates a *raw score* between two sets by summing the Tanimoto similarities between all of the interset pairs of ligands. In itself, this raw score has two shortcomings: a strong dependence on the number of ligands in the two sets and a poor discrimination between relevant and random similarities. To overcome these weaknesses, the mean and standard deviation of raw scores obtained with random sets of ligands were modeled to functions of the product of the sets' sizes, allowing us to calculate a *z*-score that is free of set size bias. Emphasis was also given to ligands across the two sets that have a strong similarity score by applying a similarity threshold below which pairwise similarities were no longer contributing to the raw score. The distribution of the *z* scores obtained for a range of these thresholds is modeled by an extreme value distribution. The threshold resulting in the best fit corresponds to the best signal/noise discriminator and was subsequently employed in all set comparisons. The final score was expressed as an expectation value (*E*-value), that is, a probability of observing a given *z*-score using random data: the smaller the *E*-value, the stronger the relationship between two ligand sets is expected to be.

Relating Ligand Sets through Bayesian Models

The global set similarity as calculated by SEA is built up from the contribution of individual pairs of related molecules. No attempt to capture the common molecular components responsible for the activity of the ligands is made. In an effort to capture information common to a structure—activity relationship series, we directly compared Bayesian models. Such models measure the contribution of a given bit in a fingerprint for a specific outcome, typically activity or inactivity against a biological target, in a training set of known active and inactive molecules. A candidate compound is scored by summing the probabilistic weights of the bits in that compound's fingerprint. A simple extension of this method enables the quantitative comparison of sets. All of the sets were first combined in a single database; for each set, a training subgroup was created by considering all of its molecules as "active" and the rest of

the molecules as "inactive". Bayesian weights were calculated for each bit by taking the logarithm of the Avidon weights;[24–26] the resulting weights formed a vector of the length of the bit-string representing the entire set. A quantitative measure of the similarity between two ligand sets was obtained by calculating the Pearson product-moment correlation[27] between their corresponding Bayesian vectors.

Sequence Comparison

A total of 193 sets of the MDDR database and 840 sets of the WOMBAT database could be mapped to the sequence of their target; the sequence similarity among them, expressed as *E*-values, was determined using PSI-BLAST.[28]

Correlation between Distance Matrices

Distance matrices were obtained by calculating the pairwise similarity between all possible sets in a database. Every row in the matrix was combined into one global array of scores. The correlation between these arrays was calculated using the Spearman rank-order correlation coefficient, which uses the ranks and thus enables the comparisons of matrices with scores lying in different ranges.[29]

Threshold Networks Comparisons

Distance matrices were converted into *threshold networks* by generating a vertex for each ligand set and assigning an edge between two vertices if the corresponding set similarity was below (better than) a threshold *E*-value for SEA; vertices that were not connected by any edge were removed from the network. A given *E*-value is comparable across different fingerprints because *E*-values are normalized; the Pearson scores, however, are not normalized and cannot be quantitatively compared. The number of edges considered in threshold networks derived from Bayesian distance matrices was set to match the number of edges of their SEA counterpart. Only the edges corresponding to the highest Pearson scores were kept; all isolated vertices were removed.

Small-World and Scale-Free Networks

On the basis of similar work in biology,[30] we looked at how *small-world* and how *scale-free* the threshold networks were. Small-world networks,[31] commonly known as six degree of separation networks, are characterized by their *aVerage path length (L)* and their *clustering coefficient (C)*. The former is defined as the length of the shortest path to connect two vertices, averaged over all pairs of nodes (eq 2). The latter measures the average probability that two vertices with a common parent will be connected, that is, the average connectedness of local neighbohoods (eq 3):

$$L=\frac{1}{N(N-1)/2}\sum_{j>i}\lambda_{ij} \tag{2}$$

where N is the number of vertices and λ_{ij} is the number of edges in the shortest path between vertex i and vertex j;

$$C=\frac{1}{N}\sum_{i}\frac{n_i}{N_i(N_i-1)/2} \tag{3}$$

where N_i is the number of neighbors connected to vertex i. The maximum possible number of edges between these neighbors is thus $N_i(N_i - 1)/2$, while n_i denotes the actual number of edges that exist among these neighbors.

If k is the average number of edges per vertex, random networks have an average path length $L_{random} \sim \ln(N)/\ln(k)$ and a clustering coefficient $C_{random} \sim k/N$, while for regular lattices (all vertices have the same number of edges), $L_{regular}=N(N+k-2)/[2k(k-1)]$ and $C_{regular} = 3(k-2)/[4(k-1)]$.[32] A network is said to be small-world if $L_{random} \leq L \ll L_{regular}$ and $C_{random} \ll C \leq C_{regular}$.

These properties hold true for networks in which everyvertex can reach any other vertex. In our threshold networks, there were islands composed of few vertices that were separated from the other components of the network (the number of islands increases as the *E*-value threshold becomes more stringent). The average path length was hence not calculated by dividing the sum of the shortest paths by $N(N-1)/2$ (eq 2) but by the actual number of vertex pairs that could be connected.[33] The resulting L may marginally underestimate the actual average path had our databases sampled a larger number of targets.

Scale-free networks[34] are characterized by the connectivity distribution, estimated by the frequency $P(k)$ of vertices of degree k. If a network is small-world and $P(k) \sim k^{-\gamma}$, with γ values typically between 2 and 4, then the network is considered scale-free.

RESULTS

A startling result from our initial work on pharmacological networks was the observation that networks based on ligand similarities differed greatly from those based on the sequence identities among their targets.[4] We wanted to quantify this difference and investigate how sensitive it was to the representation of chemical information. We began, therefore, by calculating ligand-based pharmacological networks using seven representations of chemical information, most of which are widely used in chemoinformatics. As initial questions, we asked how similar the networks were on the basis of the different molecular fingerprints, how sensitive they were to the exact identities of the ligands used to define the sets, and how similar the chemoinformatics networks were to sequence-based bioinformatics networks of the same targets.

The seven molecular representations that we chose were Daylight, Unity, MDL Keys, ECFP_4, FCFP_4, CATS, and FEPOPS.[14,18,19] The first five fingerprints are based on two-dimensional molecular topology and represent molecules as bit-strings where the presence or absence of a chemical substructure is denoted by the status of one or several particular bits (see Methods). The CATS descriptor encodes the histogram of through-bond distances between pairs of pharmacophoric atom types, and the FEPOPS descriptor represents molecules by the physicochemical properties of the four feature points that result from the clustering of the three-dimensional coordinates of the atoms. We calculated fingerprints for the 65 367 molecules in the subset of the MDDR database that can be assigned to particular molecular targets. A molecular target is a specific macromolecule such as dihydrofolate reductase (DHFR), for which 216 ligands are annotated in the MDDR database; there are 249 molecular targets in the Schuffenhauer subset of the MDDR database. We then calculated the relationships among each of the molecular targets through the similarities of their ligand sets, using either SEA[4] or the method comparing Bayesian models. In SEA, the Tanimoto similarities were calculated for every interset pair of ligand descriptors, and all similarities that were over a certain threshold value were summed. This raw score was then corrected for the similarity we would expect at random and expressed as an expectation value. In the Bayesian approach, a weight was assigned to each bit of the fingerprint by comparing the number of times it was set in the ligands of a

receptor versus the number of times it appeared in all of the molecules of the database. The resulting vector of weights was used as a descriptor for this set of ligands. The similarity between two ligand sets was quantified by measuring the Pearson correlation between their two vectors (see Methods). This method requires the fingerprint to be binary and could not be used with the FEPOPS descriptors.

Correlation between the Ligand-Based Similarity Matrices

The ligand sets were compared seven times—one for each fingerprint—and the similarities among them were compared fingerprint to fingerprint. This meant calculating a 249-square matrix, one for every ligand set (and hence target), for each descriptor. To compare any two networks, we could simply compare the square matrices (Figure 2). For instance, when using the Daylight fingerprints to represent ligand information, the most similar ligand set to the DHFR ligands is that of glycinamide ribonucleotide formyltransferase (GART) with an expectation value of 8.63×10^{-79} (Table 1). Correspondingly, the GART ligand set is most similar to DHFR and then to the thymidylate synthase (TS) ligand set, with an *E*-value of 5.19×10^{-66}; the TS ligands are the third most similar set to the DHFR ligands (*E*-value of 6.55×10^{-48}). All of these sets are highly related when the ligands are represented by Daylight fingerprints. We compare these similarities for the same ligand sets when the ligands are represented by ECFP_4 fingerprints (Table 1). With these fingerprints, the DHFR ligand set has an *E*-value of 1.31×10^{-264} to the GART ligand set and remains its nearest neighbor; and the second nearest neighbor of GART remains the TS set, with an *E*-value of 5.35×10^{-256}. The TS ligand set, represented by ECFP_4 fingerprints, is no longer the third most similar set to the DHFR ligands but is now the second most similar set, with an *E*-value of 8.52×10^{-142}. Whereas the exact expectation values between these sets, represented by Daylight or ECFP_4 fingerprints, differ, they are related by rank order, which is probably a more informative criterion. We compared the full matrices of similarities to one another, using Spearman rank-order correlation coefficients to quantify monotonic order similarities (Table 2). Each matrix, and hence each pharmacological network, was similar when using the five topology-based fingerprints, with Spearman rank-order coefficients varying from a low 0.783 (Daylight-fingerprint network vs MDL Keys-fingerprint network) to a high 0.940 (Daylight-fingerprint network vs Unity-fingerprint network), where a coefficient of +1 would have been perfect correlation, 0 a complete lack of correlation, and −1 a perfect inverse correlation. Topology-fingerprint-based networks were also correlated to the CATS and FEPOPS networks. Here, the Spearman coefficients were lower, ranging from 0.530 (MDL Keys-fingerprint network vs FEPOPS-descriptor network) to 0.622 (ECFP_4-fingerprint networks vs CATS-fingerprint networks). The worst Spearman coefficient, 0.365, was obtained by comparing the CATS and FEPOPS networks but is still significant ($\alpha \ll 0.05$), as illustrated by the Spearman value, 0.002, obtained between randomized networks. Similar correlations were observed between the matrices when the similarity between the ligand sets was quantified with the Bayesian method rather than the SEA method (Supporting Information, Table S2A), or when the WOMBAT database was considered instead of the MDDR database (Supporting Information, Tables S2B and S2C). Lastly, taking these comparisons to a final step, we investigated how the seven networks—one for each fingerprints—compared when calculated by the SEA or the Bayesian approach. These two methods differ greatly—SEA represents the entire molecule and corrects for a random background, whereas the Bayesian method looks for common substructural features—and are related mostly by both considering ligand information. As might be expected, the seven SEA-based and the seven Bayesian-based networks were much less correlated than the seven networks were when compared within each method. The Spearman coefficients varied from 0.300 to 0.428 and from 0.203 to 0.518 for the MDDR and the WOMBAT databases, respectively. Still, despite the fact that the similarity matrices are not, *a priori*, expected to correlate to each other, their rankings were related at a statistically significant level.

Correlation between Chemoinformatics and Bioinformatics Networks

Of the 249 ligand sets in the MDDR database, 193 were linked to the sequence of their corresponding protein target. The 193-square bioinformatics matrix was obtained by calculating the PSI-Blast[28] similarity between each pair of protein sequences, and the bio- and chemoinformatics matrices were compared by calculating the Spearman rank-order correlation coefficient. None of the chemoinformatics matrices, irrespective of the choice of ligand-set comparison method or molecular representation, were correlated to the sequence-based matrix; Pearson values varied from −0.228 to −0.027 (Table 3). With the WOMBAT database, 840 of the 1183 ligand sets were associated with a protein sequence, resulting in an 840-square sequence-based matrix of PSI-Blast scores. Here too, chemo- and bioinformatics matrices were very different, with Spearman coefficients between matrices ranging from 0.009 to 0.017 (Supporting Information, Table S3). This difference is, retrospectively, sensible. Many of the targets in the MDDR are members of large superfamilies that are all related by sequence but recognize unrelated ligands. Thus, more than 40% of the ligand sets in the MDDR are associated with a GPCR, which all derive from a common ancestor and have long regions of sequence identity, such as the transmembrane helices, that are not intimately linked to ligand recognition. The heat map of the sequence-based similarity matrix is correspondingly densely filled (Figure 2A), while the heat maps of the ligand-based similarity matrices are much sparser (Figure 2B and C). On the other hand, the chemoinformatics methods recognize pharmacologically relevant relationships that are often obscure to bioinformatics approaches, owing to ligand-set similarity. For instance, the $5HT_3$ ionotropic and the metabotropic $5HT_4$ serotonergic receptors are related by ligand-based methods (Figure 1 and Figure 2), as are the ionotropic glutamate receptors (AMPA, NMDA, and Kainate) and the glutamate metabotropic receptors, for example, mGluR4 (Supporting Information, Table S1). By sequence and structure, of course, there is little relationship between these ion channels and GPCRs.

Percentage of Overlapping Edges of the Threshold Networks

Another metric of similarity consists of asking how many relationships (graph edges) are shared among the ligand sets (vertices) when the networks are calculated with the different fingerprints (in the last section, we compared rank ordering of similar sets; here, we ask how many sets are considered “related” over a given threshold of similarity). We made this comparison using threshold networks which were calculated using the ligand-set similarity scores of the square matrices. Every ligand set, and hence molecular target, was represented by a vertex, and two vertices were connected by an edge if the two ligand sets were more similar than a given threshold (Figure 1). The degree of agreement between the nearest neighbor lists of the ligand sets was measured by the percentage of overlapping edges in the threshold networks. Considering the MDDR ligand sets compared to one another using SEA, all of the topology-fingerprint-based networks had between 69.9% and 90.4% overlapping edges (Table 4). The percentage of common edges of these networks varied between 38.9% and 51.9%, with the CATS-fingerprint network and between 27.4% and 35.9% with the FEPOPS-descriptor network. Similar observations were obtained with the WOMBAT ligand sets (Supporting Information, Table S4), with percentages of common edges varying between 69.5% and 85.8% for networks where topology-based fingerprints were used, and an average of 42.8% and 30.7% edge-overlap between the CATS-fingerprint and FEPOPS-descriptor networks, respectively.

The chemoinformatics threshold networks were also compared to the sequence-based bioinformatics networks (Table 5). With the MDDR database, the percentage of common edges averaged 23.5% and 28.4% between the sequence-based networks and the ligand-based networks calculated by SEA and the Bayesian method, respectively (Table 5). In contrast, 40.2% of the edges in the SEA-based threshold networks were also found in the Bayesian-based threshold networks. This trend is more pronounced when comparing the WOMBAT sequences and ligand sets; the average percentage of overlapping edges is 16.9% and 11.3%

between the sequence-based threshold networks and the SEA- and Bayesian-based threshold networks (Supporting Information, Table S5), but the average percentage of common edges between SEA- and Bayesian-based networks was 36.2%. These trends were preserved when the *E*-value threshold was decreased $10^{-10} \rightarrow 10^{-20} \rightarrow 10^{-50}$ or increased to 1.

Consistency of the Top Hits

It could be argued that the most important similarities are those at the very top of any list. Minimum spanning trees derived from the ligand-based similarity matrices connect the most similar neighbors together. Clusters of known pharmacological target families may be observed consistently in these networks irrespectively of the set comparison method or fingerprints considered (Figure 1). Hence, it seemed useful to quantify how the rankings of the most similar sets to any given query set changed as we looked at different fingerprints. We measured the consistency (or inconsistency) of the nearest neighbors of a query set by calculating the percentage of time the nearest neighbor of a ligand set in one network was also in the list of top-three and top-five nearest neighbors in another. When the MDDR ligand sets were compared using SEA, the nearest neighbors in the topology-fingerprint-based networks are mostly found in the top-three nearest neighbors of any other topology-fingerprint-based network with values varying from a low 67.9% (Daylight-fingerprint networks vs MDL Keys-fingerprint networks) to a high 89.6% (ECFP_4-fingerprint network vs FCFP_4-fingerprint network) (Supporting Information, Table S6). The top hits of these five same networks are found in the top-three nearest neighbors of the CATS- and FEPOPS-fingerprint networks in, at worst, 57.1% and 49.0% of the cases, respectively. The percentage of top nearest neighbors obtained with the CATS fingerprint also found in the top-three nearest neighbor list with the FEPOPS descriptor was 47.8%. On average, more than two of the three (68.34%) closest neighbors to a given target in one network were also found among the closest three neighbors of another. Similar results were observed with the WOMBAT ligand sets, with an average of 68.0% of the nearest neighbors of one network also found in the closest three neighbors of another (Supporting Information, Table S6). Meanwhile, the closest neighbors in the bioinformatics networks were found in, at most, 24.7% and 23.2% of the closest three neighbors among the chemoinformatics networks with the MDDR and the WOMBAT databases, respectively (Supporting Information, Table S7). This overlap between the sequence-based and chemoinformatic neighbors did not exceed 28% when the closest five nearest neighbors were considered. In contrast, the closest neighbors of the SEA-based ligand networks were also found in at least 60% of the closest neighbors of the Bayesian method.

Effectiveness of the Different Fingerprints

With the above analysis suggesting that the chemoinformatics networks are robust and thus in some sense meaningful, we wondered if any one of these fingerprints was better than the others for the purpose of network similarity. We looked at the effectiveness of the different representations of the chemical information using a 10-fold cross-validation experiment. Each ligand set with 50 compounds or more in the MDDR database was randomly divided into 10 subsets. A test database was built by grouping one subset for each ligand set, while the remaining nine subsets formed the training database. The procedure was repeated 10 times so that each subset was used once in the test database. Each compound in the test database was scored against each ligand set in the training database, measuring the effectiveness of a particular fingerprint by the extent to which true ligands received higher scores than decoys using the area under the receiver operating characteristic curve (ROC-AUC) as a criterion. A ROC-AUC of 1 indicates a perfect discrimination, while a value of 0.5 denotes an absence of discrimination. A total of 1790 (179 sets × 10 subsets) ROC-AUC values were computed for each fingerprint. The average ROC-AUC was calculated for every fingerprint, and a fingerprint was declared more effective if it had a higher average ROC-AUC value. The most effective descriptor was ECFP_4 (Table 6). In order of decreasing effectiveness, irrespective of whether

the SEA or the Bayesian methods were used, the different molecular representations rank as follow: ECFP_4 > FCFP_4 > Daylight > Unity > MDL Keys > CATS.

Effectiveness of the Different Methods

Using the exact same set of 10-fold validation experiments, we compared the effectiveness of the SEA and the Bayesian approaches. SEA was found to be more effective than the Bayesian method, irrespective of the fingerprints used to encode the molecules (Table 6). In two cases, when the ECFP_4 and FCFP_4 fingerprints were used, the effectiveness of the two methods was almost equivalent with average ROC-AUC values of 0.987 and 0.979 and of 0.984 and 0.978, respectively.

Properties of the Threshold Networks

In many disciplines, the structure of complex networks has sparked considerable debate.[30,35,36] Two classes of networks are particularly relevant: small-world networks[31] and scale-free networks.[34] The properties of these two classes are important because they affect how the structure of a network evolves with its size. Small-world networks have average path lengths (L) that increase logarithmically with the number of vertices, while preserving a significant local neighborhood (C). In practice, a network is said to be small-world if its average path length and clustering coefficient lie between the estimated values calculated for random and regular networks with similar properties (see Methods).[32,35] All of the chemoinformatics threshold networks were found to be small-world, as their average path lengths and clustering coefficients were between those of their corresponding random and regular networks (Figure 3). In contrast, the sequence-based network was not small-world; its clustering coefficient was higher than that of its corresponding regular networks. In small-world networks, it is possible to connect any two vertices through just a few links, underlying the presence of large central hubs that lead to smaller archipelagos.

Small-world networks fall into three more classes: *scale-free* networks, *broad-scale* networks, and *single-scale* networks.[37] Scale-free networks emerge in the context of a growing network, where new vertices connect preferentially to the more highly connected vertices in the network. A network can only have scale-free properties if it already has small-world properties. Networks are labeled scale-free if the vertex connectivity, that is, the frequency of the number of edges per vertex, follows a power law distribution (see Methods). This was the case, irrespective of the fingerprint used, for the MDDR and WOMBAT threshold networks (E-value $\leq 10^{-10}$) obtained from the SEA comparison of the ligand sets with coefficients of determination (R^2) varying from 0.70 to 0.88 and from 0.78 to 0.85, respectively (Figure 4 and Supporting Information, Figures S1 and Figure S2). No such relationship between the number of edges per vertex and its frequency could be observed with the bioinformatics networks. A close inspection of the figures shows that the connectivity distribution of the chemoinformatics networks starts with a power law regime followed by a sharp cutoff that is characteristic of broad-scale networks, also known as “truncated” scale-free networks. Broad-scale networks are scale-free networks where the addition of edges was limited for some reason. As before, varying the thresholds used in the construction of these networks from $1 \rightarrow 10^{-10} \rightarrow 10^{-20} \rightarrow 10^{-50}$ had little effect on the properties.

DISCUSSION

Biological targets may be related by their ligands, leading to connections unanticipated by bioinformatics similarities. As intriguing as these chemoinformatics associations are, they are unsupported by formal theory, unlike those based on bioinformatics networks. To investigate the stability of chemoinformatics networks, we varied the fingerprints encoding ligand information, varied the precise ligands used to represent targets, and varied the method by

which ligands sets are related. Three key points emerge. First, the chemoinformatics and bioinformatics networks differ substantially. Second, the chemoinformatics networks are robust to perturbations in ligand representation and identity. Third, the chemoinformatics networks are well-behaved for the pharmacological targets by network theory.

The sequence-based bioinformatics networks and the ligand-based chemoinformatics networks differed substantially: no rank-order correlation between ligand-set-based and sequence-based matrices was observed (Table 3), and heat maps that were densely filled by sequence similarity were sparse by ligand similarity (Figure 2). Naturally, the two metrics do not always disagree; there are cases where high sequence similarity implies high ligand-set similarity between receptors. For instance, thrombin and trypsin, two serine proteases, are similar by sequence, with an *E*-value of 3.0×10^{-94}, and their inhibitor sets are also similar, with an *E*-value of 2.38×10^{-85} using SEA and the ECFP_4 fingerprints. What was surprising to us is that such correspondences were the exception and not the rule. There were many more cases where targets highly related by sequence were unrelated by ligands, and many cases where receptors unrelated by sequence were highly related by ligands. For instance, the opioid receptors are related to many serotonergic receptors; both are GPCRs with many structural similarities. Their ligands bear little relationship, however. On the other side, the metabatropic and ionotropic glutamate receptors bear no sequence similarity, the one being GPCRs, the other ion channels, but their ligands are often highly related (Supporting Information, Table S1). Overall, only 20–30% of the targets were highly related by both sequence and ligand similarity (Tables 3 and Supporting Information, Table S7).

The differences between ligand-set similarity and sequence identity may be readily explained. Sequence identity is measured across an entire protein, whereas ligand similarity is local. The GPCRs, for example, are conserved in overall sequence and most share a common ancestry, but many recognize unrelated ligands. High sequence identity among receptor superfamilies leads to the dense heat map matrix for drug targets (Figure 2) and to the highly clustered sequence-based networks of pharmacological targets (Figure 3 and Figure 4). These dense relationships and non-natural networks are no reflection on the quality of the bioinformatics metrics but simply the domain of proteins they have been asked to relate—those that are drug targets. Indeed, previous studies have shown that bioinformatics networks based on the sequence similarity of the binding sites alone are less densely clustered and more closely resemble small-world and scale-free networks.[38] Still, it does appear that linkages provided by ligand information will often be more apposite to pharmacological interests than those provided by protein sequence, however construed.

Differences between chemoinformatics and bioinformatics networks would be less interesting if the chemoinformatics relationships were unstable to chemical representation; there is no single consensus or “right” way to represent chemical information, and many different chemical fingerprints have been proposed. One way to investigate this is to ask how networks change when we vary molecular representation and how we compare it. There was generally good agreement between the networks when the molecules were represented by five topology-based fingerprints (Daylight, Unity, MDL Keys, ECFP_4, and FCFP_4). This agreement was supported by multiple metrics, including rank correlation of the similarity matrices (Table 2), percentage of overlapping edges in the threshold networks (Table 4 and Supporting Information, Table S4), and consistency of the nearest neighbors (Supporting Information, Table S6). Admittedly, agreement substantially decreased when comparing these topology-based networks with networks based on CATS or FEPOPS descriptors, which represent abstracted features of the molecules. This may simply highlight the different uses for which the fingerprints were designed. The CATS and FEPOPS fingerprints are thought to be less effective than other representations for virtual screening applications[19,39] but, because they are fuzzier, are better suited to finding new scaffolds.

The chemoinformatics networks were surprisingly stable even to the method of comparing the chemical information. The two methods used to compare the ligand sets, SEA and the Bayesian method, differ substantially in their approach to quantify set similarity: SEA perceives local instances of similarity between sets by considering every pair of interest ligands explicitly, whereas the Bayesian method collapses all the set's information and thus captures the global commonalities between sets. There is no reason why the two networks *should* be related. Nevertheless, 60% of the most related targets of one method were among the top three hits of the other (Supporting Information, Table S7). The targets most related by SEA and Bayesian analysis resembled each other much more closely than did either compared to targets related by sequences. The similarity of these networks is perhaps clearest from a comparison of their minimum spanning trees (Figure 1). Whereas the details of the wiring of these networks between different methods or fingerprints differ, the known pharmacological families typically clustered similarly in each of the chemoinformatics networks. This suggests that the ligand information itself, despite our inability to fully capture it in any single fingerprint or compare it any single "best" way, is the fundamental basis for the relationships found among the receptors. Any good way of representing ligand information may likely afford similar relationships among targets.

A final surprise was that the ligand-based networks appeared to be small-world and broad-scale (Figure 3 and Figure 4). These are properties shared by networks relating many natural phenomena and human activities, such as those mapping Internet connectivity,[34,40,41] social relationships,[31,42] or commerce,[43] all of which show a high degree of self-organization. The topology of chemoinformatics networks thus obeys well-defined rules and appears natural. This was unexpected because the drugs and reagents upon which the ligand-based networks are based are themselves contrivances of human invention and have no history of development in the natural world. Their status as well-behaved and apparently "natural" networks has implications for their structure and future growth. Being small-world implies that any two ligand sets can be connected by hopping through the chemistry of only a few other sets, which is to say that new drug classes are likely to appear as neighbors of already established classes. Also, the ligands for this putative new drug target are more likely to adhere to a set that already has many connections. In some senses, the polypharmacology observed for many ligands is reflected in the broad-scale aspect of these networks, where there is a high chance of having a highly connected vertex compared to random networks.

It is appropriate, in closing, to return to the observation that is arguably most relevant to medicinal chemistry: associations between targets are most naturally drawn on the basis of ligand similarities, which will often be more informative, pharmacologically, than those drawn on the basis of protein sequence or structure. These ligand-based networks also have unexpected connections that suggest off-target effects; these may be directly tested. Whereas not all of these predictions will hold up, a key result of this work is that they are soundly based in the information content of the ligands themselves and are not peculiarities of how we represent them.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

Supported by NIH grant GM71896 and a National Science Foundation graduate fellowship (to M.J.K.). J.H. was partly supported by the 6th Framework Program of the European Commission. We are grateful to MDL Information Systems Inc. for the MDDR database; Sunset Molecular Discovery LLC for the WOMBAT database; Daylight Chemical Information Systems Inc. and OpenEye Scientific Software for software support; Jeremy Jenkins for the generation of the FEPOPS descriptors. We thank Kristin E. Coan and Michael Mysinger for reading this manuscript.

REFERENCES AND NOTES

1. Izrailev S, Farnum MA. Enzyme classification by ligand binding. Proteins: Struct., Funct., Bioinf 2004;57:711–724.
2. Vieth M, Higgs RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. Biochim. Biophys. Acta 2004;1697:243–257. [PubMed: 15023365]
3. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL. Global Mapping of Pharmacological Space. Nat.Biotechnol 2006;24:805–815. [PubMed: 16841068]
4. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Iriwin JJ, Shoichet BK. Relating Protein Pharmacology by Their Ligands. Nat. Biotechnol 2007;25:197–206. [PubMed: 17287757]
5. Johnson, MA.; Maggiora, GM. Concepts and Applications of Molecular Similarity. New York: John Wiley; 1990.
6. Frye SV. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chem. Biol 1999;6:R3–R7. [PubMed: 9889153]
7. Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat. ReV. Genet 2004;5:262–275. [PubMed: 15131650]
8. Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J. Chem. Inf. Model 2006;46:1124–1133. [PubMed: 16711732]
9. The MDL Drug Data Report Database is available from MDL Information Systems, Inc. at http://mdl.com. [accessed Feb 2008].
10. Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, TI. WOMBAT: World of Molecular Bioactivity. In: Oprea, TI., editor. Chemoinformatics in Drug DiscoVery. New York: Wiley-VCH; 2004. p. 223-239.
11. Schuffenhauer A, Zimmermann J, Stoop R, van der Vyver JJ, Lecchini S, Jacoby E. An Ontology for Pharmaceutical Ligands and Its Application for in Silico Screening and Library Design. J. Chem. Inf. Model 2002;42:947–955.
12. The Daylight Toolkit is available from Daylight, Inc. at http://www.daylight.com. [accessed Oct 2007].
13. Unity is available from Tripos, Inc. at http://www.tripos.com. [accessed Feb 2008].
14. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL Keys for Use in Drug Discovery. J. Chem. Inf. Model 2002;42:1273–1280.
15. Pipeline Pilot is available from SciTegic, Inc. at http://www.scitegic.com. [accessed Feb 2008].
16. Fechner U, Franke L, Renner S, Schneider P, Schneider G. Comparison of correlation vector methods for ligand-based similarity searching. J. Comput.-Aided Mol. Des 2003;17:687–698. [PubMed: 15068367]
17. Schneider G, Neidhart W, Giller T, Schmid G. “Scaffold-Hopping” by Topological Pharmacophore Search: A Contribution to Virtual Screening. Angew. Chem., Int. Ed 1999;38:2894–2896.
18. Jenkins JL, Glick M, Davies JW. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. J. Med. Chem 2004;47:6144–6159. [PubMed: 15566286]
19. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. Org. Biomol. Chem 2004;2:3256–3266. [PubMed: 15534703]
20. Willett P, Barnard JM, Downs GM. Chemical Similarity Searching. J. Chem. Inf. Comput. Sci 1998;38:983–996.
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol 1990;215:403–410. [PubMed: 2231712]
22. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U.S.A 1990;87:2264–2268. [PubMed: 2315319]
23. Pearson WR. Empirical statistical estimates for sequence similarity searches. J. Mol. Biol 1998;276:71–84. [PubMed: 9514730]

24. Avidon VV, Arolovich VS, Kozlova SP, Piruzyan LA. Statistical Study of Information File on Biologically Active Compounds. II. Choice of Decision Rule for Biological Activity Prediction. Khim. Farm. Zh 1978;12:88–93.
25. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J. Chem. Inf. Model 2006;46:462–470. [PubMed: 16562973]
26. Xia X, Maliski EG, Gallant P, Rogers D. Classification of kinase inhibitors using a Bayesian model. J. Med. Chem 2004;47:4463–4470. [PubMed: 15317458]
27. Sheskin, DJ. Handbook of Parametric and Nonparametric Statistical Procedures. Vol. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC; 2003.
28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 1997;25:3389–3402. [PubMed: 9254694]
29. Siegel, S.; Castellan, NJ. Nonparametric Statistics for the BehaVioral Sciences. New York: Mc-Graw Hill; 1988.
30. Grigorov MG. Global properties of biological networks. Drug DiscoVery Today 2005;10:365–372. [PubMed: 15749285]
31. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature 1998;393:440–442. [PubMed: 9623998]
32. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phys. ReV. E: Stat., Nonlinear, Soft Matter Phys 2002;65:061910.
33. Albert R, Barabási A-L. Statistical mechanics of complex networks. ReV. Mod. Phys 2002;74:47.
34. Barabási A-L, Réka A. Emergence of Scaling in Random Networks. Science 1999;286:509–512. [PubMed: 10521342]
35. Greene LH, Higman VA. Uncovering network systems within protein structures. J. Mol. Biol 2003;334:781–791. [PubMed: 14636602]
36. Arita M. The metabolic world of Escherichia coli is not small. Proc. Natl. Acad. Sci. U.S.A 2004;101:1543–1547. [PubMed: 14757824]
37. Amaral LAN, Scala A, Barthélémy M, Stanley HE. Classes of small-world networks. Proc. Natl. Acad. Sci. U.S.A 2000;97:11149–11152. [PubMed: 11005838]
38. Zhang Z, Grigorov MG. Similarity networks of protein binding sites. Proteins: Struct., Funct., Bioinf 2006;62:470–478.
39. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M. Bridging chemical and biological space: “target fishing” using 2D and 3D molecular descriptors. J. Med. Chem 2006;49:6802–6810. [PubMed: 17154510]
40. Adamic LA, Lukose RM, Puniyani AR, Huberman BA. Search in power-law networks. Phys. ReV. E: Stat., Nonlinear, Soft Matter Phys 2001;64:046135.
41. Yook S-H, Jeong H, Barabási A-L. Modeling the Internet’s large-scale topology. Proc. Natl. Acad. Sci. U.S.A 2002;99:13382–13386. [PubMed: 12368484]
42. Barabási A-L, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A 2002;311:590–614.
43. Kogut B, Walker G. The Small World of Germany and the Durability of National Networks. Am. Sociol. ReV 2001;66:317–335.

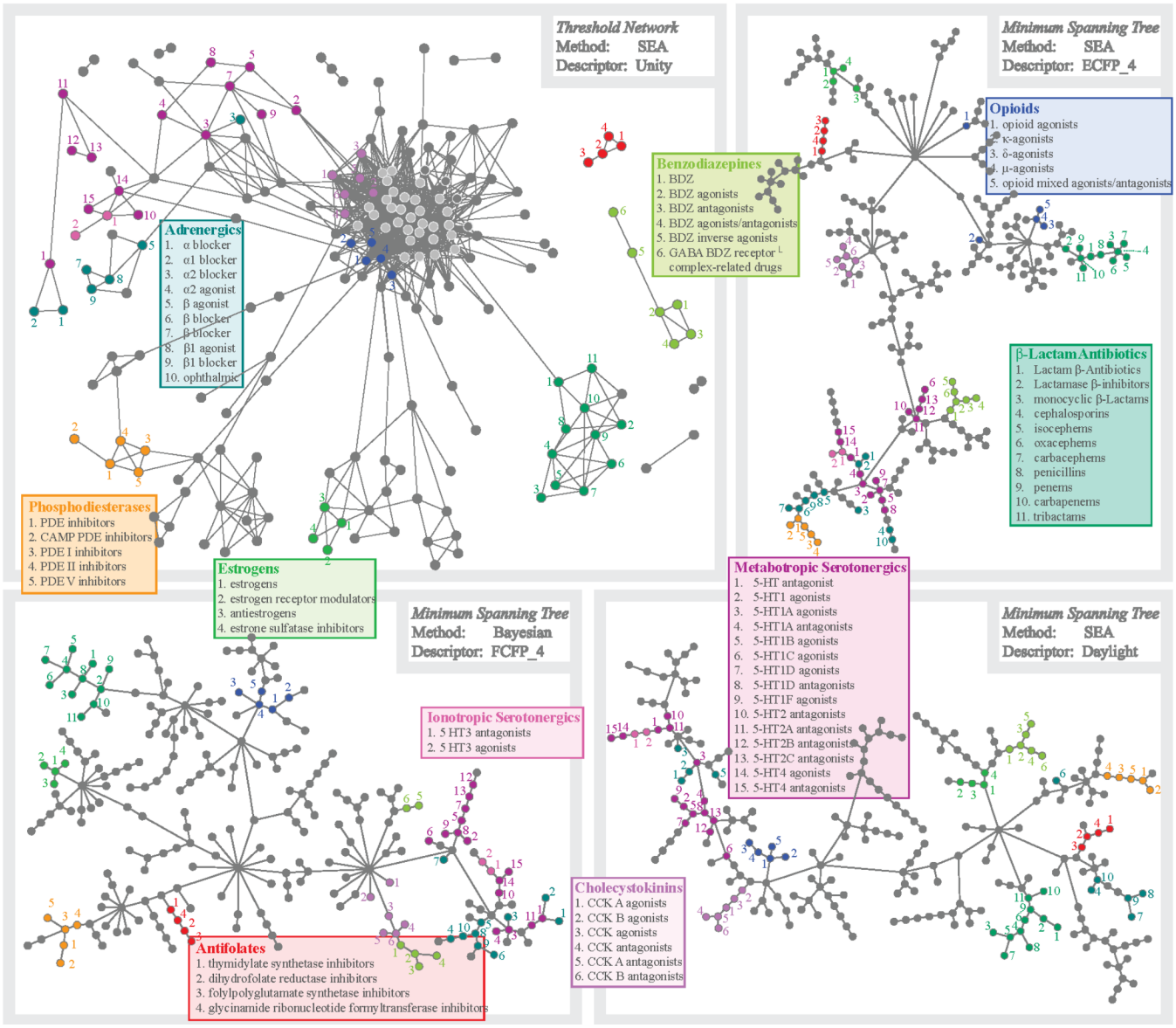

Figure 1.

Threshold network and minimum spanning trees of drug targets in the MDDR. Ligand sets are linked by edges according to a minimum level of similarity (E-value $\leq 10^{-10}$) or by which sets are most related (minimum spanning trees). The networks obtained using different ligand set comparison methods and different descriptors are depicted; the ligand sets corresponding to several protein families are highlighted to illustrate the clustering that emerges naturally from all of these networks.

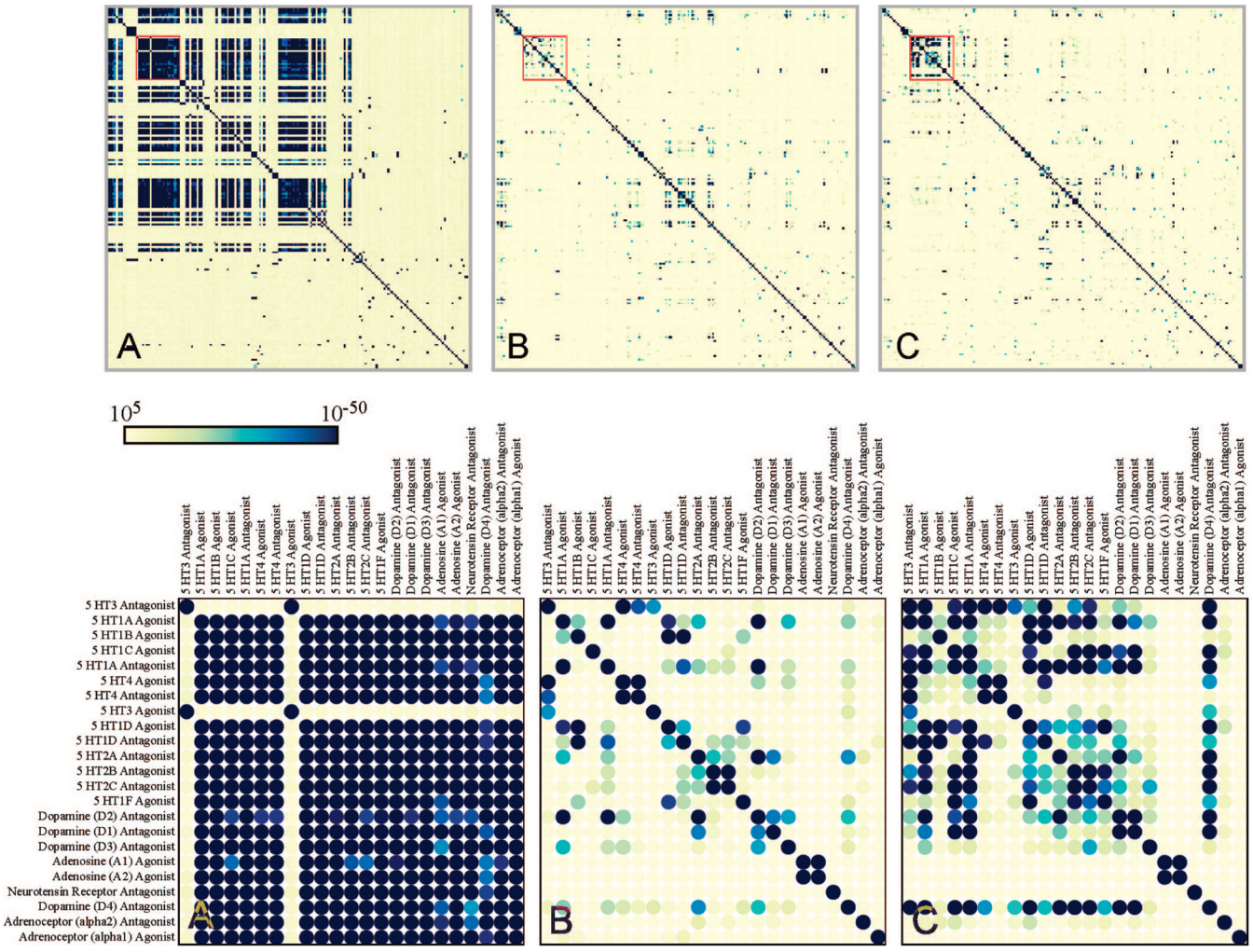

Figure 2.

Similarity heat maps of 249 MDDR targets calculated using (A) PSI-Blast from the sequences, (B) SEA with ECFP_4 fingerprints, and (C) Bayes with Unity fingerprints. Dark blue regions indicate high similarity (E-value $\leq 10^{-50}$), whereas light yellow regions indicate low similarity (E-value $\geq 10^{5}$). The lower row shows blowups of the highlighted region in the overall maps.

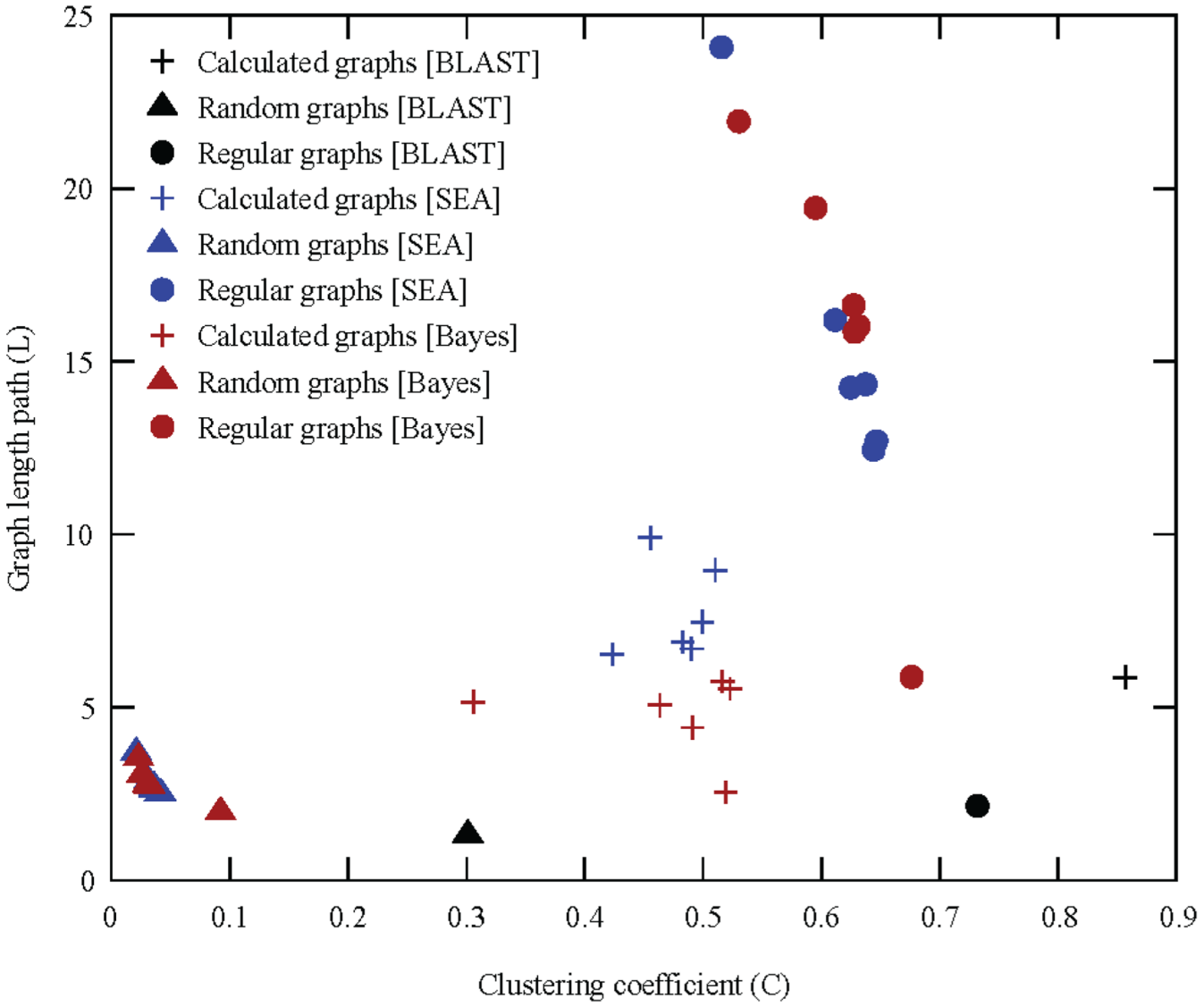

Figure 3.

The "small-world" nature of the chemoinformatic networks. Average path length (*L*) versus clustering coefficient (*C*) for 249 MDDR targets using an *E*-value threshold of 10^{-10} and all different descriptors. Black marks correspond to path lengths from the sequence similarity network, blue to path lengths from the SEA network, and red marks to the path lengths from the Bayesian networks. Triangles represent the values calculated for random graphs, circles those calculated for regular graphs, and crosses the actual observed values for each method and descriptor.

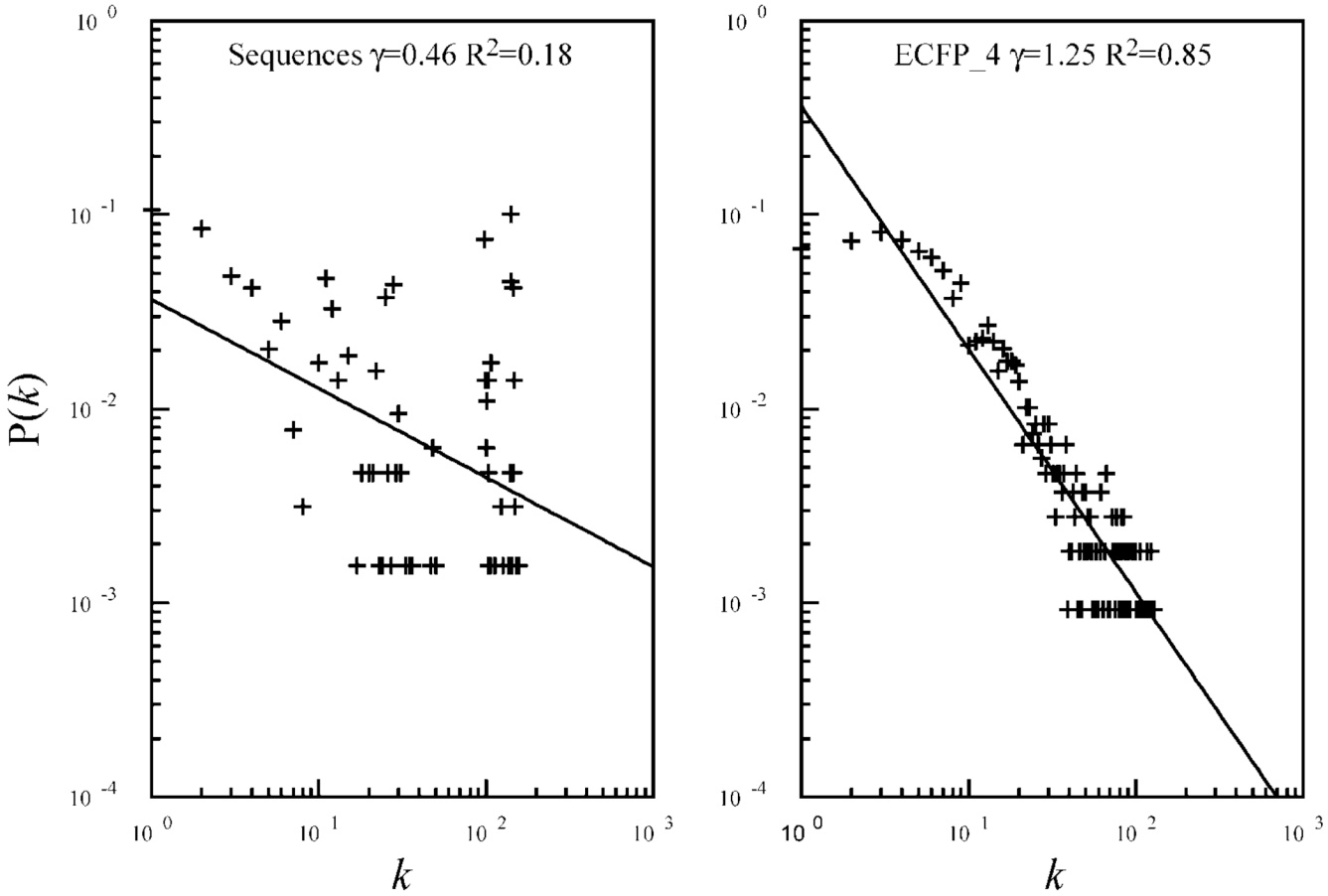

Figure 4.
The chemoinformatics networks are “broad-scale”. Frequency of the number of edges per vertex (*P(k)*) versus number of edges per vertex (*k*) calculated with the WOMBAT threshold networks (*E*-value < 10^{-10}) where the ligand sets were compared with the SEA method.

Table 1

An Example of Ligand Set Similarities and Their Variation with Fingerprint Type[a]

DHFR Is the Reference Ligand Set					
Daylight fingerprint			**ECFP_4 fingerprint**		
rank	**ligand-set annotation**	***E*-value**	**rank**	**ligand-set annotation**	***E*-value**
1	GART	8.6 E–79	1	GART	1.3 E–264
2	folylpolyglutamate synthetase	4.4 E–48	2	thymidylate synthetase	8.5 E–142
3	thymidylate synthetase	6.6 E–48	3	folylpolyglutamate synthetase	7.5 E–132
GART Is the Reference Ligand Set					
Daylight fingerprint			ECFP_4 fingerprint		
rank	ligand-set annotation	*E*-value	rank	ligand-set annotation	*E*-value
1	DHFR	8.6 E–79	1	DHFR	1.3 E–264
2	thymidylate synthetase	5.2 E–66	2	thymidylate synthetase	5.4 E–256
			3	folylpolyglutamate synthetase	3.6 E–09

[a]Ligand sets similar to dihydrofolate reductase (DHFR) and glycinamide ribonucleotide formyltransferase (GART) ligands (*E*-value ≤ 1). SEA was used to compare the sets of ligands; the sets were drawn from the MDDR. A Tanimoto cutoff of 0.55 for the Daylight fingerprints and of 0.35 for the ECFP_4 fingerprints was used.

Table 2

Correlations among Ligand-Set Similarities Using Seven Different Fingerprints to Represent the Ligands[a]

	Unity	MDL Keys	ECFP_4	FCFP_4	CATS	FEPOPS
Daylight	0.940	0.783	0.896	0.902	0.585	0.545
Unity		0.790	0.892	0.894	0.586	0.537
MDL Keys			0.829	0.811	0.593	0.530
ECFP_4				0.938	0.622	0.596
FCFP_4		average = 0.705 ± 0.171			0.599	0.578
CATS		randomized = 0.002 ± 0.011				0.365

[a]Spearman rank-order correlation coefficients were calculated for the 249 × 249 square matrices that relate ligand sets. MDDR ligand sets were compared using SEA.

Table 3

Correlations between Targets Related by Sequence and hose Related by Ligand-Set Similarity[a]

	PSI_Blast[b] vs SEA	PSI_Blast[b] vs BAY	SEA[c] vs BAY
Daylight	−0.168	−0.136	0.387
Unity	−0.178	−0.153	0.379
MDL Keys	−0.228	−0.156	0.313
ECFP_4	−0.155	−0.074	0.401
FCFP_4	−0.172	−0.050	0.428
CATS	−0.027	−0.052	0.300
FEPOPS	−0.030		
average	−0.137	−0.104	0.368

[a] Spearman rank-order correlation coefficients were calculated etween sequence and ligand-set similarity matrices.

[b] A total of 93 MDDR targets for which specific sequences could be assigned.

[c] A total of 249 MDDR targets from the Schuffenhauer ontology.[11]

Table 4

Percentage of Common Edges between Threshold Networks (E-value $\leq 10^{-10}$) Calculated Using Different Chemical Fingerprints (Using SEA and the 249 MDDR Targets)

	Unity	MDL Keys	ECFP_4	FCFP_4	CATS	FEPOPS
Daylight	87.1	72.5	77.9	77.2	40.1	27.4
Unity		69.9	77.3	78.1	38.9	29.4
MDL Keys			75.4	80.4	51.9	28.6
ECFP_4				90.4	41.3	35.9
FCFP_4		average = 56.4% ± 22.8			42.0	35.3
CATS		randomized = 5.6% ± 1.1				27.4

Table 5

Percentage of Common Edges between Threshold Networks (E-value $\leq 10^{-10}$) Calculated Using Different Fingerprints and Methods of Calculating Sequence and Target Information: PSI-Blast, SEA, and the Bayesian Method[a]

	PSI_Blast vs SEA	PSI_Blast vs BAY	SEA vs BAY
Daylight	24.0	27.9	46.1
Unity	23.3	27.9	39.9
MDL Keys	28.2	33.0	30.4
ECFP_4	26.7	28.9	54.2
FCFP_4	28.4	24.3	44.0
CATS	17.5	28.3	26.4
FEPOPS	16.0		
average	23.5%	28.4%	40.2%

[a]The MDDR sets were again used.

Table 6

Self-Recognition by Ligand Sets in a 10-Fold Validation Calculation[a]

	SEA		Bayesian Method	
	mean	SD	mean	SD
Daylight	0.975	0.014	0.945	0.029
Unity	0.973	0.014	0.933	0.036
MDL Keys	0.965	0.018	0.904	0.049
ECFP_4	0.987	0.010	0.979	0.015
FCFP_4	0.984	0.011	0.978	0.015
CATS	0.926	0.035	0.862	0.064

[a] Average ROC-AUC values using the different descriptors for the SEA and the Bayesian methods. Higher values indicated better ability of the test subsets to recognize the same ligand family in the training sets.