Chapter 8

Off-Target Networks Derived from Ligand Set Similarity

Michael J. Keiser and Jérôme Hert

Summary

Chemically similar drugs often bind biologically diverse protein targets, and proteins with similar sequences or structures do not always recognize the same ligands. How can we uncover the pharmacological relationships among proteins, when drugs may bind them in defiance of bioinformatic criteria? Here we consider a technique that quantitatively relates proteins based on the chemical similarity of their ligands. Starting with tens of thousands of ligands organized into sets for hundreds of drug targets, we calculated the similarity among sets using ligand topology. We developed a statistical model to rank the resulting scores, which were then expressed in minimum spanning trees. We have shown that biologically sensible groups of targets emerged from these maps, as well as experimentally validated predictions of drug off-target effects.

Key words: SEA, Expectation value, Target network, Polypharmacology, Off-targets

1. Introduction

How similar are two proteins? Typically, proteins are compared using bioinformatics approaches based on sequence or structure. While these methods quantify historical protein divergence, drugs and other small molecules often bind to targets that are unrelated from an evolutionary standpoint *(1, 2)*. For example, the enzymes thymidylate synthase, dihydrofolate reductase, and glycinamide ribonucleotide formyltransferase have no substantial sequence identity or structural similarity but they all recognize folic acid derivatives and are inhibited by antifolates. Similarly, the drug methadone binds both the μ-opioid receptor, a GPCR, and the structurally unrelated *N*-methyl-D-aspartate receptor, an ion channel. Polypharmacology, the ability of chemically similar

Edgar Jacoby (ed.), *Chemogenomics*, Methods in Molecular Biology, vol. 575
DOI 10.1007/978-1-60761-274-2_8,

drugs to bind biologically diverse proteins, has inspired recent efforts to find protein relationships by means other than their sequence or structure *(3–5)*.

The Similarity Ensemble Approach (SEA) considers proteins from a chemocentric point of view, relating them through the chemical similarity of their ligands *(6)*. The idea is that similar molecules have similar biological profiles *(7)* and bind similar targets *(8, 9)*. This technique links hundreds of ligand sets–and correspondingly their protein targets–together in minimal spanning trees where biologically related proteins cluster together as an emergent property (*see* **Fig. 1**). These networks are robust *(10)* and may be used to predict off-target effects *(6)*. The similarities among ligand sets may reveal the pharmacological relationships of the targets whose actions they modulate.

How does SEA work? An overview of the different stages is available in **Fig. 2**. The similarity between two ligand sets is first approximated by summing the similarity scores of molecule pairs across the sets (*see* **Fig. 2b**). In itself, the resulting *raw score* is not a good estimate of the overall similarity of the sets, as it does not discriminate relevant similarities from random and depends on the number of ligands in each set. SEA corrects for these shortcomings via a statistically determined threshold–pairs of molecules that score below it are discarded and do not contribute to the overall set similarity. We then convert the raw score to a size-bias-free z-score using the mean and standard deviation of raw scores modeled from sets of random molecules. Finally, we express the similarity score between two sets as an *E*-value, i.e., the probability of a given z-score that high or better to be observed from random data. Small *E*-values, then, reflect relationships between ligand sets that are stronger than would be expected by random chance alone.

2. Materials

1. A reference database of chemical structures, annotated by therapeutic indication or mechanism of action. For the purpose of illustration, we used the *MDL Drug Data Report* (MDDR) *(11)* which contains 65,367 molecules organized in 249 sets (*see* **Note 1**).
2. A molecular descriptor generator to encode the structural information of the compounds. We obtained the best results with 2-dimensional fingerprints based on topology of the molecules such as the 2,048-bit default Daylight or 1,024-bit folded Scitegic ECFP_4 descriptors (*see* **Note 2**).
3. A similarity coefficient, such as the Tanimoto coefficient (*see* **Note 3**).

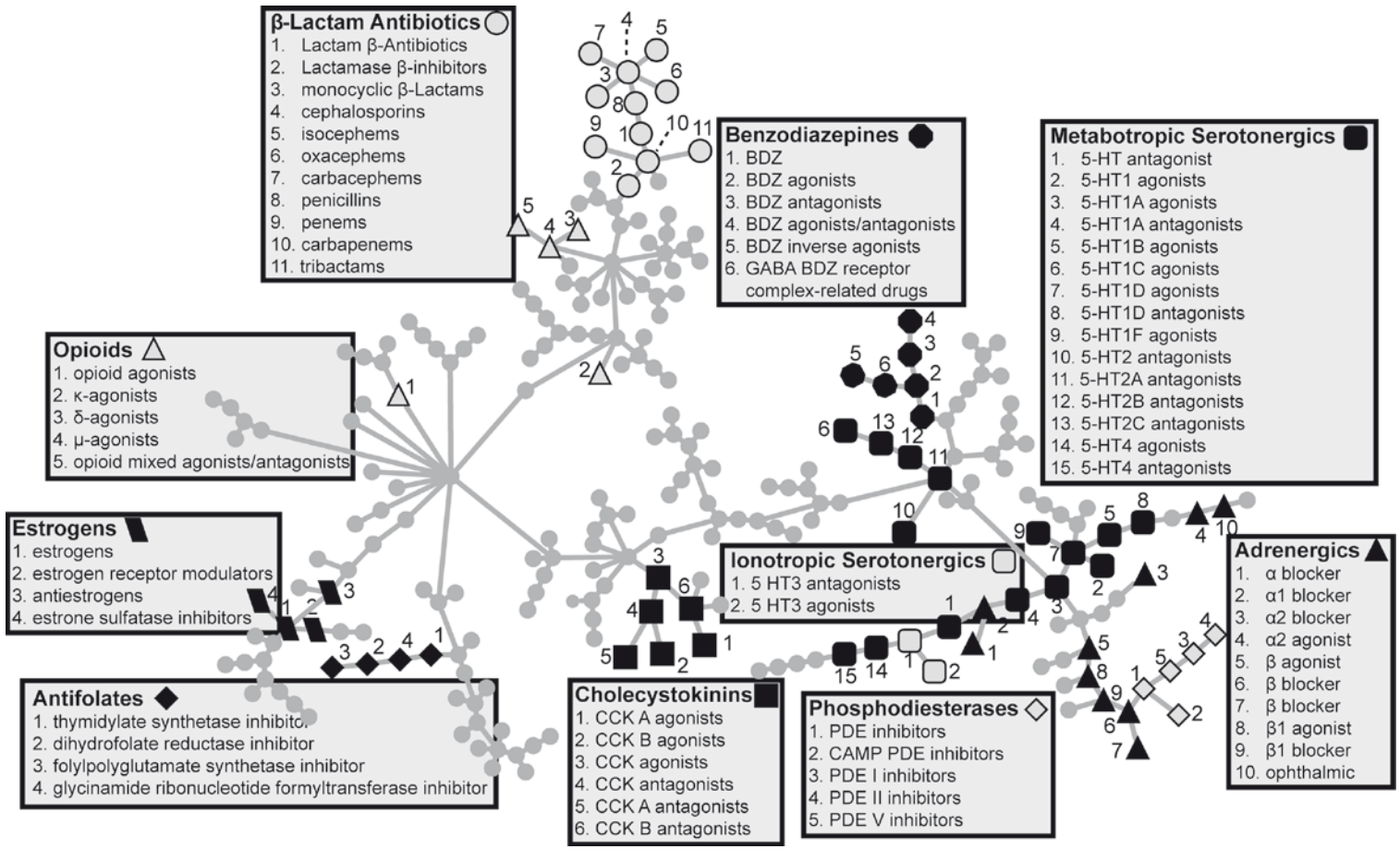


Fig. 1. Pharmacological network of the MDDR drug targets. Each vertex represents a ligand set and hence a protein target. The vertices are linked together by their SEA *E*-values (*edges*) and organized into a minimum spanning tree. Several protein families are highlighted to emphasize the natural clustering that emerges.

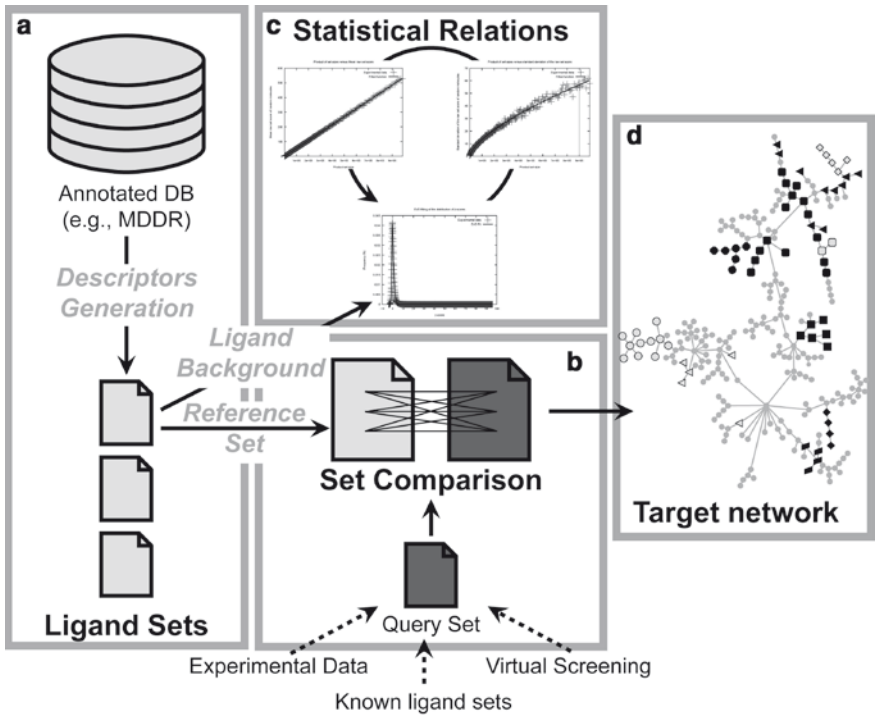


Fig. 2. Method overview: Ligand sets derived from existing databases (**a**) are used in set-wise comparisons (**b**) against a query set, the result of which is quantified by the statistical model inferred from that reference database (**c**). The generated probabilistic data can be used to construct chemical mappings of the ligand sets and correspondingly the biological targets (**d**).

4. Calculating the parameters of the reference database requires a fitter program to calculate nonlinear regressions (*see* **Note 4**).
5. Building a similarity network requires a graph visualization program, such as Cytoscape *(12)*.

3. Methods

SEA quantifies the similarity among sets of compounds which may be organized by the targets they modulate, the therapeutic indications they address, their activity in a high-throughput screening campaign, or a variety of other criteria. So far, we have focused on sets organized by targets, but SEA can be used with other annotations.

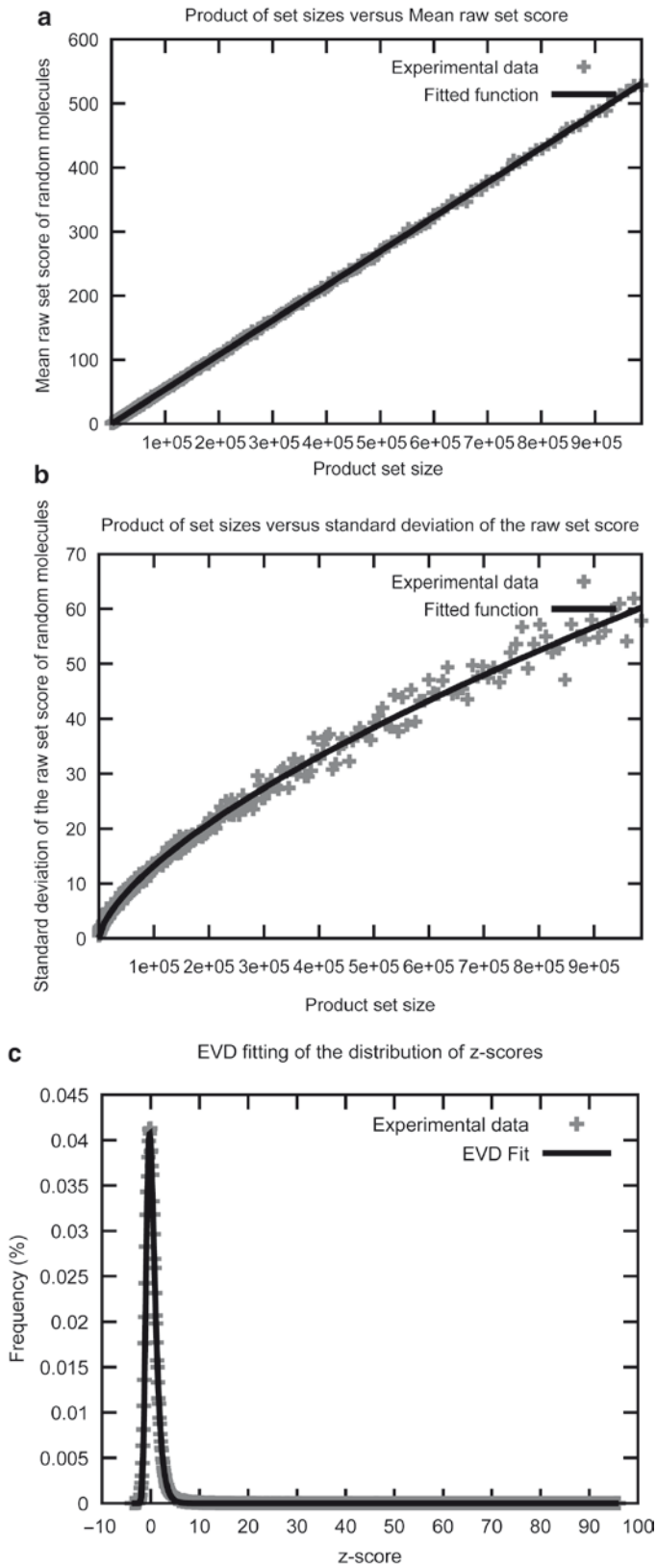
Before comparing any sets with SEA, the parameters of the background database–generally the one containing the sets one wishes to compare to–need to be calculated. While this step is computationally intensive, it is only required once for a given database, molecular descriptor, and similarity coefficient (*see* **Subheading 3.1**). Once the optimal threshold t_i and the formulae

of the mean y_μ and standard deviation y_σ as a function of the product of the sets' sizes ($|a| \times |b|$) have been determined, SEA can be applied to quantify set similarity (*see* **Subheading 3.2**).

3.1. Calculating the Parameters of the Reference Database

In this section, we generate thousands of randomly populated pairs of ligand sets and determine the uncorrected similarity among them. We use these "random" similarities to build an empirical model of background chemical similarity. The particular choice of chemical database will determine the type of background: KEGG molecules will yield a metabolic background, whereas ZINC molecules will produce drug- or lead-like backgrounds (depending on the exact subset used). It is preferable to choose as large a database as possible; those in excess of 100,000 molecules are often ideal.

1. Choose minimum and maximum set sizes $s_{\min}$ and $s_{\max}$ for sampling, such that they will be representative of molecule sets annotated in the database (*see* **Note 5**).
2. Sample at least 1,000 integers s_i from the range ($s_{\min} \times s_{\min}$) to ($s_{\max} \times s_{\max}$) (*see* **Note 6**).
3. For each product of sets' sizes s_i, calculate all its integer factors f_i, such that $s_{\min} \leq f_i \leq s_{\max}$.
4. For each s_i, choose 30 of its f_i at random and construct two sets a and b, consisting of f_i and s_i/f_i molecules, respectively, randomly selected from the background molecule database (*see* **Note 7**).
5. For each pair of sets a and b, calculate standard chemical similarities $c_{a,b}$ for each pair of ligands across the sets using your previously chosen chemical similarity descriptor and coefficient.
6. For t_i, where $0 \leq t_i < 1$ with step size 0.01, calculate a "raw score" $r_{a,b}(t_i)$ equal to the sum of all $c_{a,b}$ where $c_{a,b} > t_i$. Store all calculated $r_{a,b}(t_i)$, along with the sizes of sets a and b (*see* **Note 8**).
7. For each t_i, plot all $r_{a,b}(t_i)$ scores vs. the product of set sizes a and b, e.g., plot all points ($|a| \times |b|$, $r_{a,b}$). There should be 100 plots (*see* **Note 9**), each corresponding to a particular choice of t_i.
8. For each plot, use the nonlinear fitter to determine the mean expected random chemical similarity (*see* **Fig. 2c** and **Fig. 3a**). Typically, an equation of the formula $y_\mu = mx^n + p$ will be appropriate (*see* **Note 10**).
9. For each plot, bin the data by the x-axis values, such that each bin ideally has no fewer than five data points. Given the previously fitted y_μ, calculate the standard deviation of each bin with Laplacian correction, and fit the resulting standard deviation points nonlinearly (*see* **Fig. 2c** and **Fig. 3b**). Again, $y_\sigma = qx^r + s$ will typically be appropriate.


a
Product of set sizes versus Mean raw set score
Experimental data
Fitted function
Mean raw set score of random molecules
Product set size
b
Product of set sizes versus standard deviation of the raw set score
Experimental data
Fitted function
Standard deviation of the raw set score of random molecules
Product set size
c
EVD fitting of the distribution of z-scores
Experimental data
EVD Fit
Frequency (%)
z-score

10. For each plot, use the fitted y_μ and y_σ to transform all original points ($|a| \times |b|$, $r_{a,b}$) to their z-scores $z_{a,b} = (r_{a,b} - y_\mu(|a| \times |b|))/y_\sigma(|a| \times |b|)$ (*see* **Note 11**). Construct a histogram of these z-scores.
11. For each histogram, nonlinearly fit the data to Gaussian and extreme value type I (EVD) distributions (*see* **Note 12**, **Fig. 2c**, and **Fig. 3c**).
12. Based on goodness of fit, such as each fit's observed-vs.-expected χ^2 value, select the threshold choice t_i, such that the histogram best fits an EVD instead of a Gaussian distribution (*see* **Note 13**).
13. Record the chosen t_i and that t_i's formulae for y_μ and y_σ. These values comprise the random background model. All other plots, histograms, and formulae may be discarded at this point.

3.2. Calculating Set-Wise Similarity Ensembles

To calculate the set-wise similarity among sets of ligands, we reuse much of the machinery developed to calculate background models and extend it to calculate *E*-values. By exhaustively comparing all pairs of sets across two collections (databases), we can then rank the top hits for any particular ligand set.

In practice, a ligand set should not comprise fewer than ten ligands, unless you intend to compare it against large sets only. For instance, it would not be statistically reliable to compare two sets of five ligands each, but a set of five ligands compared against a set of thirty should be acceptable. Although the particular choice of set size should depend on the diversity of ligands within a set, a good rule of thumb is to build sets such that the product of the set sizes will be no less than 100 (e.g., the product of set sizes is 25 for the five-by-five case, and 150 for the five-by-thirty case mentioned earlier).

1. To calculate similarity ensembles, choose two collections of sets C_a and C_b to compare (*see* **Note 14**).
2. For each set a and b from collections C_a and C_b, respectively, calculate $r_{a,b}(t_i)$ as previously described using only the optimal threshold t_i from the background model. Be sure to use the actual molecule structures annotated for each set.
3. Transform each $r_{a,b}(t_i)$ to z-score $z_{a,b}$ as described in **Subheading 3.1, step 10.**

Fig. 3. Statistical models: **a** Correlation between the product of sets' sizes and the mean of the raw score. The fitted function typically corresponds to an equation of the formula $y_\mu = mx^n + p$ with $n = 1$. **b** Correlation between the product of sets' sizes and the standard deviation of the raw score. The fitted function typically corresponds to an equation of the formula $y_\sigma = qx^r + s$, with $0.6 < r < 0.7$. **c** Distribution of the z-scores obtained from random data using ECFP_4 fingerprints, with a similarity score threshold (t_i) of 0.57 and fitted to an extreme value distribution.

4. Transform each $z = z_{a,b}$ to p-value $P(Z > z) = 1 - \exp(-e^{-z\pi/\mathrm{sqrt}(6)-\Gamma'(1)})$, where $\Gamma'(1)$ is the Euler–Mascheroni constant ($\approx$0.577215665) (*see* **Note 15**).
5. Optionally, the p-value may be transformed to a BLAST-like E-value by calculating $E(z) = P(Z > z) \times n_{db}$, where n_{db} = the number of set-vs.-set comparisons made when comparing all sets from collection C_a against all sets from collection C_b. Typically, $n_{db} = |C_a| \times |C_b|$.
6. For each set a, rank all sets b_i from C_b by their E-value, where values approaching zero are the best scores (*see* **Note 16**).

3.3. Building a Similarity Network

A similarity network is a graphical view of the E-value relationships among all ligand sets in a particular database (*see* **Note 17**). If these ligand sets represent particular drug targets, for instance, it is a visualization of the significant chemical similarity present among these targets (*see* **Fig. 1**).

1. Calculate the similarity ensemble E-values between all sets a_i and a_j from C_a versus itself (*see* **Note 18**), as previously described.
2. The resulting matrix of E-values defines a strongly connected graph, where each node corresponds to a molecule set and each edge to the E-value between two sets (*see* **Note 19**).
3. We use Kruskal's algorithm *(13)* to construct a minimum spanning tree (MST):
   a. Create a set S_{tree} that initially contains all individual nodes, unconnected. We refer to elements of S_{tree} as "trees."
   b. Create a set S_e that contains all possible edges e_i (E-values).
   c. While S_e is not empty
      i. Remove the minimum-weighted (best) edge e_{min} from S_e.
      ii. If e_{min} connects two existing trees t_a and t_b in S_{tree}.
         1. Remove t_a and t_b from S_{tree}, connect them into a single new tree t_{ab} using e_{min}, and add t_{ab} back into S_{tree}.
      iii. Else, discard e_{min}.
   d. When the algorithm finishes, S_{tree} will contain only one tree, which is the graph's MST.

4. Notes

1. Examples of other freely or commercially available annotated chemogenomics databases include *WOMBAT*, *KEGG*, and *DrugBank*. Note, however, that SEA can be used with any kind of annotation and is not limited to ligand-target association.

2. For efficiency, the steps in **Methods** will be faster if fingerprints are precalculated and stored for each molecule.
3. While it is not technically necessary, we assume that the similarity coefficient is normalized from 0.0 to 1.0. If not, choose appropriate bounds for the range of t_i thresholds discussed in **Subheading 3.1, step 6**.
4. The open-source Scientific Python (SciPy) package *(14)* provides a least-squares optimizer that can be used for fitting nonlinear regressions.
5. If you are unsure of appropriate values, use $s_{min} = 10$ and $s_{max} = 300$.
6. More than 1,000 points may be sampled, but in our experience this does not yield a substantial difference in the final model.
7. If there are fewer than 30 distinct factors f_i for a particular integer s_i, randomly sample from the available f_i 30 times. Sampling more than 30 points is also acceptable, depending on the diversity of the background database and computational resources.
8. These raw scores are the "random" similarities that form the background model. Besides the choice of similarity descriptor and coefficient, the threshold t_i is the only settable major SEA parameter. By sampling across the range of t_i choices, we will be able to determine an optimal choice of t_i in later steps.
9. For the steps plotting these data (and later, the histograms), you need not actually draw out the full plots. All that is strictly necessary is that your data are formatted appropriately for input into your chosen fitter. Using SciPy, for instance, it is enough to store these data points in internal arrays.
10. In our experience, the mean raw score fit y_μ has always been linear.
11. The z-score is the number of standard deviations by which a particular raw score exceeds the expected mean.
12. You may use the "norm" and "gumbel_r" SciPy data types for Gaussian and extreme value type I distributions, respectively.
13. There is currently no formal justification for choosing the t_i threshold, but this approach is consistent and enriches for a BLAST-like background probability distribution. Some experiments also suggest that this choice is reasonable, as thresholds derived from retrospective cross-fold analysis are identical or close to the threshold t_i (unpublished).
14. One such collection may be built from the annotated molecular structure database. The second may be the exact same collection (for symmetric comparisons), or derived from a different database of annotated molecules.

15. This formula converts EVD z-scores to their *p*-values, where the *p*-value expresses the probability of finding a z-score that strong or better, by random chance alone.
16. An *E*-value of 1 or higher is not statistically significant. The similarity between two sets becomes significant when it is at least one order of magnitude smaller than random chance alone, i.e., 10^{-1}. Sets that are highly similar have *E*-values «10^{-50}, although there is no single cutoff for *E*-value significance. The SEA Search tool at http://sea.docking.org may also be used check the accuracy of the z-scores and *E*-values calculated in **Subheading 3.2**.
17. While there are many appropriate graph-theoretic approaches, we have chosen an MST. An MST is a selection over all graph edges (*E*-values) such that the resulting tree links all nodes (ligand sets) at lowest "cost" to the network as a whole. For example, an edge with an *E*-value approaching zero has a lower cost to the tree than one with an *E*-value of 1. The resulting MST will preferentially include only those edges with the smallest *E*-values. It may be interpreted as a simplified view of higher-dimensional chemical similarity space.
18. These instructions apply only to symmetric collection comparisons, e.g., $C_a = C_b$.
19. You may either (a) use Cytoscape to filter out all edges above an E-value threshold of your choice, or (b) construct a global MST.

Acknowledgments

M.J.K is supported by a National Science Foundation graduate fellowship. J.H. is supported by the sixth Framework Program of the European Commission. We are grateful to MDL Information Systems Inc. for the MDDR database; Daylight Chemical Information Systems Inc.; and OpenEye Scientific Software for software support. We thank John J. Irwin for reading the manuscript and Brian K. Shoichet for mentoring.

References

1. Roth, B., Sheffler, D., and Kroeze, W. (2004) Magic shotguns versus magic bullets: Selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353–359.
2. Paolini, G., Shapland, R., van Hoorn, W., Mason, J., and Hopkins, A. (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805–815.
3. Nidhi, Glick, M., Davies, J.W., and Jenkins, J.L. (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **46**, 1124–1133.

4. Izrailev, S., and Farnum, M.A. (2004) Enzyme classification by ligand binding. *Proteins* **57**, 711–724.
5. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., and Bork, P. (2008) Drug target identification using side-effect similarity. *Science* **321**, 263–266.
6. Keiser, M., Roth, B., Armbruster, B., Ernsberger, P., Irwin, J., and Shoichet, B. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197–206.
7. Johnson, M.A., and Maggiora, G.M. (1990) Concepts and Applications of Molecular Similarity. John Wiley, New York.
8. Frye, S.V. (1999) Structure–activity relationship homology (SARAH): A conceptual framework for drug discovery in the genomic era. *Chem. Biol.* **6**, R3–R7.
9. Jacoby, E., Schuffenhauer, A., and Floersheim, P. (2003) Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect.* **16**, 93–102.
10. Hert, J., Keiser, M., Irwin, J., Oprea, T., and Shoichet, B. Quantifying the relationships among drug classes. *J. Chem. Inf. Model* **48**, 755–765.
11. The MDL Drug Data Report Database is available from MDL Information Systems, Inc. (Accessed at http://www.mdl.com.)
12. Shannon, P., Markielm, A., Ozierm, O., et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.
13. Kruskal, J. (1956) On the shortest spanning subtree and the traveling salesman problem. *Proc. Amer. Math. Soc.* **7**, 48–50.
14. SciPy: Open Source Scientific Tools for Python. (2001) (Accessed at http://www.scipy.org.).