Pharmacological networks derived from the similarity of ligand-sets

Jérôme Hert and Brian K. Shoichet

University of California San Francisco

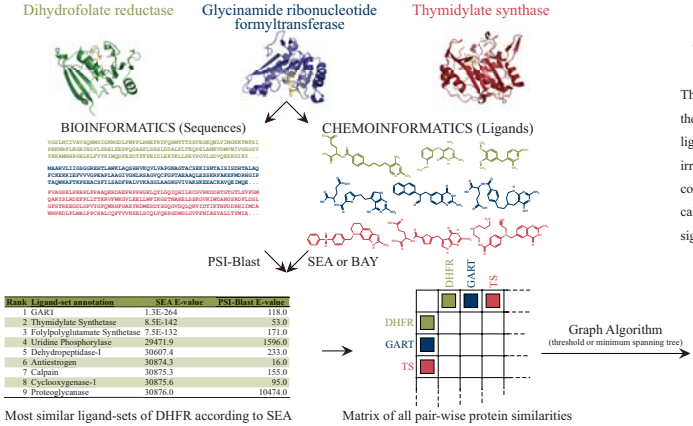
Department of Pharmaceutical Chemistry, University of California San Francisco, 1700 4th St, San Francisco, CA, 94158. Email: hert [at] cgl [dot] ucsf [dot] edu

SUMMARY

Pharmacological networks chart the relationships between drug targets; each protein is represented by a vertex and each edge denotes a link between two proteins. Such networks are usually derived from biological information where protein sequences or structures are compared using bioinformatics tools. An alternative and complementary chemo-centric approach consists of using the similarity between the ligands as a proxy to the pharmacological relationship between two proteins. Here we explore the difference between bioinformatics and chemoinformatics-based networks and investigate how the choice of a particular chemical information representation affects their robustness. We carried out extensive comparisons of sequence-based and ligand-set-based networks using the MDL Drug Data Report (MDDR) and the World Of Molecular BioActiviTy (WOMBAT) databases and seven different ways of representing molecules (Daylight, Unity, MDL Keys, ECFP_4, FCFP_4, CATS and FEPOPS). Three key points emerged from these comparisons:

1. The bioinformatics-based and chemoinformatics-based networks are different.
2. The chemoinformatics-based networks are robust.
3. The chemoinformatics-based networks are pharmacologically relevant and lead to testable predictions.

FROM SEQUENCES AND LIGANDS TO NETWORKS

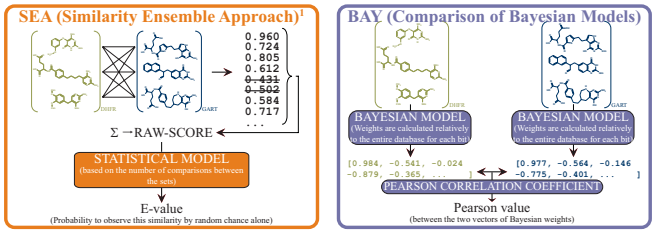


Rank	Ligand-set annotation	SEA E-value	PSI-Blast E-value
1	GART	1.3E-264	118.0
2	Thymidylate Synthetase	8.5E-142	53.0
3	Folylpolyglutamate Synthetase	7.5E-132	171.0
4	Uridine Phosphorylase	29471.9	1596.0
5	Dehydropeptidase-1	30607.4	233.0
6	Antiestrogen	30874.3	16.0
7	Calpain	30875.3	155.0
8	Cyclooxygenase-1	30875.6	95.0
9	Proteoglycanase	30876.0	10474.0

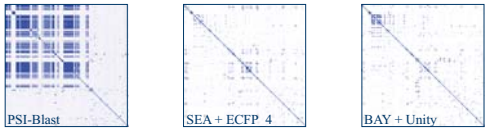
Most similar ligand-sets of DHFR according to SEA

Matrix of all pair-wise protein similarities

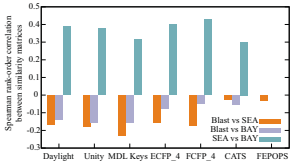
ESTIMATING THE SIMILARITY BETWEEN LIGAND-SET



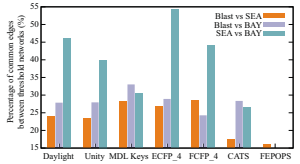
BIOINFORMATICS AND CHEMOINFORMATICS NETWORKS ARE DIFFERENT



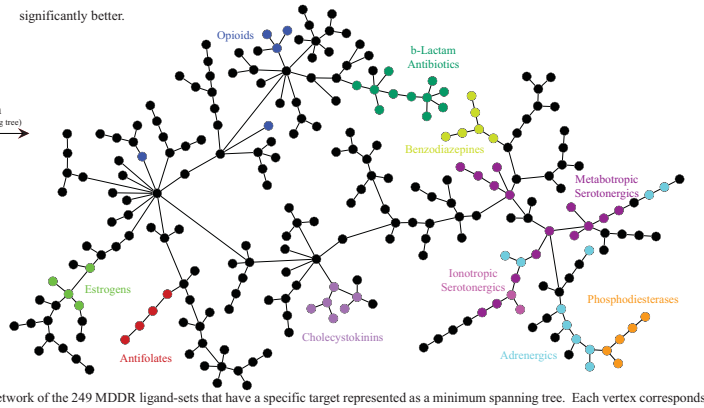
Similarity matrices of 193 proteins in the MDDR represented as heatmaps where dark blue squares correspond to high similarity (E-value < 10^{-5}) and white squares to low similarity (E-value > 10^{5}). Many proteins are related when sequences are compared using PSI-Blast (left) while substantially less proteins are highly similar when ligand-sets are compared with SEA (middle) or BAY (right).



The Spearman rank-order correlation coefficient between the sequence-based similarity matrix and the ligand-set-based similarity matrices is very poor irrespective of the choice of the fingerprint. In contrast, the correlation between the ligand-set-based matrices calculated with the SEA and BAY approaches are significantly better.

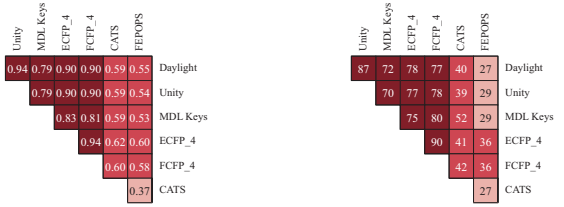


Bioinformatics and chemoinformatics threshold networks (E-value < 10^{-10}) share between 20 and 30% of their edges. With 2 out of 3 fingerprints, the percentage of common edges is significantly higher between SEA-similarities based networks and BAY-similarities based ones.



Network of the 249 MDDR ligand-sets that have a specific target represented as a minimum spanning tree. Each vertex corresponds to a ligand-set (and hence to a protein); when sets of ligands are considered similar (here using SEA with the ECFP_4 fingerprints) an edge is drawn between them. Clusters of pharmacologically related proteins appear as an emergent property of the technique; no biological information other than the ligands was used to create this network.

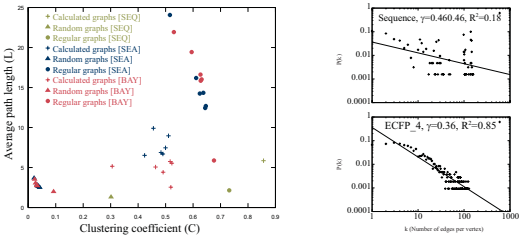
CHEMOINFORMATICS NETWORKS ARE ROBUST



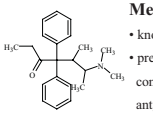
Spearman rank-order correlation coefficient between matrices calculated with SEA and the different fingerprints. The matrices are highly correlated when topology-based fingerprints are used to represent the molecules (Daylight, Unity, MDL Keys, ECFP_4, FCFP_4), less so when CATS or FEPOPS descriptors are used.

Percentage of common edges in the threshold networks (E-value < 10^{-10}) calculated with SEA and the different fingerprint. The five topology-based networks have more than 70% of the edges in common. This percentage decreases when the networks are calculated using the CATS and the FEPOPS descriptors.

CHEMOINFORMATICS NETWORKS ARE PHARMACOLOGICALLY RELEVANT

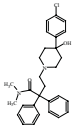


By network theory, the threshold networks based on ligand-set similarity are more natural than those based on sequence identity. Chemoinformatics networks are small-world because their clustering coefficient is higher than that of random networks while its average path length is moderately bigger than the one calculated for random networks (see left graph). Bioinformatics network do not have these properties (the clustering coefficient is too high). Chemoinformatics networks also have a distribution of number of edges per vertex similar to that of broad-scale ("truncated" scale-free) networks (see right graph). The main property of scale-free networks is to continuously grow with new edges connecting preferentially to highly connected vertices.



Methadone[1]

- known μ-Opioid
- predicted (and experimentally confirmed) Muscarinic M3 antgonist



Loperamide[1]

- known μ-Opioid
- predicted (and experimentally confirmed) Neurokinin NK2 antagonist

REFERENCES

1. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Iriwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by their ligands. Nature Biotechnology 2007, 25, 197-206.