Knowledge Discovery in Pharmaceutical Drug Transport using Emerging Graph Patterns

Guillaume Poezevara[1], Alban Lepailleur[1], Ronan Bureau[1]
Lancelot Lemoine[1], Bertrand Cuissart[2], Bruno Crémilleux[2]
Sylvain Lozano[3], Alban Arrault[3] and Philippe Vayer[3]

[1]Centre d' Études et de Recherche sur le Médicament de Normandie (CERMN UPRES EA-4258 FR CNRS INC3M), Université de Caen Basse-Normandie, France
[2]Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen (GREYC CNRS UMR 6072), Université de Caen Basse-Normandie, France
[3]Technologie Servier, 27 rue Eugène Vignat, 45000 Orléans, France

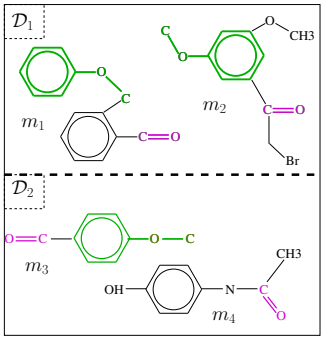
email: [1,2]{firstname.lastname}@unicaen.fr, [3]{firstname.lastname}@fr.netgrs.com

Introduction

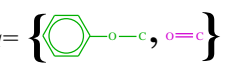
The clinical importance of efflux and uptake transporters in drug disposition is widely acknowledged, and membrane transporter anomalies are the basis for certain clinical disorders. Particularly, the Organic Cation Transporter 2 (OCT_2) is a renal transporter that plays a key role in disposition and renal clearance of most currently prescribed drugs. Computational models could predict undesirable effects that are based on drug transporter interactions, and statistical models like quantitative structure-activity relationships and pharmacophores have been proposed. Several works have been conducted in chemoinformatics to extract the frequent substructures from a dataset of graphs and to link them to a biological activity. However, in the recent years studies have given more attention to the discovery of patterns that are significant, like the emerging patterns [Auer and Bajorath, 2006, Sherhod et al., 2012], than simply frequent. For example, we showed that the Emerging Graph Patterns (EGPs) are very useful in predictive toxicology [Poezevara et al., 2011] as EGPs can be linked to toxicophores.

EGPs as potential OCT_2 inhibitors

$\mathcal{D}$: a chemical dataset partitioned into $\mathcal{D}_1$ and $\mathcal{D}_2$.



- a *fragment* denominates a connected molecular substructure,
- a *graph pattern* denominates a set of fragments.

Example: $q = \{$  $\}$

- *extension*: $ex_{\mathcal{D}}(q) = |\{m \in \mathcal{D} \mid q \sqsubseteq m\}| = 3$
- *frequency*: $fr_{\mathcal{D}}(q) = \frac{ex_{\mathcal{D}}(q)}{|\mathcal{D}|} = \frac{3}{4} = 0.75$
- *growth rate*: $\rho_{\mathcal{D}_1}(q) = \frac{fr_{\mathcal{D}_1}(q)}{fr_{\mathcal{D}_2}(q)} = \frac{1}{0.5} = 2$

EGP : a graph pattern is an *Emerging Graph Pattern (EGP)* from $\mathcal{D}_2$ to $\mathcal{D}_1$ iff:

i) its frequency in $\mathcal{D}$ exceeds a minimum threshold f_{min},
ii) its growth rate from $\mathcal{D}_2$ to $\mathcal{D}_1$ exceeds a minimum threshold ρ.
iii) it is not included in any other EGP with the same extension in $\mathcal{D}$,
iv) it doesn't contained any fragment which is a substructure of another fragment in this pattern.

When the dataset is partitioned according to the OCT_2 inhibition, EGPs may correspond to OCT_2 inhibitors.

Filtering the EGPs with the Fisher's exact test

In order to reduce the set of EGPs while keeping the contrast information, each pattern q is studied according to its partitioning into $\mathcal{D}_1$ and $\mathcal{D}_2$. The Fisher's exact test is used to keep only patterns which are statistically significant.

Hypothesis H_0: Being inhibitor of OCT_2 is independent on contain q.

Molecules	Inhibitors	Non-inhibitors
Contain q	$ex_{\mathcal{D}_1}(q)$	$ex_{\mathcal{D}_2}(q)$
Don't contain q	$\lvert\mathcal{D}_1\rvert - ex_{\mathcal{D}_1}(q)$	$\lvert\mathcal{D}_2\rvert - ex_{\mathcal{D}_2}(q)$

If the corresponding *pvalue* is under a fixed error threshold (α), q is dependent on the OCT_2 inhibition, thus q is kept. Otherwise it is pruned.

Materials and methods

Dataset:
- Retrieved from [Kido et al., 2011] (244 inhibitors and 663 non-inhibitors),
- extraction of statistically significant EGPs for OCT_2 inhibition.

Protocol:
- A five folds cross-validation procedure ($f_{min} = 5\%$ and $\alpha = 1\%$),
- *a molecule is classified as inhibitor if it contains at least one EGP*,
- Results are given in average over the five folds.

Quantitative assessment of the EGPs

Extraction measurements: (i) the number of fragments and patterns (nb) (ii) the average number of atoms in the fragments or the average number of fragments in the patterns ($avgs$) and (iii) the average growth rate of fragments and patterns ($avg\rho$).

Classification measurements: Increasing the minimum growth rate allows to build a ROC Curve to study the error rate on false inhibitor classification in function of the success rate on true inhibitor classification. The deduced Area Under ROC Curve is a measure of the quality of the classifier on the learning set (auc_L) and on the testing set (auc_T).

Fisher Test	Fragments					Patterns				
	nb	$avgs$	$avg\rho$	auc_L	auc_T	nb	$avgs$	$avg\rho$	auc_L	auc_T
No	$4.35e^3$	9.2	1.52	0.766	0.702	$9.57e^5$	9.02	1.58	0.791	0.714
Yes	$1.28e^3$	10	2.02	0.763	0.708	$1.15e^5$	9.63	2.23	0.809	0.723

The Fisher exact test divides by 4 the number of fragments and by 9 the number of patterns. The resulting EGPs have a better average growth rate: it is respectively multiplied by 1.32 and 1.41 for fragments and patterns. As the results in a classification context are almost the same, we can deduce that the Fisher exact test selects the most significant EGPs.

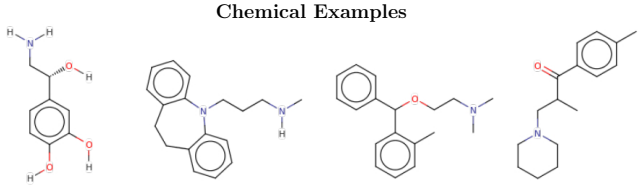
Expert analysis

An analysis of substructure fragments over-represented in inhibitors particularly indicated the importance of positive charge for ligand binding to OCT_2. A first illustration deals with the impact of the category of the amine.

Emerging Fragments

$\rho = 0.6$ $\rho = 1.19$ $\rho = 2.76$ $\rho = 3.91$

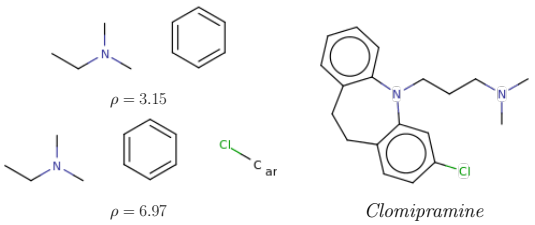
Chemical Examples



Norepinephrine *Desipramine* *Orphenadrine* *Tolperisone*

The corresponding discovered patterns show a straight relation between the number of substituents on the nitrogen and the inhibitory potential of the compound: the growth-rate value increases from primary amine to teriary amine. Besides a bulkier substituent on the amine seems to also increase the inhibitory potential. The study also emphasized the importance of hydrophobicity as a significant determinant in the binding of inhibitors to OCT_2.

Emerging Graph Patterns **Chemical Example**



$\rho = 3.15$ $\rho = 6.97$ *Clomipramine*

The strength of our method corresponds to the discovery of combinations of fragments. For example, a tertiary amine alone has a growth-rate value of 2.76. Associated with a benzene ring, the growth-rate increases to 3.15 and to 6.97 by adding a chlorinated aromatic ring.

Conclusion

In this work, we proved the effectiveness of using the information provided by the occurrences of the most significant EGPs in a context of OCT_2 inhibition. The fisher exact test can be used to reduce the number of EGPs while keeping the most significant patterns. We also showed that the evolution of the growth rate values can give new keys to understand the influence of the environment around a chemical function. These results strongly advocate the potential of this approach to predict pharmaceutical drug transport.

References

[Auer and Bajorath, 2006] Auer, J. and Bajorath, J. (2006). Emerging chemical patterns: A new methodology for molecular classification and compound selection. *Journal of Chemical Information and Modeling*, 46(6):2502–2514.

[Kido et al., 2011] Kido, Y., Matsson, P., and Giacomini, K. M. (2011). Profiling of a prescription drug library for potential renal drug–drug interactions mediated by the organic cation transporter 2. *Journal of Medicinal Chemistry*, 54(13):4548–4558.

[Poezevara et al., 2011] Poezevara, G., Cuissart, B., and Crémilleux, B. (2011). Extracting and summarizing the frequent emerging graph patterns from a dataset of graphs. *J. Intell. Inf. Syst.*, 37(3):333–353.

[Sherhod et al., 2012] Sherhod, R., Gillet, V. J., Judson, P. N., and Vessey, J. D. (2012). Automating knowledge discovery for toxicity prediction using jumping emerging pattern mining. *Journal of Chemical Information and Modeling*, 52(11):3074–3087.