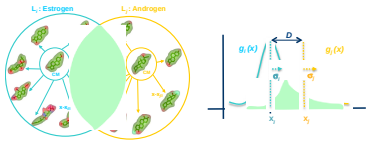
Gaussian ensemble screening (GES): A new approach to polypharmacology and virtual screening

Violeta I. Pérez-Nueno, Vishwesh Venkatraman, Lazaros Mavridis, David W. Ritchie
Orpailleur Team, INRIA Nancy - Grand Est

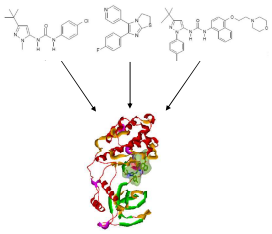


LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications),
INRIA Nancy – Grand Est, 615 rue du Jardin Botanique, 54506 Vandoeuvre-lès-Nancy, France

Polypharmacology

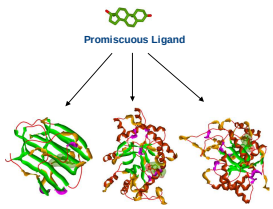
Polypharmacology (Drug selectivity)

Multiple drugs bind to a given target (promiscuous targets)



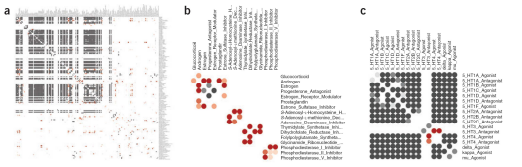
Promiscuous Target

A given drug binds to more than one target (promiscuous ligands)



Previous work

Relate receptors to each other quantitatively based on the similarity in the:



Ligand space (chemical fingerprints)

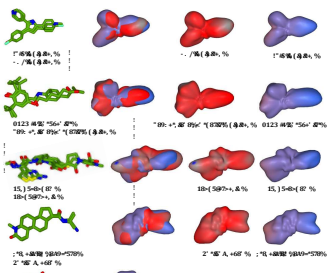
Sequence space

Binding pocket space (pharmacophoric descriptors)

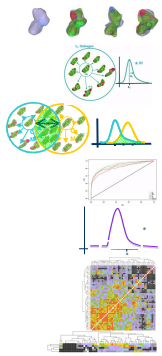
•Keiser *et al. Nature Biotechnol.* **2007**, *25*, 197-206. Similarity Ensemble Approach (SEA) relates proteins based on the set-wise chemical similarity among their ligands.
•Vidal & Mestres. *Mol. Inf.* **2010**, 29, 543. PHRAG, FPD, SHED molecular descriptors.
•Weskamp *et al. Proteins* **2009**, *76*, 317-330. Similarity amongst binding pockets extracted by LIGSITE algorithm.
•Milletti, F.; Vulpetti, A. *J. Chem. Inf. Model.*, **2010**, *50*, 1418–143. Binding pocket comparison using four-point pharmacophoric descriptors based on GRID.

Our approach

Gaussian Ensemble Screening (GES): 3D spherical harmonic (SH) shape-based approach which compares molecular surfaces and predicts quantitatively the relationships between drug classes very fast and efficiently.



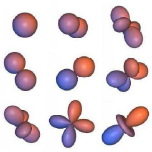
Methodology



1. Calculating SH consensus shapes and center molecules
2. Ligand set representations
3. Gaussian ligand set comparisons
4. Finding the best clustering threshold
5. Gaussian p-values
6. MDDR polypharmacology interaction matrix
7. Examples of strongly related targets

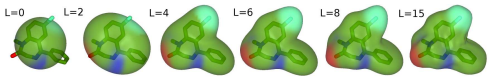
1. Calculating spherical harmonic shapes

Surface shapes are represented as radial distance expansions of the molecular surface with respect to the center of the molecule.



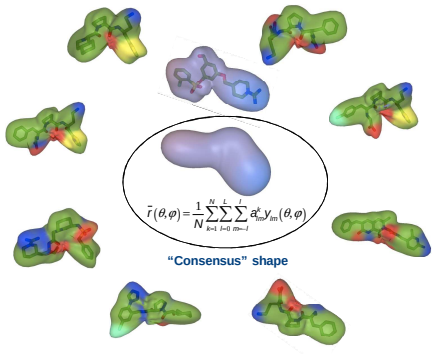
- **Real SHs:** $y_{lm}(\theta,\phi)$
- **Coefficients:** a_{lm}
- **Encode radial distances from origin as SH series…**
- **Solve coefficients by numerical integration…**

$$r(\theta,\phi) = \sum_{l=0}^{15} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta,\phi)$$



Ritchie, D.W. and Kemp, G.J.L. *J. Comp. Chem.* **1999**, 20, 383–395.

2. Calculating SH consensus shapes and center molecules



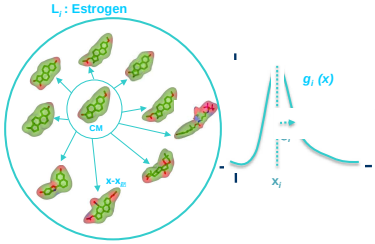
"Consensus" shape

Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

3. Ligand set representations

The idea is to represent a cluster of molecules as a Gaussian distribution with respect to a selected centre molecule (CM).

- Calculate SH molecular surfaces of each ligand in each ligand set and superpose them.
- Calculate the center molecule (CM) of the ligand set and the normalised SH distance (1-Similarity Score) between that of the CM and each cluster member.
- Assuming that these distances follow a Gaussian distribution, each cluster may be represented as a probability density function $g_i(x)$



$$g_i(x) = \sqrt{\frac{1}{2\pi\sigma_i^2}} \cdot e^{\frac{(x-x_i)^2}{2\sigma_i^2}}$$

σ : SD of the member distances

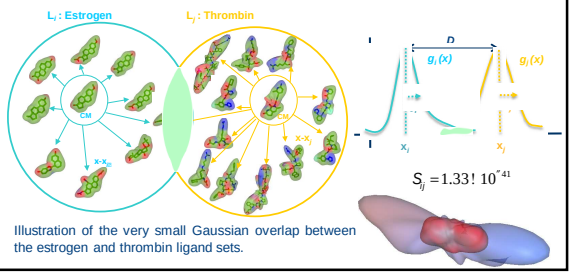
An illustration of a Gaussian ligand set cluster.

4. Gaussian ligand set comparisons

By considering the SD of the member distances as the Gaussian width of a distribution, we calculate a "distance" (*D*) between two clusters, *i* and *j*, and normalizing the distance term we can write it as a Hodgkin-like similarity score *Sij* between two distributions.

$$S_{ij} = \frac{2\int g_i(x)\cdot g_j(x)\,dx}{\int g_i(x)^2\,dx + \int g_j(x)^2\,dx} \qquad S_{ij} = \frac{2^{3/2}\cdot\left(\frac{a\cdot b}{a+b}\right)^{3/2}\cdot e^{-\left(\frac{ab}{a+b}\right)x_{ij}^2}}{\left(a^{3/2}+b^{3/2}\right)}$$

a=1/2σ_i²
b=1/2σ_j²
x_ij: distance between the CMs of clusters *i* and *j*



$S_{ij} = 1.33!\ 10^{-41}$

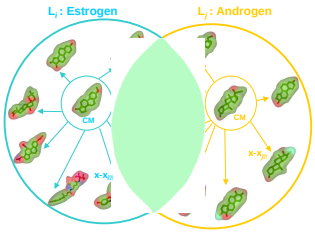
Illustration of the very small Gaussian overlap between the estrogen and thrombin ligand sets.

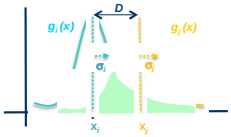
9

4. Gaussian ligand set comparisons

The similarity between drug classes can be calculated rapidly and reliably by calculating the Gaussian overlap between pairs of such clusters.

Thus, it is straight-forward to calculate all-against-all cluster comparisons. It is worth noting that our cluster similarity score depends only on the similarity of pairs of centre molecules and the SDs of their respective clusters. It does not depend on the number of members of each cluster .





$S_{ij} = 0.57$

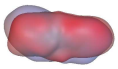
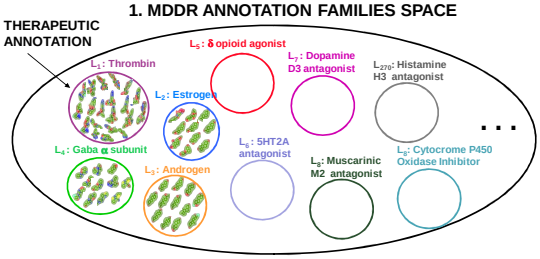


Illustration of the large Gaussian overlap between the estrogen and androgen ligand sets.

10

4. Gaussian ligand set comparisons

1. MDDR ANNOTATION FAMILIES SPACE

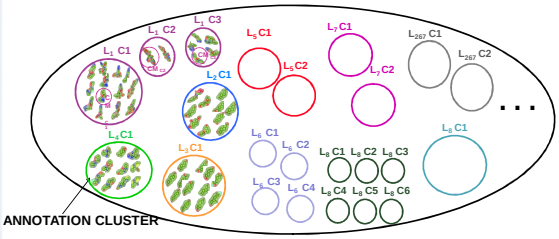


We applied the approach to 270 specific therapeutic annotations in MDDR. Ligands which share an annotation define a set of functionally related molecules which we call a "ligand set". MDDR annotations are quite general and were primarily derived from the patent literature. A given annotation may thus contain a diverse set of compounds with a wide range of affinities.

11

4. Gaussian ligand set comparisons

2. MDDR ANNOTATION SHAPE CLUSTERS

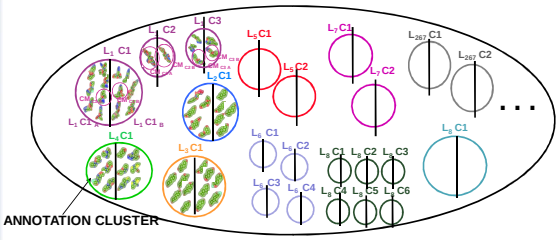


In order to eliminate outliers, we used the CAST clustering algorithm to cluster the members of each annotation using their PARAFIT Tanimoto similarity scores. We then calculated the consensus shape and the center molecule for each cluster, and we eliminated any cluster members beyond 1.5 standard deviations (SDs) from the corresponding CM.

12

5. Finding the best clustering threshold

2. MDDR ANNOTATION SHAPE CLUSTERS

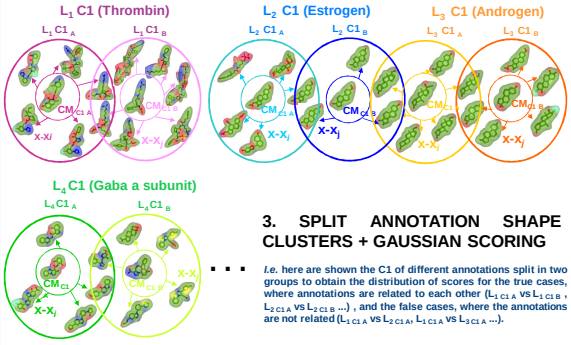


We clustered each annotation according to Parafit Shape Tanimoto using different similarity thresholds: 0.6, 0.65, 0.675, 0.7, 0.8, 0.85.

Each ligand set was randomly split into two almost equally sub-clusters, and all- vs-all clustering was performed with the aim of split and reassemble the split clusters correctly.

13

5. Finding the best clustering threshold



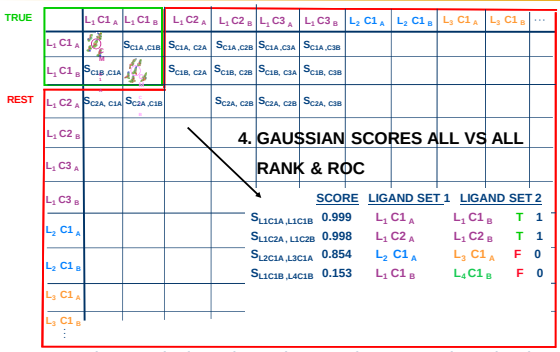
3. SPLIT ANNOTATION SHAPE CLUSTERS + GAUSSIAN SCORING

I.e. here are shown the C1 of different annotations split in two groups to obtain the distribution of scores for the true cases, where annotations are related to each other ($L_{1\,C1\,A}$ vs $L_{1\,C1\,B}$, $L_{2\,C1\,A}$ vs $L_{2\,C1\,B}$ ...) , and the false cases, where the annotations are not related ($L_{1\,C1\,A}$ vs $L_{2\,C1\,A}$, $L_{1\,C1\,A}$ vs $L_{3\,C1\,A}$ ...).

If we can split and reassemble clusters of molecules that we know they are related, then we can identify interesting relationships between clusters of molecules that we don't know they are related

14

5. Finding the best clustering threshold



4. GAUSSIAN SCORES ALL VS ALL

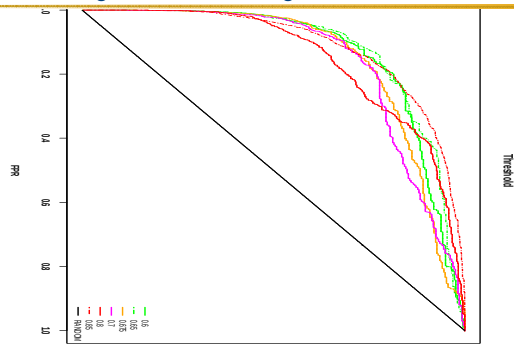
RANK & ROC

	SCORE	LIGAND SET 1	LIGAND SET 2		
$S_{L1C1A\,,L1C1B}$	0.999	$L_1\,C1_A$	$L_1\,C1_B$	T	1
$S_{L1C2A\,,L1C2B}$	0.998	$L_1\,C2_A$	$L_1\,C2_B$	T	1
$S_{L2C1A\,,L3C1A}$	0.854	$L_2\,C1_A$	$L_3\,C1_A$	F	0
$S_{L1C1B\,,L4C1B}$	0.153	$L_1\,C1_B$	$L_4\,C1_B$	F	0

We produce a matrix of Gaussian Overlap Scores for true target classes (members of the same annotation cluster) and the rest (members supposed not to be related).

15

5. Finding the best clustering threshold



ROC curves obtained for the range of similarity thresholds of 0.6, 0.65, 0.675, 0.7, 0.8, 0.85. Using a PARAFIT shape Tanimoto value of 0.65 gave the best early performance AUC(5%,10%). Hence, 0.65 was chosen as the appropriate for shape-based clustering.

16

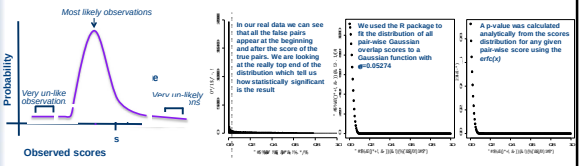
6. Gaussian p-values

In order to transform a list of cluster similarity scores into a more meaningful list of probabilities, a statistical model was developed.

Each Gaussian similarity score was transformed into a probability value, or "p-value", from the observed distribution of scores. For a Gaussian distribution, it can be shown that the probability of finding at random from the distribution some value X greater than a given value x is given by:

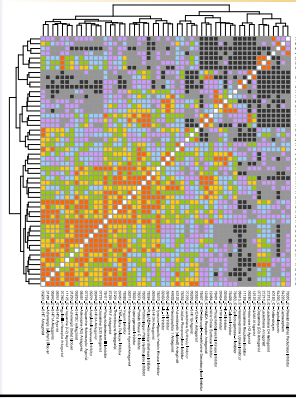
$$p(X > x) = \int_x^{\infty} f(t)dt = erfc(x) \qquad p(S > s) = erfc\left(\frac{S}{\sqrt{2\sigma^2}}\right)$$

where $f(t)$ is the standard normalized Gaussian probability density function and $erfc(x)$ is the complementary error function. For a normalized distribution of scores, we obtain $p(S > s)$



***A "p-value" for a given score, s*, is the probability of finding at random from the distribution some other score, *S*, which is greater than *s*.**

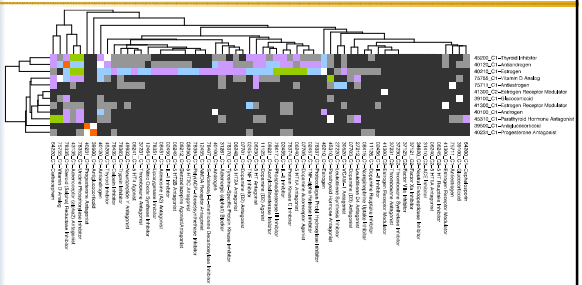
7. MDDR polypharmacology interaction matrix



- **$p \leq 5\times10^{-60}$**
- **$5\times10^{-60} < p \leq 5\times10^{-50}$**
- **$5\times10^{-50} < p \leq 5\times10^{-40}$**
- **$5\times10^{-40} < p \leq 5\times10^{-30}$**
- **$5\times10^{-30} <$ p-value $\leq 5\times10^{-20}$**
- **$5\times10^{-20} < p \leq 5\times10^{-10}$**
- **$p > 5\times10^{-10}$**
- **No score for a ligand set and itself**

MDDR polypharmacology interaction matrix for the top 50 ligand set relationships found from the all against all comparison of the 270 MDDR specific annotations

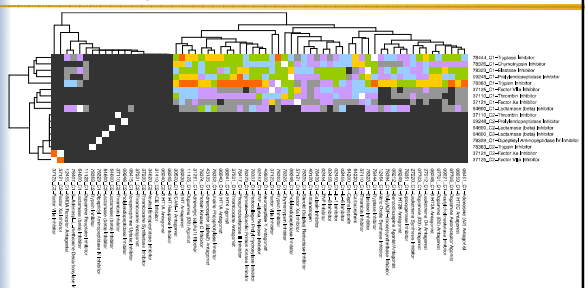
8. Nuclear hormone receptors



A subset of the MDDR interaction matrix involving several **nuclear hormone receptors**.

Antiglucocorticoids and *progesterone antagonists* are identified as promiscuous (orange).

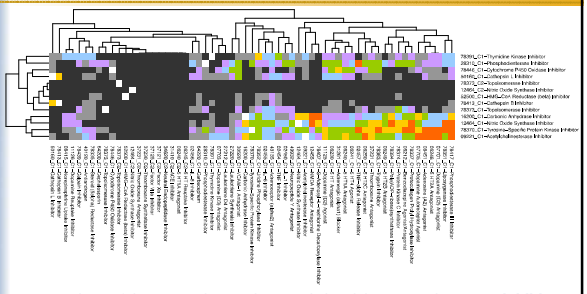
8. Serine proteases



A subset of the MDDR interaction matrix involving several **serine proteases**.

Coagulation factors Xa and *VIIa inhibitors* are identified as promiscuous, as well as *trypsin inhibitors* (orange).

8. Enzyme inhibitors

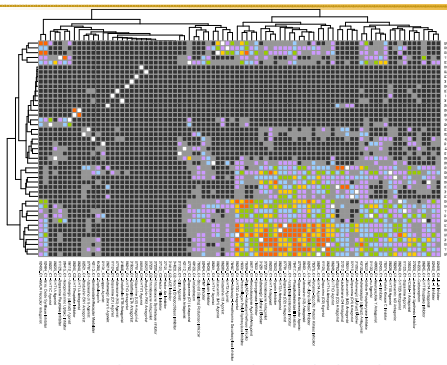


A subset of the MDDR interaction matrix involving several **enzyme inhibitors**.

Acetylcholinesterase inhibitors and *Tyrosine Specific protein kinase inhibitors* are identified as promiscuous (orange).

21

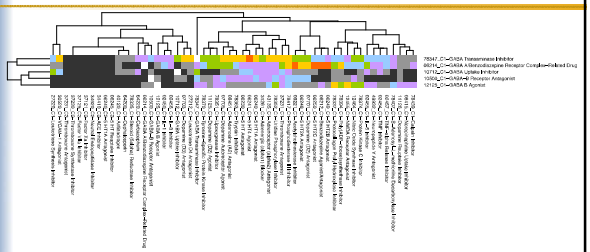
8. GPCRs



5HT2A and 5HT2C antagonists, and *dopamine D2 agonists and antagonists* are identified as promiscuous **GPCRs** (orange).

22

8. Ions channels

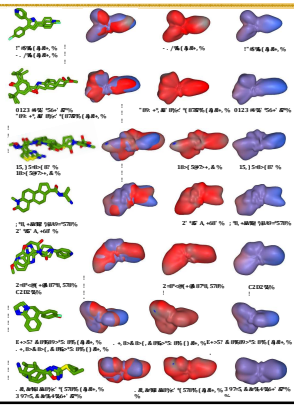


A subset of the MDDR interaction matrix involving several **ion channels**.

GABA A/Benzodiazepine receptor complex are identified as promiscuous (orange).

23

8. Examples of strongly related targets



Examples of SH shape superpositions of the CMs of some of the strongly related targets found in the selected MDDR subset.

Overall, we find interesting relationships between targets such as GABA A and tyrosine-specific protein kinase, ACE and neutral endopeptidase, thromboxane antagonist and thromboxane synthetase inhibitor, dopamine reuptake inhibitor and norepinephrine uptake inhibitor, *ie*, **whose dual inhibitors have been experimentally confirmed.**

GES also detects other relationships previously predicted by SEA and subsequently confirmed in vitro by Keiser *et al.* such as serotonin reuptake inhibitors acting also as β-blockers, 5 HT re-uptake inhibitors and adrenergic β-blockers.

24

Conclusions

- We have presented a new 3D shape-based approach for predicting and quantifying drug promiscuity by correlating Gaussian clusters of ligand SH shapes.
- The method has been validated using drug ligand sets of the MDDR and has been demonstrated to be effective in identifying drug families which are known to have related MDDR activity classes.
- Our results show that GES provides an efficient way to measure the similarity between clusters of arbitrary numbers of members.
- The examples shown in this study demonstrate that GES is a useful way to study polypharmacology relationships, and it could provide a novel way to propose new targets for drug repositioning.

25

Acknowledgements

- INRIA Nancy - Grand Est
- FP7 Marie Curie IEF Fellowship (DOVSA 254128)

Papers: http://www.loria.fr/~pereznue/
http://www.loria.fr/~ritchied/

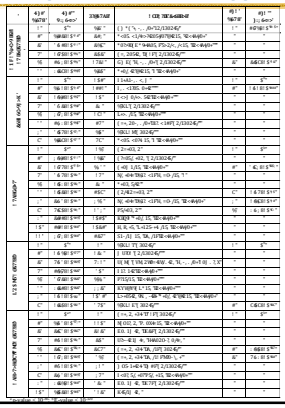
ParaSurf + ParaFit: http://www.ceposinsilico.de/

26

Thank you!

27

Comparison with SEA approach



28