USING MATCHED SERIES TO DECIDE WHAT COMPOUND TO MAKE NEXT

Noel M. O'Boyle,[1] Jonas Boström,[2] Roger A. Sayle[1], Adrian Gill[2]

[1] *NextMove Software Ltd, Cambridge, UK* [2] *AstraZeneca, Mölndal, Sweden*

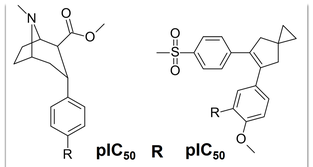
J. Med. Chem. **2014**, *57*, 2704

Summary

Large amounts of activity data across a broad set of targets are available either publicly (ChEMBL) or internally within a pharmaceutical company. We describe a method that exploits these data to predict R-groups that will improve binding affinity [1].

What is a matched molecular series?

A ***matched (molecular) series*** describes a set of molecules with the same scaffold but different R groups at a particular position [2]. The related term, ***matched pair***, corresponds to a matched series consisting of just two molecules.



pIC_{50}	R	pIC_{50}
7.00	H	8.30
7.68	F	8.00
8.51	Cl	7.77
8.77	Br	8.89
[Br>Cl>F>H]		[Br>H>F>Cl]

Enriched activity orders

For an ordered matched series, there are *N!* ways of arranging the R-groups. An interesting and useful observation is that for a particular matched series, certain activity orders of the R-groups may be preferred. These preferences are more pronounced for longer series.

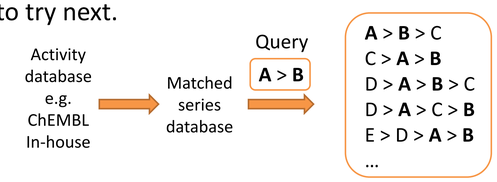
For example, using pIC50 data from ChEMBL there are 982 ordered matched series involving H, F, Cl, and Br, arranged in 24 possible ways:

		Count	Enrichment	
1.	Br > Cl > F > H	230	5.62	← most preferred order
3.	H > F > Cl > Br	69	1.69	← most preferred order "reversed"
24.	Br > H > F > Cl	9	0.22	← least preferred order

The origin of these preferences is assumed to be the existence of particular binding site environments that occur again and again across multiple targets. Using these order preferences, we can predict whether a particular R-group is likely to increase activity given an observed activity order.

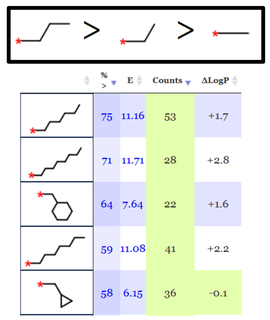
The Matsy algorithm: predicting R-groups likely to improve activity

Given a query matched series, the Matsy algorithm searches an activity database for all R-groups that have been measured along with the query, and calculates the percentage of times each R-group increased the activity beyond the most active R-group in the query. The R-groups with the highest percentages are presented as the best candidates to try next.

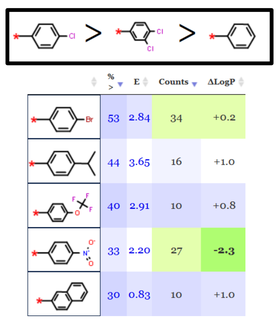


R Group	Observations	Obs that increase activity	% that increase activity
D	3	3	100
E	1	1	100
C	4	1	25
...	...		...

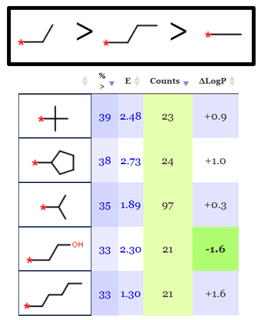
A large-scale retrospective test using ChEMBL16 showed that the longer the series the greater the predictive power [1].



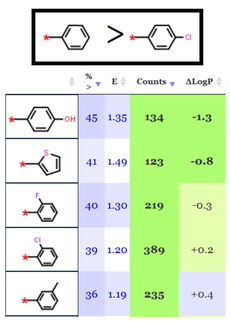
%	E	Counts	ΔLogP
75	11.16	53	+1.7
71	11.71	28	+2.8
64	7.64	22	+1.6
59	11.08	41	+2.2
58	6.15	36	-0.1



%	E	Counts	ΔLogP
53	2.84	34	+0.2
44	3.65	16	+1.0
40	2.91	10	+0.8
33	2.20	27	-2.3
30	0.83	10	+1.0



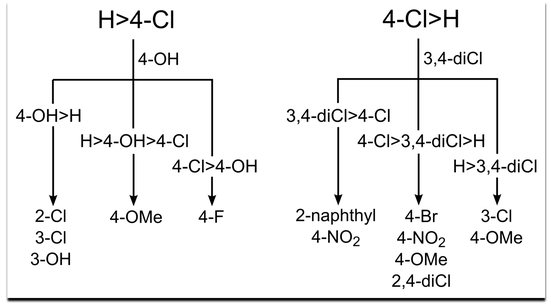
%	E	Counts	ΔLogP
39	2.48	23	+0.9
38	2.73	24	+1.0
35	1.89	97	+0.3
33	2.30	21	-1.6
33	1.30	21	+1.6



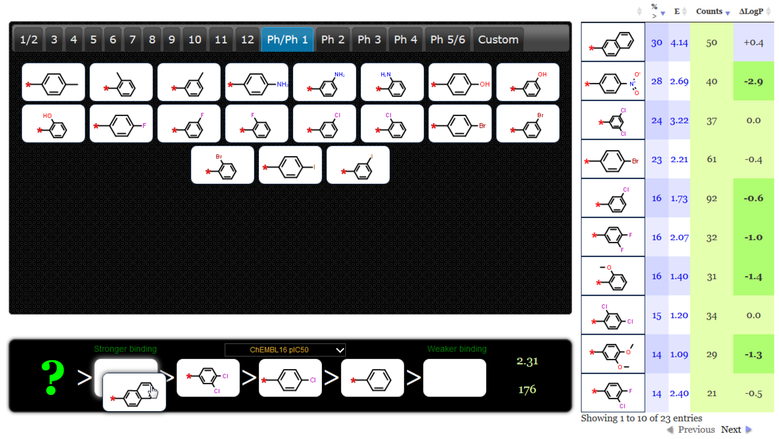
%	E	Counts	ΔLogP
45	1.35	134	-1.3
41	1.49	123	-0.8
40	1.30	219	-0.3
39	1.20	389	+0.2
36	1.19	235	+0.4

Updating the Topliss Tree

In a landmark paper [3], Topliss described a decision tree that guides a medicinal chemist to the most potent analogue of a substituted phenyl ring. Topliss based his tree on a rational analysis of the activity order. Using the Matsy algorithm, we have created a similar tree that is based on observed experimental results in ChEMBL.



Matsy GUI



Bibliography

1. O'Boyle, N. M.; Boström, J.; Sayle, R. A.; Gill, A. **Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity.** *J. Med. Chem.* **2014**, *57*, 2704.
2. Wawer, M.; Bajorath, J. **Local Structural Changes, Global Data Views: Graphical Substructure–Activity Relationship Trailing.** *J. Med. Chem.* **2011**, *54*, 2944.
3. Topliss, J. G. **Utilization of Operational Schemes for Analog Synthesis in Drug Design.** *J. Med. Chem.* **1972**, *15*, 1006.

NextMove Software

www.nextmovesoftware.com
@nmsoftware

NextMove Software Limited
Innovation Centre (Unit 23)
Cambridge Science Park
Milton Road, Cambridge
UK CB4 0EY