

# 2D Control

## Emulating the Algorithms in a Chemist's Head

Roger Sayle

OpenEye Scientific Software



OpenEye Scientific Software

OpenEye CUP IX, Santa Fe, March 2008

# Dead Guy Quote

- Clarke's Third Law:  
"Any sufficiently advanced technology is indistinguishable from magic".
- Gehm's Corollary  
"Any technology distinguishable from magic is insufficiently advanced".

Sir Arthur Charles Clarke, CBE  
16 December 1917 – 19 March 2008

# Motivation #1

- What is the null hypothesis of property prediction?
- Lookup values in a table, and refuse to extrapolate/interpolate to compounds that haven't been measured.
- Submission #126
  - CUP08026 dichlorobenil
  - a.k.a. 2,6-dichlorobenzonitrile, published as -5.22.
  - In most solvation training sets (Kollman, Trular...)
- Lowest median error submission at SAMPL: 0.51



# Motivation #2

- How well do off-the-self atom additive solvation energy models do?
- Implementation of ALOGS, Ghose 1999.
- Implementation has  $R^2$  of 0.9898 to original paper, fits training set to  $R^2$  of 0.9466 vs.  $R^2$  of 0.9561 claimed in the original paper.
- 70 atom types. 265 compound training set.
- The strict atom typing of the method results in only 6 out of 63 structures being covered by the parameterization.
- Median Error: 2.43,  $R^2$ : 0.54



# Reference

Vellarkad N. Viswanadhan, Arup K. Ghose, U. Chandra Singh and John J. Wendoloski,  
**“Prediction of Solvation Free Energies of Small Organic Molecules: Additive-Constitutive Models Based on Molecular Fingerprints and Atomic Constants”**,  
*Journal of Chemical Information and Computer Science (JCICS)*, Vol. 39, No. 2, pp. 405-412, 1999.



# Motivation

- Are 2D Knowledge-based approaches competitive with docking and 3D virtual screening methods?
- Use of (easily) publically available information.
- All literature/Google searches performed prior to seeing/downloading the challenge data set.
- No target-specific expertise.
- What can be discovered about decoys from browsing through the compound database.
- Murzin/CASP and Bradshaw/HoEI philosophy.



# Protocol Part 1

- For each target, extract a small representative sample (about 32-35 compounds) of active compounds from the available literature.
- Triage the compound sets into three categories; probably active, probably decoy and unknown.
- The “probably active” subset is based upon very high similarity to a known active or functional group known to be important/critical for binding.
- The “probably decoy” subset is based on eliminating inorganics, hydrocarbons or similar suspicious functionality, or missing a feature present in all of the prototype actives.

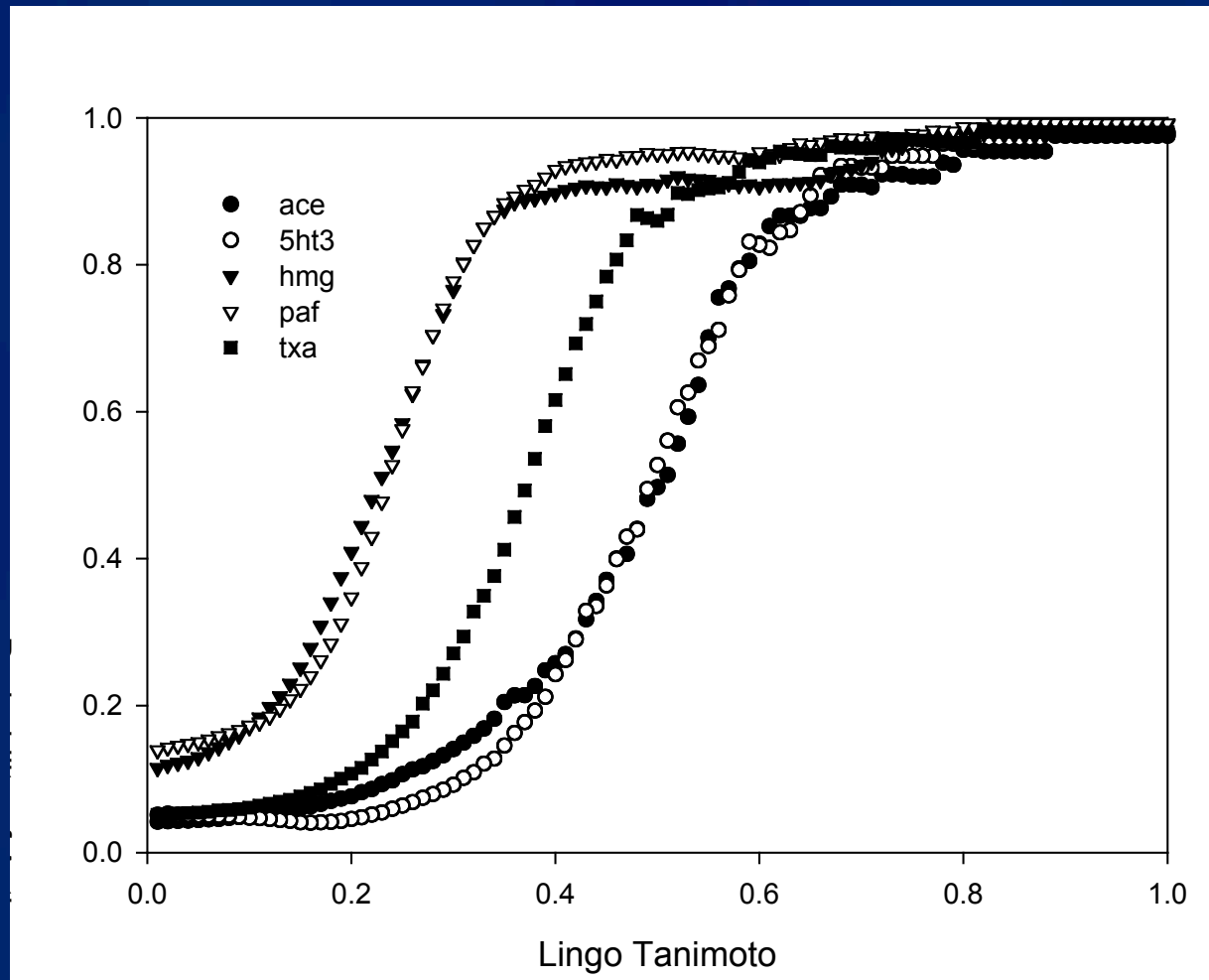


# Protocol Part 2

- Finally, within each triage partition, sort the compounds by 2D similarity to the representative actives.
- The chosen 2D similarity method was an implementation of Merck's TOPOSIM, which was thought to have better global properties than LINGOs or MDL fingerprints.
- Each similarity was normalized to a Z-score (ignoring the top 100 matches) to account for significance of a match.
- The maximum normalized similarity (Z-score) was considered the similarity to the actives.



# Why Z-Scores?



# Red Flags

- Strange Element (not in C,H,N,O,F,P,S,Cl,Br,I)
- Inorganics (must contain C)
- Hydrocarbons (must contain other than C,H)
- Nitrogenous (must contain N)
- Has "Functional" atom...
  - "[#6]", "\*S(=\*)(=\*)\*", "o", "s", "n-\*", "[n+]" etc...
- Not an obvious Urokinase decoy
  - #16,#779,#945,#1463,#4855,#5106,#5990



# Protocol Validation

- The choice of LINGOs vs. TOPOSIM and the use of normalized Z-scores, rather than the raw unnormalized similarities was based upon looking at the enrichment (dilution) of the “probably decoys”.
- Normalized TOPOSIM had fewer “probably decoys” (red flags) in the top 50 results than the other method combinations.



# Triage Summary

- **JNK3**
  - 12546 database compounds
  - 32 prototype compounds
  - 11 green flags (0.09%)
  - 6177 red flags (49.23%)
- **UROK**
  - 8351 database compounds
  - 35 prototype compounds
  - 131 green flags (1.57%)
  - 4372 red flags (53.35%)



# JNK3 Rank (12546 cmpds)

active.34 jnk.vs.1-8441	@3	active.15 jnk.vs.1-11633	@928
active.32 jnk.vs.1-4688	@4	active.10 jnk.vs.1-7876	@952
active.29 jnk.vs.1-505	@5	active.26 jnk.vs.1-5668	@956
active.31 jnk.vs.1-1740	@6	active.11 jnk.vs.1-7873	@1036
active.30 jnk.vs.1-4683	@7	active.3 jnk.vs.1-7895	@1037
active.35 jnk.vs.1-2876	@9	active.8 jnk.vs.1-7885	@1065
active.2 jnk.vs.1-2766	@11	active.12 jnk.vs.1-9254	@1198
active.7 jnk.vs.1-2254	@59	active.5 jnk.vs.1-7879	@1284
active.36 jnk.vs.1-7960	@129	active.18 jnk.vs.1-7877	@1468
active.21 jnk.vs.1-7880	@157	active.6 jnk.vs.1-7883	@1618
active.19 jnk.vs.1-7886	@257	active.23 jnk.vs.1-7957	@1660
active.24 jnk.vs.1-7890	@301	active.14 jnk.vs.1-7872	@1920
active.20 jnk.vs.1-2189	@304	active.4 jnk.vs.1-7915	@2035
active.25 jnk.vs.1-2191	@496	active.9 jnk.vs.1-7882	@2162
active.17 jnk.vs.1-7893	@555	active.13 jnk.vs.1-7875	@2229
active.37 jnk.vs.1-3165	@649	active.22 jnk.vs.1-7916	@3056
active.1 jnk.vs.1-3465	@473	active.16 jnk.vs.1-7870	@3116
active.15 jnk.vs.1-11633	@928	active.27 jnk.vs.1-11856	@3558
active.10 jnk.vs.1-7876	@952	active.38 jnk.vs.1-12331	@3812
active.26 jnk.vs.1-5668	@956	active.28 jnk.vs.1-2971	@4566

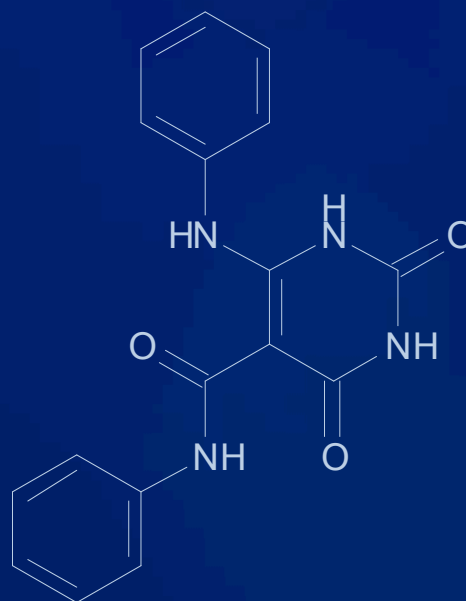


# JNK3 Hitlist

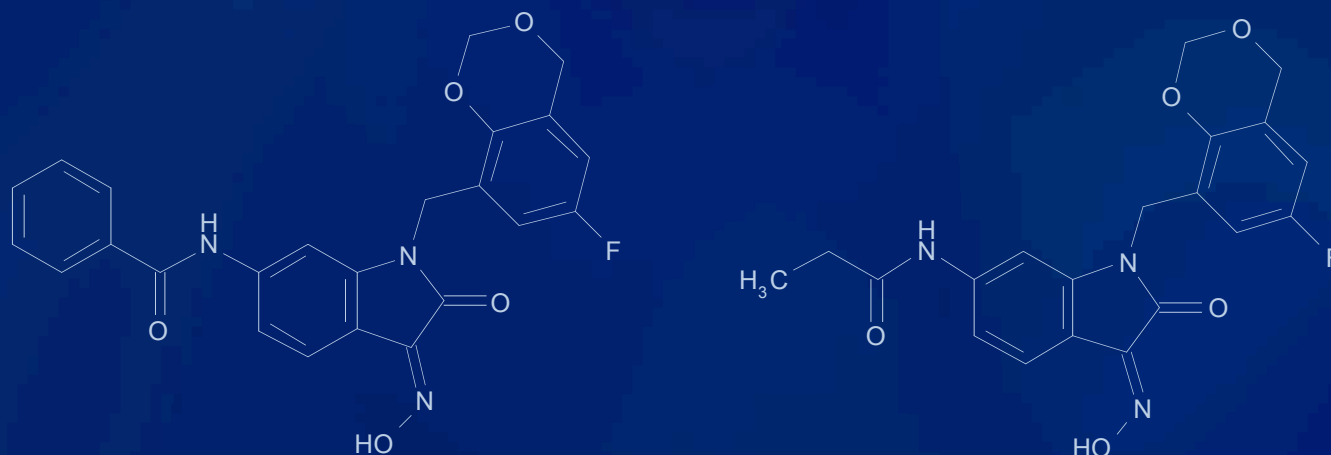
	<u>Cmpd</u>	<u>Class</u>	<u>Proto</u>	<u>Sim</u>	<u>Z-Score</u>	<u>Decoy</u>
1.	#3693	good	13	1.000	6.718	
2.	#3367	???	11	0.700	4.427	
3.	#8441	active	11	0.694	4.369	#8442
4.	#4688	active	11	0.681	4.219	#4689
5.	#505	active	11	0.671	4.110	#504
6.	#1740	active	11	0.649	3.866	#1741
7.	#4683	active	11	0.643	3.797	#4682
8.	#1775	good	11	0.630	3.657	
9.	#2876	active	11	0.595	3.274	#2875
10.	#2877	shape	11	0.595	3.274	
11.	#2766	active	11	0.592	3.237	



# JNK3 #3693 (@1)



# JNK3 #3367 (@2)



Anthony M. Manning and Roger J. Davis, "Targeting JNK for Therapeutic Benefit: From Junk to Gold?", *Nature Reviews in Drug Discovery*, Vol. 2, No. 7, pp. 503-598, July 2003.



# JNK #2971 (@4566)



# UROK Green Flags

- Arylcarboxamidines
  - $[ND1]\sim C(\sim [ND1])c([aD2])[aD2]$
- Other amidines
  - $[ND1]\sim C(\sim [ND1])C[CH2]$
- Guanidines
  - $[ND1]\sim C(\sim [ND1])\sim N$
- Except Obvious Decoys
  - #16, #779, #945, #1463, #4855, #5106, #5990



# UROK Rank (8351 cmpds)

active.2 uk.vs.1-2011	@1	good.5 uk.vs.1-2358	@5
active.10 uk.vs.1-3737	@2	good.8 uk.vs.1-4383	@37
active.5 uk.vs.1-2134	@3	good.4 uk.vs.1-5796	@40
active.18 uk.vs.1-572	@14	good.9 uk.vs.1-2032	@65
active.8 uk.vs.1-4499	@15	good.6 uk.vs.1-6275	@77
active.6 uk.vs.1-7569	@18	good.7 uk.vs.1-4692	@90
active.12 uk.vs.1-6411	@21	good.1 uk.vs.1-366	@171
active.3 uk.vs.1-5192	@23	good.2 uk.vs.1-3990	@395
active.13 uk.vs.1-7194	@26	good.3 uk.vs.1-1239	@638
active.17 uk.vs.1-7201	@30		
active.11 uk.vs.1-2354	@60	inactive.2 uk.vs.1-521	@12
active.4 uk.vs.1-6277	@70	inactive.1 uk.vs.1-3296	@13
active.9 uk.vs.1-3713	@72	inactive.5 uk.vs.1-8076	@24
active.15 uk.vs.1-2598	@73	inactive.7 uk.vs.1-1570	@33
active.16 uk.vs.1-5551	@78	inactive.6 uk.vs.1-3992	@47
active.14 uk.vs.1-6748	@91	inactive.3 uk.vs.1-3195	@57
active.7 uk.vs.1-5556	@94	inactive.4 uk.vs.1-4344	@83
active.1 uk.vs.1-2414	@524		

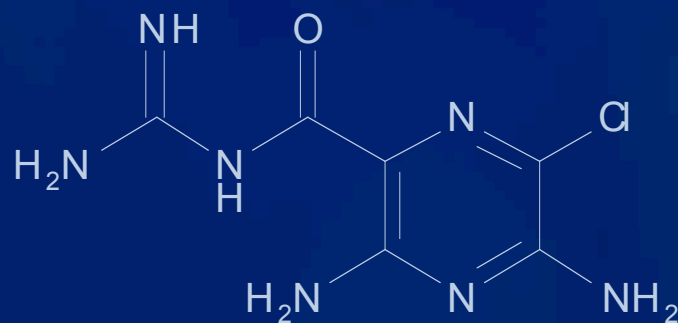


# UROK Hitlist

	<u>Cmpd</u>	<u>Class</u>	<u>Proto</u>	<u>Sim</u>	<u>Z-Score</u>
1.	#2011	active	3	1.000	17.133
2.	#3737	active	6	1.000	11.544
3.	#2134	active	4	0.647	9.395
4.	#7966	dud	25	0.544	8.635
5.	#2358	active	4	0.628	8.042
6.	#488	dud	25	0.545	7.454
7.	#1305	dud	4	0.523	7.233
8.	#2898	schrod	5	0.515	6.944
9.	#4822	dud	5	0.505	6.770
10.	#6937	dud	3	0.555	5.899



# UROK #2011 (@1)

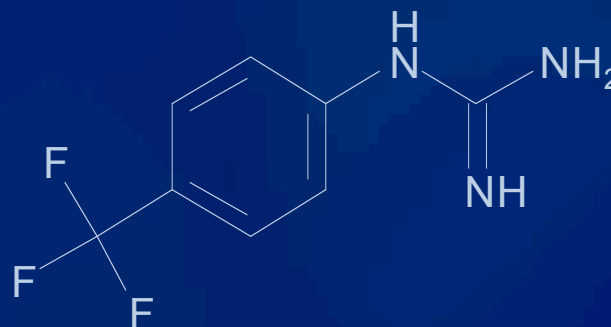
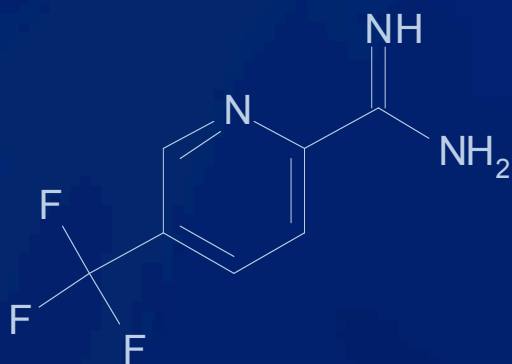


Amiloride

Paul Fish et al., "Selective Urokinase-Type Plasminogen Activator Inhibitors. 4. 1-(7-Sulfonamidoisoquinolyl)guanidines", J. Med. Chem., Vol. 50, No. 10, pp. 2341-2351, 2007.  $K_i=7,000\text{nM}$



# UROK #7966 (@4)



Paul Fish et al., "Selective Urokinase-Type Plasminogen Activator Inhibitors. 4. 1-(7-Sulfonamidoisoquinolyl)guanidines", J. Med. Chem., Vol. 50, No. 10, pp. 2341-2351, 2007.  $K_i=6,000\text{nM}$



# UROK #2414 (@524)



# Post Mortem

- No affinity information was used, and all prototypes were considered equal.
- The approximate number of actives wasn't used, as I didn't discover the counts were published until after my entries were submitted.
- Impressive screening against uncolor decoys.
- Biggest problems were clash and dud decoys.



# 20/20 Hindsight: JNK3 All

JNK3		raw		triage	
			norm		norm
lingos	rognan	0.49	0.65	0.73	0.81
	schrod	0.49	0.38	0.59	0.48
	dud	0.37	0.43	0.47	0.54
lingos2	rognan	0.60	0.50	0.76	0.77
	schrod	0.55	0.39	0.61	0.44
	dud	0.52	0.45	0.59	0.50
toposim	rognan	0.83	0.52	0.88	0.60
	schrod	0.60	0.32	0.61	0.32
	dud	0.51	0.31	0.53	0.33



# 20/20 Hindsight: JNK3 Good

JNK3		raw		triage	
			norm		norm
lingos	rognan	0.36	0.65	0.62	0.81
	schrod	0.37	0.41	0.46	0.49
	dud	0.23	0.47	0.32	0.56
lingos2	rognan	0.36	0.33	0.53	0.59
	schrod	0.30	0.22	0.39	0.31
	dud	0.28	0.25	0.37	0.35
toposim	rognan	0.68	0.54	0.77	0.62
	schrod	0.42	0.31	0.42	0.31
	dud	0.35	0.31	0.36	0.32



# 20/20 Hindsight: UROK All

UROK		raw		triage	
			norm		norm
lingos	rognan	0.83	0.80	0.89	0.90
	schrod	0.78	0.77	0.88	0.87
	dud	0.72	0.63	0.80	0.78
lingos2	rognan	0.93	0.89	0.94	0.92
	schrod	0.90	0.86	0.91	0.89
	dud	0.81	0.67	0.83	0.78
toposim	rognan	0.72	0.40	0.93	0.93
	schrod	0.66	0.46	0.89	0.90
	dud	0.77	0.36	0.86	0.77



# 20/20 Hindsight: UROK Good

UROK		raw		triage	
			norm		norm
lingos	rognan	0.73	0.73	0.78	0.83
	schrod	0.69	0.69	0.75	0.76
	dud	0.61	0.59	0.66	0.66
lingos2	rognan	0.92	0.85	0.94	0.90
	schrod	0.85	0.80	0.86	0.82
	dud	0.69	0.54	0.68	0.62
toposim	rognan	0.60	0.32	0.89	0.85
	schrod	0.54	0.38	0.78	0.81
	dud	0.68	0.26	0.74	0.63



# Final Thoughts

- From an N=2 sample (SAMPL), data fusion is simpler than I thought.
- The 2D similarity method of choice depends upon the local vs. global properties of the target compounds.



# Acknowledgements

- Geoff Skillman
- Anthony Nicholls
- All the folks that provided challenge systems.

