The University Of Sheffield.

Choosing the Right Similarity Measure

John Holliday, University of Sheffield, UK

The University Of Sheffield.

Overview

- Bias fusion of similarity coefficients
- Machine learning approach
- Design your own coefficient
- Fusion of fingerprint pathlengths
- Non-hierarchical k-modes algorithm

The University Of Sheffield.

Similarity Coefficients

- Originally used 22 coefficients
- Results of searches clustered to identify similar coefficients
- 13 identified as unique
- Relative performance of each appears to be size dependent

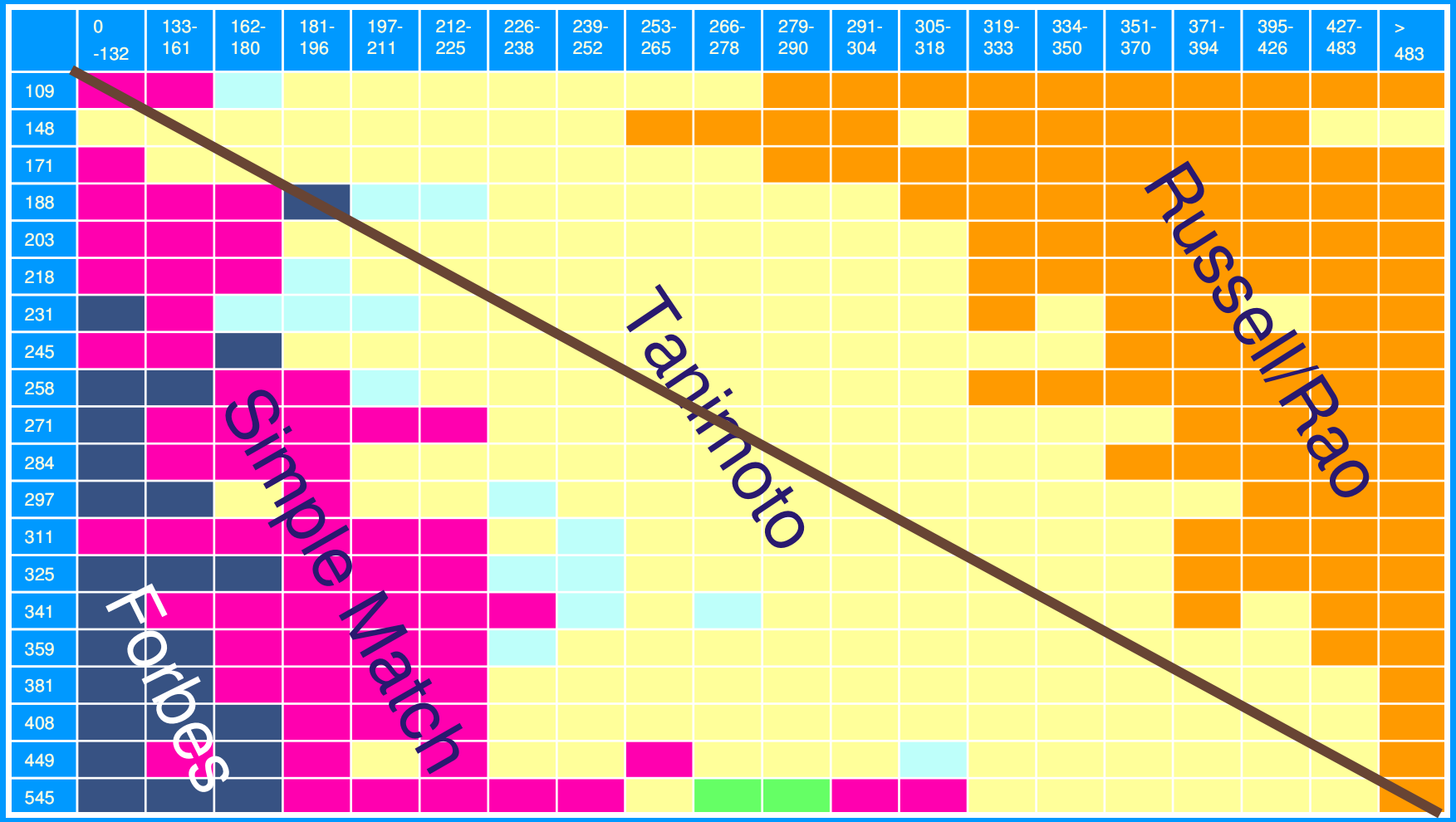
The University Of Sheffield.

Size Dependency

- MDDR sorted by bit density
- Divided into 20 equal partitions
- One compound from middle of each partition used as query
- All 13 coefficients used
- Best performing coefficient deduced for each partition

The University Of Sheffield.

Size dependency



The University Of Sheffield.

Size dependency

Size Range	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
0-100	1	0	31	1	1	1	21	1	5	1	8	1	1
101-150	15	0	93	28	13	14	72	13	8	18	33	18	25
151-200	91	6	157	135	83	68	155	79	16	97	114	95	113
201-250	158	22	83	175	123	90	117	123	19	137	113	136	150
251-300	162	89	49	139	155	142	66	155	83	148	125	151	141
301-350	211	214	9	130	224	224	21	225	206	207	175	207	188
351-400	107	189	0	41	130	152	2	131	181	111	83	111	88
401-450	18	124	0	5	35	59	0	35	113	23	18	24	12
451-500	1	78	0	0	12	20	0	12	72	3	4	4	0
>500	0	47	0	0	0	6	0	0	44	0	0	0	0

Retrieval (top 5%) of Antihypertensives - 200 bits

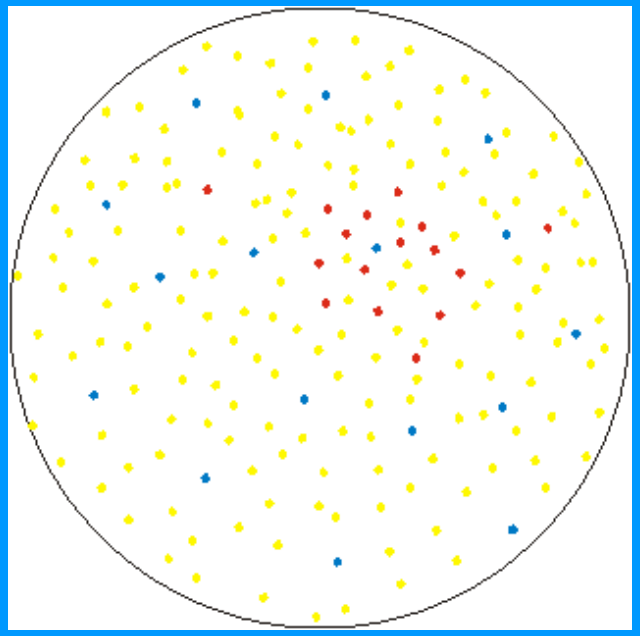
The University Of Sheffield.

Data Fusion

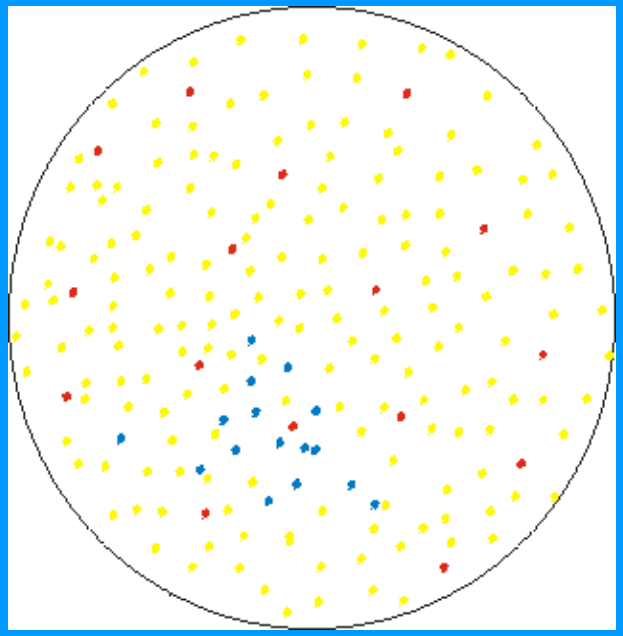
- Combine rankings from two or more coefficients
- Rankings combined by MAX or SUM
- Has shown to improve performance
- Choice of coefficients not obvious
- Size dependent & Class dependent

The University Of Sheffield.

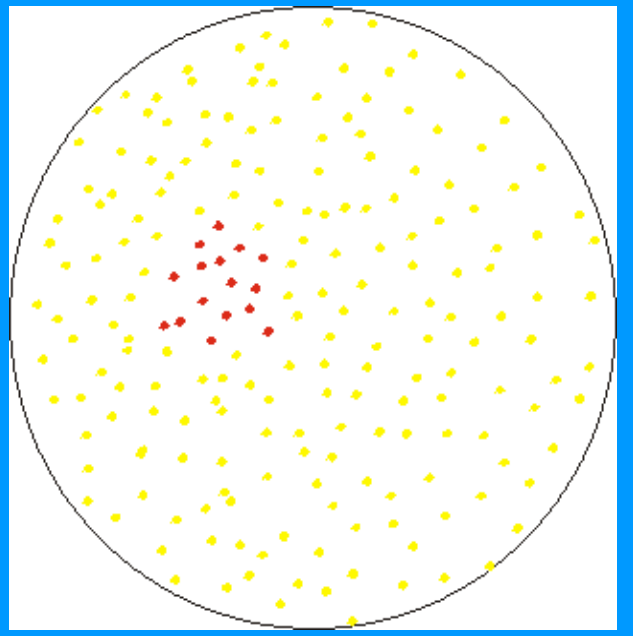
Aims



Russell Space



Forbes Space



Combined Space

Red = Class A, Blue = Class B, Yellow = bulk of DB

The University Of Sheffield.

Biasing coefficient selection

- Using four complementary coefficients:

Forbes	Simple Match	Tanimoto	Russell/Rao
$\frac{na}{(a+b)(a+c)}$	$\frac{a+d}{n}$	$\frac{a}{a+b+c}$	$\frac{a}{n}$

- Various weighting schemes used to combine these
  - based on previous search results

The University Of Sheffield.

Size dependency

Size Range	Tan	Rus	SM	For
0-100	1	0	31	21
101-150	15	0	93	72
151-200	91	6	157	155
201-250	158	22	83	117
251-300	162	89	49	66
301-350	211	214	9	21
351-400	107	189	0	2
401-450	18	124	0	0
451-500	1	78	0	0
>500	0	47	0	0

Retrieval (top 5%) of Antihypertensives - 200 bits

The University Of Sheffield.

Weighted Fusion

- F1 Equal weights - SUM
- F2 Equal weights - MAX
- F3 Number of dominant size ranges - SUM
- F4 Number of dominant size ranges - MAX
- F5 Manually-selected weights
- F6 1.0 for target weight, decreasing by 10% away from this

The University Of Sheffield.

Weighted Fusion

Class	Tan	F1	F2	F3	F4	F5	F6
43200	**13**	1.08	1.0	1.0	1.0	1.0	1.0
1200	**7**	1.0	1.0	1.0	1.0	2.0	1.0
75000	**68**	1.0	1.0	1.03	1.0	1.1	1.01
27200	**79**	0.92	0.97	1.0	1.0	0.94	0.97
6200	**109**	0.99	1.0	1.02	1.0	0.83	1.0
72	**73**	1.01	1.01	1.0	0.99	1.01	1.01
7000	**41**	1.56	1.8	1.22	1.0	1.2	1.2
9200	**68**	0.94	0.87	1.0	1.0	1.0	1.0
75000	**39**	1.15	1.13	1.15	1.15	1.03	1.1
2000	**34**	1.03	0.94	1.0	1.0	1.03	1.03
9200	**29**	1.48	1.41	1.07	1.0	1.17	1.03
27200	**216**	1.05	1.04	1.04	0.99	1.01	1.01
75000	**89**	1.0	0.99	0.99	1.0	0.96	0.96
6200	**92**	0.99	0.9	0.97	0.97	1.0	0.97
70000	**234**	0.9	0.83	0.96	1.0	1.0	0.99
31000	**19**	1.11	1.0	1.05	1.21	1.0	1.05
37200	**53**	1.13	1.09	1.23	1.02	1.09	1.15
68000	**245**	0.7	0.62	0.82	0.87	1.0	1.0
2000	**32**	1.38	1.5	1.06	1.0	1.0	1.0

The University Of Sheffield.

Machine Learning Approach

- To identify optimum weights for combining coefficients for a given active class
- Training sets of 1000 compounds
  - 70-100 actives
  - Rest made up of random database cmpds

The University Of Sheffield.

Machine Learning Approach

- Use actives as queries for each weighted combination
  - Search using every active
  - Search using modal fingerprint
- Weight combination controlled by
  - GA
  - Systematic approach in 4% steps
- Fitness function = Median rank position

The University Of Sheffield.

Modal Fingerprint

	1 0 0 1 0 1 1 0 1 0 0 0 1 1 0 1 0 0 1
	1 0 0 0 1 0 1 1 1 1 1 0 1 0 1 1 0 0 1 0
	0 0 1 0 0 1 0 1 1 0 1 0 1 0 0 1 0 1 1 0
	0 1 1 1 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1
	1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 0 0 0 0 0
40% threshold	1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
60% threshold	1 0 1 1 0 1 0 0 1 0 1 0 0 0 1 0 0 0 0 0
80% threshold	0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0

The University Of Sheffield.

Training set results

Summary of Systematic Results for Fusion (Median)						Median Ranks for Individual Coefs.			
Class	**TanWt**	**RusWt**	**SMWt**	**ForWt**	**Results**	**Tan**	**Rus**	**SM**	**For**
64220	0.20	0.32	0.48	0.00	**38.65**	39.61	41.35	43.58	86.49
78413	0.24	0.20	0.04	0.52	**138.72**	160.65	294.86	151.92	151.50
12200	0.00	0.00	0.20	0.80	**296.87**	349.68	496.74	309.14	297.05
7707	0.00	0.68	0.32	0.00	**47.75**	48.98	49.31	54.67	59.25
44200	0.00	1.00	0.00	0.00	**202.58**	265.12	**202.58**	495.19	472.39
80499	0.00	0.00	0.92	0.08	**193.56**	292.28	566.47	194.12	199.57
59210	0.52	0.00	0.00	0.48	**81.50**	97.51	116.88	100.93	92.80
31281	0.00	0.04	0.96	0.00	**105.65**	188.01	489.38	105.67	134.13
52503	0.00	0.00	1.00	0.00	**215.91**	312.60	514.66	**215.91**	250.12
42710	0.04	0.96	0.00	0.00	**91.49**	95.37	93.44	162.21	168.48

The University Of Sheffield.

Test Set Results

Number of Actives on the Top 500					
Class	**Cmpd**	**Tan**	**W1**	**W2**	**W3**
64220	143075	32	31	12	**32**
64220	188743	33	**34**	25	**34**
78413	154230	6	**6**	**6**	4
78413	195947	4	**6**	**6**	**7**
12200	186494	4	**4**	**4**	**4**
12200	174953	4	3	3	1
7707	215004	42	**42**	40	**42**
7707	213232	38	29	**40**	**41**
44200	223448	8	**8**	**8**	7
44200	214248	16	**16**	**16**	**16**
80499	197635	4	**4**	**4**	**4**
80499	257429	5	**5**	**5**	**5**
59210	183938	22	**23**	**22**	**23**
59210	227061	3	**3**	2	**3**
31281	154907	18	**20**	**32**	**31**
31281	143339	24	**30**	**34**	**32**
52503	248597	11	**11**	**11**	**11**
52503	207515	9	**9**	**9**	8
42710	214762	27	**27**	**27**	**27**
42710	200021	7	6	**8**	**8**

W1: Fusion with equal weightings

W2: Fusion with weights from trained + modal

W3: Fusion with weights from trained

The University Of Sheffield.

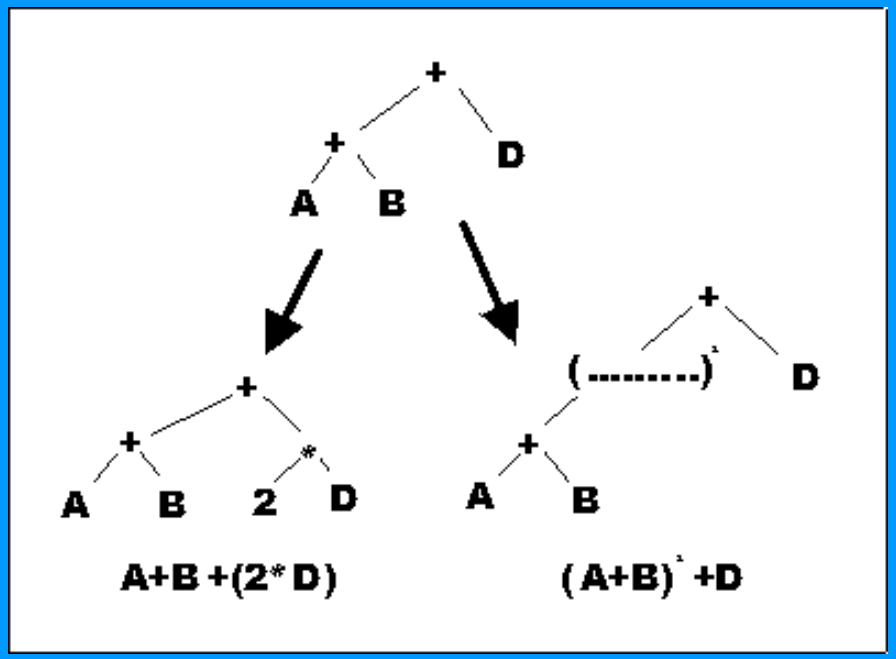
Four Complementary Coefficients

Forbes	Simple Match	Tanimoto	Russell/Rao
$\frac{na}{(a+b)(a+c)}$	$\frac{a+d}{n}$	$\frac{a}{a+b+c}$	$\frac{a}{n}$

The University Of Sheffield.

Formula Derivation

Decision tree method



The University Of Sheffield.

Formula Derivation

$$\frac{n^{l_1}(\pm i_1 a \pm i_2 b \pm i_3 c \pm i_4 d)^{m_1}(\pm i_5 a \pm i_6 b \pm i_7 c \pm i_8 d)^{m_2}}{n^{l_2}(\pm i_9 a \pm i_{10} b \pm i_{11} c \pm i_{12} d)^{m_3}(\pm i_{13} a \pm i_{14} b \pm i_{15} c \pm i_{16} d)^{m_4}}$$

- Driven by GA
- $l_{1\text{-}2} = 0$ or 1; $i_{1\text{-}16} = 0$, 1, 2 or 3; $m_{1\text{-}4} = 0$, 1 or ½
- Uses a 58 bit bitstring
- Same fitness function & training regime as before
- Tests included to remove erroneous formulae
- May require simplification
- Ranges are difficult to deduce

The University Of Sheffield.

Formula Results

Class	Actives	Search Results (top 500)	Tan (top 500)	Best Formula
64220	143075	32	32	(-3a+3b-2d) / (-3b-c)
64220	188743	33	33	
78413	154230	5	6	(-3b+2c+3d) / (b-2c+2d)(-3a-3c-d)
78413	195947	4	4	
12200	186494	4	4	n(-2a)(-3a-3c+d) / (3a-3b-c+2d)(-3a-b-d)
12200	174953	2	4	
80499	197635	3	4	sqrt(c+3d) / (-a-b-c-3d)
80499	257429	4	5	
59210	183938	23	22	(2a-3b+3d)(-3a-b-3c) / (3a)
59210	227061	3	3	

$$\log_{10} \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$$

The University Of Sheffield.

Fusion of Pathlengths

- MDDR database characterised by Daylight and BCI fingerprints in sets of different pathlength
- BCI – atom sequences of length 2-3, 4-5, 6-7, 8-9
- Daylight – pathlengths 2-4, 5-7, 8-10, 11-13, 14-16, 17-19, 20-22, 23-25, 26-28, 29-31

The University Of Sheffield.

Fusion of Pathlengths

- 20 compounds each from 11 active classes
- Fusion of all possible 2, 3, and 4 sets for BCI
- Fusion of all possible 2 and 3 sets for Daylight

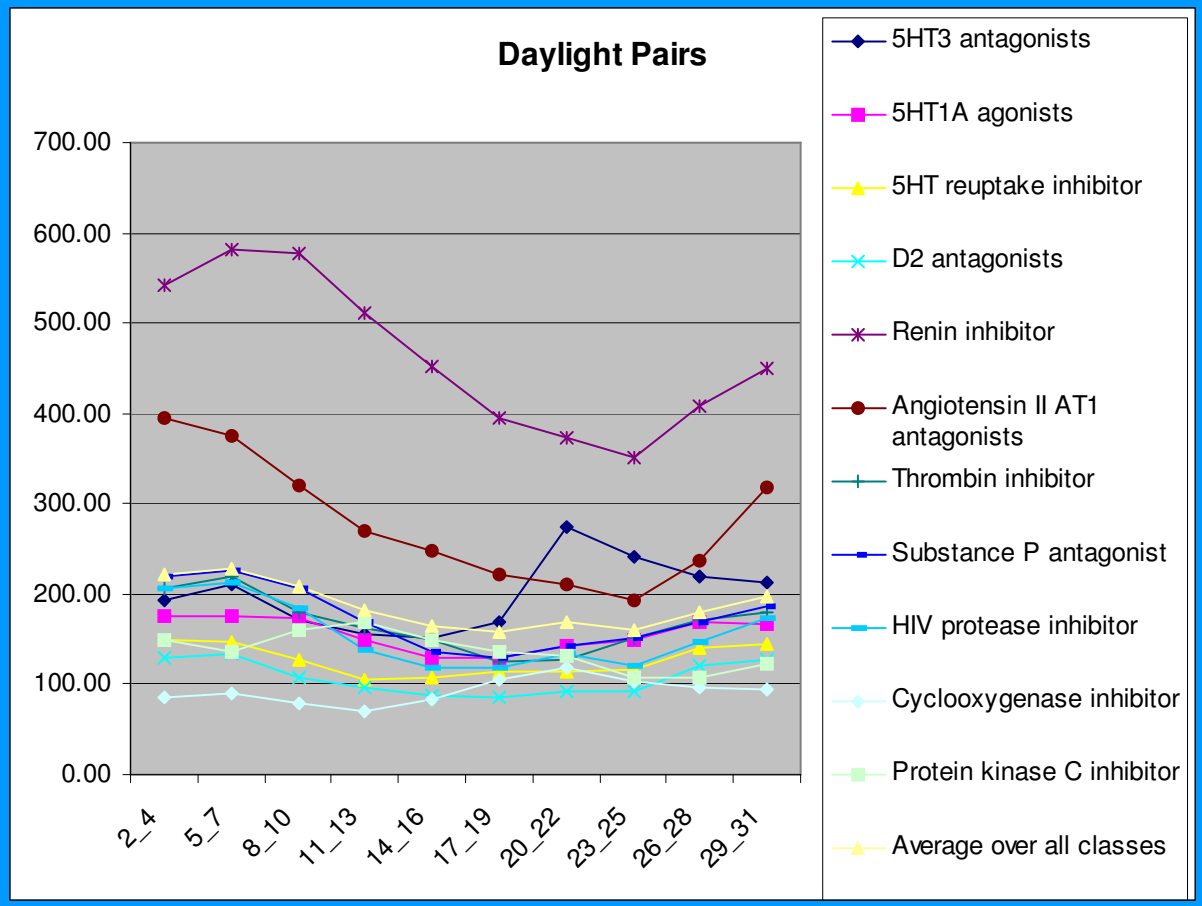
The University Of Sheffield.

Daylight Results

Class	Full	Double		Triple	
5HT3 antagonisis	23.60	5-7/20-22	32.63	8-10/17-19/29-31	?
5HT1A agonists	20.88	8-10/26-28	22.45	2-4/8-10/26-28	23.77
5HT reuptake inhibitor	21.81	2-4/5-7	19.68	2-4/5-7/29-31	20.71
D2 antagonists	20.53	2-4/5-7	18.79	2-4/5-7/29-31	18.95
Renin inhibitor	79.23	5-7/8-10	87.77	2-4/5-7/8-10	85.36
Angiotensin II AT1 antagonists	60.63	2-4/5-7	59.45	2-4/5-7/8-10	57.72
Thrombin inhibitor	30.82	2-4/5-7	31.21	2-4/5-7/29-31	29.72
Substance P antagonist	28.09	5-7/8-10	32.54	2-4/5-7/8-10	32.94
HIV protease inhibitor	35.73	2-4/5-7	34.13	2-4/5-7/8-10	33.75
Cyclooxygenase inhibitor	10.10	17-19/20-22 20-22/23-25	12.76	17-19/20-22/23-25	12.98
Protein kinase C inhibitor	31.63	8-10/11-13	22.69	2-4/8-10/11-13	25.38

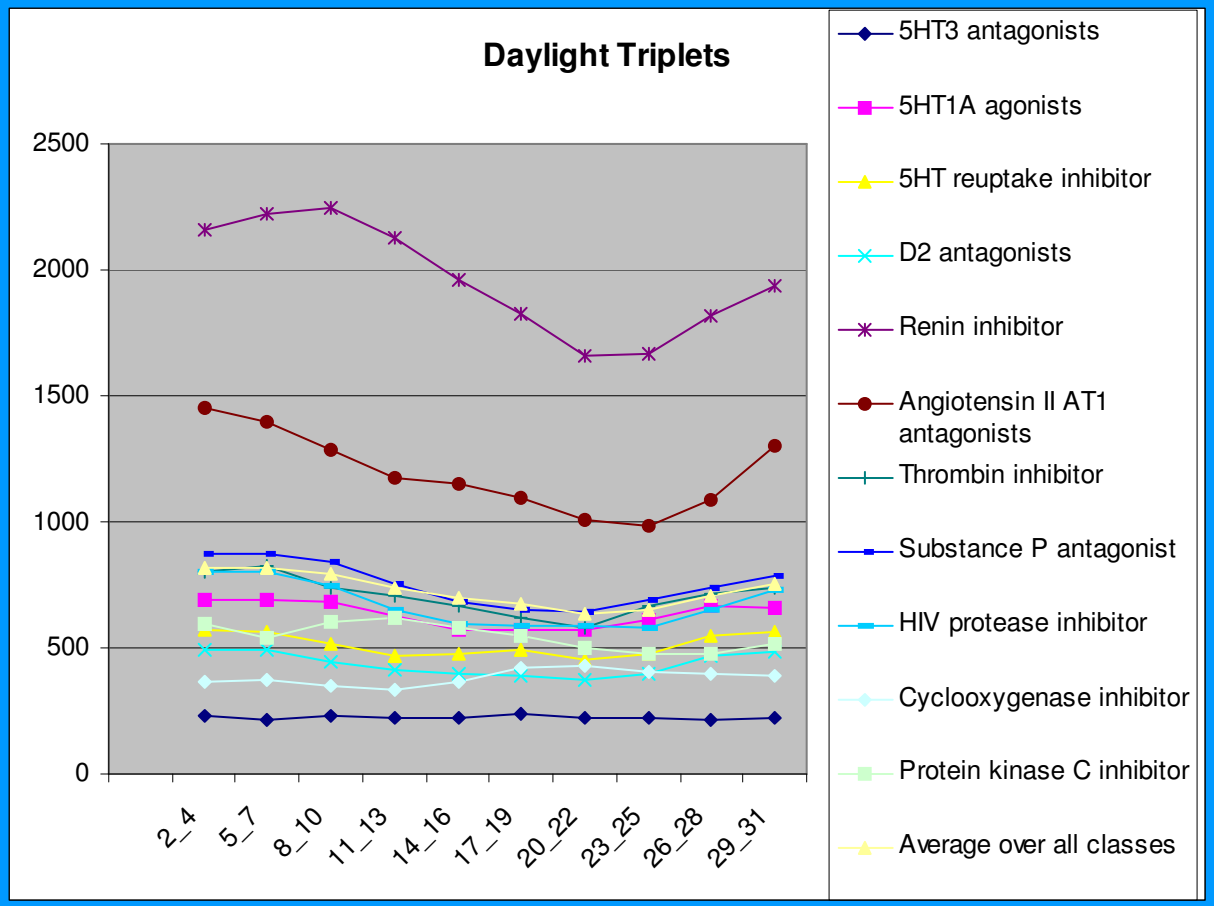
The University Of Sheffield.

Daylight Pairs - contributions



The University Of Sheffield.

Daylight Triplets - contributors



The University Of Sheffield.

K-means clustering

- Non-hierarchical, relocation
- Initial set of k seeds as cluster centre
  - Assign compounds to cluster centre
  - Recalculate cluster centre
  - Repeat until no relocation of compounds

The University Of Sheffield.

K-modes clustering

- Uses association coefficient - Tanimoto
- Modes instead of means for clusters
- Frequency based update method
- Need to optimise modal threshold

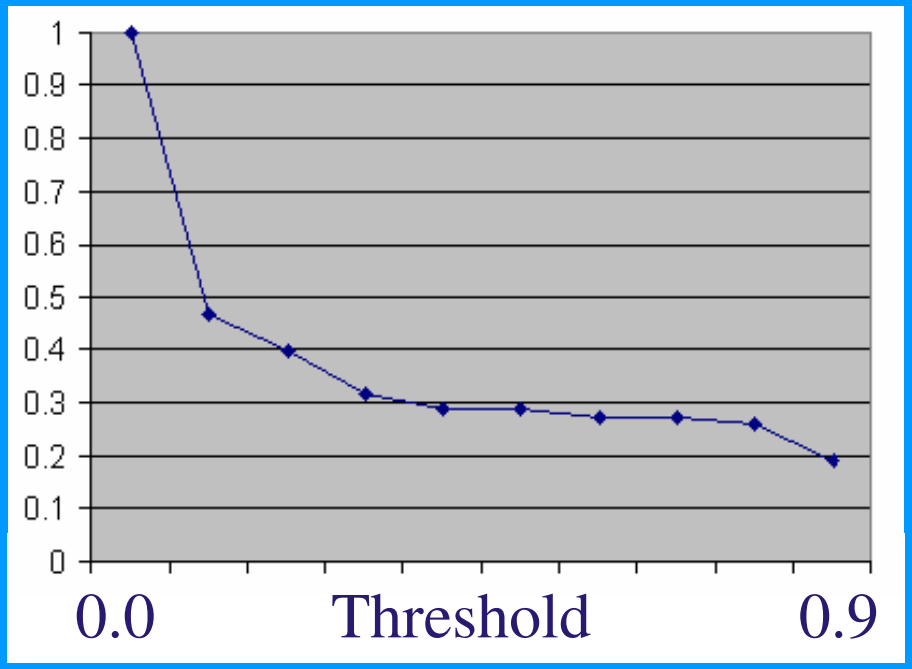
The University Of Sheffield.

Identifying Multiple Classes

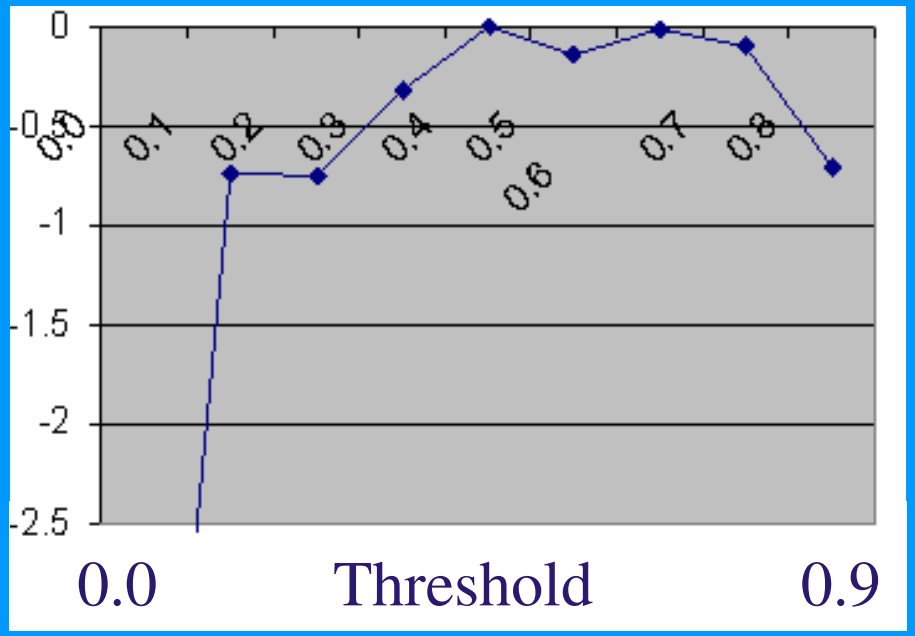
- To identify multiple classes and deduce optimum threshold
- 150 classes from MedChem02
- Modal fingerprints for each class generated at 0%, 10%,…,90% threshold
- Ratio of bits set at each threshold to bits set at 0% plotted

The University Of Sheffield.

Iron Chelator – 42 analogues



Bits at threshold vrs Bits at 0%



First derivative
Shows two peaks indicating two classes

The University Of Sheffield.

Using class seeds to cluster

- 150 classes form MedChem02 database
- Modal fingerprint generated for each class
- Used as seeds for k-modes algorithm
- Also used random seeds
- Repeated using 300 and using no relocation
- No improvement in cluster performance
- Repeated again using 250, 500, 1K, 2K, 4K random seeds
  - Considerable improvement observed

The University Of Sheffield.

Summary

- Methods for improving searches are possible
- Class-based methods
  - Weighted fusion of coefficients
  - Tailored coefficients
  - Different pathlengths
- Use of modal fingerprint

The University Of Sheffield.

Acknowledgements

- University of Sheffield
  - Jenny Chen, Peter Willett, Kay Busari, Jerome Hert
- Daylight
  - John Bradshaw, Jack Delany
- Others
  - BCI, Tripos, MDL, NCI, Current Drugs, Wolfson Foundation, Royal Society