Tanimoto similarity
For the training set, used for building the COMBINE model of thrombin, a matrix of Tanimoto similarity values were generated. The Tanimoto similarity is a value between 0 and 1 based 2D structure information of two ligands. These values can be computed by a GUI for CDK written by Rajarshi Guha:
/http://cdk.sourceforge.net/api/org/openscience/cdk/fingerprint/Fingerprinter.html/
/http://cheminfo.informatics.indiana.edu/~rguha/code/java/cdkws.html#sim
/http://cheminfo.informatics.indiana.edu/~rguha/code/java/sim.html

In JMP6 this matrix can be used for cluster analysis and a dentogram can be generated. Unfortunately, the GUI is just working for 100 smiles strings with a maximum of 100 characters, but most of the smiles of the other training and test sets have more characters.

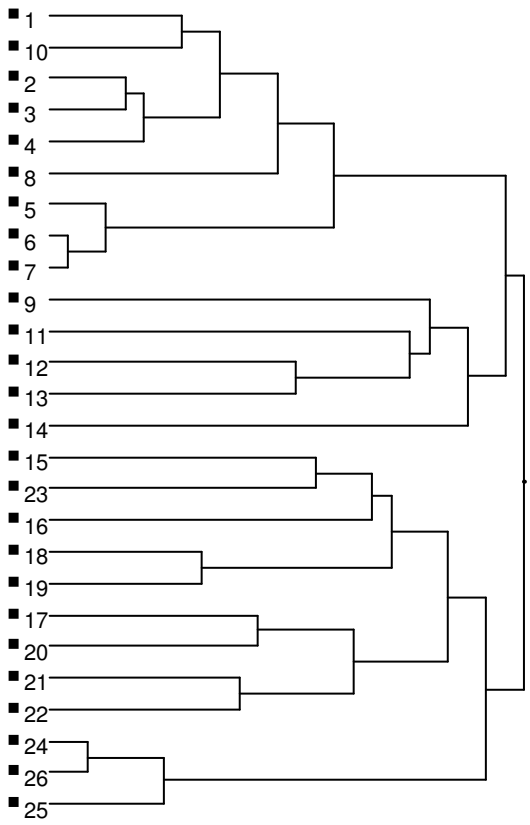
thrombin training set (C_thrombin)

21	A04	1
10	A12	2
16	A13	3
20	A14	4
3	A15	5
5	A16	6
8	A17	7
2	A18	8
288	A19	9
289	A20	10
291	A22	11
292	A23	12
293	A24	13
295	A26	14
299	A30	15
300	A31	16
301	A32	17
302	A33	18
303	A34	19
304	A35	20
305	A36	21
306	A37	22
307	A38	23
308	A39	24
309	A40	25
310	A41	26

Hierarchical Clustering

Method = Ward

Dendrogram



Clustering History

Number of Clusters	Distance	Leader	Joiner
25	0,25139319	6	7
24	0,30216318	24	26
23	0,42105732	5	6
22	0,44715073	2	3
21	0,71057822	2	4
20	1,06991258	24	25
19	1,23208773	1	10
18	1,26699584	18	19
17	1,34436756	1	2
16	1,61189059	21	22
15	1,88937238	17	20
14	2,35225582	1	8
13	2,41256098	12	13
12	2,59840059	15	23
11	2,80089374	1	5
10	2,94740989	17	21
9	3,06634850	15	16
8	3,34030603	15	18
7	3,60828107	11	12
6	3,73859530	9	11
5	4,29617137	15	17
4	6,37525209	9	14
3	8,40301001	15	24
2	8,48847689	1	9
1	18,78190421	1	15