Identifying maximally enriched scaffolds in HTS data sets

Martin Packer

19th June 2007

AstraZeneca

Enriched scaffold perception

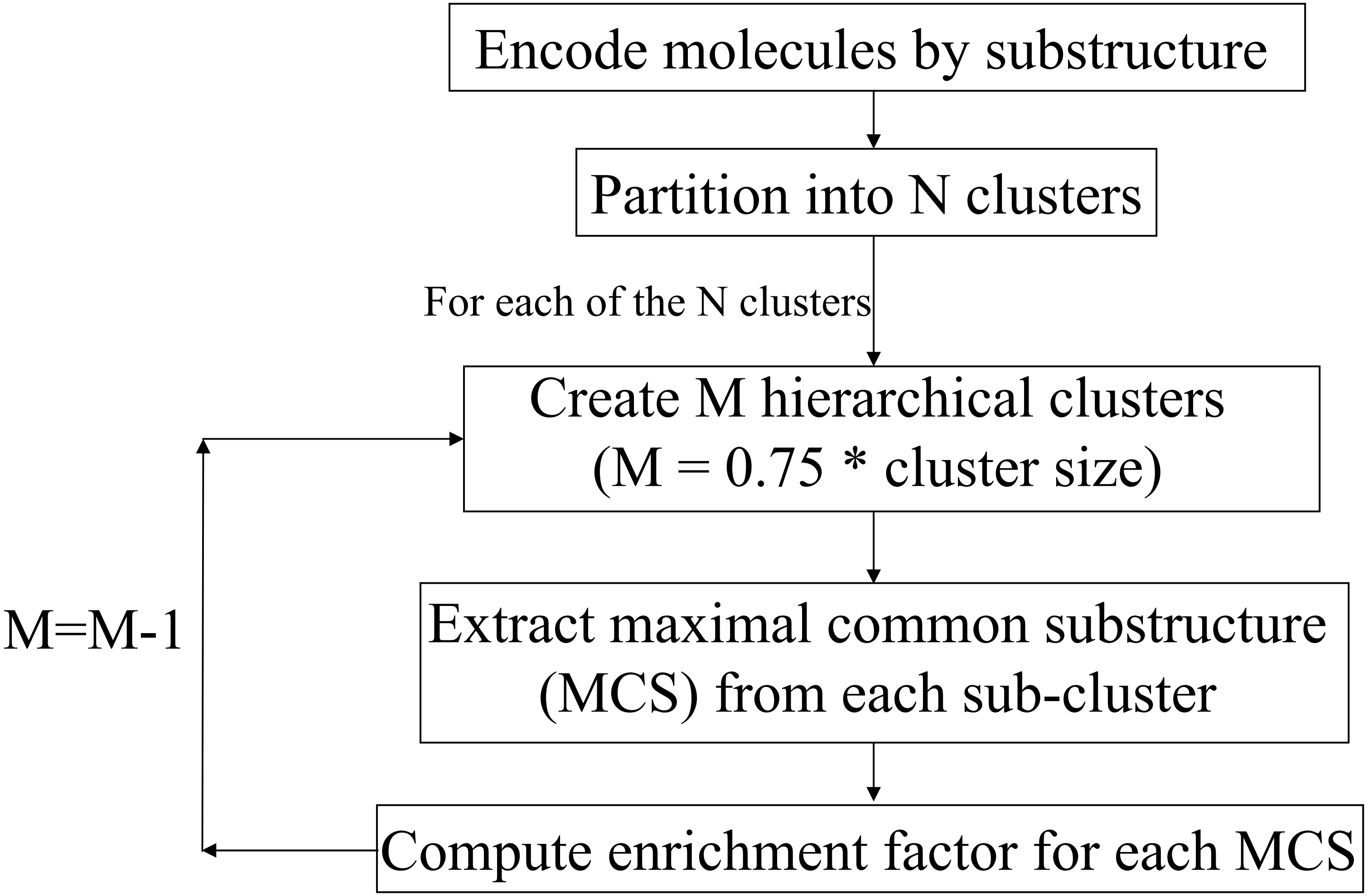
AIM

Identify scaffolds which are maximally enriched relative to activity; avoid bias from initial cluster definitions.

METHODOLOGY

Use automated scaffold perception to search clusters systematically; report those which are significantly enriched.

Enriched scaffold perception - workflow



Encode molecules by substructure

Partition into N clusters

For each of the N clusters

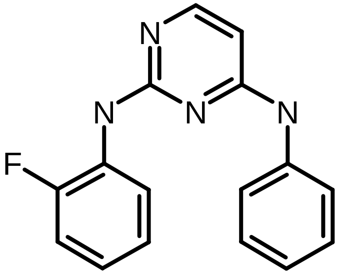
Create M hierarchical clusters (M = 0.75 * cluster size)

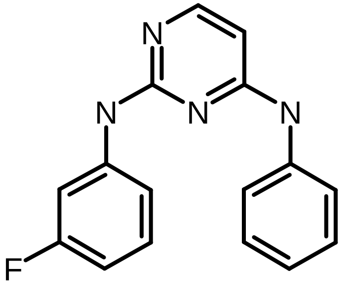
Extract maximal common substructure (MCS) from each sub-cluster

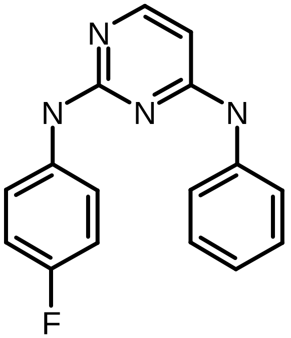
Compute enrichment factor for each MCS

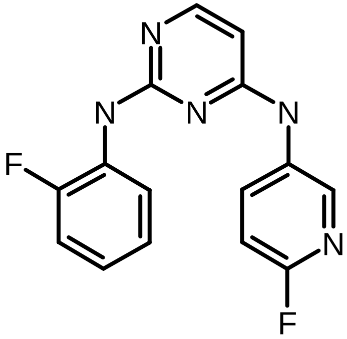
M=M-1

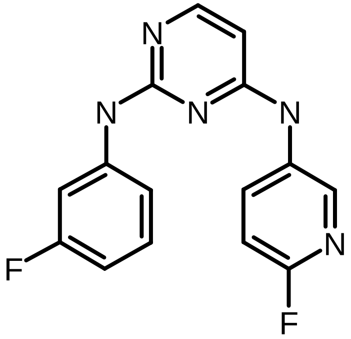
An illustration – defining scaffolds

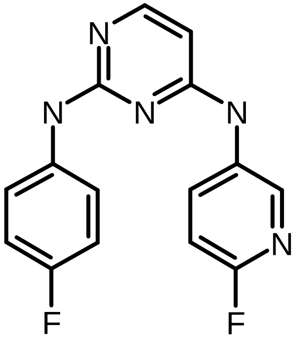


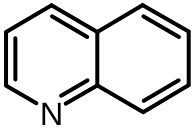


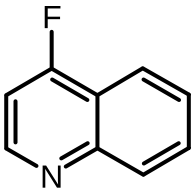


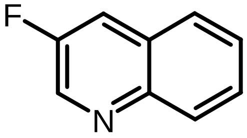




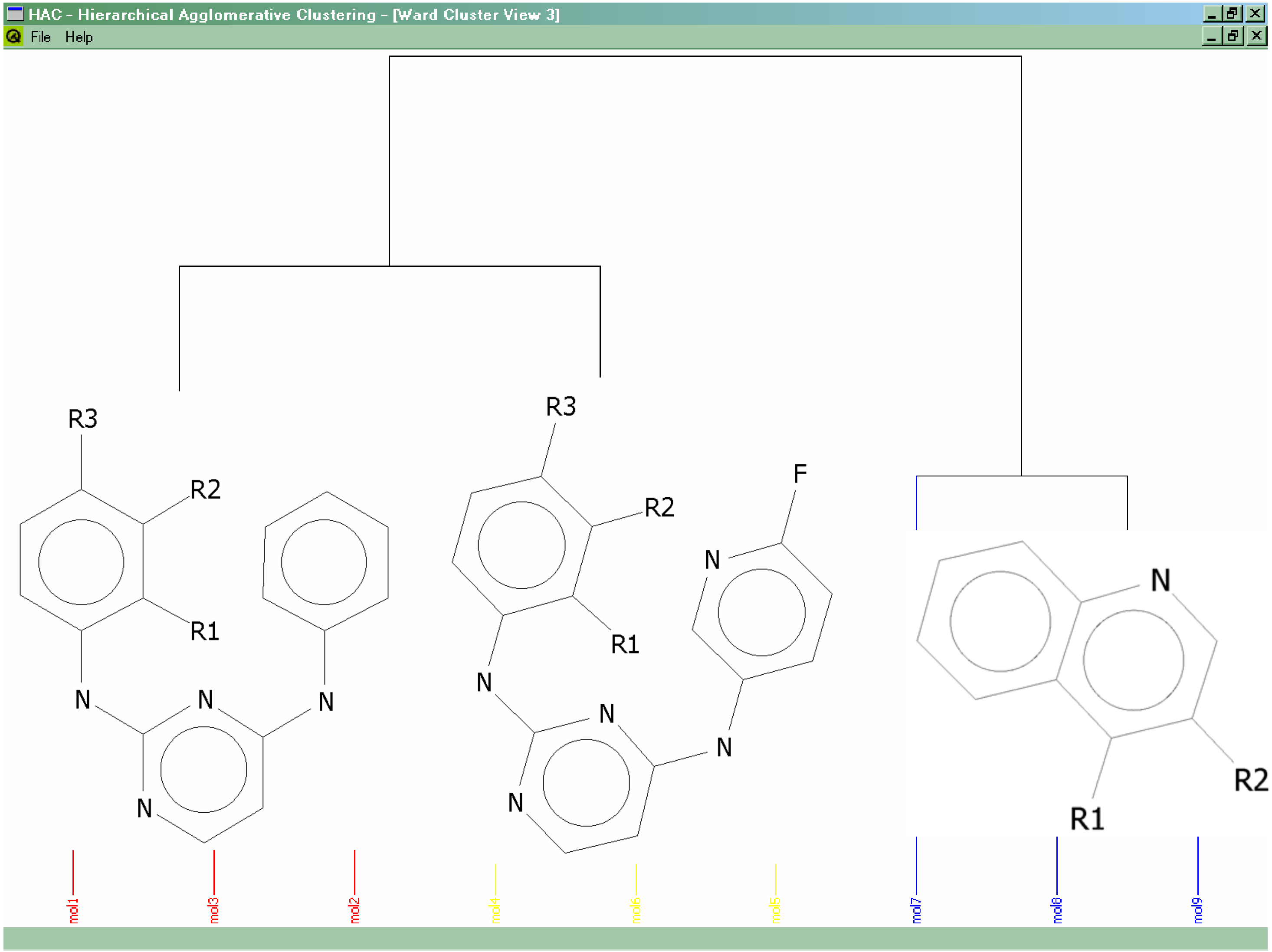




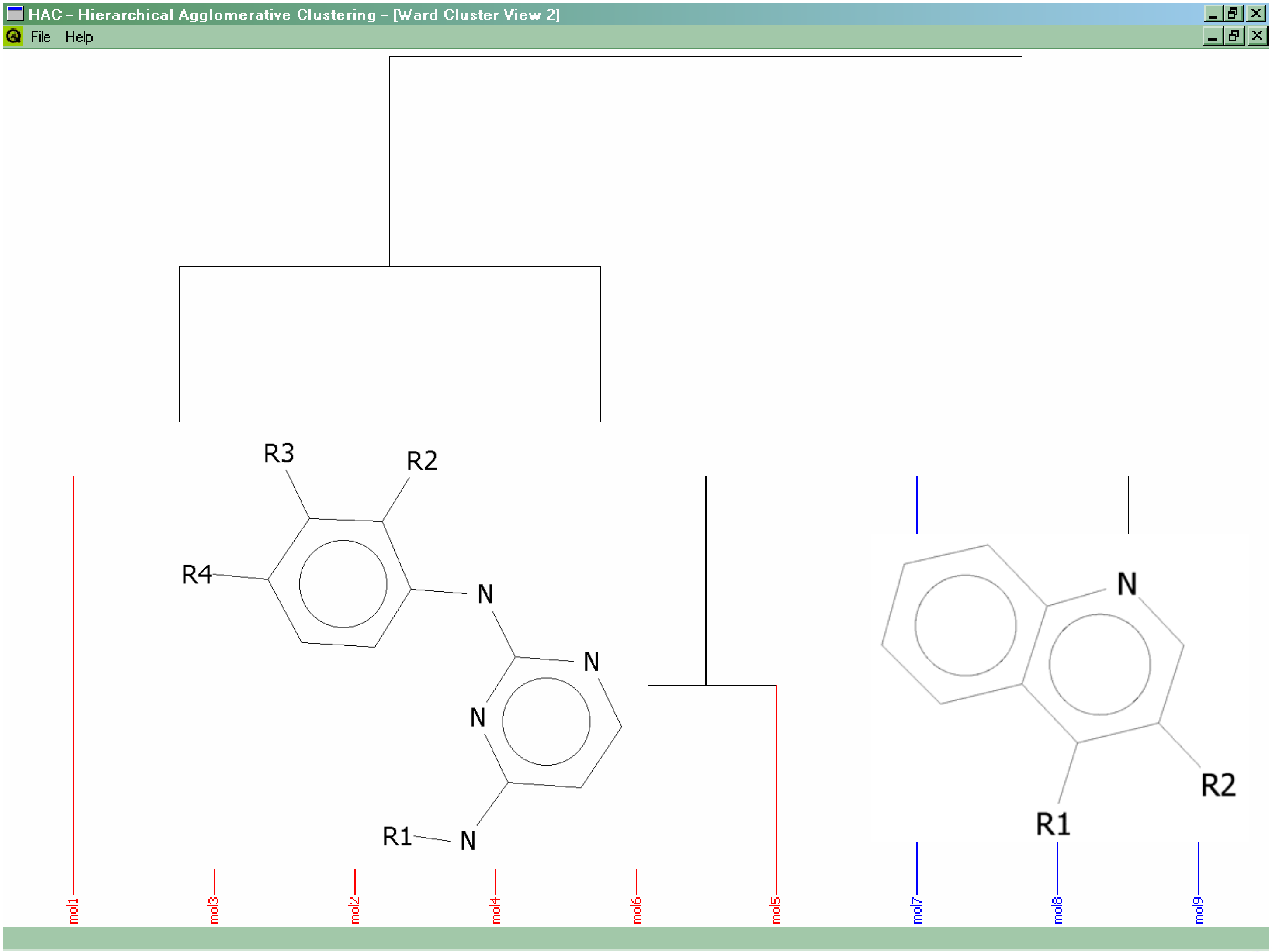




Three clusters

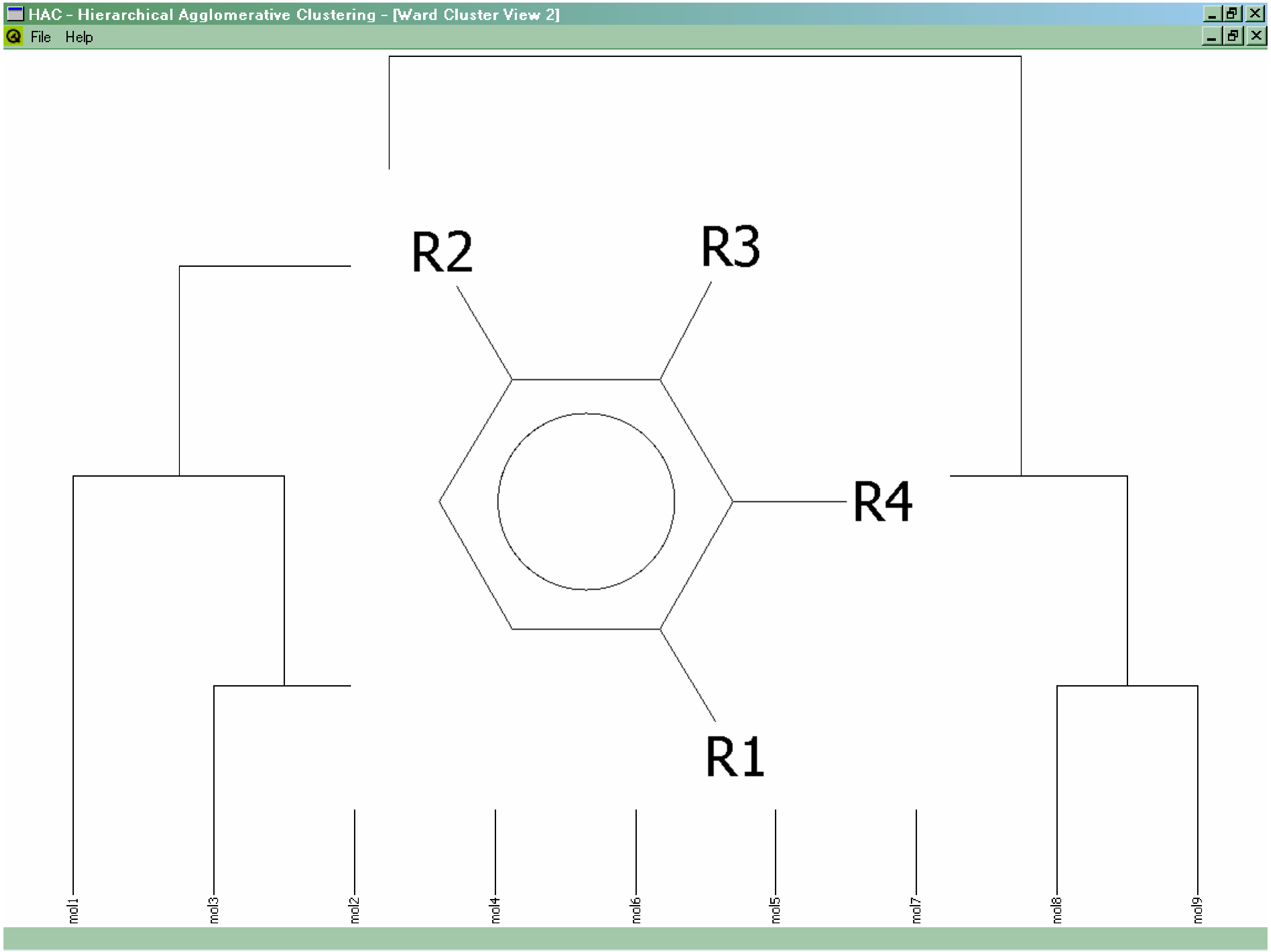


Two clusters



HAC - Hierarchical Agglomerative Clustering - [Ward Cluster View 2]
File
Help
R3
R2
R4
N
N
N
R1
N
N
R1
R2
mol1
mol3
mol2
mol4
mol6
mol5
mol7
mol8
mol9

One cluster



HAC - Hierarchical Agglomerative Clustering - [Ward Cluster View 2]
File
Help
R2
R3
R4
R1
mol1
mol3
mol2
mol4
mol6
mol5
mol7
mol8
mol9

Definition of enrichment

$$\boxed{Enrichment = \frac{N_{Scaffold}^{Actives}}{\left\langle N^{Actives} \right\rangle}}$$

$N_{Scaffold}^{Actives}$ = Number of actives containing scaffold

$\left\langle N^{Actives} \right\rangle$ = Expected number of actives

$\left\langle N^{Actives} \right\rangle$ is defined by binomial distribution

Binomial cumulative distribution, F

- Select N compounds at random from the HTS data
- P^A - probability of selecting an active

$$P^A(X) = F\left(X \mid N, P^A\right)$$

- $P^A(X)$ - probability of selecting X actives

Inverse binomial function

The inverse binomial distribution function will give the expected number of actives in a random selection of N compounds

$$\left\langle N^{Actives} \right\rangle = F^{-1}\left(N^{Total}, P^{A}, \alpha\right)$$

N^{Total} is the total number of compounds which contain a particular scaffold.

α – significance level

Confidence and climate change

The IPCC "Summary for Policymakers" uses the following definitions for confidence limits:

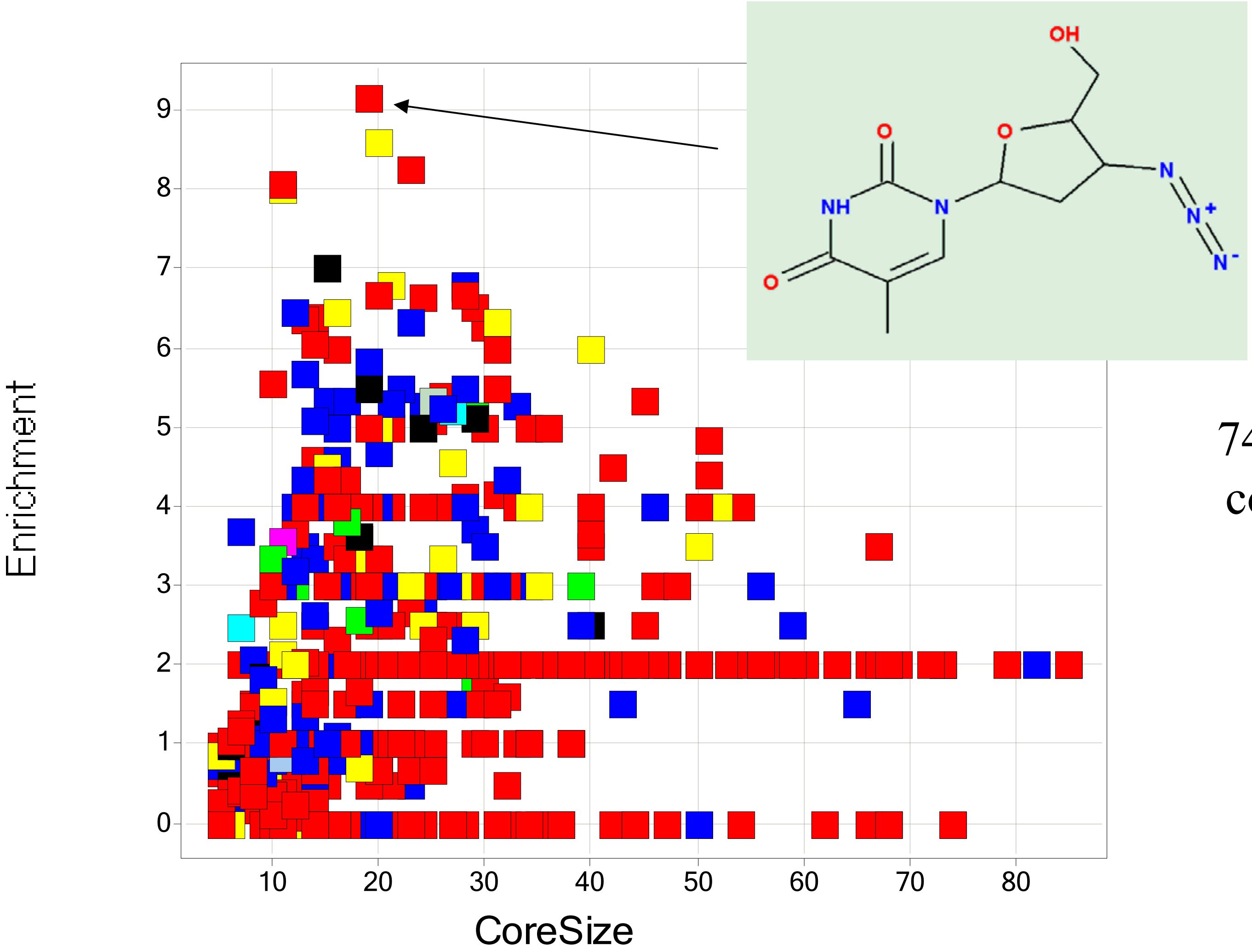
- *Virtually certain* > 99 per cent probability of occurrence
- *Extremely likely* > 95 per cent
- *Very likely* > 90 per cent
- *More likely than not* > 50 per cent
- *Extremely unlikely* < 5 per cent

We want to be "virtually certain" that scaffolds are genuinely enriched – i.e. $\alpha < 0.01$

NCI HIV dataset

- Extracted from PubChem in March 2007
- 41440 compounds with activity data in NCI AIDS antiviral assay (link)
- 1485 confirmed actives
- 276 compounds with “Unspecified activity” removed from dataset

NCI HIV – enriched scaffolds

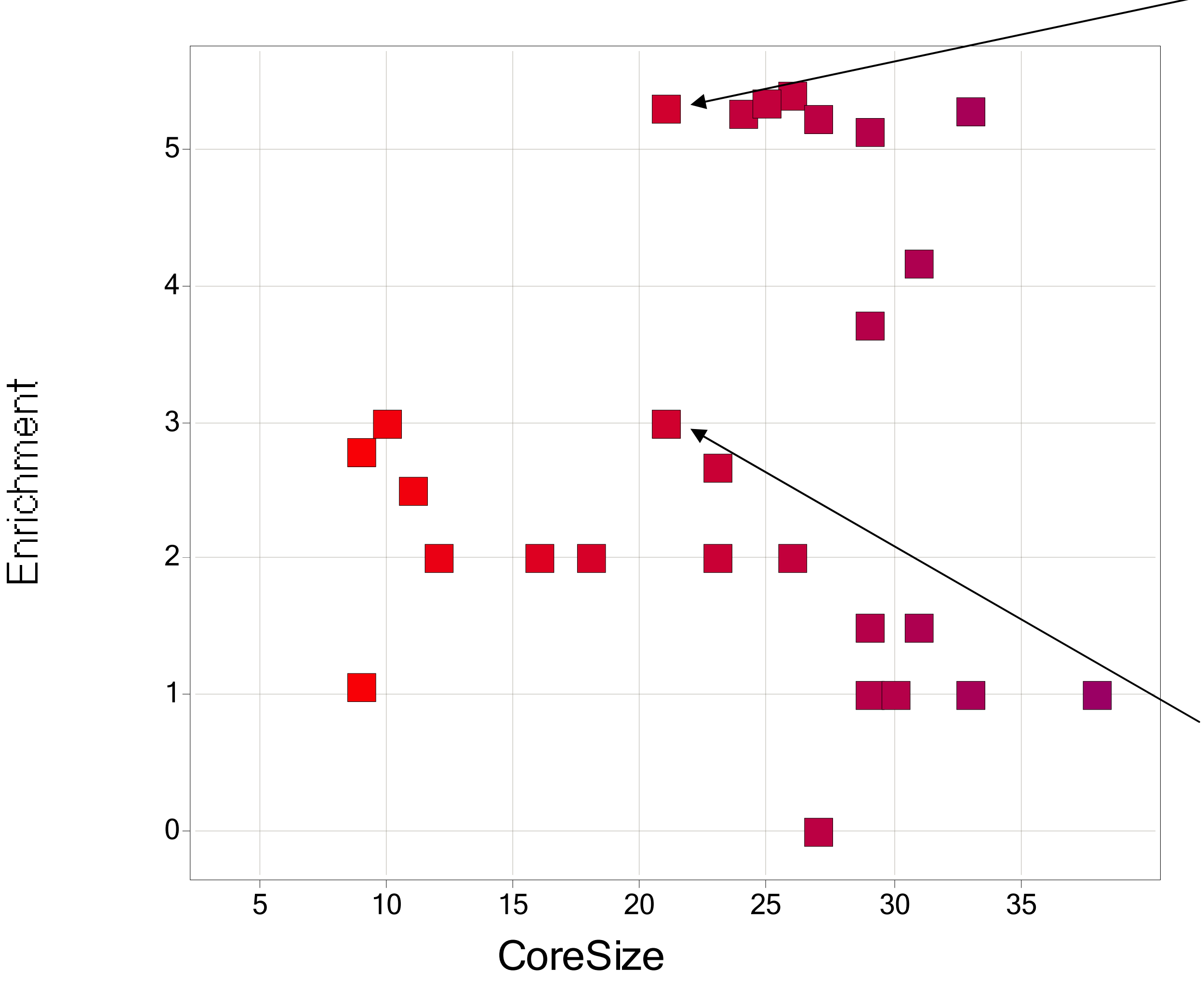


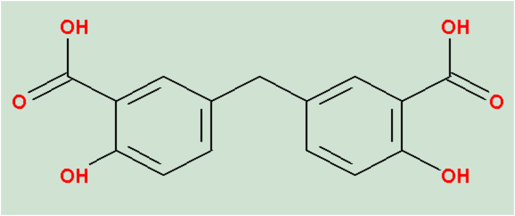
74 compounds tested contain this scaffold

↓

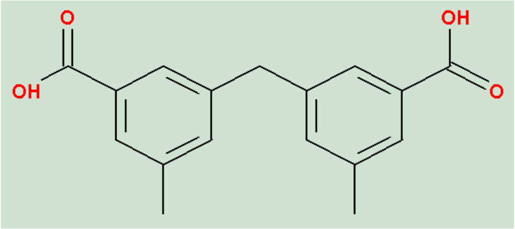
64 are active

Emergent SAR – cosalane scaffold





Substitution meta to the acid groups reduces enrichment factor.



Born under a bad sign?

In a study of 10,674,945 residents of Ontario, based on hospital admissions data:

> "…Sagittarians had a higher probability of humerus fracture…"

The more tests you make, the higher your chance of generating a spurious result.

P.C. Austin et al. **"Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health"** J. Clin. Epid. (2006) **59(9)** 964-969.

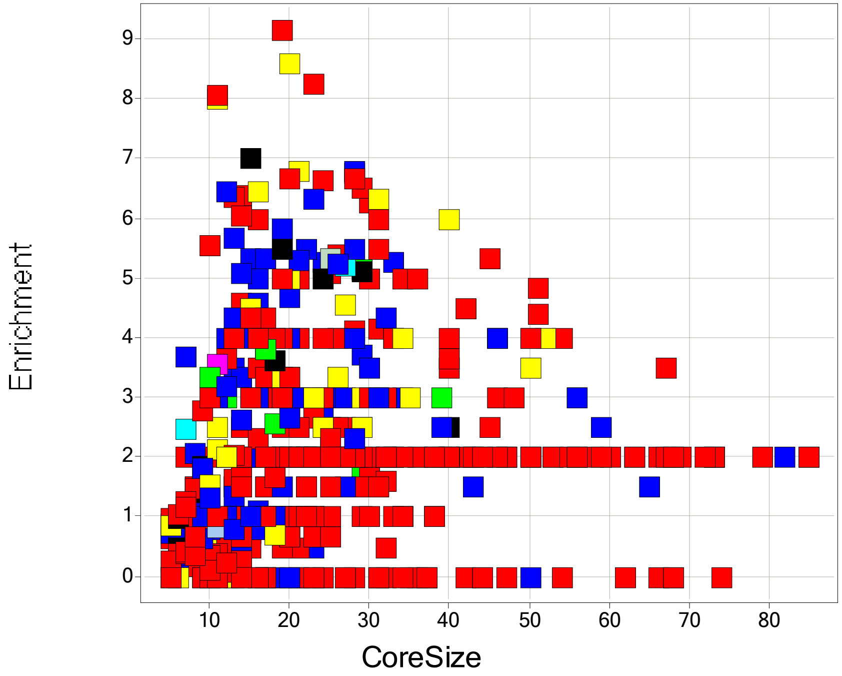
Bonferroni correction

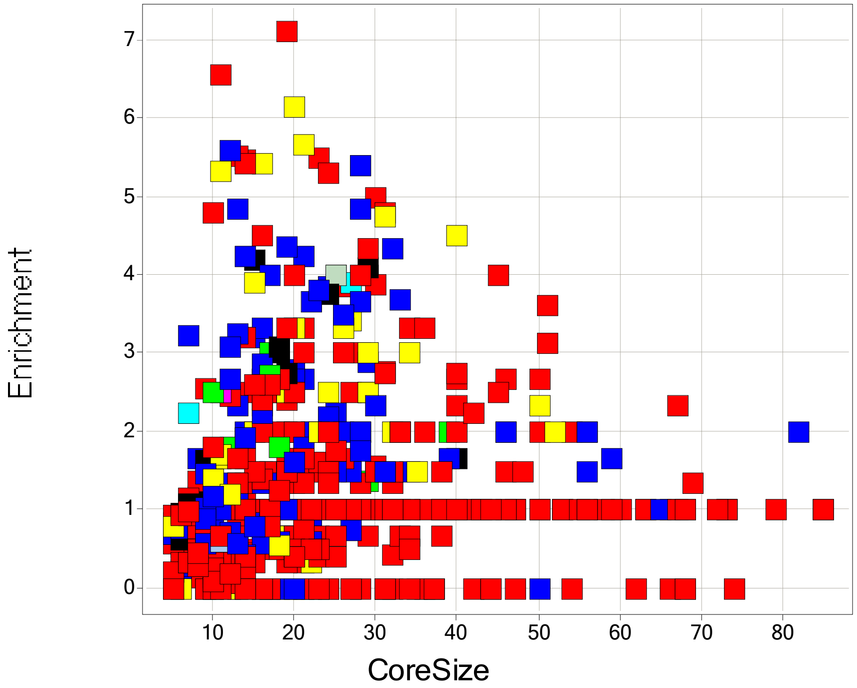
$$\alpha[\text{per test}] = \frac{\alpha[\text{per family of tests}]}{\text{N}}$$

N is the number of tests performed within a family of tests (e.g. testing each star sign against a data set requires N = 12).

When we partition into clusters and extract a scaffold, α is corrected for the number of partitions made.

NCI HIV enriched scaffolds – effect of Bonferroni correction



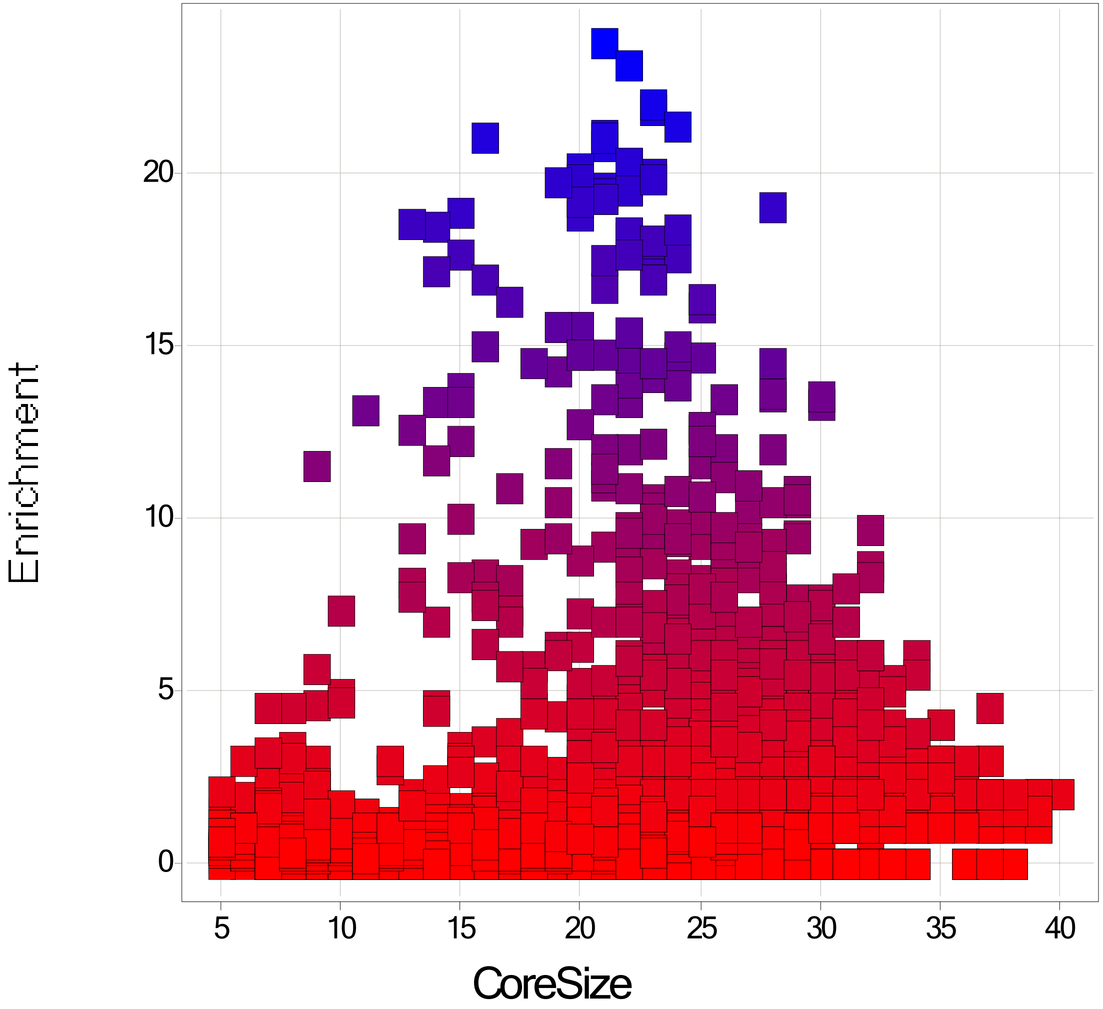


Enrichment values are reduced; the large set with enrichment of 2 now fall to 1, suggesting that these are not significant scaffolds.

What about larger data sets?

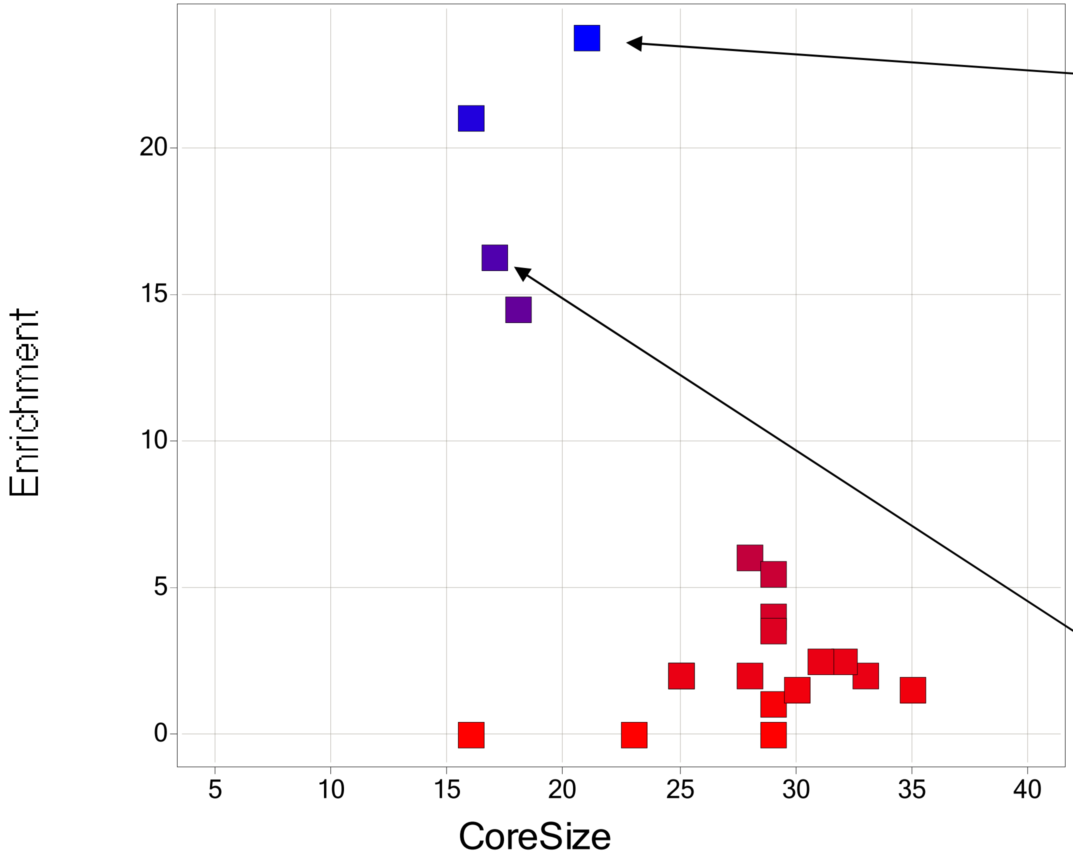
- In-house kinase HTS
- 540,000 compounds tested
- 6737 actives: > 30% inhibition at 10 uM
- Actives partitioned into 200 clusters

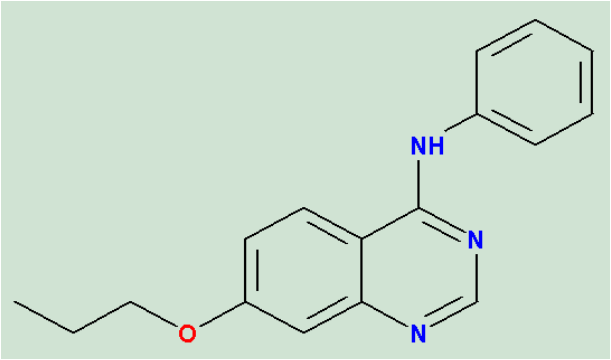
Enriched scaffolds – kinase target



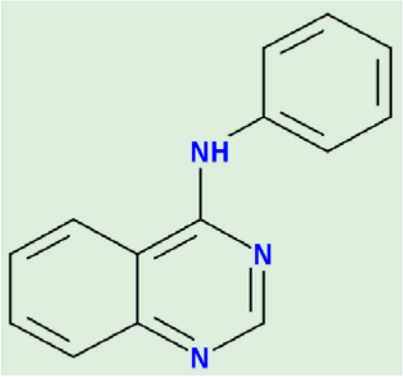
- Enrichments much higher than for HIV set
- AZ collection contains series which target kinases
- Overall bell-shaped plot is also observed in most other cases.

Emergent SAR for quinazoline scaffold





Substituent at the 7-position enhances enrichment.



Conclusions

- Enriched scaffolds can be mined from very large data sets
- Exhaustive, hierarchical approach ensures that maximally enriched scaffolds are located
- SAR emerges from HTS data
- In conjunction with other tools, scaffolds can inform chemical decision making in hit explosion

Technical details

- Scaffold perception – OEChem toolkit

- Data analysis and clustering – MATLAB

- Similarity searching and cluster visualisation – AZ in-house toolkit (Dave Cosgrove)