

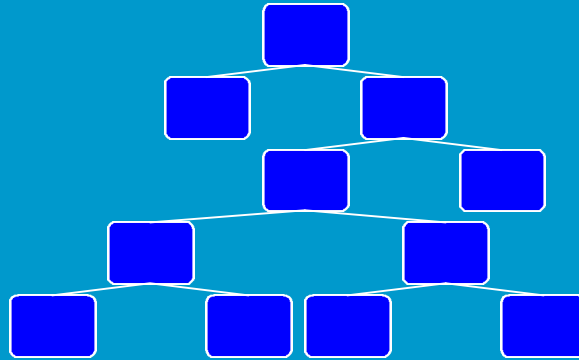
QSAR's using HTS data

Gavin Harper

GlaxoWellcome

gh75680@ggr.co.uk

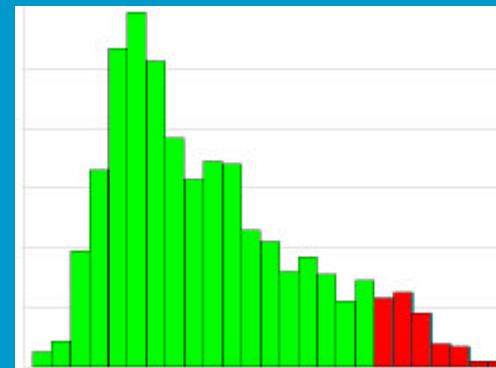
Recursive Partitioning



- Descriptors (continuous/discrete/binary)
 - splits based on descriptors
- Responses (continuous/discrete/binary)
 - try to make splits to “separate out” different kinds of response

HTS screening data

- Response
 - continuous (percent inhibition)
 - could be “made” binary (active/inactive).
- Descriptors
 - we choose to use binary structural descriptors (atom pairs and topological torsions)
- SCAM - use continuous response



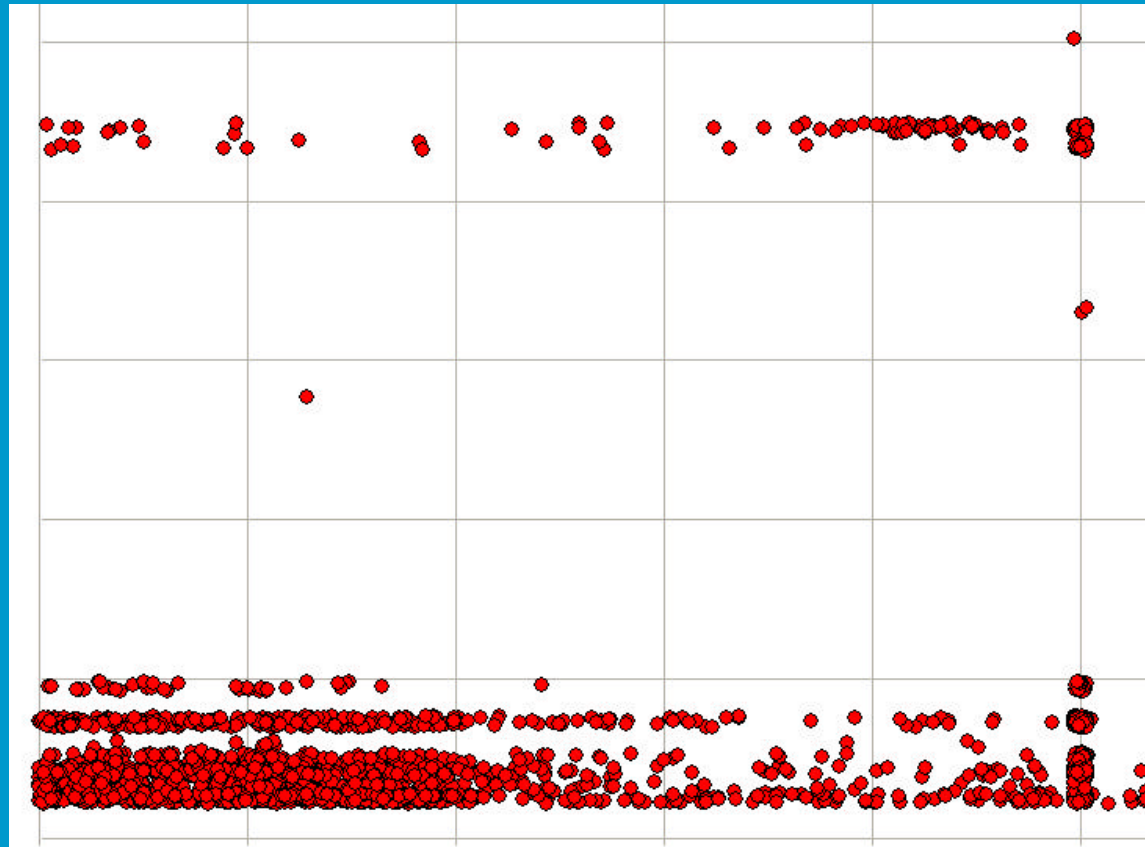
SCAM split-rule (t-test)

$$t = \frac{\bar{y}_{\text{right}} - \bar{y}_{\text{left}}}{\text{variance term}}$$

- Dependent on MAGNITUDE of difference
 - Less likely to pick up weakly active compounds

Most actives “unexplained”

Predicted Activity
(0-100%)



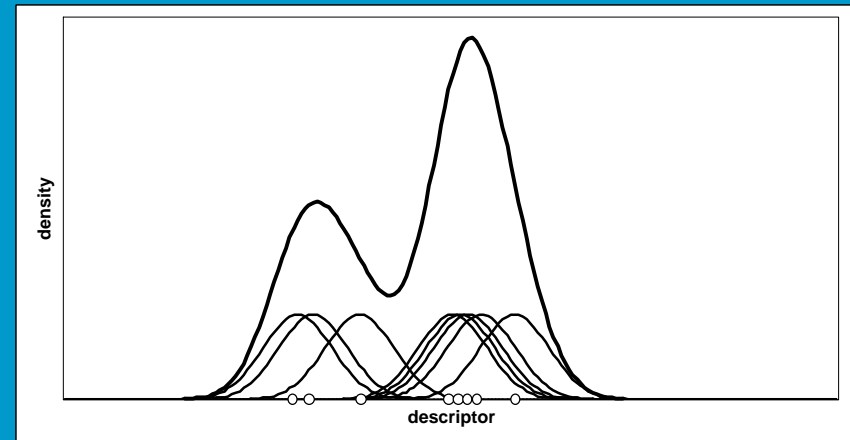
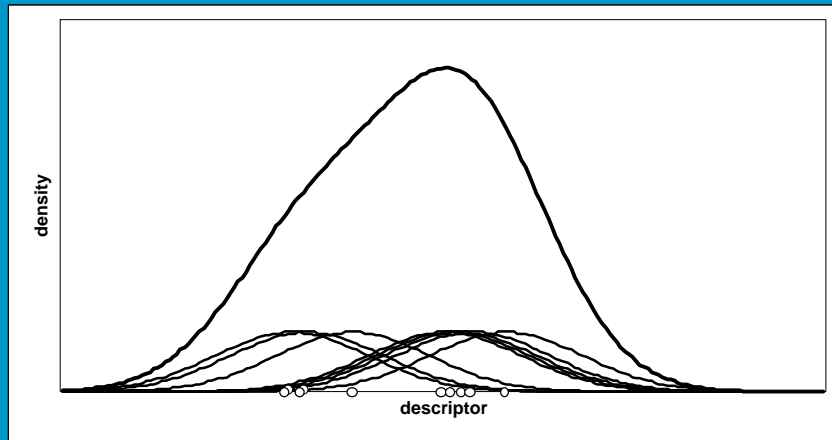
Measured Activity (50-100%)

HTS analysis using binary kernel discrimination

- Kernel density estimate of parent distribution at \mathbf{x} is:

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{\lambda}(\mathbf{x}, \mathbf{x}_i)$$

K_{λ} is the *kernel (density) function*; λ is a *smoothing parameter* that affects the range of influence of each point \mathbf{x}_i in the sample

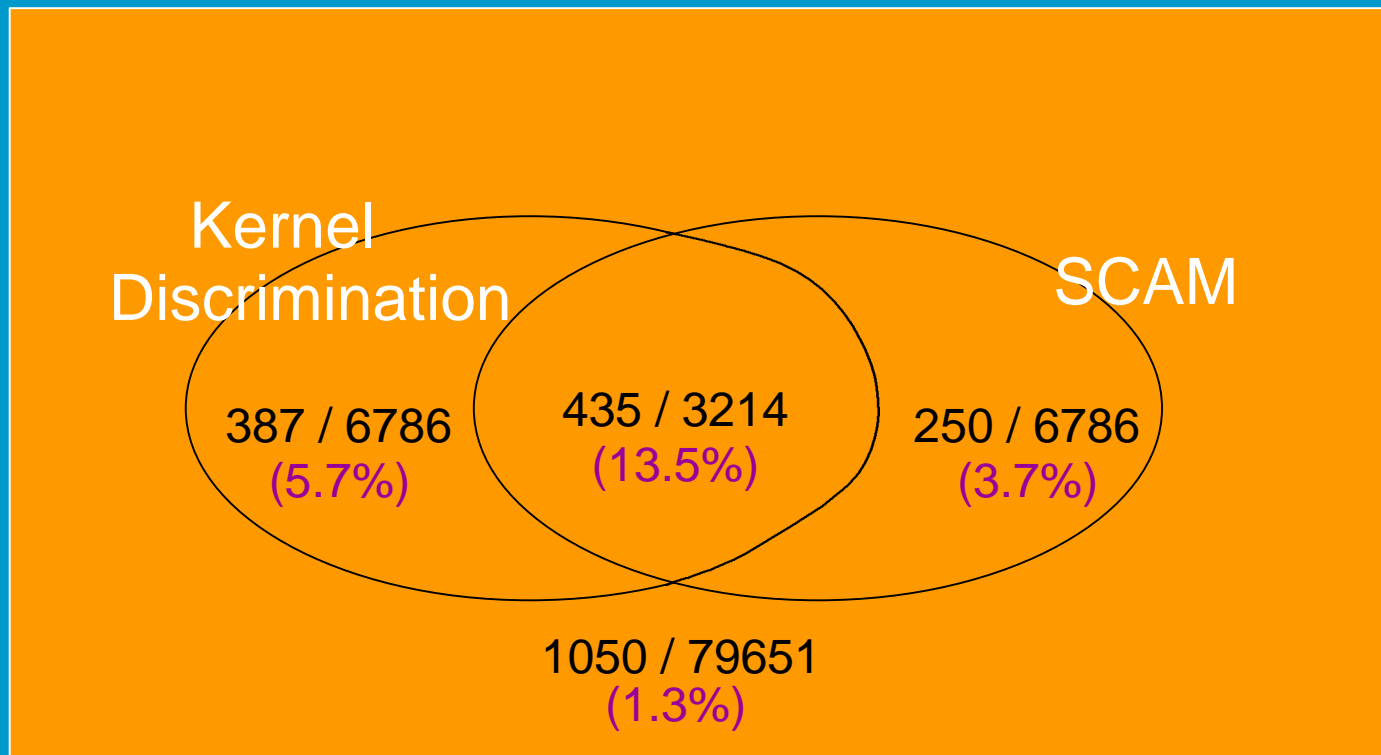


- Aitchison and Aitkin form of K_{λ} for binary data:

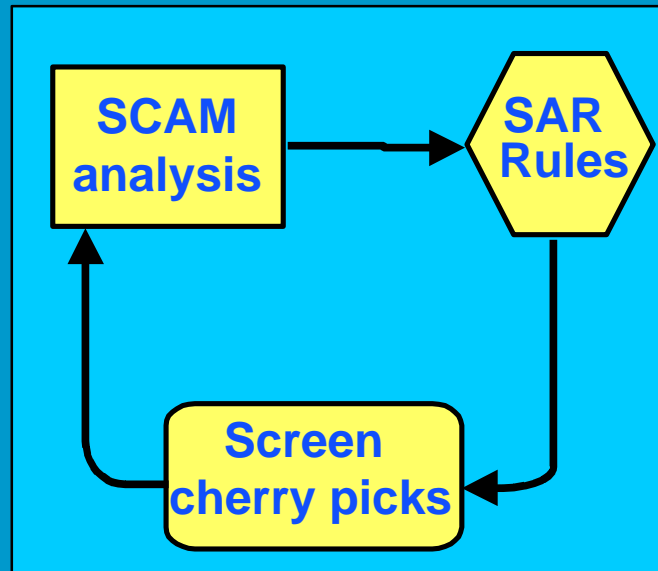
$$K_{\lambda}(\mathbf{x}, \mathbf{x}_i) = I^{M-d} (1-I)^d$$

– \mathbf{x}, \mathbf{x}_i are vectors of length M differing at d positions

Other Methods Find Other Things



Iterative Screening



- WILL focus on “active regions” identified
- WON’T necessarily find all lead series
 - may lose diversity of hits

QSAR on HTS data

- **CAN**
 - Highlight some of the obvious data features
 - Provide a window on the data
 - Be used to improve “hit rate”
- **CANNOT**
 - Spot all the patterns that are there
 - Replace domain knowledge / “intuition”
 - Get the most out of the data

General thoughts on HTS



- We are not trying to find as many “hits” as we can
- We are out to produce a diversity of progressable hits
- This is a SEARCH for the best (few) leads possible
- This is NOT a numbers game - but most quantitative methods are designed to maximize the expected number of hits
- In a sequential search, the overall success depends on gaining information about the space of compounds, as well as finding hits.

Thanks to:

- Andrew Leach, Darren Green, Andy Whittington
 - GW Computational Chemistry, UK.
- Stan Young, Chris Keefer, Deborah Jones-Hertzog
 - GW Chemoinformatics, US.
- John Gittins, Brian Ripley
 - Dept of Statistics, University of Oxford.