gsk GlaxoSmithKline

Design of a Compound Screening Collection

Gavin Harper

Cheminformatics, Stevenage

In the Past...

- Scientists chose what molecules to make
- They tested the molecules for relevant activity

Now...

- We often screen a whole corporate collection
  - 10^5-10^6 compounds

- But we choose what’s in the collection

- If the collection doesn’t have the right molecules in it
  - we fail

“Screen MORE”

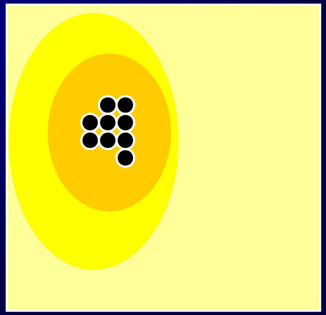
- Everything’ll be fine
- We’ll find lots of hits

- Not borne out by our experience

How do I design a collection? - 1

- Pick the right kind of molecules
  - hits similar biological targets
  - computational (in-silico) model predicts activity at right kind of target for given class of molecules
  - exclude molecules that fail simple chemical or property filters known to be important for “drugs”

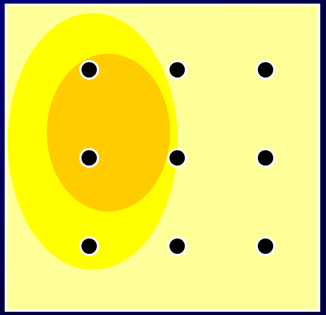
- FOCUS!



How do I design a collection? - 2

- Cover all the options
- Pick as “diverse” a set of molecules as possible
- If there’s an active region of chemical space, we should have it covered

- DIVERSE SELECTION
  - – opposite extreme to focused selection



Basic Idea of Our Model

- Relate biological similarity to chemical similarity
- Use a realistic objective
  - maximize number of lead series found in HTS
- Build a mathematical model on **minimal** assumptions

⇒ How does our collection perform now in HTS?

- relate this to our model

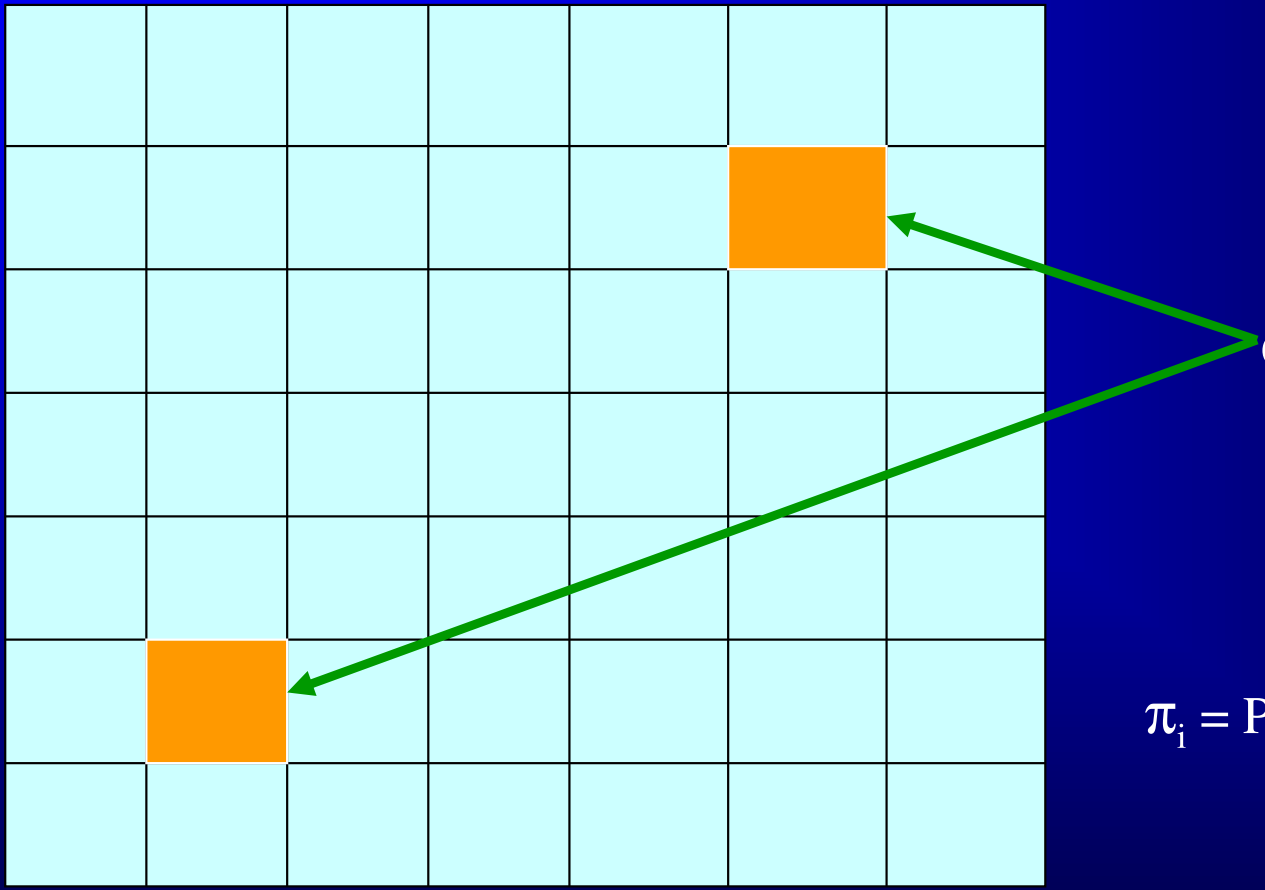
⇒ Learn what we need to make/purchase for HTS to find more leads

A “simple” model

- Chemical space is clustered (partitioned)
  - there are various possible ways to do this

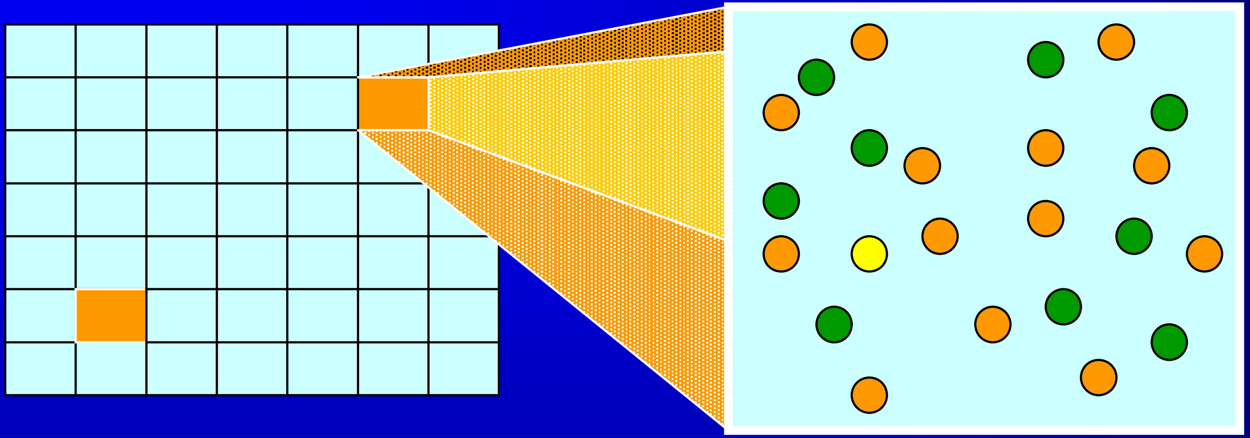
- For a given screen, each cluster *i* has
  - a probability π_i that it contains a lead
- If we sample a random compound from a cluster containing a lead, the compound has
  - a probability α_i that it shows up as a hit in the screen
- If we find a hit in the cluster, that’s enough to get us to the lead

And in pictures...



clusters containing leads

π_i = Pr(box i is orange)



🟢 Hit
🟠 Non-Hit
🟡 Lead

α_i = Pr(dot is green)

Constrained Optimization Problem

- Suppose that we want to construct a screening collection of fixed size *M*

- To maximize expected number of lead series found we have to

$$
\textbf{(P)}\qquad
\begin{array}{ll}
\text{Maximize} & \displaystyle\sum_{i=1}^{p} \pi_i [1-(1-\alpha_i)^{N_i}] \\[2ex]
\text{subject to} & \displaystyle\sum_{i=1}^{p} N_i = M
\end{array}
\qquad N_i \geq 0 \quad (i = 1, \ldots, p)
$$

Solution

$$N_i = \begin{cases} \dfrac{\ln\lambda - \ln\pi_i - \ln(-\ln(1-\alpha_i))}{\ln(1-\alpha_i)} & \text{whenever this is} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

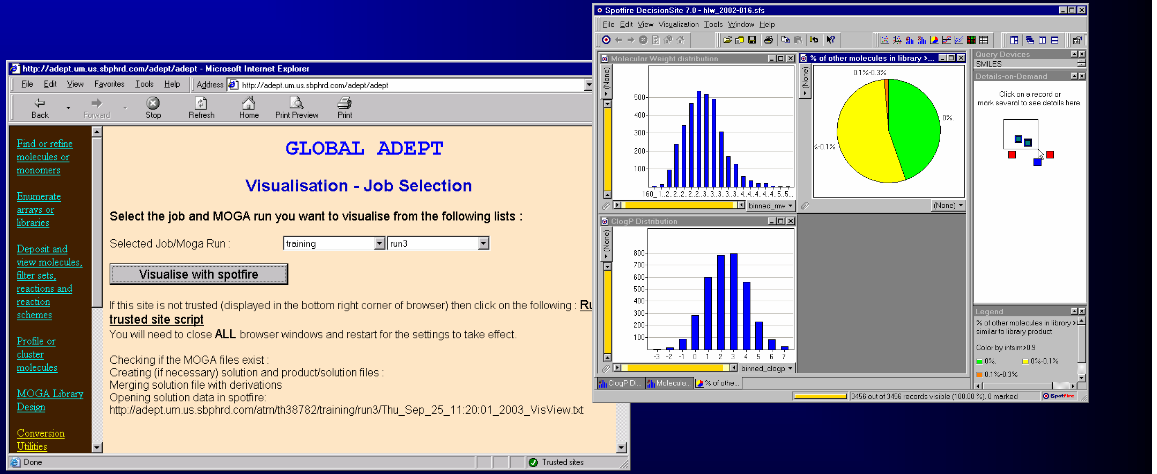
- If we know very little (α_i,π_i equal for all i)
  - select the same number from each cluster - **diversity solution**
- If e.g. we know some clusters are far more likely than others to contain leads for a target
  - select compounds only from these clusters - **focused solution** (filters)
- But we also have a solution for all the situations in between, where there is a balance between diversity and focus

Immediate Impact

- Improved “diversity” score

$$D(\{N_i\}_{i=1}^{p}) = \sum_{i=1}^{p} [1-(1-\alpha)^{N_i}]$$

- Use in assessing collections for acquisition
- We have integrated this score into our Multi-Objective Library Design Package



http://adept.um.us.sbphrd.com/adept/adept - Microsoft Internet Explorer

GLOBAL ADEPT

Visualisation - Job Selection

Select the job and MOGA run you want to visualise from the following lists :

Selected Job/Moga Run : training run3

Visualise with spotfire

If this site is not trusted (displayed in the bottom right corner of browser) then click on the following : **Ru trusted site script**
You will need to close **ALL** browser windows and restart for the settings to take effect.

Checking if the MOGA files exist :
Creating (if necessary) solution and product/solution files :
Merging solution file with derivations
Opening solution data in spotfire:
http://adept.um.us.sbphrd.com/atm/th38782/training/run3/Thu_Sep_25_11:20:01_2003_VisView.txt

Spotfire DecisionSite 7.0 - hlw_2002-016.sfs

Click on a record or mark several to see details here.

% of other molecules in library > similar to library product

Color by intsim>0.9

3456 out of 3456 records visible (100.00 %), 0 marked

* Gillett et al., *J. Chem. Inf. Comp. Sci.* **2002**, *42*, 375-385.

What value should α take?

- Determining a value of α is important. We can cluster molecules using a variety of methods.
- Fortunately, there is a recent paper from Abbott which answers this question
- In 115 HTS assays, with a TIGHT 2-D clustering, $\alpha \sim 0.3$
  - consistent: mostly varies between 0.2 and 0.4
- This agrees well with our experience
- In practice we use this (Taylor-Butina) clustering with radius 0.85 and using Daylight fingerprints

* Martin et al., *J. Med. Chem.* **2002**, *45*, 4350-4358.

- **A consistent value of α is necessary, irrespective of cluster**
- **Otherwise, very difficult to parameterise model accurately**

The Rights of a Molecule

- Every molecule has the right to be treated equally
  - The probability of similar biological activity at similarity x should be the same, independent of bit density (or any other global properties)

- Our limited experience suggests larger molecules may be less likely than small molecules to be active using our 0.85-radius clustering
- Needs further exploration
  - But would we expect this to happen?

Recent papers: bit density vs similarity

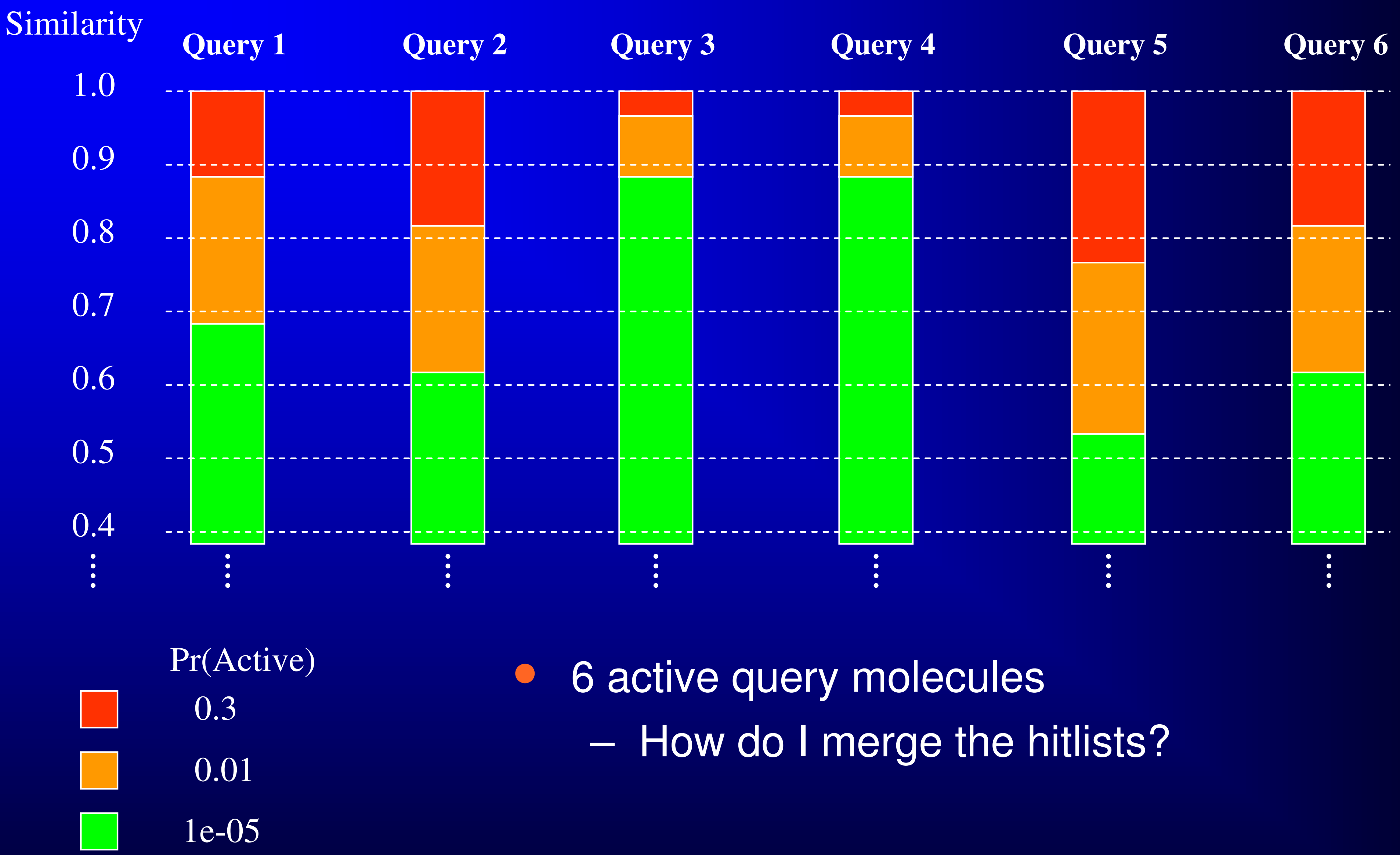
- Flower: JCICS 48, 379-386 (1998)
- Fligner et al. Technometrics 44, 110-119 (2002)*
- Holliday et al. JCICS 43, 819-828 (2003)

- * In Fligner et al., they propose a simple random model.
  - Compare 2 molecules of same bit density:
  - Under model, expected Tanimoto similarity is approx p/(2-p)
    - where p is proportion of bits set
  - More dense bit strings
    - ➢ higher Tanimoto similarity

But it doesn’t just matter for my model!

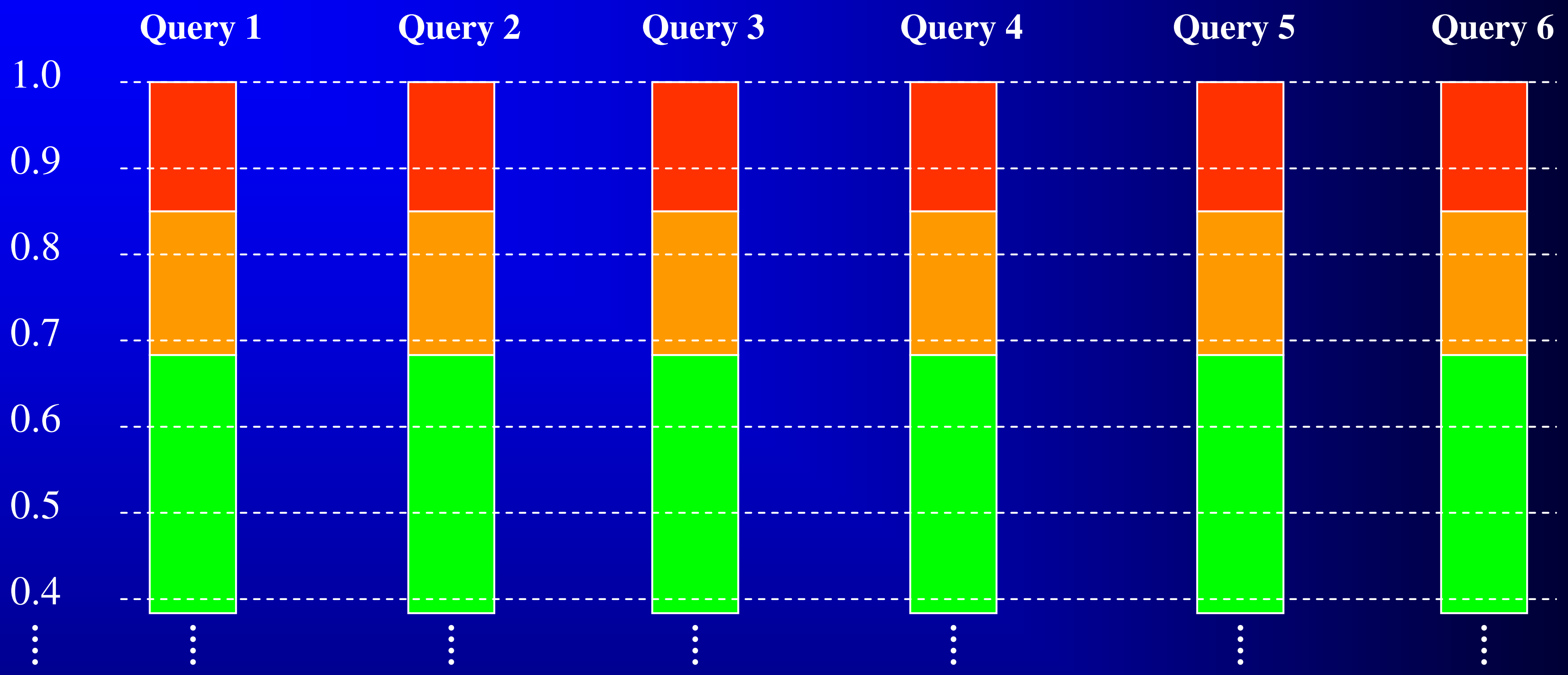
- Papers were mainly concerned with dissimilarity problems
  - Easier to find low bit density compounds with near-zero similarity to existing compounds
    - Sequential dissimilarity-based selection bias

- But consider similarity searching with multiple queries.



- 6 active query molecules
  - How do I merge the hitlists?

Life would be easier if…



- Finally of course
  - Use “the model” to work out which molecules to actually screen
  - It won’t just be the top n if they’re all highly similar to each other

Applications

- Compound acquisition
- Library design
- Strategic Decision-Making Tool
  - Resource allocation - what to buy, what to make.
  - What targets to screen
- Prioritisation of hits in virtual screening
  - Similarity searching
  - Pharmacophore searching?
  - Docking?
- Others?...

Acknowledgements

- Stephen Pickett
- Darren Green
- Jameed Hussain
- Andrew Leach
- Andy Whittington

* Harper et al., *Combinatorial Chemistry and High Throughput Screening* **2004,** *7*, 63-70.