Searching Large Sets of Virtual Libraries in Real Time using Modal Fingerprints

Jens Loesel, Pfizer Sandwich
Nathan Kuroczycki, University Sheffield
Ian G Johnson, University Bath

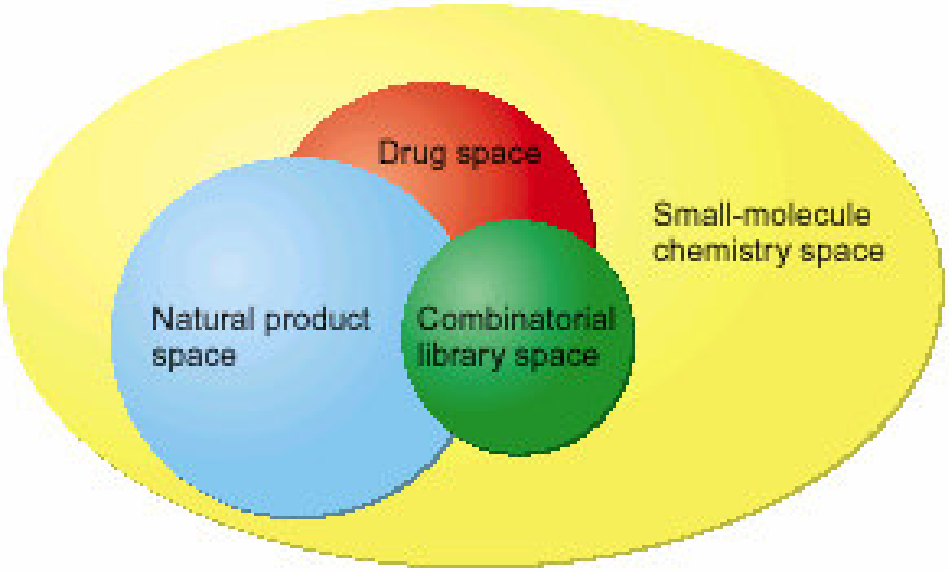
Overview

- Why do we need a fast search method for virtual libraries ?
- How can fast searching of virtual libraries be achieved ?
  - a proof of concept study
- Summary

Development of CombiChem

- Combinatorial Chemistry has become a routine method in pharmaceutical industry
- The portfolio of combinatorial libraries already done or could be done is growing
- Need to follow up non-combinatorial or competitor compounds with combinatorial chemistry
- No single person any longer has an overview of all libraries possible
- Need for a new and fast search method that spans multiple virtual libraries

The needle in the Haystack



- Where do you start searching if your available combinatorial space is > 1 trillion compounds ?

Image Rose&Stevens 2003

The challenge

- We need a simple search method for our portfolio of several hundred libraries
- Based on one or more starting compounds
- Fast results – chemists wait while search is in progress

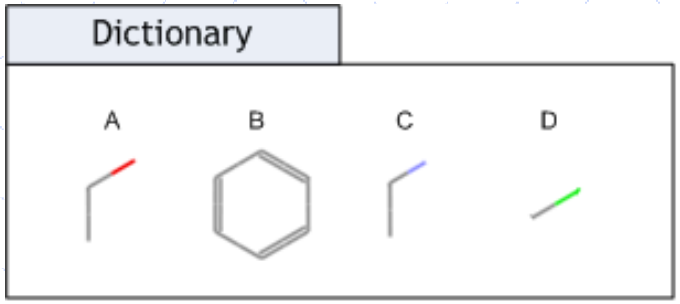
A pragmatic solution

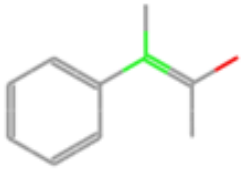
- Technology for an exhaustive search to deliver the closest compound exists
  - but searching and storing > 1 trillion compounds isn't feasible
- Return libraries/protocols instead of compounds
  - Speed up
  - Natural way in which a chemist follows up a lead
- Return a selection
  - give the chemist some choice
- Approximate solution
  - get it right > 8 / 10 cases

Library descriptors

- Summarizing complete libraries in form of a single descriptor offers multiple advantages
  - Minimal storage overheads
  - Potential use of existing software
  - Number of libraries << number of compounds
- The only question is – how much information do you lose – is such a descriptor good enough for daily work

Are Modal Fingerprints the solution?





1	1	0	1

Training set of actives:

mol 1: 1001011100001010
mol 2: 0011010000011000
mol 3: 1101010001111101
mol 4: 1011011101010010
mol 5: 0100111100011101

→

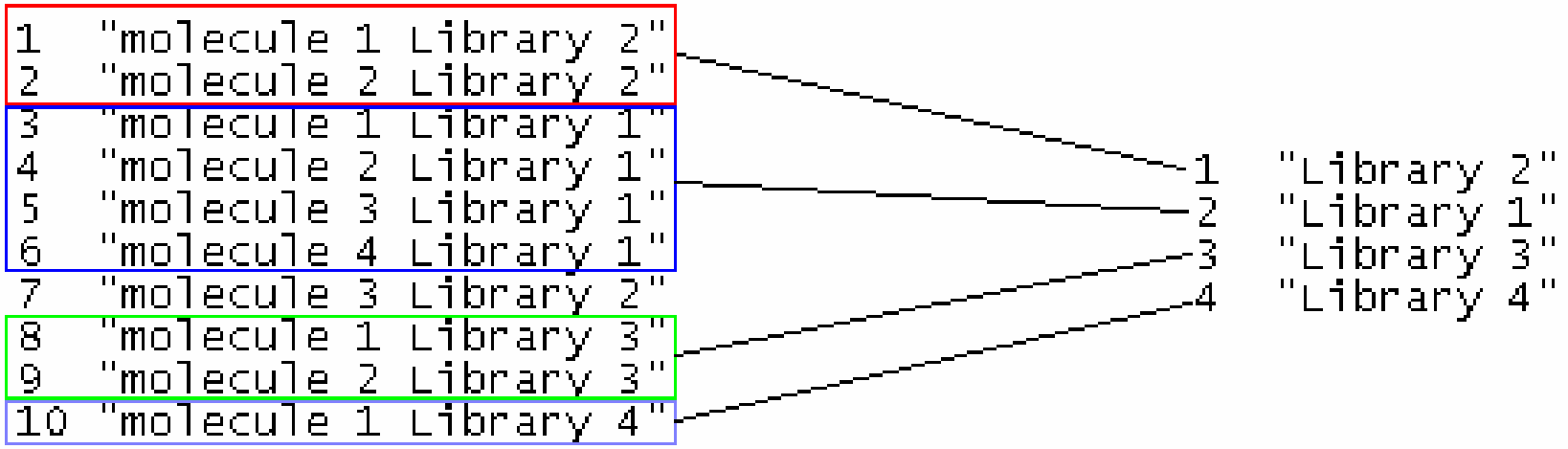
Modal at 40%:

1111011101011111

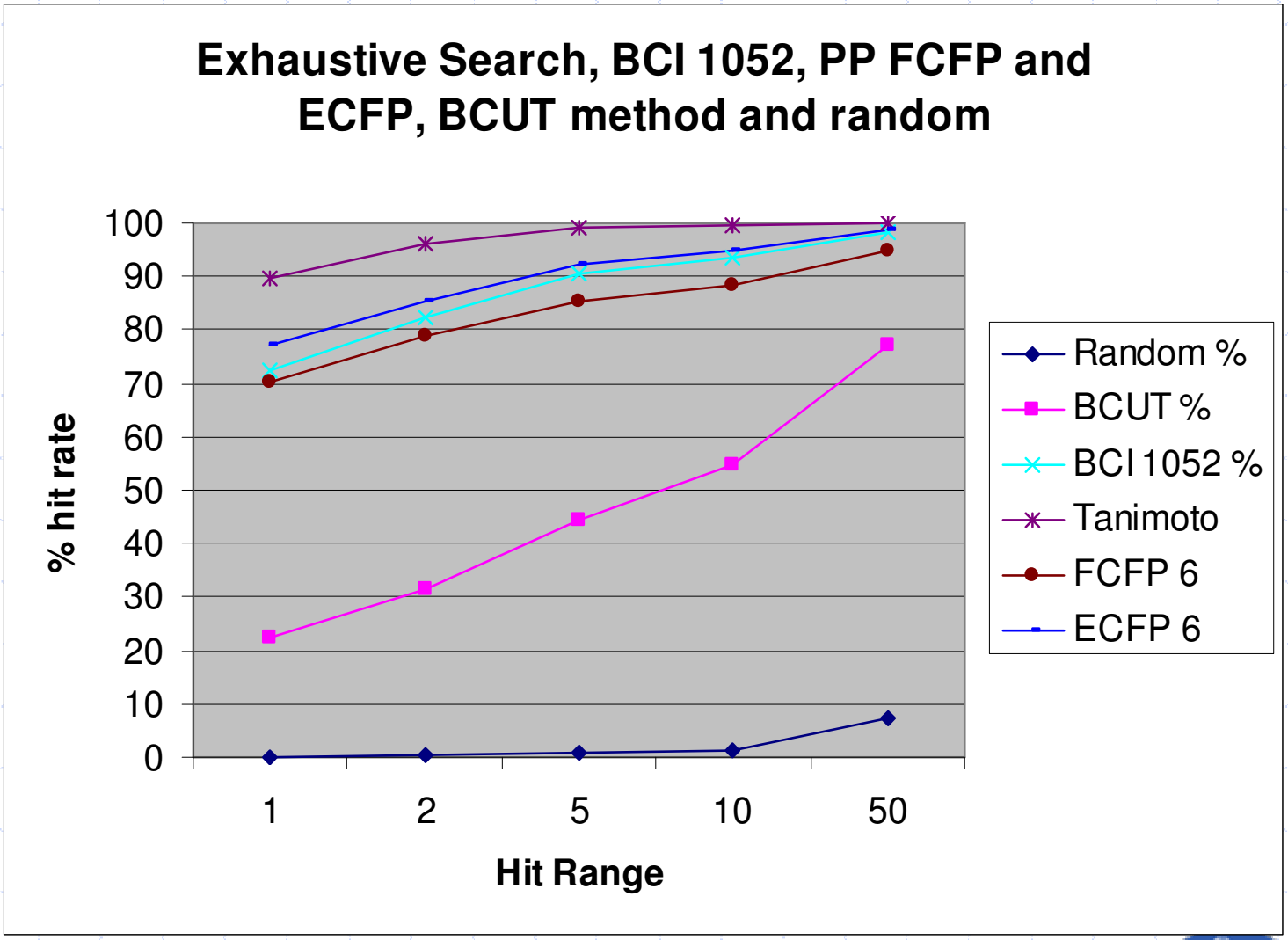
Testing the success rate

- Generate different library descriptors for several hundred synthesized libraries > 200 compounds
- Pick a random compound out of each library and use as search probe
  - Do this 10 times over to minimize random effect
- Do a similarity search between probe and libraries
  - Use similarity score to ranking the libraries
- Compare the results to an exhaustive search

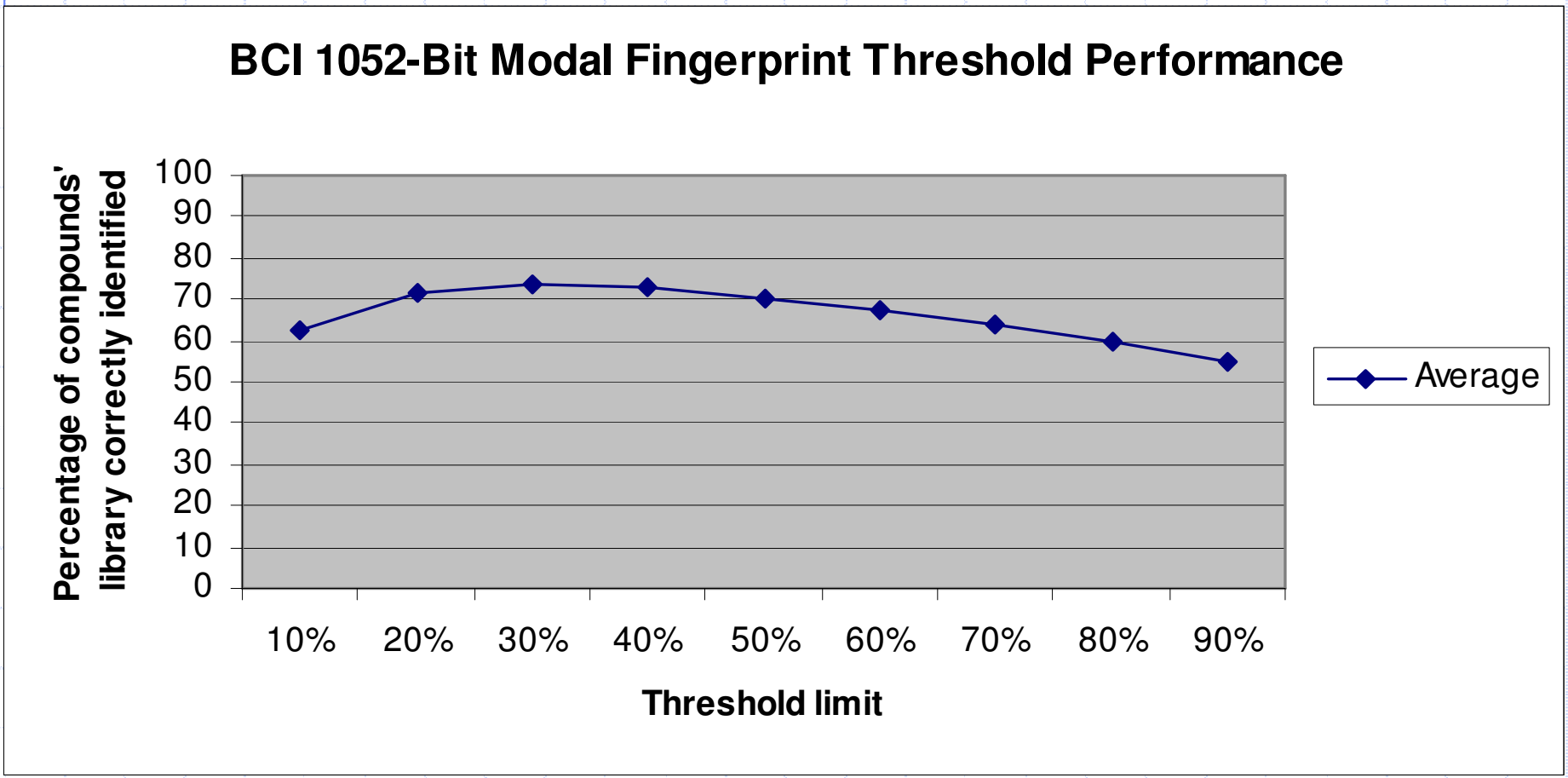
Ranking libraries in the exhaustive search



Recall rates for correct library



Dependence on cut-off



Good enough for daily work ?

- Hit rate for exhaustive search < 90% for top hit
- Best modal fingerprint reaches 94.8% for top 10 hits
- Other top hits are often related
  - Possibility for library hopping
  - Choice for the chemist
  - Manual inspection yields on average > 3 libraries the chemist likes
  - Several hits not obvious from Markush structure / reaction scheme

Summary

- Libraries descriptors have been generated and shown to be successful in library searches
- Search time proportional #libraries
  - allows one to search a virtual library space of > 1 trillion in a sub second timeframe
- The increased search speed outweighs the information loss compared to exhaustive searches
- The descriptors are stable
  - choice of fingerprints, compounds chosen and % cut-off show only minor differences in performance

Outlook

- Follow on search with ‘classic’ methods
- Comparison with fingerprints for Markush structures would be of interest
- Inter-library clustering and similarity searches possible
- Application of modal fingerprints to other sets of molecules
  - Clusters, gene classes, ???