# Use of Chemical Similarity in Drug Discovery

Darko Butina
ChemoMine Consultancy

# Basic principle of medicinal chemistry

- Similar structures exhibit similar behaviour
  - If interaction with 5-HT receptors needed, design ‘5-HT-like’ molecule
- While it is not surprising when two very similar molecules show very similar activity, when they do NOT, one have possibly found the famous ‘methyl’ substituent that made uM hit into nM lead (or most likely the other way around)
  - or check experimental data

# Calculating Chemical Similarity

- Two parameters needed
  - Descriptors (fingerprints)
  - Similarity index
- Objective of this talk is use of
  - Daylight fingerprints
  - Tanimoto similarity index

# References to Daylight Fingerprints and Tanimoto index

- MUG ’98: Adding non-structural data into fingerprints (John Bradshaw)
- Tanimoto similarity index

$$c/(a+b-c)$$

Where

a is number of bits on in molecule A

b is number of bits on in molecule B

c is number of bits in common between A and B

Two molecules that are identical have Tanimoto = 1, while if nothing in common, Tanimoto = 0

# Basic science behind Daylight fingerprints

- The fingerprinting algorithm examines the molecule and generates the following:
  - Pattern for each atom
  - Pattern representing each atom and its NN (plus the bonds that join them)
  - ..continuing with path up to 7 bonds away (default)
- Not possible to assign a particular bit to each pattern, as with structural keys
  - each pattern serves as a seed to a pseudo-random generator (‘hashed’) and set of bits produced is added using logical OR to the fingerprints

# Two important parameters to set ‘properly’

- How many bonds away (path) to generate patterns
- Size of the fingerprints
  - 64, 128, 256, 512, **1024, 2048**, ….

# dbclus clustering algorithm

- Darko Butina, J.Chem.Inf.Comput.Sci., 1999, 39, 747-750
- Clustering algorithm based on Daylight fingerprints and Tanimoto similarity index
- Clustering is driven by tanimoto similarity to be achieved ***within*** the clusters as the only argument to the algorithm
- Used throughout this presentation, including its use as
  - similarity matrix calculator,training/test set selection, design of representative sets, data pre-processing before QSAR work, quality control of experimental in vitro data

# History of development of dbclus

- Need for design of representative set for GW (GSK now) for 200,000 molecules in the liquid store
- Use of Jarvis-Patrick algorithm have produced clusters of variable quality, in terms of similarity within the clusters, depending on the parameters used
- J-P clustering needed lot of human intervention to obtain best results
- ***Medicinal chemists have responded well to 'relationship' between given tanimoto level and overall structural similarity in Merlin searching***

# Basic principle of dbclus algorithm:

‘many Merlin searches in an automated manner with Tanimoto similarity index controlling clustering algorithm’

# Key steps in dbclus

- From the first molecule in the list, to the last do:
  - First molecule becomes centroid ‘0’
    - Calculate T-similarity for each molecule to centroid_0
      - If T-similarity >= given tanimoto (T) – label molecule as cluster_0 member
        - Else get next molecule
    - Go back to top of the file and find first molecule that has not been ‘labelled’
      - Label that molecule as centroid_1
        - Go down the list and if molecule not labelled do as above
    - When finished
      - All molecules that have not been labelled mark as singletons

# dbclus output

name cl_0 0.9
name cl_0 0.95
name cl_0 (2)
name cl_1 0.9
name cl_1 0.95
name cl_1 0.99
name cl_1 (3)

The bottom two lines in the output report number of clusters and singletons found for the set at the given T

# Order independent dbclus algorithm

- Molecules with largest numbers of nearest neighbours (NN) are potential cluster centroids
- Step 1 in dbclus:
  - Calculate half similarity matrix for the whole set (at given T-level)
  - Sort set by number of NN that each molecule have – largest at the top
- Start dbclus algorithm

# Final points about dbclus

- Similarity within cluster achieved
  - At least T-range between cluster centroid and its members
  - Maximum distance between ay two members within the cluster is $(\text{T-range})^2$
- Current version of dbclus will cluster 100,000 molecules in 8 minutes (single processor laptop running RedHat v8.0)

# Rule of thumb with Daylight fingerprints and Tanimoto

- T-range 0.9 – 0.95 will put together Me, Et, Pr – ‘SAR type subsets’
- T-range 0.7 for representative sets
  - Detects ‘chemotypes’
    - Indoles
    - Steroids
    - piperidines

# Slide with clusters at 0.9

- Linux2windows
  - Prado → .ps
  - ps2pdf (multi pages display for windows ad printout)

# Daylight fingerprints default settings

- Size:
  - Start at 2048 and fold down till density 30-40%
- Path:
  - From 0 – 7 bonds

# Two key patterns to look for when calibrating Daylight fingerprints

- Very long aliphatic chains
  - Possible solution: increase path length
- Multi ring systems (> 3)
  - Because of ALL paths through molecule are considered and hashed, ring systems tend to dominate fingerprints
    - Loss of resolution in functional group contribution
      - Acids and amines very similar in T-terms
  - Possible solution
    - Increase fingerprints size (2048 to 4096)
      - Check how many bits ON
      - If >50% of fingerprint size – increase size (-b -c)
    - Increase path settings

# Applications

- Representative sets (main clusters)
- Design of training/test set
- Data pre-processing (dmso solubility and P450)
- Qc
- High throughput (fall positives/negatives)

# Design of representative sets

- Set of major structural classes representing company's depository to go through all in vitro screening
  - Unbiased set with maximum overall chemical diversity
- Co-clustering in house depository or project set with the chemistry available 'outside'
  - Either buying in more of the similar (mixed clusters) or
  - Buying in compounds coming from 'pure' external clusters

# Designing sets for very low throughput experimental screens

- Difficult decision when deciding how to chose 20 out of 100 compounds
  - 20 most active compounds (historically preferred choice by medicinal chemists) is usually set coming from the single chemotype and with very low chemical diversity
    - If one compound fails, say clearance, chances are all of them will fail too
  - Cluster at 0.2 or 0.3 will give maximum chemical diversity that exists within that set
    - If there is a compound in that set that it behaves differently – good chance of finding it within the first round of testing
- Other methods for experimental design are available

# Design of training/test sets for QSAR work

- If no test set used and only with cross-validated R2
  - LOO or LMO
  - Desperate measures and should be only used when dealing with very small sets (20-50 data points)
- Random selection
- Selection algorithms, like Kennard-Stone
- Clustering, like dbclus
  - Cluster centroids and singletons ➔ training set
  - Cluster members ➔ test set
    - Will guarantee sampling across chemical diversity and no ‘holes’ in chemical structure terms

# Data pre-processing before QSAR work

- Example 1:
  - Building model for 2-3% DMSO-buffer solubility
    - Standard in industry for HT solubility screen
    - Solvent system for most in vitro screens – good model could help with detecting false positives/negatives due to solubility
  - Me, Et and Bu analogues (high T-similarity) soluble at 40 uMol, while Pr 200 uMol – check experimental data

# Example 2

- Good quality ic50 screen data for 2c9 (P450 enzyme class)
  - Couple of compounds with MeO-Ar having ic50 at around 6, while phenol analogue at 8 (T-range around 0.9)
    - Known mechanism for 2c9 substrates suggests importance of proton transfer
    - Tight cluster with large spread of IC50 makes sense
    - This will also suggest to develop descriptors that can differentiate phenols and ethers

# QC application

- Company has registration system that registers different salts of the same parent structure under different RegNo
- Database containing RegNo and in vitro data
  - Convert all structures to parent and unique smiles
  - Run clustering at 1.0 to check whether
    - Same compound already tested
    - If yes, do they have same experimental values

# HTS – false positives/negatives

- Check for very tight clusters (0.95) that have great variance in observed experimental data
  - Large set of close analogues (combinatorial library approach) all showed activity in HTS screen
    - One of them re-made and fully characterised, but tested inactive
    - Before re-synthesis of the rest of the series, initial HTS sample re-analysed and starting material identified as an active component

# Danger of ‘chemotype’ terminology

- R1-Ph-R2
- Virtual library naming will have usually have aromatic core part as the key ‘code-label’
- Task:
  - profile the virtual library and design smallest library to make, that will have optimum profile in activity and ADME terms

# Influence of R1 and R2

- If R1 and R2 are simple substituents, the task will be also simple, but
- If R1 and R2 are much more complex than core part, in terms of functional groups or ring systems, the following will happen:
  - R1 and/or R2 will become dominant components influencing SAR
  - It will have an effect of ‘chemotype hopping’
  - It will most likely have very different SAR

# Comparison of ‘goodness’ of different chemotypes

- Based on ‘core’ bit will force sampling **within** each chemotype and conclusions that library A has better overall profile from library B – it will ignore potentially large chemical space that is in ‘reality’ dominated by R1 and/or R2
- Sampling of the chemical space defined by the all of the virtual space, regardless of the ‘core-codes’, i.e sampling in the whole molecules space will give best chance of finding compounds with best profiles

# find_analogue code

- Store daylight fingerprints in a corporate database
- Any hit from HTS or IC50 screen – look for closest analogues already made or available

# In conclusion

- How to get best out of Daylight fingerprints and Tanimoto similarity index
- Very powerful tool in automated pattern recognition of large datasets
- Chemical similarity, even at a very simple level, has very important part to play in drug discovery