Daylight

Chemical
Information
Systems, Inc.

Enterprise-level cheminformatics.

Fingerprints, similarity and clustering

Summer school 2004

Documentation references

http://www.daylight.com/dayhtml/doc/theory/theory.finger.html

http://www.daylight.com/dayhtml/doc/cluster/index.html

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Reasoning by analogy

- Reasoning by analogy is a very powerful concept.
- Given two objects are similar in some way, it is probable that they will be similar in some other related way.
- In chemistry, this sort of reasoning allowed Mendeleev to construct the periodic table, without a knowledge of atomic structure.
- **"I began to look about and write down the elements with their atomic weights and typical properties, *analogous* elements and *like* atomic weights on separate cards, and this soon convinced me that the properties of elements are in periodic dependence upon their atomic weights."**

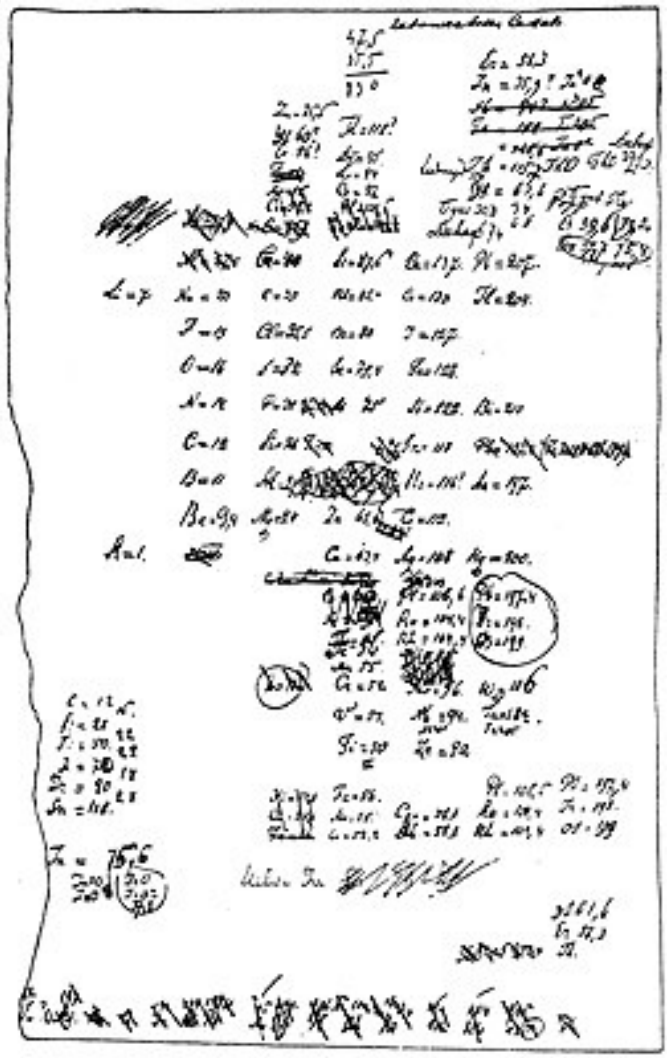
  --*Mendeleev*, Principles of Chemistry, *1905, Vol. II*

Daylight

Chemical Information Systems, Inc.

Enterprise-level cheminformatics

Brainstorm



Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Mendeleev's periodic table

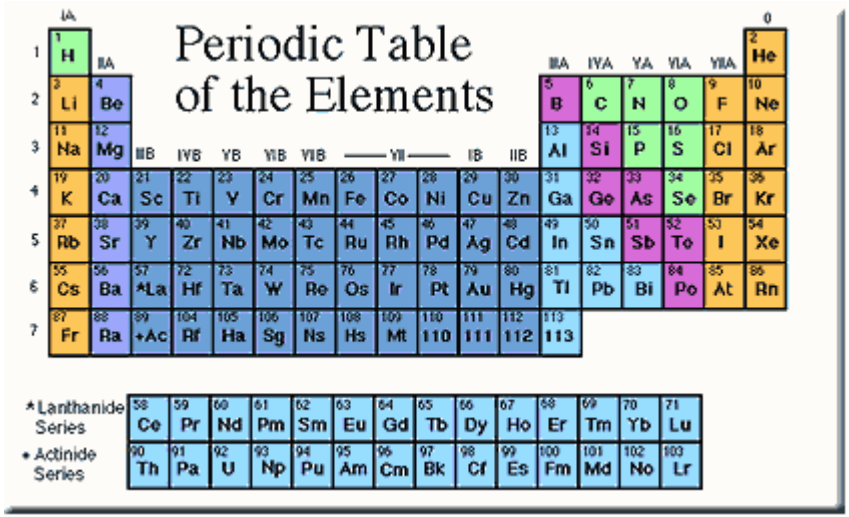
TABELLE II

REIHEN	GRUPPE I. — R^2O	GRUPPE II. — RO	GRUPPE III. — R^2O^3	GRUPPE IV. RH^4 RO^2	GRUPPE V. RH^3 R^2O^5	GRUPPE VI. RH^2 RO^3	GRUPPE VII. RH R^2O^7	GRUPPE VIII. — RO^4
1	H=1							
2	Li=7	Be=9,4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27,3	Si=28	P=31	S=32	Cl=35,5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59, Ni=59, Cu=63.
5	(Cu=63)	Zn=65	—=68	—=72	As=75	Se=78	Br=80	
6	Rb=85	Sr=87	?Yt=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104, Pd=106, Ag=108.
7	(Ag=108)	Cd=112	In=113	Sn=118	Sb=122	Te=125	J=127	
8	Cs=133	Ba=137	?Di=138	?Ce=140	—	—	—	— — — —
9	(—)	—	—	—	—	—	—	
10	—	—	?Er=178	?La=180	Ta=182	W=184	—	Os=195, Ir=197, Pt=198, Au=199.
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	—	—	
12	—	—	—	Th=231	—	U=240	—	— — — —

Figure 2.5 Dmitri Mendeleev's 1872 periodic table. The spaces marked with blank lines represent elements that Mendeleev deduced existed but were unknown at the time, so he left places for them in the table. The symbols at the top of the columns (e.g., R^2O and RH^4) are molecular formulas written in the style of the 19th century.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Modern periodic table



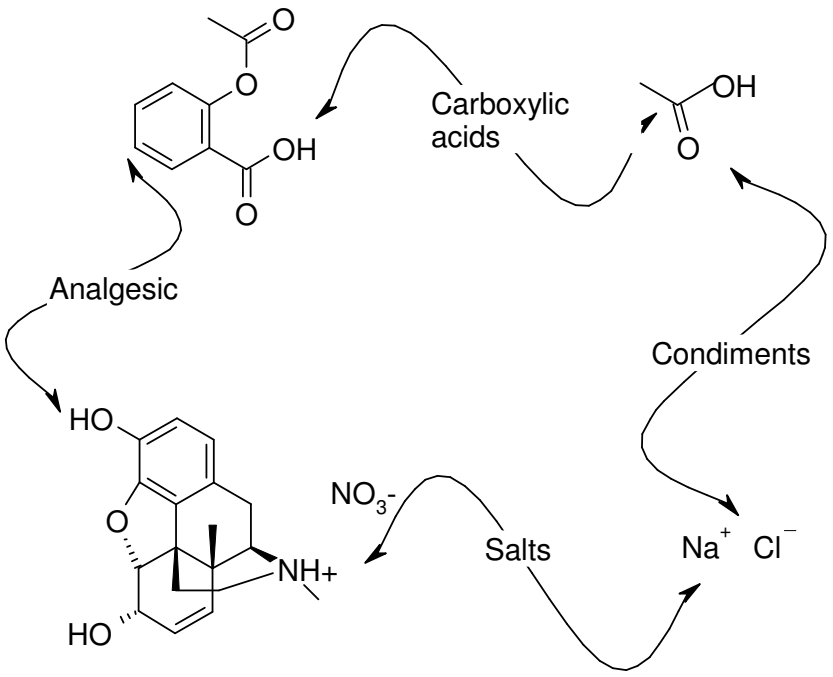
Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

The problem

- There are two implicit aspects to saying that two objects are similar.
  - How are the objects described?
  - How is the relationship, between the two sets of descriptors, measured?
- In chemistry there are two main classes of descriptor
  - Structure based.
  - Property based.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Different spaces



Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Fingerprints and feature keys

- The default object descriptor for molecules in Daylight is structure based.
- There are two main types of structure based descriptions.
  - Feature keys
    + These map well to observations and to the class nature of organic chemistry.
    - However they require you know the classes up front to set the keys.
    - Potentially there are a large number of possible features.
  - Fingerprints
    + These are graph based so do not rely on *a priori* classification.
    + It is possible to pack them into a fixed width, irrespective of number of features.
    - There is no simple relationship between the pattern and the feature.

Daylight
Chemical
Information
Systems, Inc.

Enterprise-level cheminformatics.

Daylight fingerprints

- Starting with each atom, traverse all paths, branches, and ring-closures up to a certain depth (typically 8). For each substructure, derive a hash-like number from unique, relatively-prime, order-dependent contributions of each atom and bond type. Critical properties of this number are that it is reproducible (each substructure produces a single number) and its value and graph are not correlated (a linear congruential generator is used to insure this).
- Map each resulting number into a large range (typically 2K-64K) to produce a redundant, large-scale, binary representation of the substructural elements. The resultant "fingerprint" contains a large amount of information at a low density.
- Iteratively "fold" the fingerprint by OR-ing the fingerprint in half until the bit-density reaches a minimum required value or until the fingerprint reaches a minimum allowable length. The resulting fingerprint now has a high information density with a minimal (and controllable) information loss.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

OK. So what does that mean?

- For example, the molecule OC=CN would generate the following patterns:
  - *0-bond paths:*C O N
  - *1-bond paths:*OC C=C CN
  - *2-bond paths:*OC=C C=CN
  - *3-bond paths:*OC=CN
- The list of patterns produced is exhaustive: *Every* pattern in the molecule, up to the pathlength limit, is generated. For all practical purposes, the number of patterns one might encounter by this exhaustive search is infinite, but the number produced for any *particular* molecule can be easily handled by a computer.

Enterprise-level cheminformatics.

Health warning

- Fingerprints (and also feature keys) were designed to act as filters in substructure and superstructure searches.
- If molecule A is a substructure of molecule B, all the patterns that exist in the fingerprint of molecule A must be present in the fingerprint of molecule B.
- In a fingerprint, created as described, all parts of the molecule are treated equally. Aliphatic carbon has the same weight as aromatic arsenic.
- Whilst the folding paradigm works well for filtering, in a similarity search the value is directional (more later)

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Fingerprints are not…

- Representations of high dimensional Cartesian space.
- Appropriate input for a neural network for QSAR.
- Unique
  - Try

```
thorlist medchem02demo          \
| grep 'FP<'                    \
| sort                          \
| uniq –c                       \
| sort –nr                      \
| more
```

  - There is less duplication with unfolded fingerprints.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

But not all my molecule matters

- One of the advantages of the Daylight approach to fingerprinting is that you do not need represent all of the molecule.
  - The algorithm sets bits for *substructures*
- Substructures in the molecule can be fingerprinted exclusively e.g.
  - Fragments only
  - Rings only
  - No aliphatic carbon chains
- These can be generated *via* the demo code provided and compared in similarity searches in DayCart™ or in merlin as an exercise.
  - cat myfile.tdt | addfp –FRAGMENT –RINGS –NO_C_CHAINS –MINBITS 2048

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Two's company

- The similarity of two fingerprints is a function of the bits in common between two structures.
  - This is returned by the toolkit function dt_fp_commonbitcount()

- This comparison is modulated by the bits which are unique to each of the fingerprints.
- These relationships can be visualised as Venn diagrams

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Similarity coefficients

- Over the years several coefficients have been developed to provide a normalised scale of similarity.
- All are f(a,b,c,d) where
  - **a** = count of on-bits unique to fingerprint A
  - **b** = count of on-bits unique to fingerprint B
  - **c** = count of on-bits common to both fingerprints A and B
  - **d** = count of off-bits common to both fingerprints A and B
- A list of the common ones are here
- The most common coefficient is that due to Tanimoto, but others are now being seriously investigated and are available.
- Given the nature of Daylight fingerprints it is inappropriate to use measures with the common off-bits **d,** as this value can be arbitrarily altered by adjusting the size.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Asymmetric similarity coefficients

- There are two ways to ask the similarity question
  - How alike are A *and* B (symmetric)
  - How like is A *to* B (asymmetric)
- Asymmetric similarity has the idea of a prototype.
  - We may ask how like is the UK to the USA (prototype)
- In the chemical world this corresponds to similarity as a superstructure or as a substructure.
- Daylight has implemented this via the Tversky coefficient where α and β are adjustable parameters to reduce the effect of the unique bits

$$\frac{c}{\alpha * a + \beta * b + c}$$

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Similarity searching

- The user identifies a *target structure* or set of structures from which a (modal) fingerprint can be derived.
- This *target fingerprint* is compared with a whole set of other fingerprints, be they in a database under merlin or Oracle®, or a file.
- A selection of compounds is made where the fingerprint comparison exceeds a certain value, or the whole list is ordered.
- If a bioactive target is searched for, then the top-ranked molecules, or *nearest neighbours* are also likely to possess that activity.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Similar Property Principle

- This has become known as the Similar Property Principle in Life Sciences which states that…

  Molecules which are structurally similar are likely to have similar properties.

  M.A. Johnson and G.M. Maggiora (eds) *Concepts and Applications of Molecular Similarity* (John Wiley, New York, 1990)
- Clearly this is a restatement of the Analogy Principle discussed earlier.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Three's a crowd

- The process of taking a large set of objects and partitioning them into subsets such that objects, within a set, are more *like* each other than they are *like* objects in other sets, is known as clustering.
- If we take our ordered lists for all possible targets then in the same way that a pair of compounds is said to be similar if they contain a proportion of the same substructures (shared bits = **c**), compounds can be grouped if they share a *proportion of nearest neighbours.*
- This grouping by proportion of shared nearest neighbours is an appropriate algorithm for Daylight non-parametric descriptors and is the basis of the Jarvis-Patrick clustering algorithm.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Clustering algorithms

- There are many algorithms available for clustering objects.
  - Agglomerative
  - Divisive
  - Hierarchical
  - Non-hierarchical
  - Parametric
  - Non-parametric
- Which algorithm to use depends on the nature of the descriptor for the object and to a lesser extent the measure of pair-wise similarity

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Daylight's clustering algorithms

- Currently Daylight makes available 3 non-parametric non-hierarchical clustering algorithms.
  - Jarvis-Patrick
  - Sphere exclusion
  - K-modes
- Users can take the similarity matrix and use packages such as SAS
- Other vendors which do not have databasing capability also read Daylight fingerprints and tdt's as input into their clustering packages e. g. BCI

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Jarvis-Patrick Clustering

- The full documentation at http://www.daylight.com/dayhtml/doc/cluster/index.html is recommended reading.
- The method, as published (R.A. Jarvis and E.A. Patrick, Clustering using a similarity method based on shared nearest neighbours, *IEEE Transactions on Computers* C-22 (1973) 1025-1034) works like this:
  - For each item, find its **J** nearest neighbours. This requires $O(N^2)$ CPU time, but needs to be done only once. The Daylight implementation is closer to O(NlogN) generally.
  - Two structures cluster together if:
    (a) They are in each other's list of **J** nearest neighbours,

    and

    (b) **K** of their **J** nearest neighbours are in common.

Daylight implementation

- This method is implemented in the Clustering Package as the programs **nearneighbors** and **jarpat**.
- Removing clustering requirement (a) usually results in improved clustering due to a more exhaustive search but at a high cost in speed.
- Partially relaxing this requirement, i.e. only requiring that one must be in the other's list, approximates the more exhaustive search and runs even faster than the published method.
- **jarpat** provides all three methods.
- Daylight does *not* implement the more stringent requirements that the ranking of the near neighbours should match.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Advantages of Jarvis-Patrick

- The Jarvis-Patrick algorithm appears to be an ideal method for clustering chemical structures:
  - The same results are produced regardless of input order (almost!!)
  - It's a non-parametric method
  - Cluster resolution can be adjusted (**J**,**K**) to match a particular need
  - Autoscaling is built into the method
  - It will find tight clusters embedded in loose ones
  - It is not biased towards globular clusters
  - The clustering step is very fast
  - Overhead requirements are relatively low

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

So why don't people like J-P

- The Jarvis-Patrick algorithm appears to be an non-ideal method for clustering chemical structures:
  - The same results are produced regardless of input order (almost!!)
  - It's a non-parametric method
  - Cluster resolution can be adjusted (**J**,**K**) to match a particular need
  - Autoscaling is built into the method
  - It will find tight clusters embedded in loose ones
  - It is not biased towards globular clusters
  - The clustering step is very fast
  - Overhead requirements are relatively low

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

A note on singletons

- In a parametric world singletons are thought of as outliers, ‘distant’ from other members of the set
- In the non-parametric world the idea of singletons is not necessarily so intuitive as every object has the same number of neighbours.
- Singletons i.e. objects that fail to cluster, can arise from two causes in Jarvis-Patrick corresponding to the two parameters **J** and **K.**
  - If the object has none of its **K** neighbours in common with any other object it will remain a singleton.
  - If there are **j** neighbours in common, when **j** < **J** then it too will fail to cluster.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Running Jarvis Patrick

- The Jarvis-Patrick clustering method is implemented in the Clustering Package as the programs **nearneighbors** and **jarpat**.
- The near neighbour search is the slow step and is typically done only once.
- Clustering with **jarpat** is relatively fast but requires that appropriate clustering parameters are supplied.
- The program **jpscan** is provided to assist in selection of clustering parameters

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

nearneighbors

- **nearneighbors** reads a Thor datatree file containing fingerprint data, copying its input to output, adding a *"Nearest Neighbours"* (NN) dataitem after each selected fingerprint. This program uses a bunch of computational optimizations to beat $O(N^2)$ for most chemical data sets, but it's still CPU-intensive.
- **nearneighbors** can take advantage of multiple CPUs on some multiprocessing machines. This option (-NUM_PROCESSES) controls the number of child processes which get spawned on these machines. Using multiple processes decreases the overall processing time linearly with increased CPUs.
- **mergeneighbors** allows near neighbours lists generated on the same input fingerprint files to be merged. This is extremely useful for processing of large databases.
- **nearneighbors** can be stopped and restarted at will and the intermediate lists can be easily merged.
- Currently we do not support non-shared memory multi-CPU environments.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

jpscan and jarpat

- **jpscan** and **jarpat** both perform Jarvis-Patrick clustering based on nearest neighbours (NN) data. Both programs use two Jarvis-Patrick clustering parameters: the number of neighbors to examine and the number required to be in common.
- **jpscan** repeatedly clusters data using all possible parameter combinations up to a given limit (typically set to the list length, default is 16) and outputs tables of statistics intended to help in selecting a pair of parameters appropriate to the problem at hand.
- **jarpat** requires that the parameters be specified and outputs the clustering results.
- It is advisable to run **jpscan** and examine its output before running **jarpat**.
- Both programs also allow control of the way the clustering search is done :
  - as published (the default)
  - an exhaustive search (only useful for very small data sets)
  - a faster search which approximates the exhaustive search (recommended).

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

jarpat

- **jarpat** provides two (nonexclusive) methods for dealing with singletons:
  - rescuing singletons
  - writing them out to a separate file.
  - If singleton rescue is used (option -RESCUE_SIZE), rescued singletons will appear in clusters to which they are rescued.
  - If a singleton file is generated (option -SINGLETON_FILE), it may be fed back to **nearneighbors** and then reclustered.
- **jarpat** provides an additional processing option which is not part of the original Jarvis-Patrick algorithm.
  - This option (-NN_BEST_THRESHOLD) allows the preprocessing of the neighbours lists as follows:
    - the best neighbour (excluding itself) for each structure is compared with the threshold value. If the best neighbour has a similarity lower than the specified threshold, then the structure is marked as a singleton and is excluded from the clustering. This is a useful way to discover very tight(?) clusters within a dataset.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

showclusters and listclusters

- **showclusters** and **listclusters** read cluster (CL) and fingerprint (FP) dataitems in a Thor datatree (e.g. those written by **jarpat**).
- **showclusters** produces summaries and tables suitable for textual display or printing.
- **listclusters** reformats cluster data in a way suitable for processing by other programs. Both programs are able to sort structures by cluster and compute the intra-cluster statistics.
- Cluster results to be passed on to any other program should be processed by **listclusters** first. Aside from computing *intra*-cluster statistics and removing temporary data items, **listclusters** sorts and renumbers the clusters in a more useful, less arbitrary manner than is done by **jarpat**. By default, **listclusters** writes its output in Thor datatree format, but SMILES formatted output can be also be generated. The latter is more useful for DayCart™ users.
- Although **showclusters** does exactly the same sorts and statistical computations as **listclusters**, it offers a number of summary displays and output formatting options specific to textual presentation. **showclusters**' output uses only printable ASCII (and newline) and is suitable for use in virtually any environment.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

More on clustering

- With the next release, all the different similarity measures will become available in **nearneighbors**.
- The issue of ties is dealt with in Jarvis-Patrick
- Two new clustering algorithms will be offered
  - Sphere exclusion
  - K-modes
- Both of these new methods are very fast, and can make use of user defined similarity measures.

Daylight
Chemical Information Systems, Inc.
Enterprise-level cheminformatics.

Practical exercises

- No practical sessions have been scheduled for this module.
- However given the fundamental importance of these concepts to chemoinformatics, please take time out to read and understand the relevant chapters in the documentation and recent developments at http://www.daylight.com/meetings/mug04/Delany/clustering.html