



BENCHMARKING AND VALIDATION OF JCHEM ECFP AND FCFP FINGERPRINTS

Roger Sayle, NextMove Software Ltd, Cambridge, UK
roger@nextmovesoftware.co.uk

1. Abstract

The cornerstone of pharmaceutical chemistry is Crum Brown's observation that similar compounds have similar therapeutic benefits. Cheminformatics tries to capture this insight by defining measures of similarity between the computer representations of two molecules, with the hope of capturing a medicinal chemist's intuitive sense of "likeness", and thereby correlate with bioactivity. This poster evaluates the chemical similarity measures offered by ChemAxon on a standard reference benchmark. Any such benchmark must by necessity be flawed; the similarity between two molecules is influenced by the framework by which they are compared [2]. However a robust similarity measure should typically perform better on such benchmarks, whilst a weaker model of chemical similarity would be expected to perform worse (on average).

2. Briem & Lessel Benchmark

The benchmark employed in this evaluation is the commonly used Briem and Lessel benchmark [1]. This test assesses a method's ability to identify near neighbours with the same biological activity from decoy molecules and molecules with different biological activities. Five classes of active compounds are used: 40 ACE inhibitors, 49 TXA antagonists, 110 HMG-CoA reductase inhibitors, 133 PAF antagonists and 48 5HT3 antagonists. In addition to these 380 active compounds, the data set contains 573 "random" MDDR compounds, for a total of 953 molecules. The benchmark proceeds by determining the 10 nearest neighbours for each of the 380 active compounds. The query is not considered a neighbour of itself. The score for each activity class is the fraction of these neighbours that have the same activity as the query. Finally, the overall score is the average of the score of the five activity classes.

3. Fingerprint Methods

Historically, the similarity method underlying ChemAxon's JChem search engine relied upon Chemical Fingerprints ("CF"). These are path-based fingerprints similar to Daylight fingerprints, which allow a number of variants depending upon parameters for the number of bits in the fingerprint, the longest bond path to encode and the number of bits set by each path. The "Marvin FP" below uses the generatemd defaults of 1024 bits, paths of up to 7 bonds, and 3 bits per path. The "JChem FP" below uses the JChemManager defaults of 512 bits, paths up to length 6 and 2 bits per path.

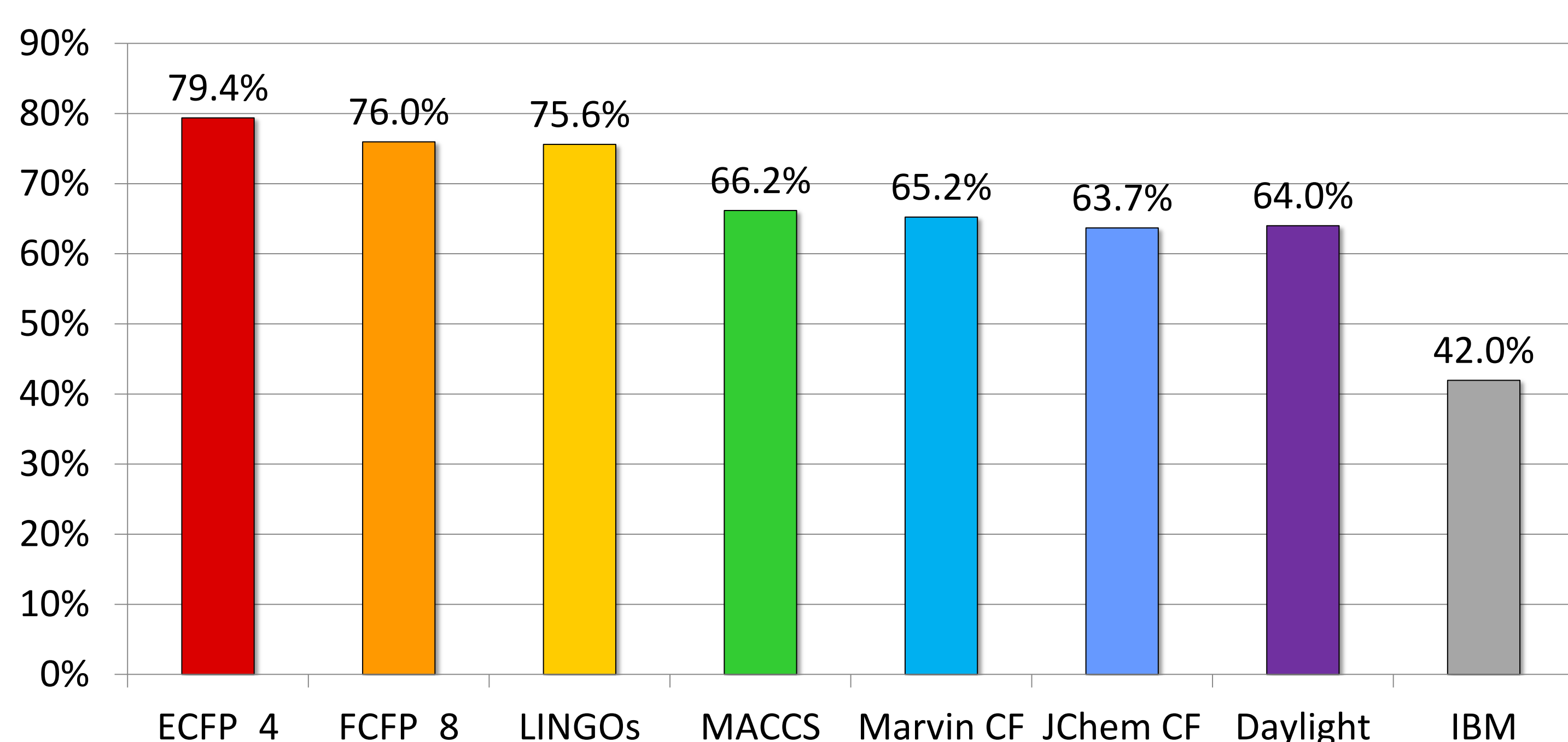
Recently, in v5.4, ChemAxon has added support for ECFP and FCFP fingerprints originally introduced by Scitegic, now Accelrys [8]. These are termed "ECFP_4" and "FCFP_8" below indicating the ChemAxon implementation with diameter parameters of 4 bonds and 8 bonds respectively.

For reference comparison to other methods, also shown are LINGOs similarity [4], MACCS 166-bit keys [3], Daylight fingerprints and IBM's patented InChI-based chemical similarity (US20080004810A1) as used in their SIMPLE product [7].

4. Tanimoto Coefficient

Many ways of comparing similarity between binary fingerprints have been discussed in the literature [9]; generally the best performing of these is the Tanimoto coefficient, $T = |X \cap Y| / |X \cup Y|$. This definition has almost magical properties, normalizing the differences between two feature sets by their sizes, intuitively "the fraction in common". Experimentally this correlates well to the chemical and biological notion of what makes two molecules similar.

5. Evaluation Results



6. Fingerprint Saturation

A common failing with binary fingerprints is caused by their inability to represent the number of times a feature (such as a path or substructure) occurs. The fingerprints for decane (C_{10}), undecane (C_{11}) and dodecane (C_{12}) are typically identical, as are those for many protein and DNA sequences. A more powerful representation that solves these issues is to replace occurrence bits with counts, turning binary fingerprints into occurrence histograms.

LINGOs similarity achieves better results on the Briem & Lessel benchmark by using counts instead of bits. However, as described in the Continuous Tanimoto section below, care has to be taken to use a suitable similarity measure for comparing histograms.

ChemAxon have announced that the upcoming release of JChem, version 5.5, will support ECFP and FCFP fingerprints with counts.

7. Continuous Tanimoto

Although there is universal agreement on how the Tanimoto coefficient should be interpreted for binary values, its application to continuous values, such as histogram counts, has been implemented differently by different authors [4,5]. Consider the two alternate definitions T_0 and T_1 given below.

$$T_0(x, y) = \frac{\sum_i^N x_i y_i}{\sum_i^N (x_i^2 + y_i^2 - x_i y_i)} \quad T_1(x, y) = \frac{\sum_i^N \min(x_i, y_i)}{\sum_i^N \max(x_i, y_i)}$$

Both definitions agree for binary valued vectors, and are guaranteed to return increasing fractional values between zero and one. Notice however that for $x = \{3\}$ and $y = \{4\}$, then $T_1 = 3/4 = 0.75$ but $T_0 = 12/13 \sim 0.923$.

In experiments with LINGO's histograms, T_1 was found to be superior (producing an improvement of $\sim 0.9\%$) whereas T_0 actually made the results worse (by $\sim 3\%$).

8. Conclusions

- ChemAxon's Chemical Fingerprints perform comparably with other path and feature-based fingerprints (including MACCS 166-bit keys, Daylight fingerprints and PubChem/CACTVS fingerprints). All these methods perform equivalently.
- ECFP fingerprints, originally developed by Scitegic/Accelrys and as recently implemented by ChemAxon, perform exceptionally well on the standard Briem and Lessel benchmark.
- The announced ECFP histograms would be anticipated to set new records in 2D chemical similarity.

9. Acknowledgements

To Miklos Vargyas and Alex Allardyce for the invitation to present a poster at the ChemAxon UGM, to Peter Kovacs for JChem ECFP support and rapid bug fixing, and to AstraZeneca and Vertex Pharmaceuticals for their interest in 2D similarity.

10. Bibliography

- Hans Briem and Uta F. Lessel, "In vitro and in silico Affinity Fingerprints: Finding Similarities beyond Structural Classes", *Perspectives in Drug Discovery and Design*, Vol. 20, pp. 231-244, 2000.
- Robert D. Brown and Yvonne C. Martin, "Use of Structure-Activity Data to Compare Structure-based Clustering Methods and Descriptors for Use in Compound Selection", *JCICS*, Vol. 36, No. 3, pp. 572-582, 1996.
- Joseph L. Durant, Burton A. Leland, Douglas R. Henry and James G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery", *JCIM*, Vol. 42, pp. 1273-1280, 2002.
- J. Andrew Grant, James A. Haigh, Barry T. Pickup, Anthony Nicholls and Roger A. Sayle, "Lingos, Finite State Machines and Fast Similarity Searching", *JCIM*, Vol. 46, No. 5, pp. 1912-1918, 2006.
- Thierry Kogel, Ola Engkvist, Niklas Blomberg and Sorel Muresan, "Multifingerprint Based Similarity Searches for Targeted Class Compound Selection", *JCIM*, Vol. 46, No. 3, pp. 1201-1213, 2006.
- Steven W. Muchmore, Derek A. Debe, James T. Metz, Scott P. Brown, Yvonne C. Martin and Philip J. Hajduk, "Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping", *JCIM*, Vol. 48, No. 5, pp. 941-948, 2008.
- James Rhodes, Stephen Boyer, Jeffrey Kreule, Ying Chen and Patricia Ordonez, "Mining Patents using Molecular Similarity Search", *Pacific Symposium on Biocomputing*, Vol. 12, pp. 304-315, 2007.
- David Rogers and Mathew Hahn, "Extended Connectivity Fingerprints", *JCIM*, Vol. 50, No. 5, pp. 742-754, 2010.
- Peter Willet, John M. Barnard and Geoffrey M. Downs, "Chemical Similarity Searching", *JCICS*, Vol. 38, No. 6, pp. 893-996, 1998.