

# Virtual screening in chemoinformatics: the role of data fusion.

Peter Willett, University of Sheffield

Chemoinformatics techniques are used in pharmaceutical and agrochemical industries to search databases of chemical structures. One of the most common techniques is similarity searching, which normally involves representing molecules by fingerprints that describe the fragment substructures present in a molecule, and then comparing fingerprints to find those database molecules that are most similar to a user-defined reference structure. This paper commences by giving a brief introduction to chemoinformatics [1] and then focuses on the calculation of chemical similarity and the use of data fusion to combine the results of multiple similarity searches [2, 3]. A detailed comparison of a large number of similarity coefficients, both on their own and when combined, demonstrates that the Tanimoto coefficient is the method of choice for the computation of fingerprint-based similarity, despite possessing some inherent biases related to the sizes of the molecules that are being sought. Group fusion involves combining the results of similarity searches based on multiple reference structures and a single similarity measure. We demonstrate the effectiveness of this approach to screening, and also describe an approximate form of group fusion, turbo similarity searching, that can be used when just a single reference structure is available.

1. Leach, A.R. and Gillet, V.J. (2003) *An Introduction to Chemoinformatics*, Kluwer
2. Willett, P. "Similarity-based virtual screening using 2D fingerprints." *Drug Discovery Today*, **11**, 2006, 1046-1053.
3. Willett, P. "Enhancing the effectiveness of ligand-based virtual screening using data fusion." *QSAR and Combinatorial Science*, **25**, 2006, 1143-1152.