

## Fourth Joint Sheffield Conference on Chemoinformatics June 18-20, University of Sheffield, UK

Dr. Wendy A. Warr  
Wendy Warr & Associates  
wendy@warr.com | <http://www.warr.com>

---

This conference (<http://cisrg.shef.ac.uk/shef2007/>), sponsored by the Chemical Structure Association Trust (<http://www.csa-trust.org/>) and the Molecular Graphics and Modelling Society (<http://www.mgms.org/>) and organized by the University of Sheffield chemoinformatics research group (<http://www.shef.ac.uk/is/research/groups/chem/index.html>), takes place every three years in the year preceding the Noordwijkerhout International Chemical Structures Conference. This fourth conference was attended by 230 delegates, the number being set by the size of the dining facilities at Chatsworth House, the site of the superb conference outing where we were shown some of the state rooms and enjoyed an excellent dinner. It is clear that the meeting has become very popular since all places were taken by the end of the early registration period, with delegates coming from Australia, Austria, Belgium, Cyprus, Denmark, France, Germany, Hungary, India, Iran, Ireland, Italy, Netherlands, Poland, Serbia, Spain, Sweden, Switzerland, the Ukraine, the United Kingdom and the United States.

Twenty-four papers were presented, in sessions entitled structure-based design, new algorithms and techniques, deriving structure-activity relationships, clustering, and QSAR and ADMET. More than sixty posters were presented. In this report I am summarizing only the QSAR-related papers, which means I am obliged to omit some of the material that I myself found most interesting. It is a shame to have to ignore, for example, the excellent paper by Andy Good on the defects of enrichment studies in the comparison of virtual screening (i.e., docking) tools. It is unfortunate that I have to gloss over controversial comments from Anthony Nicholls (“docking sucks”, and “you cannot calculate binding energy”) and his attack on Richards and Ballester’s Ultrafast Shape Recognition. Indeed, Nicholls’ own paper was controversial in itself.

Nicola Richmond of GlaxoSmithKline presented a fast, novel, graph-matching algorithm, based on the comparison of distance degree sequences. The algorithm matches pairs of nodes, one from each graph, by solving the linear assignment problem. The graph similarity is then given by the minimum cost associated with the optimal set of matching pairs of nodes. By representing molecules as 2D topological pharmacophores, Richmond has adapted the algorithm to rank a corporate collection against a query molecule of interest, and to cluster the ranked list into groups of compounds that have identical chemical graphs. The clustering component has a useful visualization facility. The highest ranked compounds correspond to the analogues of the query; families of “lead hops” follow. This unsupervised approach is not a substitute for substructure search but it is fast and it may produce a new template around which a chemist can search. It can follow GSK’s automated high throughput screening process to recover not only families of compounds on which to build structure activity relationships, but also hits missed by high throughput screening (HTS).

Enriched scaffolds in HTS data sets can be identified by clustering on substructure and then extracting the maximal common substructure (MCS) for each cluster. However, if clustering is performed without reference to the assay data, the resulting scaffolds are unlikely to show optimal

enrichment for the assay in question. Martin Packer of AstraZeneca has developed a method for locating scaffolds with high enrichment factors, using a hierarchical search strategy. Molecules encoded by substructure are partitioned into  $N$  clusters and for each cluster,  $M$  hierarchical clusters are generated. The MCS is extract from each cluster and an enrichment factor is computed. The enrichment factor is calculated for each maximal common substructure. The procedure is iterated by setting  $M = M - 1$ . The method was applied to a 540,000-compound in-house kinase data set and 6737 actives were partitioned into 200 clusters. The AstraZeneca collection contains lots of kinase series; so a Bonferroni test was applied to correct for the chance of generating a spurious result. The hierarchical nature of the search means that structure-activity relationships emerge for the most enriched scaffolds. Emergent SAR was found for a quinazoline scaffold: substitution at the 7-position enhanced enrichment.

Bob Clark of Tripos Discovery Informatics has been looking for answers to “the alignment problem”. When no structure is available, researchers must fall back on pharmacophore matching or comparative molecular field analysis (CoMFA). Unfortunately, ligand binding often induces structural changes that significantly reduce the usefulness of apoprotein structures for docking and scoring. In such cases it is often better to dock into the binding site of a ligand-protein complex from which the ligand has been extracted *in silico*. Even when a naïve protein structure is suitable for docking, ligands can provide critical information about the location of the relevant binding site. Moreover, interactions with specific binding site residues illuminated by bound ligands have been successfully used to direct docking and to tailor scoring functions to specific target proteins. An extreme version of this is the use of docking to align molecules for CoMFA. Clark displayed lots of  $q^2$  values and models docked with Surflex but I found it hard to extract a take-home message from this talk.

Ansgar Schuffenhauer and his colleagues at Novartis have published a Pareto analysis of methods for classification of chemical structures by scaffold<sup>[1]</sup>. Rule-based methods such as that of Bemis and Murcko<sup>[2]</sup> scale linearly with the number of structures since the classification process is done individually for each molecule and incremental update are possible. The classes created by such methods are more intuitive to chemists than those produced by clustering and other methods. Schuffenhauer described a variation on Bemis and Murcko’s molecular frameworks. His hierarchical classification method<sup>[3]</sup> uses molecular frameworks as the leaf nodes of a scaffold tree. By iterative removal of rings, scaffolds forming the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less characteristic, peripheral rings are removed first, e.g., in order of precedence:

- Keep macrocycles with at least twelve atoms
- Choose the parent scaffold having the smallest number of acyclic linker bonds
- Retain bridged rings, spiro rings, and nonlinear ring fusion patterns in preference
- Remove rings of sizes 3, 5, and 6 first
- Remove rings with the least number of heteroatoms first.

Highlighting by color intensity is used to show the fraction of active compounds containing a scaffold: this immediately identifies those branches of the scaffold tree, which contain active molecules. Schuffenhauer concluded that chemical series are not always equivalent to biological activity classes and what is actually desirable is continuous change in biological activity with the chemical variation in a chemical series.

Evotec have applied a spectral clustering method to 2D structures<sup>[4]</sup> and have found it particularly useful in the analysis of screening data. It provides a means to quantify the degree of intermolecular similarity within a cluster and the contribution that the features of a molecule make to a cluster. These two criteria can be used to arrange molecules into clusters of chemically related molecules and quantify inter-cluster relationships so that the resultant classification scheme appears intuitive from a medicinal chemistry perspective. Mark Brewer presented applications of the method to, for example, a data set of 125 COX-2 inhibitors.

Johann Gasteiger of the University of Erlangen-Nürnberg showed how modeling of chemical reactions can help in drug discovery. For example, in lead discovery and lead optimization, an estimate of synthetic accessibility can be useful. Gasteiger's team has devised a scoring method that rapidly evaluates synthetic accessibility of structures based on structural complexity, similarity to available starting materials, and assessment of strategic bonds where a structure can be decomposed to obtain simpler fragments<sup>[5]</sup>. These individual components are combined to give an overall score of synthetic accessibility by an additive scheme. A demonstration of the system (SYLVIA) is available at [http://www.molecular-networks.com/online\\_demos/sylvia](http://www.molecular-networks.com/online_demos/sylvia).

Modeling metabolism is also possible. To this end, XENIA, an in-house CYP450 database, has been developed at the University of Erlangen-Nürnberg. MetaboGen systematically generates all metabolites of a drug, applying a set of the most important phase I reactions. A demonstration of the ISOCYP Web service for prediction of the predominant P450 isoform<sup>[6]</sup> is available at [http://www.molecular-networks.com/online\\_demos/cyp450/](http://www.molecular-networks.com/online_demos/cyp450/). Molecular Networks also supplies the biochemical pathways database, BioPath (<http://www.mol-net.de/biopath/index.html>).

Markus Wagener of Organon presented a novel, rule-based method, SyGMa (Systematic Generation of Metabolites) that predicts potential metabolites of a given parent structure. The method is based on reaction rules derived from metabolic reactions that occur in man, reported in Elsevier MDL's Metabolite database (<http://www.mdl.com>). The database was filtered (to remove assumed metabolites, incomplete and large structures etc.) to give 7307 biotransformations as a training set. Reaction templates were encoded as SMIRKS and reaction probabilities were calculated based on training set statistics. The predicted metabolites are ranked according to the empirical probability score. Evaluation of the method demonstrated a significant enrichment of true metabolites at the top of the ranking list. The current rule set covers about 70% of the human *in vivo* data of the Metabolite database. To gain an understanding of the nature of the reactions, a similarity analysis of the reaction types was performed using difference fingerprints<sup>[7]</sup> calculated by subtracting fingerprints generated from atom environments<sup>[8]</sup>. SPE<sup>[9]</sup> was used to project the reaction space. Wagener gave some examples of SyGMa, including the pathway for buspirone. Predictions from SyGMa are used at Organon to plan experiments aimed at experimental metabolite identification and to suggest labile sites amenable to optimization by medicinal chemistry.

Metabolism was also the topic of a paper by Anton Schwaighofer of Fraunhofer FIRST. His team, idalab of Berlin, and Bayer Schering Pharma (BSP) have jointly developed machine learning tools to predict the metabolic stability of compounds from drug discovery projects at BSP. They used experimental metabolic stability data from four different *in vitro* assays. They compared a variety of machine learning approaches in terms of performance, difficulty of the model selection procedures, interpretability, and how the "domain of applicability" can be checked. They concluded that Gaussian Process classification has specific benefits. The effort required for model selection is minimal, so fully automatic re-training is possible. Also, the probabilistic output

is easy to interpret and shows almost ideal properties. Competing methods achieve similar performance, but need more careful tuning by an expert. The models developed were validated on recent project data at BSP: the best models are highly accurate and are able to identify the domain of applicability correctly. These models are fully integrated in the working environment at BSP and a tool for automatic regular retraining of the models is currently being implemented. A paper has been submitted to *J. Chem. Inf. Model.*

The initial part of Joelle Gola's talk consisted of generalities about local and global models and the features of BioFocusDPI's product Admense Interactive ([http://www.biofocus.com/In\\_silico\\_optimization](http://www.biofocus.com/In_silico_optimization)). Later on, Gola gave what approached a sales pitch about the proprietary algorithm Glowing Molecule, which highlights problematic regions within potential drugs responsible for deficiencies in ADMET properties. Gola intended to present case studies of his company's new automatic techniques for model building. These enable non-computational scientists to capture and share the knowledge contained in their experimental data by building local models for individual chemical series, iteratively improving models as more data are generated and using new models to predict properties for new chemical structures. Gola did describe the Gaussian Process method at the heart of this work<sup>[10]</sup> but he ran out of time while rushing through his three examples.

Lastly, Damjan Krstajic, of the Research Centre for Cheminformatics in Serbia described yet another answer to the challenge of designing a system that will cope with constant influx of new information. Discovery Bus aims to automate QSAR modeling without sacrificing the quality of predictions. It is an implementation of Competitive Workflow, novel software architecture, implemented using autonomous software agents<sup>[11]</sup>. All possible combinations of components are explored leading to exhaustive evaluation of potential solutions. The idea is that we cannot know in advance which technique or approach to use in solving a QSAR problem, but if we apply most of the well known techniques and approaches, then we will have an explosion in the number of models, but we will also end up with multiple good solutions among them. A related poster, by David Leahy and co-workers at Newcastle University, described a multi-objective reverse QSAR search agent called Forager. This has been developed to search for non-dominated solutions to the research target profile definition of a new drug within a complex descriptor space, where the search heuristics are provided by multiple QSAR models. Forager uses a modified Particle Swarm Optimization algorithm.

I have reported on only 10 out of 24 presentations. It is hoped that another report will appear in the CSA Trust newsletter (<http://cisrg.shef.ac.uk/shef2007/>). Abstracts for all the papers and posters are online at <http://cisrg.shef.ac.uk/shef2007/conference.htm> and readers are encouraged to seek out the many interesting papers for which I have not had space to comment. All in all, this was an excellent meeting. It is a shame that it could not accommodate more delegates, but if it did so, perhaps some of the more useful interactions and discussions would be impeded. I shall make a point of booking early for the fifth incarnation of this event.

## References

1. Schuffenhauer A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and Rule-based Classifications of Chemical Structures Evaluated in the Biological Activity Space. *J. Chem. Inf. Model.* **2007**, *47*, 325-336.
2. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.
3. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel S.; Koch M. A.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47-58.
4. Brewer, M. Development of a Spectral Clustering Method for the Analysis of Molecular Datasets. *J. Chem. Inf. Model.* **2007**, in press.
5. Boda, K.; Seidel, T.; Gasteiger, J. Structure- and Reaction-based Evaluation of Synthetic Accessibility. *J. Comput.-Aided Mol. Des.* Published online February 9, 2007.
6. Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-based Models for the Isoform Specificity of Cytochrome P450 3A4, 2D6, and 2C9 Substrates. *J. Chem. Inf. Model.* **2007**, in press.
7. Zhang, Q. -Y.; Aires-de-Sousa, J. Structure-based Classification of Chemical Reactions without Assignment of Reaction Centers. *J. Chem. Inf. Model.* **2005**, *45*, 1775-1783.
8. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708-1718.
9. Agrafiotis, D. K.; Xu, H. A Self-organizing Principle for Learning Nonlinear Manifolds. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*(25) 15869-15872.
10. Obrezanova, O.; Csányi, G.; Gola, J.; Segall, M. Gaussian Processes: a Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, in press.
11. Cartmell, J.; Enoch, S.; Krstajic, D.; Leahy, D. E. Automated QSPR through Competitive Workflow. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 821-833.

---

Note: All organization names, websites and URLs, products, trademarks and/or service marks etc. mentioned in this article are intellectual property of their respective owners. Strand Life Sciences fully acknowledges the rights of respective owners over such intellectual property.