From the 8th to the 10th November 2009, the Chemistry-Information-Computers (CIC) division of the German Chemical Society (GDCh) has invited the chemoinformatics and modeling community to Goslar, Germany to participate in the 5th German Conference on Chemoinformatics (GCC 2009). The international symposium addressed a broad range of modern research topics in the field of computers and chemistry. The focus was on recent developments and trends in the fields of Chemoinformatics and Drug Discovery, Chemical Information, Patents and Databases, Molecular Modeling, Computational Material Science and Nanotechnology. In addition, other contributions from the field of Computational Chemistry were welcome.

The conference was opened traditionally with a “Free-Software-Session” on Sunday afternoon right before the official conference opening at 5 pm including three talks about the Open Source projects Bingo, Dingo and OrChem. In parallel the “Chemoinformatics Market Place” took place including software tutorials by Chemical Computing Group, Helmholtz-Center Munich and the Cambridge Crystallographic Data Center.

The scientific program was opened by an evening talk giving an overview on the field of Systems Chemistry (Günter von Kiedrowski). In addition, the program included six plenary lectures (Eberhard Voit (USA), Knut Baumann (Germany), Thomas Kostka (Germany), Anthony J. Williams (USA), Karl-Heinz Baringshaus (Germany), Christoph Sotriffer (Germany)), 17 general lectures as well as 54 poster presentations.

Besides the scientific program a special highlight of the conference were the FIZ-CHEMIE-Berlin 2009 awards on Monday afternoon (Figure 1). The CIC division awards this price each year to the best diploma thesis and the best PhD in the field of Computational Chemistry. The price for the PhD thesis was awarded to Dr. José Batista from the group of Prof. Dr. Jürgen Bajorath, University of Bonn for his dissertation “Analysis of Random Fragment Profiles for the Detection of Structure-Activity Relationships”. The award for the best diploma thesis has gone to Frank Tristram from the group of Dr. Wolfgang Wenzel, Karlsruher Institute of Technology with the title “Modellierung der Hauptkettenbeweglichkeit in der rechnergestützten Medikamentenentwicklung”.

**FREE SOFTWARE SESSION PRESENTATIONS**

**F1**

**Bingo from SciTouch LLC: chemistry cartridge for Oracle database**

Dmitry Pavlov*, Mikhail Rybaklin, Boris Karulin
SciTouch LLC, Budapestskaya st. 23-2-82, RU-192212 St. Petersburg, Russia
E-mail: dpavlov@scitouch.net

Journal of Cheminformatics 2010, 2(Suppl 1):F1

Bingo is a data cartridge for Oracle database that provides the industry’s next-generation, fast, scalable, and efficient storage and searching solution for chemical information. Bingo seamlessly integrates the chemistry into Oracle databases. Its extensible indexing is designed to enable scientists to store, index, and...
search chemical moieties alongside numbers and text within one underlying relational database server. For molecule structure searching, Bingo supports 2D and 3D exact and substructure searches, as well as similarity, tautomer, Markush, formula, molecular weight, and flexmatch searches. For reaction searches, Bingo supports reaction substructure search (RSS) with optional automatic generation of atom-to-atom mapping. All of these techniques are available through extensions to the SQL and PL/SQL syntax. Bingo also has features not present in other cartridges, for example, advanced tautomer search, resonance substructure search, and fast updating of the index when adding new structures. The presentation itself you can download from our site: http://opensource.scitouch.net/downloads/bingo-cic.pdf

### References
1. [http://opensource.scitouch.net/indigo/dingo].
2. [http://www.daylight.com/smiles/].
4. [http://www.daylight.com/smiles/].

#### F3
**OrChem**
Mark Rijnbeek
Protein and Nucleotide Database Group, Chemoinformatics and Metabolism Team, EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
E-mail: markr@ebi.ac.uk

**Background:** Registration, indexing and searching of chemical structures in relational databases is one of the core areas of chemoinformatics. However, little detail has been published on the inner workings of search engines and their development has been mostly closed-source. We decided to develop an open source chemistry extension for Oracle, the de facto database platform in the commercial world.

**Results:** Here we present OrChem, an extension for the Oracle 11G database that adds registration and indexing of chemical structures to support fast substructure and similarity searching. The chemoinformatics functionality is provided by the Chemistry Development Kit. OrChem provides similarity searching with response times in the order of seconds for databases with millions of compounds, depending on a given similarity cut-off. For substructure searching, it can make use of multiple processor cores on today’s powerful database servers to provide fast response times in equally large data sets.

**Availability:** OrChem is free software and can be redistributed and/or modified under the terms of the GNU Lesser General Public License as published by the Free Software Foundation. All software is available via http://orchem.sourceforge.net.

---

**ORAL PRESENTATIONS**

**O1** Systems chemistry: from chemical self-replication to trisoligo-based nanconstruction

Günter von Kiedrowski
Lehnshüll für Organische Chemie I - Bioorganische Chemie, Ruhr-Universität, Universitätsstrasse 150, 44780 Bochum, Germany
E-mail: kiedro@rub.de
Journal of Cheminformatics 2010, 2(Suppl 1):O1

Self-replication is one of the major principles without life could not exist. The emergence of self-replicating systems on the early earth is generally believed to have taken place before the advent of instructed protein synthesis based on a complex translation machinery. Whether the origin of self-replication is identical to the origin of the hypothetical RNA world or whether it existed at an earlier stage of evolution is an open question that has stimulated chemists to search for chemical systems capable of making copies of itselves via autocatalytic reactions. As self-replication means autocatalysis plus information transfer, the reaction products must necessarily be able to store more structural information than their precursors. Templating as a means to transfer structural information has been exploited since the first successful example of a chemical self-replicating system almost two decades ago [1]. Today we have a broad variety of such systems employing oligonucleotides, peptides, and small organic molecules as templates and autocatalytic [1-3], cross-catalytic [4], collectively autocatalytic and non-autonomous (stepwise) schemes of self-replication [5].

A link between the chemical self-replication and nanotechnology was first pointed out by G.M. Whitesides in his debate with Drexler. The ribosome is an example for a nanomachine which may be viewed as a three-dimensionally defined array of 51 modular proteins positioned by the rRNA scaffold. Biomimetic approaches towards the 3D-nano-scaffolding of modular functions may be based on trisoligonucleotides [6], viz. synthetic 3-arm junctions in which the 3’-ends of three oligonucleotides are connected by a suitable linker. We report on trisoligos as building blocks for the noncovalent construction of nanoobjects. Kinetic control - applied by rapid cooling during self-assembly - was found to favor small and defined nano-structures instead of large polymeric networks [6]. Maximal instruction was employed as design principles to generate noncovalent 3D-nano-objects in which both, the topology and the geometry is defined [7]. Examples include dodecahedral nano-scaffolds composed from 20 trisoligos each bearing three individually defined sequences [8]. It was demonstrated that the connectivity information in the nano-scaffold junctions can be copied by chemical means [9]. Chemical copying schemes may be seen in conjunction with "surface-promoted replication and exponential amplification of DNA analogues" (SPREAD) [5]. Applications of such scaffolds include the positioning of modular functions such as multidentate thioether-based gold cluster labels (RUBiGold) [10] which have been tailoroned for monoconjugatability and thermostability.

My lecture will introduce chemical self-replication and multicomponent assembly as facets of systems chemistry [3,11,12] - a nascent field which understands itself as the bottom-up pendant of systems biology towards synthetic biology [14]. The field is clearly inspired by the origin-of-life problem but goes beyond traditional prebiotic chemistry in its mission towards a quantitative dynamic understanding of complex reaction networks with autocatalytic components. Software tools like our SimFit and kinetic NMR titration [2] may significantly help to decipher signatures of interesting dynamic phenomena such as self-replication, chiral symmetry-breaking, and metabolic autocatalysis found in organic [13] and biomolecular [12] reaction systems.
References
14. MoU.

O2
The role of systems modeling in drug discovery and predictive health
Eberhard O Voit
Integrative BioSystems Institute and Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, 313 Ferst Drive, Atlanta, GA 30332-0535, USA
E-mail: eberhard.voit@bome.gatech.edu
Systems biology is the result of a confluence of recent advances in molecular biology, engineering, and the computational sciences. It can loosely be categorized into experimental and computational systems biology. Experimental high-throughput methods, assisted by robotics, image analysis, and bioinformatics, have been used in the drug industry for quite a while, and current screening tests for drug efficacy and toxicity regularly involve genomic, proteomic, and molecular modeling approaches. By contrast, the role of computational methods of biological systems analysis is still emerging. This presentation focuses on computational systems modeling and its increasingly important role at several junctures of the drug development pipeline. Examples to be discussed include mathematical models for receptor dynamics, pharmacokinetics, and metabolic and signaling pathway analysis. In the context of the latter, Biochemical Systems Theory is proposed as a highly advantageous default framework for model design, diagnostics, manipulation, and system optimization. The development of dynamic models for complex disease processes permits the straightforward inclusion of methods for custom-tailoring models, which is a key step toward personalized medicine and predictive health [1-5].

References

O3
Molecular bioactivity extrapolation to novel targets by support vector machines
Gerard JP Van Westen1,2, JK Wegner3, AP Uzman1, HWT Van Viljmen2, A Bender3
1Amsterdam Center for Drug Research, Einsteinweg 55, 2333 CC, Leiden, The Netherlands; 2Tibotec, Gen De Witterlaan L 118 3, 2800 Mechelen, Belgium
E-mail: gerard@lca.di.ledenumv.nl
The early phases of drug discovery use in silico models to rationalize structure activity relationships, and to predict the activity of novel compounds. However, the performance of these models is not always acceptable and the reliability of external predictions - both to novel compounds and to related protein targets - is often limited. Proteochemometric modeling [1] adds a target description, based on physicochemical properties of the binding site, to these models. Our proteochemometric models [2] are based on Scitegic circular fingerprints on the compound side and on a customized protein fingerprint on the target side. This protein fingerprint is based on a selection of physicochemical descriptors obtained from the AAindex database. Through PCA we selected a number of physicochemical properties which are hashed in a fingerprint using the Scitegic hashing algorithm. We compared this fingerprint to a number of protein descriptors previously published, including the Z-scales, the FASGAI and the BLOSUM descriptors. Our fingerprint performs superior to all of these. In addition, we show that proteochemometric models improve external prediction capabilities. In the case of classification this leads to models with a higher specificity when compared to conventional QSAR. In the case of regression our models show an average lower RMSE of 0.12 log units when based on a pIC50 output variable compared to conventional QSAR.

References

O4
Representation and searching of biomolecules
Josepeth L Durant1, WL Chen, BD Christie, DL Grier, BA Leland, JG Nourse
Symyx Technologies, 2440 Camino Ramon, San Ramon, California, USA
Journal of Cheminformatics 2010, 2(Suppl 1):O4
Biomolecules present challenges to chemical information systems designed for small molecules. Their sizes, up to tens of thousands of atoms, overwhelm representation/storage/searching solutions built on explicit chemical representation of the structures. But biomolecules are largely made up of many repeats of a limited number of building-block molecules, a fact which has been used to provide a compressed representation for biomolecules using templates for the building blocks. We have adopted a modified template-based representation for biomolecules. Our primary interest is in the chemically modified portions of biomolecules, for which we choose to use explicit chemistry. These areas of explicit chemistry are then embedded in the template- compressed, unmodified portions of the full biomolecule. The regions containing explicit chemistry are indexed, and thus can be structure searched with good performance. A limited number of residues surrounding explicit chemistry regions are included in the index for searching the context of these explicit regions. By using explicit chemistry to represent modified regions we can search across classes of modifications for common features. For example a single substructure search query will find green fluorescent protein, and its histidine, phenylalanine and tryptophan analogs.
Templates are stored with the structure providing a self-contained file format. The use of NEMA keys allows templates from different structures to be compared, and allows storage of structures containing a canonical list of templates. The residues have defined attachment points, allowing automated traversal of a protein backbone, or location of non-backbone bonds to residues. We will present example structures and structural queries highlighting capabilities of our representation.
Cross-validation was originally invented to estimate the prediction error of a mathematical modelling procedure. It can be shown that cross-validation estimates the prediction error almost unbiasedly. Nonetheless, there are numerous reports in the chemoinformatic literature that cross-validated figures of merit cannot be trusted and that a so-called external test set has to be used to estimate the prediction error of a mathematical model. In most cases where cross-validation fails to estimate the prediction error correctly, this can be traced back to the fact that it was employed as an objective function for model selection. Typically each model has some meta-parameters that need to be tuned such as the choice of the actual descriptors and the number of variables in a QSAR equation, the network topology of a neural net, or the complexity of a decision tree. In this case the meta-parameter is varied and the cross-validated prediction error is determined for each setting. Finally, the parameter setting is chosen that optimizes the cross-validated prediction error in an attempt to optimize the predictivity of the model. However, in these cases cross-validation is no longer an unbiased estimator of the prediction error and may grossly deviate from the result of an external test set. It can be shown that the "amount" of model selection can directly be related to the inflation of cross-validated figures of merit. Hence, the model selection step has to be separated from the step of estimating the prediction error. If this is done correctly, cross-validation (or resampling in general) retains its property of unbiasedly estimating the prediction error. Matter of factly, it can be shown that data splitting into a training set and an external test set often estimates the prediction error less precisely than proper cross-validation. It is this variability of prediction errors, which depends on test set size, that causes seemingly paradox phenomena such as the so-called "Kubinyi's paradoxon" for small data sets.

A unified approach to the applicability domain problem of QSAR models  
Dragos Horvath*, Gilles Marcou, Alexandre Varnek  
Laboratoire d'Info Chimie, Université de Strasbourg - CNRS, Institut de Chimie, 4 rue Blaise Pascal, 67000 Strasbourg, France  

The present work proposes a unified conceptual framework to describe and quantify the important issue of the Applicability Domains (AD) of Quantitative Structure-Activity Relationships (QSARs). AD models are conceived as meta-models designed to associate an untrustworthiness score to any molecule M subject to property prediction by a QSAR model. Untrustworthiness scores or "AD metrics" are an expression of the relationship between M (represented by its descriptors in chemical space) and the space zones populated by the training molecules at the basis of model μ. Scores integrating some of the classical AD criteria (similarity-based, box-based) were considered in addition to newly invented terms, such as the dissimilarity to outlier-free training sets and the correlation breakdown count.

A loose correlation is expected to exist between this untrustworthiness and the error affecting the predicted property. High untrustworthiness does not preclude correct predictions, inaccurate predictions at low untrustworthiness must be imperatively avoided. This kind of relationship is characteristic for the Neighborhood Behavior (NB) problem: dissimilar molecule pairs may or may not display similar properties, but similar molecule pairs with different properties are explicitly "forbidden". Therefore, statistical tools developed to tackle this latter aspect were applied, and lead to a unified AD metric benchmarking scheme.

A first use of untrustworthiness scores resides in prioritization of predictions, without need to specify a hard AD border. Moreover, if a significant set of external compounds is available, the formalism allows optimal AD borderlines to be fitted. Eventually, consensus AD definitions were built by means of a nonparametric mixing scheme of two AD metrics of comparable quality, and shown to outperform their respective parents.
based on InChI-layers. While the InChI ansatz supports only heteroatom-
tautomerism, we suggest an extension regarding carbon atoms too. Whereas with other tautomer generating algorithms the hydrogen shifts are based on pattern-rules, we try to overcome the rule constriction and evolve a more common solution. The advantage of our approach is quite simple. Due to the avoidance of a rule system with its necessity for exceptions to the rules, we can apply our solution to any kind of tautomerism definition. We set up a Branch-and-Bound approach, which is optimized to generate a complete enumeration of all tautomers, with regard to a certain definition, from any structure. With few and easy decisions like symmetry detection, we avoid a lot of calculation overhead. Decisions with significant influence on the algorithm efficiency are made as early as possible. We have set up several kinds of tautomer definitions and derived a stable definition covering the major kinds of prototropic tautomerism. Furthermore we analyzed, what expenditure of time for large databases (case study: more than 70,000 entries) is needed to investigate which structures have tautomers and which not for more than 99% of the database entries. This study has been financially supported by the EU project OSIRIS (IP, contract no. 037017).

O9 KnowledgeSpace - a publicly available virtual chemistry space
Carten Detering, H Clausen, M Gastreich, C Lemmen
BioSolveIT GmbH, An der Ziegelei 79, 53757 Sankt Augustin, Germany

Virtual high throughput screening of in-house compound collections and vendor catalogs is a validated approach in the quest for novel molecular entities. However, these libraries are small compared to the overall synthetizable number of compounds from validated “wet” chemical reactions in pharma companies or the public domain. In order to overcome this limitation, we designed a large virtual combinatorial chemistry space from publicly available combinatorial libraries that gives access to billions of synthetically accessible compounds. Together with FTrees, a fuzzy similarity calculator, the researcher has a means of searching this KnowledgeSpace for analogues to one or several query molecules within a few minutes. The resulting compounds not only exhibit similar properties to the query molecule(s), but also feature an annotation through which of the synthetic routes these molecules can be made. Results can expected to be diverse, based on FTrees scaffold hopping capabilities, and provide ideas for hit follow-up into novel compound classes. In this contribution we present the design and properties of the KnowledgeSpace and other in-house chemistry spaces that build on the same strategy as well as validation of results and a number of successful applications including prospective results.

O10 3D pharmacophore alignments: does improved geometric accuracy affect virtual screening performance?
Gerhard Wolber1, T Seidel1, F Bendix2
1University of Innsbruck, Institute of Pharmacy, Dept. Pharmaceutical Chemistry, Inrmain 52, 6020 Innsbruck, Austria; 2IntLigand GmbH, Mariahilferstrasse 74/11, A-1070 Vienna, Austria
E-mail: gerhard.wolber@uibk.ac.at

Virtual screening using three-dimensional arrangements of chemical features (3D pharmacophores) has become an important method in computer-aided drug design. Although frequently used, considerable features (3D pharmacophores) has become an important method in virtual screening using three-dimensional arrangements of chemical entities. However, these libraries are small compared to the overall Virtual high throughput screening of in-house compound collections and vendor catalogs is a validated approach in the quest for novel molecular entities. However, these libraries are small compared to the overall synthetizable number of compounds from validated “wet” chemical reactions in pharma companies or the public domain. In order to overcome this limitation, we designed a large virtual combinatorial chemistry space from publicly available combinatorial libraries that gives access to billions of synthetically accessible compounds. Together with FTrees, a fuzzy similarity calculator, the researcher has a means of searching this KnowledgeSpace for analogues to one or several query molecules within a few minutes. The resulting compounds not only exhibit similar properties to the query molecule(s), but also feature an annotation through which of the synthetic routes these molecules can be made. Results can expected to be diverse, based on FTrees scaffold hopping capabilities, and provide ideas for hit follow-up into novel compound classes. In this contribution we present the design and properties of the KnowledgeSpace and other in-house chemistry spaces that build on the same strategy as well as validation of results and a number of successful applications including prospective results.

O11 The protein flexibility in receptor-ligand docking simulations
Frank Tristram
Institute of Nanotechnology, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

Many small-molecule drugs work by binding specifically to a target protein in the cell. It is known for over a century that both the ligand and protein receptor change their conformation in the association process, which is called the induced-fit effect. Ligand conformational change is routinely treated in methods for in-silico drug discovery, because typical drug molecules have seldom more than 100 atoms. The flexibility of proteins, which often have more than 10000 atoms, is much harder to treat and was therefore neglected in most high-speed docking methods, limiting their accuracy and predictive value. In this project we have developed an efficient numerical search procedure that succeeds to model the interaction between a flexible ligand and important flexible parts of the protein in a computationally affordable protocol. Our virtual screening software FlexScreen thus minimizes the total interaction energy of the emerging protein-ligand complex including flexible backbone regions. We have demonstrated the reliability and accuracy of this approach on several examples, for which at least two different receptor-conformations have been experimentally observed. We succeeded to predict the correct binding pose of a protein-ligand complex starting from the crystal structure of an unrelated protein conformation, where large conformational changes of the protein are necessary to bind the ligand. Correctly predicting such conformational changes makes our approach attractive for virtual screening of medium sized databases, in particular for kinases and other target receptors which have flexible binding pockets.

O12 Random molecular substructures as fragment-type descriptors
José Batista1,2
1Jado Technologies GmbH, Tatzberg 47-51, D-01307 Dresden, Germany; 2Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universitat Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany

A novel approach for analysis of structure-activity relationships for sets of active compounds is reported. Molecular similarity relationships are analyzed among compounds with related biological activity based on the evaluation of randomly generated fragment populations, Fragments are randomly generated by iterative bond deletion in molecular graphs and hence depart from those obtained by well-defined fragmentation schemes [1]. A major finding has been that for any given activity class, dependency relationships of fragment co-occurrence exists within random fragment populations. Through the analysis of these relationships the chemical information content of generated substructures can be quantified [2]. This has lead to the identification of small numbers (10-40) of so-called activity class characteristic substructures (ACCS) that have been successfully utilized in virtual screening applications [3,4].

References
Mathematical methods - both, for consumer goods and industrial innovative products with state-of-the-art molecular modeling and examples will be given how scientific computing methods contribute to mechanics and mesoscopic simulations to QSAR approaches. Carsten Wittekindt*, H Kuhn

Biomaterial coatings - a challenging task studied by the molecular fragment dynamics

A special and important task in medicinal implant technology is to prevent the material from fouling or make it compatible for tissues by coating the material with functionalized lipid biomolecules. In order to rationalize the design of such new materials structural information about the film is needed. In this presentation we give an account on the investigation of coating of the tetraether lipid GDNT on a borofluorate surface in different solvents. To gain insights into the dynamic process on the microsecond and micrometer scale we applied our recently developed Molecular Fragment Dynamics method (MFD) [1]. Aside of the size of the system and the duration of the simulation, the challenging task of the investigation lies in the divers chemical functionality of the compounds. The applied MFD algorithm is developed to resolve especially these problems. The MFD algorithm is based on the Dissipative-Particle-Dynamics method (DPD). Whereas in the DPD method the system is divided into different regions of fluid subsystems, the MFD algorithm calculates the interaction of molecular fragments. Consequently, the MFD method is well suited for the molecular modelling simulation of systems having a diverse variety of chemical functionalities thus allowing for investigation of biocoatings, polymers and tensids. Figure 1.

Reference

 Mechanistic studies on the ring-opening polymerisation of D,L-lactide with zinc guanidine complexes

The increasing shortage of resources has enforced the search for production techniques for biodegradable polymers made of renewable raw materials. An example is poly(lactide) (PLA), an aliphatic polyester, which can be obtained by the metal-catalysed ring-opening polymerisation (ROP) of lactide (Fig. 1). The monomer for PLA production is available from corn or sugar beets by a bacterial fermentation process in few steps [1]. The PLA can be either recycled or composted after use, making the use CO₂-emission-neutral. Most large-scale processes are based on the use of tin compounds as initiators [2]. However, for use in food packaging or similar applications, heavy metals are undesirable [3]. In order to substitute heavy-metal-based catalyst systems, especially zinc guanidine complexes are suited for the polymerisation due to their polymerisation activity and non-toxicity [4]. The investigations on the catalytic potential in the bulk polymerisation of lactide have shown that these compounds are able to act as initiators for lactide polymerisation with only few exceptions. Polylactides with molecular weights (Mₘₜₐ) between 20,000 and 170,000 g/mol could be obtained.

As the polymerisation activity depends strongly on the ligand structure (spacer, guanidine and amine unit), DFT studies on the complex properties have been performed [5]. A correlation between calculated partial charges on the zinc and the guanidine N atoms and the catalytic activity has been found. The guanidine is supposed to open nucleophilically the lactide ring and supersede the presence of alkoxides which is traditionally required. This crucial step in the ROP has been analysed by DFT and a mecha-nism will be presented. The deeper understanding of the catalytic transition state will enable a more efficient access to catalyst design. Figure 2.

1TU Dortmund, Fakultät Chemie, 44221 Dortmund, Germany; 2University of Paderborn, Paderborn, Germany

Figure 1 (abstract O15). ROP of D,L-lactide

Figure 2 (abstract O15). General motif for hybrid quanidines
ChemSpider - building a foundation for the semantic web by hosting a crowd sourced databasing platform for chemistry

Anthony J Williams1, Valery Tscharniko2, Sergey Golovin3, Richard Kidd4, Graham McCann5
1Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC-27587, USA; 2NCBI, NLM, National Institutes of Health, Washington, USA; 3Royal Society of Chemistry, Cambridge, UK


There is an increasing availability of free and open access resources for chemists to use on the internet. Coupled with the increasing availability of Open Source software tools we are in the middle of a revolution in data availability and tools to manipulate these data. ChemSpider is a free access website for chemists built with the intention of providing a structure centric community for chemists. It was developed with the intention of aggregating and indexing available sources of chemical structures and their associated information into a single searchable repository and making it available to everybody, at no charge.

There are tens if not hundreds of chemical structure databases such as literature data, chemical vendor catalogs, molecular properties, environmental data, toxicity data, analytical data etc. and no single way to search across them. Despite the fact that there were a large number of databases containing chemical compounds and data available online their inherent quality, accuracy and completeness was lacking in many regards. The intention with ChemSpider was to provide a platform whereby the chemistry community could contribute to cleaning up the data, improving the quality of data online and expanding the information available to include data such as reaction syntheses, analytical data, experimental properties and linking to other valuable resources. It has grown into a resource containing over 21 million unique chemical structures from over 200 data sources.

ChemSpider has enabled real time curation of the data, association of analytical data with chemical structures, real-time deposition of single or batch chemical structures (including with activity data) and transaction-based predictions of physicochemical data. The social community aspects of the system demonstrate the potential of this approach. Curation of the data continues daily and thousands of edits and depositions by members of the community have dramatically improved the quality of the data relative to other public resources for chemistry.

This presentation will provide an overview of the history of ChemSpider, the present capabilities of the platform and how it can become one of the primary foundations of the semantic web for chemistry. It will also discuss some of the present projects underway since the acquisition of ChemSpider by the Royal Society of Chemistry.

O17 Integration of chemical information with protein sequences and 3D structures

Adel Golovin, K Henrick, G Kleywegt
EMBL-EBI/PDB, Hinxton Hall, Genome Campus, Cambridge, Cambs, CB10 1SD, UK
E-mail: golovin@ebi.ac.uk


The Protein Data Bank (PDB) contains a wealth of small molecule - macro molecule complexes the study of which contribute enormously to our understanding of the interactions. However, exploiting and mining this treasure trove of data requires advanced analysis and retrieval methods that take into account both types of molecules. One such method is PDBeMotif, that has been developed by the Protein Data Bank in Europe (PDBe) at EMBL-EBI. Utilizing a relational database model at the back-end, the data structure represents a network of molecule, residue and motif interactions as well as their relative positions in the sequence and in 3D. The loader applies a number of algorithms to analyse PDB and derive structure information, such as planarity and aromaticity of the chemical compounds, hydrogen-bonds network, coordination geometry, bond types (including pi electron interactions), 3D structural motifs, sequence domains and families. It collects information about sequence features, motifs and catalytic sites from available Distributed Annotation System (DAS) resources. The web application allows for a wide variety of searches and data analysis including protein motifs with chemical fragments association, protein sites characterisation, correlating properties, hits multiple sequence and 3D alignments. The whole system is released under GPL and available with the source code from http://sourceforge.net/projects/pdbsam and on line at http://www.ebi.ac.uk/pdbe-site/PDBeMotif/

References

O18 Personalized information spaces: improved access to chemical digital libraries

O Koepler1, W-T Balke, B Köhncke, I Sens, S Tonnes
Technische Informationsbibliothek Hannover (TIB), Welfengarten,18, 30167 Hannover, Germany


Today digital libraries provide access to a vast, but largely unstructured, amount of document collections. Facing the ever increasing challenge of the information overload content providers have to focus on new ways in user-centered retrieval, not only providing tools for searching information, but tools for personalizing, managing, evaluating and working with the returned search results.

Within the ViFaChem II research project the TIB and the L3S Research Center, Hannover, have developed an enhanced retrieval platform for chemical digital libraries.

The prototypical interface processes full text and bibliographic metadata collections to create semantically enriched document collections with chemical metadata. Processing steps include text mining for chemical entities, reaction types and chemical structure reconstruction. However, the ViFaChem digital library does not only offer the classical document access via a text or chemical structure search, but also semantic search based on the faceted browsing paradigm. In particular it includes the Semantic GrowBag algorithm, which automatically creates facets for navigational access and query refinement using the relationship between documents and author keywords. Based on higher-order co-occurrences of keywords in documents these graphs hierarchically arrange all related topics dynamically with respect to the underlying content collection. Having different personal views on the document collection the user now can search and navigate through search results using individual retrieval strategies. Facets for chemical reactions, chemical entities or topics combined with an underlying ontology thus allow a semantic driven access to documents.

Combined with Web 2.0 features the interface of ViFaChem II provides the user with a new experience in searching large document collections, where the user can navigate through query results based on his personalized knowledge space.
O19

Current aspects and future trends of computer-aided rescaffolding
Karl-Heinz Barlinghaus*, Gerhard Hessler, Thomas Klabunde
Sanofi-Aventis Deutschland GmbH, CAS Drug Design, Building G 878, 65926
Frankfurt am Main, Germany

The competitive pressure in pharmaceutical industry is reflected by longer
development times and increasing development costs yielding less new
chemical entities. In particular, identification of suitable lead compounds
is one of the key challenges in early drug discovery. Next to well
established techniques like high throughput screening (HTS) and virtual
screening computer-aided rescaffolding has become an important
approach in detecting novel chemical structures [1,2].

The development of New Chemical Entities (NCEs) requires the
exploration of novel landscapes in chemical space. Rescaffolding is in
particular important in lead identification but also in lead optimization.
If a lead series cannot be optimized in multidimensional space, scaffold-
rescuing is often perceived as back-up strategy to transfer hitherto
available SAR into a new scaffold.

With a binding site or a pharmacophore in hand often de novo design
methods are applied to identify novel chemical matter. However, de novo
design differs from rescaffolding in that the goal of the first is primarily
to generate new molecules in chemical space while the latter aims to design
classified scaffolds under constraints: high similarity in the desired property
space and novelty in scaffold space. This allows jumping out of the known
region of chemical space towards new regions of chemistry resulting in
molecules with similar properties and activities but with novel frameworks.

This talk will focus on ligand-based rescaffolding by taking into account
molecules with similar properties and activities but with novel frameworks.

References
Potential of Fragment-Based De Novo Design: The Chances and Limits of
Variation. Combinatorial Chemistry & High Throughput Screening 2009,
2. Mauer H, Guba W: Recent developments in de novo design and scaffold
hopping. Current Opinion in Drug Discovery & Development 2008,

O20

The membrane bound aromatic p-hydroxybenzoic acid
oligoprenyltransferase (UbiA) - how iterative improvements lead to a
realistic structure that offers new insights into functional aspects of
prenyl transferases and terpene synthases
W Brandt*, J Kufka, D Schulze, E Schulze, F Rausch, L Wessjohann
Leibniz Institute of Plant Biochemistry, Department of Bioorganic Chemistry,
Weinberg 3, D-06120 Halle (Saale), Germany
E-mail: wbrandt@ipb-halle.de

Prenyltransferring enzymes are at the basis of the vast isoprenoid natural
product diversity. 4-Hydroxybenzoate oligoprenyltransferase of E. coli,
encoded in the gene ubiA, is a key enzyme in the biosynthetic pathway
to ubiquinone. No X-ray structure exists of this membrane protein. It
encodes a dimer. Based on the model, amino acids identified as important
for the catalytic mechanism were selectively replaced to obtain five new
mutants. All mutants were tested for their ability to form geranylated
hydroxybenzoate from geranyl diphosphate, but only the unmodified
UbiA-enzyme and to minor extent one mutant showed enzymatic activity.

These results indicate the involvement of all mutated amino acids in the
catalytic mechanism but at the same time demand a remodelling of
the previously proposed enzyme structure combining two active sites.
They have been placed into close proximity in the new model. Based on
these experimental results and structural classification of prenyl enzymes,
a highly relevant 3D-model could be developed. This model is able to
explain a wide range of substrate specificities and is in complete
agreement with the results of site directed mutagenesis [2].

Aromatic prenyl transferases, prenyl diphosphate synthases, and terpene
synthases, activate an (oligoprenyl) diphosphate to form a stabilized prenyl
cation, which undergoes a rapid nucleophilic addition. When transferred to
a nucleophile (C-C bond) and deprotonation delivers the product(s). Aromatic amino acids
have been suggested to stabilize the cation intermediate. Ab initio LMP2
calculations indicate not only stabilization of a prenyl cation by aromatic
amino acid side chains but also by a methionine side chain. This
suggestion is supported by site directed mutagenesis, bioinformatics, and
modelling studies. In addition, a new catalytic diad composed of Tyr and
Asp, respectively, is an important player for deprotonation and proton-relay in intermediates, or for the finalizing
deprotonation step of many prenyl transferring and cyclizing enzymes.

References

O21

CELLmicrocosmos 2.2: advancements and applications in modeling of
three-dimensional PDB membranes
Björn Sommer*, T Dingesen, S Schneider, S Ruben, C Garnot
1University of Bielefeld, Universitätsstraße 25, 33615 Bielefeld, Germany;
2D-85748 Garching, Germany

Background: Today, only a few programs support membrane
computation and/or modeling in 3D. They enable the user to create very
simple-structured membrane layers and usually assume a high level of
bio-chemical/-physical knowledge. The CELLmicrocosmos 2 project
developed a tool providing a simplified workflow to create membrane
(bi-)layers: The MembraneEditor (CmME).

Results: The geometry-based, scalable and modular computation concept
supports fast to more complex membrane generations. CmME is based on
the integration of two different types of PDB [1] models: Lipids are
integrated with editable perceptual distribution values and algorithms.
Proteins are inserted and aligned into the bilayer manually or automatically,
other programs is offered by extensive PDB format export settings. High
lipid densities are possible through advanced packing algorithms. Lipid
distributions can be developed by using the Plugin-Interface. Originally
not intended to change the atomic structure of the molecules due performance
issues, now it is also possible to access the atomic level for user-defined
computations. Multiple membrane (bi-)layers and microdomains are
supported as well as a reengineering function providing the re-editing of
externally simulated PDB membranes.

Conclusions: The capabilities of CmME has been extended and tested to
meet the requirements of different PDB visualization programs as well as
molecular dynamics (MD) simulation environments like Gromacs [4].

The documentation and the Java Webstart application, requiring only
an internet connection and Java 6, is accessible at: http://Cm2.
CELLmicrocosmos.org

References
Shindyalov IN, Bourne PE: The Protein Data Bank, Nucleic Acids Research
2. Turnaş V, Dozarşan Z, Simon J: Transmembrane proteins in the Protein
Data Bank: identification and classification, Bioinformatics 2004,
20(17):2964-2972.
3. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI: OPM: Orientations of
4. Lindahl E, Hess B, Spooel van den D: GROMACS 3.0: A package for
One of the main tasks of computational methods in ligand design and lead identification is the elucidation and assessment of interaction modes between small-molecule ligands and protein target structures. This generally requires to estimate the relative or absolute affinity of a protein-ligand complex from its three-dimensional coordinates. Although statistical thermodynamics would in principle provide the necessary equations to calculate free energies of binding from molecular properties, these equations are not readily amenable to computation since appropriate ensembles of the solvated systems must be generated and thoroughly sampled, which normally requires prohibitively long computing times. For the purpose of drug design, and virtual screening in particular, simpler and faster methods are needed, which are commonly referred to as scoring functions.

In general, three major classes of scoring functions can be distinguished: force-field based methods, knowledge-based potentials, and empirical scoring functions. For any of these classes, many different functions have been developed over the past years. Although large-scale and truly unbiased comparative assessments of their performance are relatively rare due to inherent difficulties in setting up appropriate test sets, the strengths and limitations of current scoring functions are fairly evident from the available data. In general, good results can be obtained in the identification or reproduction of experimentally observed binding modes. The ability to distinguish active ligands from decoys appears, at least, to be sufficient to make virtual screening a practical useful endeavour. However, the correlation with the experimental binding free energy and the possibility to quantitate the effects of small structural changes on the ligand affinity are in many cases still disappointing.

Based on the approximations used by current scoring functions, strategies for improvement can be defined. On the other hand, there are some fundamental problems, also with respect to the underlying experimental data, which suggest that the best functions presently available may already be close to the limit of what can be achieved with empirical approaches.
POSTER PRESENTATIONS

P1
CWM global search - an internet search engine for the chemist
Alexander Kos1, H-J Himmeler
Akos GmbH, Austr. 26, D-79585 Steinen, Germany
Journal of Cheminformatics 2010, 2(Suppl 1)P1

The Internet is a rich source of data and information for chemists. There are
umerous multidisciplinary databases available for free on the Internet. Some
examples of such data repositories are: PubChem, ChemSpider, eMolecules,
Drugbank, KEGG, NIST, ChemSynthesis, PharmGKB, Free patents online ...
It should be obvious that an end user is a) not aware of all the resources, and
b) has not the time to learn every user interface and is unable to
search over all of them. We provide CWM Global Search as an application
that enables to search by structure, CAS Registry Number and free text
over all these sources. Presently CWM Global Search performs searches in
30 databases and search engines accessing more than 100 million pages
that associate data with structures.

The user can submit a single query structure or several using SDFiles. In
addition to molecule searches CWM Global Search also allows to submit
reaction queries. In that case several single molecule searches are performed
for the reactants, reagents and products. This makes it easy to find
commercial suppliers and other synthesis relevant information such as
safety sheets in one query.

Searching is technically less problematic than providing the answers in a
digestible way for the user. Our first approach is to provide profiles for
searching. You can choose “Availability” if you are interested to find a
commercial supplier, or “Biology” if you are looking for biological effects.
The second help comes when we display the summary of the results. You
get a table with hyperlinks color coded by topics. If the result page of doing
a search contains a link to an MSDS, the topic ‘Safety’ is highlighted. With
profiles you limit your search to certain sources, and with topics your
answers will be ordered.

CWM Global Search is not the application for exact searches like “the melting
point of anthracene”. You will find the melting point on many pages, and the topic
“Physical Property” might help, but, in this case the link to Wikipedia gives the result quickest. Internet pages
provide us the data in unordered fashion and finding the exact answers is
time consuming. In some case like commercial suppliers we check
against an internal list if the page really displays a supplier, or, if for
instance PubChem has only a reference to ChemSpider, and ChemSpider
references again just PubChem, but nowhere you will find a supplier. We
also have to consider that many providers of the resources would not
allow us to extract data directly without leading the user to their Internet
pages. The nature of the results is fuzzy in CWM Global Search. This is an
advantage if you look for instance for biological effects, which can be
many, and/or if you want to learn why a compound could be important.

We generate both InChI names and keys for the query structure and
acceptance and usage of InChI as a mainstream standard. This new
management system is provided by means of a new organization, the
InChI Trust, an independent not-for-profit entity paid for by the
community and those who use and benefit from the InChI algorithm.

The mission of the Trust is quite simple and limited; its sole purpose is
to create and support administratively and financially a scientifically robust
and comprehensive InChI algorithm and related standards and protocols.
This presentation will describe the current technical state of the InChI
algorithm and how the InChI Trust is working to assure the continued
support and delivery of the InChI algorithm.

P2
The status of the InChI project and the InChI trust
Stephen Heller1, Alan McNaught2
1NIST, PCPD, MS-8380, 100 Bureau Drive, Gaithersburg, 20899-8380, USA;
2InChI Trust, Cambridge, UK
Journal of Cheminformatics 2010, 2(Suppl 1)P2

The IUPAC InChI/InChIKey project has evolved to the point where its
future development and promulgation require a new management
system that will provide stable and financially viable administrative
arrangements for the foreseeable future. This is necessary to give
the world-wide chemistry community that IUPAC serves the confidence
that facilities for development, maintenance and support of the InChI/InChIKey
algorithm are firmly established on an ongoing basis, and will ensure
P3
jsMolEditor: an open source molecule editor for the next
generation web
L Duan
School of Pharmacy, East China University of Science and Technology,
130 Meilong Road, Shanghai 200237, PR China
Journal of Cheminformatics 2010, 2(Suppl 1)P3

Molecule editor have become essential building blocks for modern
chemistry related databases with structural operation available. It was quite
a challenging task to input molecule structures in a web browser before
Java Applets technology was applied in chemoinformatics. Until now, Java
Applets are still the de facto standard for web based databases. With the
development of web technologies and script programming, Ajax, a Web 2.0
technology, is becoming mainstream in web development. This gradually
eliminates the need for plug-ins such as Java and ActiveX. Therefore, it is
possible to develop powerful web-based tools without plug-ins.

This poster introduces a JavaScript based molecule editor: jsMolEditor,
which had made a significant difference from other editors available on
web pages. Unlike server side editors such as PubChem editor [1],
jsMolEditor implemented all its functionality in JavaScript on the client
side. Thus it is not necessary to maintain a connection to a server during
operation. jsMolEditor works completely independently on the client side.
The JavaScript nature of jsMolEditor also made it able to run in standard
web browsers without specific plugins or virtual machines. The 2D drawing
system deals with web browser differences. By combining Canvas and
VML, jsMolEditor was able to work on all common web browsers including
Internet Explorer, Firefox, Safari, Opera and Chrome. Instead of building
ds an open source molecule editor for the next
generation web
L Duan
School of Pharmacy, East China University of Science and Technology,
130 Meilong Road, Shanghai 200237, PR China
Journal of Cheminformatics 2010, 2(Suppl 1)P3

Molecule editor have become essential building blocks for modern
chemistry related databases with structural operation available. It was quite
a challenging task to input molecule structures in a web browser before
Java Applets technology was applied in chemoinformatics. Until now, Java
Applets are still the de facto standard for web based databases. With the
development of web technologies and script programming, Ajax, a Web 2.0
technology, is becoming mainstream in web development. This gradually
eliminates the need for plug-ins such as Java and ActiveX. Therefore, it is
possible to develop powerful web-based tools without plug-ins.

This poster introduces a JavaScript based molecule editor: jsMolEditor,
which had made a significant difference from other editors available on
web pages. Unlike server side editors such as PubChem editor [1],
jsMolEditor implemented all its functionality in JavaScript on the client
side. Thus it is not necessary to maintain a connection to a server during
operation. jsMolEditor works completely independently on the client side.
The JavaScript nature of jsMolEditor also made it able to run in standard
web browsers without specific plugins or virtual machines. The 2D drawing
system deals with web browser differences. By combining Canvas and
VML, jsMolEditor was able to work on all common web browsers including
Internet Explorer, Firefox, Safari, Opera and Chrome. Instead of building
jsMolEditor in JavaScript from scratch, GWT [2] is utilized as a cross-
compiler from Java to JavaScript, which enables the reuse of existing Java
cinformat ics frameworks. jsMolEditor used a ported version of MX [3]
as the bottom layer framework to handle basic molfile I/O. Thanks to MX’s
open source license, the development of jsMolEditor saves a lot of time by
eliminating the process of “reinventing the wheel”. The poster represents
not only jsMolEditor itself, but also the building process of jsMolEditor and
the feasibility of applying the same concept to other chemistry gadgets on
web pages, such as spectrum viewers.

jsMolEditor’s source code is available at http://github.com/chemhack/
jsmoleditor/.

References
edit/]
2. GWT, Google Web Toolkit. [http://code.google.com/webtoolkit/]
3. MX Chemoinformatics Toolkit. [http://metamolecular.com/mx/]

P4
Mining public-source databases for structure-activity relationships
Bend Wende1, Fabian Böß2, Ulrike Uhl2
1Elara Pharmaceuticals, Heidelberg, Germany; 2Tripos International,
Martin-Kollar-Str. 17, München, Germany
Journal of Cheminformatics 2010, 2(Suppl 1)P4

Modeling off-target effects has become a highly important and relevant
component of the computational chemistry toolset. The presentation will
describe a new contribution that seeks to allow the extraction of 3D-
structure-activity relationships from public information starting from a
chemical structure. Several public source databases such as PubChem [1]
offering structure as well as activity information for a number of targets

have been examined for their value in extracting useful structure-activity relationships (SARs). A Topomer search [2] of public-source databases using the structures of a set of 255 marketed drugs [3] as queries yielded sets of shape- and pharmacophore similar hits. SAR-tables were constructed by collecting hits around each query structure and for a particular reported activity. A new method: quantitative series enrichment analysis (QSEA) [4] was applied to these SAR-tables to capture trends and to transform these trends into 3D-QSAR models. Overall more than 400 SAR-tables with Topomer CoMFA models were found by extracting trends from the PubChem and ChemBank [5] database. The resulting models were able to highlight the structural details of certain off-target effects of marketed drugs even in those cases where the traditional structural similarity would conclude that no off-target effect would exist. This demonstrates the usefulness of the approach in modeling off-target effects.

References
5. [http://chembank.broad.harvard.edu/].

P5
OChem - on-line CHEmical database & modeling environment
Sergii Novotarskyi1*, Iuri Sushko1, R Körner1, Anil Pandey Kumar1, Matthias Rupp1, WP Prokopenko2, Igor Tetko1
1 Institute of Bioinformatics and Systems Biology - Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, D-85764 Neuausberg, Germany. 2 Kiev, Ukraine
Journal of Cheminformatics 2010, 2(Suppl 1) 1-2

The main goal of OChem database http://qspr.eu is to collect, store and manipulate chemical data with the purpose of their use for model development. Its main features, that distinguish it from other available databases include:
1. The database is open and it is based on Wiki-style principles. We encourage users to submit data and to correct inaccurate submitted data; 2. The database is aimed at collecting high-quality data. To achieve this we require users to submit references to the article, where the data was published. The reference may include the article name, journal name, date of publication, page number, line number, etc.
3. The compound properties may vary depending on the conditions, under which they were measured; we store the measurement conditions with the data to provide the users with more accurate information about each data point.

The modeling framework is being developed to complement the Wiki-style database of chemical structures. Its main goal is to provide a flexible and expandable calculation environment that would allow a user to create and manipulate QSAR and QSPP models on-line. The modeling framework is integrated with the database web-interface that allows easy transfer of database data to the models. The web interface of the modeling environment is aimed to provide to the Web users easy means to create high-quality prediction models and estimate their accuracy of prediction and applicability domain. The developed models can be published on the Web and be accessed by other users to predict new molecules on-line. This tool is aimed to generate a new paradigm for structure activity relationship knowledge bases, making QSAR/QSPP models active, user-contributed and easily accessible for benchmarking, general use and educational purposes. The examples of the use of the database within national and EU projects will be exemplified.

References
2. [http://www.ebi.ac.uk/chembl/].
3. Rijnbeek M, Steinbeck C. An open source chemistry search engine for Oracle. in press.

P6
ChEBI: a chemistry ontology and database
Paula de Matos1, A Dekker, M Ennis, Janna Hastings, K Haug, S Turner, Christoph Steinbeck
Cheminformatics and Metabolism Team, European Bioinformatics Institute, Hinxton Cambridge, CB10 1SD, UK
Journal of Cheminformatics 2010, 2(Suppl 1) 1-6

The bioinformatics community has developed a policy of open access and open data since its inception. This is contrary to chemoinformatics which has traditionally been a closed-access area. In 2004, two complementary open access databases were initiated by the bioinformatics community, ChEBI [1] and PubChem. PubChem serves as automated repository on the biological activities of small molecules and ChEBI (Chemical Entities of Biological Interest) as a manually annotated database of molecular entities focused on ‘small’ chemical compounds. Although ChEBI is reasonably compact containing just over 18,000 entities, it provides a wide range of data items such as chemical nomenclature, an ontology and chemical structures. The ChEBI database has a strong focus on quality with exceptional efforts afforded to IUPAC nomenclature rules, classification within the ontology and best IUPAC practices when drawing chemical structures.

ChEBI is currently undergoing a period of restructuring which will allow it to incorporate the small molecule structures from (and link to) EBI’s new chemogenomics database ChEMBL [2], increasing its small molecules coverage to over 500,000 entities. We have restructured the chemical structure search facility to use Orchem [3] an Oracle chemistry plug-in using the Chemistry Development Kit [4]. The facility allows a user to draw a chemical structure or load one from a file and then execute either a substructure or similarity search. Furthermore the ChEBI text search will have extensive facilities for querying based not only on names but formula, a range of charges and molecular weight. The ability to query the ChEBI ontology and retrieve all children for a given entity will also be included.

In order to aid the distribution of ChEBI to the chemoinformatics community we have extended our export formats to include an MDL sdf format with a lighter version consisting only of compound structure, name and identifier. A complete version is available with all the ChEBI data properties such as synonyms, cross-references, SMILES and InChI. Furthermore cross-references in ChEBI have been extended to include BREnda the enzyme database, NMRSiftDB the database for organic structures and their nuclear magnetic resonance (nmr) spectra, Rhea the biochemical reaction database and IntEnz the enzyme nomenclature database.

ChEBI is available at http://www.ebi.ac.uk/chembi. References
2. [http://www.ebi.ac.uk/chembl/].
3. Rijnbeek M, Steinbeck C. An open source chemistry search engine for Oracle. in press.

P7
Comparing manual and automated extraction of chemical entities from documents
Christian Tychán1, Sorel Muresan
AstraZeneca R&D, LG CIGl, Pepparølåsle, 1, S-43183 Mölndal, Sweden
Journal of Cheminformatics 2010, 2(Suppl 1) 1-7

The chemical information landscape is changing rapidly with a yearly increase of over 1 million new compounds and more than 700,000 publications related to chemistry [1]. Exploring the chemical space covered by relevant journals and patents is a crucial step in early stage medicinal chemistry projects. Extracting chemical entities from unstructured text is a complex task and different approaches are currently used including manual extraction by expert curators, text mining supported by chemical NER or combinations thereof [2]. The chemical information and corresponding annotations are subsequently stored in relational databases allowing for complex chemical and text queries. To assess the capability of chemical NER in documents and to understand the coverage and accuracy of the underlying data we compared the chemistry extracted by manual curation (GVKBIO) and text mining (SureChem) from a small patent corpus.

• GVKBIO databases are populated with explicit relationships between compounds, assays and sequence identifiers that have been manually extracted from journals and patents on a large scale [3].
• SureChem Portal [4] is a gateway for chemical patent search on full text collections for USPTO, EPO and WO. SureChem users can perform structure and keyword searches on more than 9 million unique compounds.

We have selected a set of 250 patents covering various target classes and for which a minimum of 25 records per patents were retrieved from GVKBIO Patent database. The analysis was done using PipelinePilot protocols [5].

These initial results demonstrate the benefits and challenges of text mining for chemical information extraction from unstructured text.

References
4. [http://www.surechem.org].

P8 Embedded infrastructure for primary data in chemistry
Igor Grbavac1*, Oliver Koeppler1, S Dohmeyer-Fischer2, Gregor Fels2, Irina Sens1, J Brase1
1 German National Library of Science and Technology, Welfengarten 18, 30167 Hannover, Germany; 2Universität Paderborn, Department Chemistry, Warburgerstrasse 100, 33098 Paderborn, Germany

Researchers around the world produce vast amounts of experimental data each day. Primary (experimental) data are the key elements of research projects and scientific publications. During the scientific workflow from the experiment to the final scientific publication primary data is analysed, interpreted and condensed to a final quintessence. But until now the handling of primary data in chemistry contains no commonly approved standards concerning reusability and long-time accessibility. Predominantly there are no quality control, no assured long-term archival storage, no proofs and no exploitation of primary data and consequently there is no assurance of data given. The concept study “Konzeptstudie Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz” of the TIB, the workgroup of Prof. Fels and the FIZ CHEMIE aims to identify the key factors for an embedded infrastructure of primary data in the field of chemistry. In this infrastructure, chemical primary data shall be saved persistently in a central database, linked and be citable by the use of DOI (digital object identifier), be accessible and searchable. Figure 1.

We have carried out a questionnaire among researchers to analyse the scientific workflow and the chemical life-cycle of primary data in chemistry and to identify the requirements for processes and structures in an embedded scientific infrastructure for researchers. We have furthermore defined both a general and extended metadata scheme for the storage, citation and searching primary datasets. General metadata include elements necessary for citations while chemical metadata cover specific needs of describing and searching within the data itself.

![Figure 1 (abstract P8).](image)

P9 Prediction of the partition coefficient between air and body compartments from the chemical structure
Stefanie Stock1, R Kühne, R-U Ebert, G Schüümann
UFZ Department of Ecological Chemistry, Helmholtz Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany

For PBPK modeling, partition coefficients between tissues and environmental compartments are required. A simple approach starts with the system blood/air, fat/air, and fat/blood. Employing thermodynamic relationships, one of the three coefficients can be calculated from the other two values. With respect to available human and mammal data, modeling efforts focus on the blood/air and fat/air partition coefficient. Respective data sets from literature have been collected and evaluated. The chemical domain of the sets is presented in terms of chemical structure, complexity, and polarity. Data gaps have been identified. The blood/air and fat/air partition coefficient data sets have been applied to a validation of literature models, with particular remark on the performance for specific compound classes. A new model for the blood/air partition coefficient and a preliminary new model for the fat/air partition coefficient are presented.

The study was supported by the EU projects 2FUN (contract No. 036976) and OSIRIS (IP, contract No. 037017).

P10 Modelling dissociation constants of organic acids by local molecular parameters
Haiying Yu1, R Kühne, R-U Ebert, G Schüümann
Department of Ecological Chemistry, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, 04318, Leipzig, Germany

The dissociation constant pKa defines the ionization degree of organic compounds. It is an important parameter that affects their toxicity and environmental behavior and fate. Among the present approaches to predict pKa values of organic chemicals, increment methods show a good performance but are limited by missing values of special groups. The purpose of this study is to develop models for predicting the pKa of organic acids directly from their molecular structures. A quantum chemical method was introduced by employing local molecular parameters. Experimental pKa values of more than 1000 organic acids, including phenols, aromatic carboxylic acids, aliphatic carboxylic acids and alcohols were selected, and all molecules were optimized using the semi-empirical AM1 Hamiltonian. Simple models with several descriptors for different subsets were calibrated by multilinear regression (MLR). Substituent positions on aromatic rings were also taken into account. The obtained models yield good predictive squared correlation coefficients q2 and superior performance compared with other quantum chemical models. The prediction capability is further evaluated using cross validation. The models unravel that the potential of oxygen atom conjoint to ionizable hydrogen atom to accept extra electronic charge is the single most important contribution to the dissociation of hydrogen atom. Non-linear statistical analysis methods were also applied because of the non-linear character of the employed descriptors.

The study was supported by the China Scholarship Council and by the EU project OSIRIS (IP, contact No. 037017), which is gratefully acknowledged.

P11 Where are the boundaries? Automated pocket detection for druggability studies
Andreas Volkamer1*, T Grombacher2, Matthias Rarey1
1 University of Hamburg, Center for Bioinformatics, Bundesstr. 43, 20146 Hamburg, Germany; 2Bioinformatics, Merck Serono, Frankfurter Str. 250, 64293 Darmstadt, Germany

Computer-based prediction of protein drugability is an essential task in the drug development process. Early identification of disease modifying
targets that can be modulated by low-molecular-weight compounds can help to speed up and reduce costs in drug discovery. Recently, first methods have been presented performing a druggability estimation solely based on the 3D structure of the protein [1-3]. The essential first step for such methods is the identification of the active site. A multitude of methods exist for automated active site prediction [4-6]. However, most methods developed for automated docking procedures do not explicitly focus on the definition of the boundary of the active site. Since druggability estimates are based on structural descriptors of the active site, a precise description of the active site boundaries is vital for correct predictions.

In this work, we present a method to predict protein binding pockets and split them into subpockets such that small molecules are mostly contained within one sub-pocket. The method is based on a novel strategy to geometrically detect narrow regions in pockets. For druggability predictions, such pocket descriptions result in more meaningful structural descriptors like active site surface or volume. Moreover, if several structures from one protein are known, sub-pockets can give hints about protein flexibility and induced fit conformational changes.

Our method was evaluated on 718 proteins from the PDBbind [7] data set, as well as 5419 proteins from the scFDB [8] data set. Binding pockets are extracted from the ideal model in 94% and 93% of the datasets, respectively, and 45% of the proteins from the two datasets contain pockets which can be divided into more than one sub-pocket. In all cases one sub-pocket completely covers the co-crystallized ligand. Besides the classical overlap-measure of ligand versus predicted active site, we additionally considered the pocket coverage by the co-crystallized ligand. We found that the number of test cases with more than 30% pocket coverage rises from 20% to 74% (PDBbind) and from 28% to 63% (scFDB), respectively, when considering sub-pockets.

References

P12
Protein negative/positive cooperatively binding to zwitterionic/anionic vesicles
F Torrens*, G Castellano
Institut Universitari de Ciència Molecular, Universitat de València, Edifici C\nInstituts de Paterna, P. O. Box 22085, 46071 València, Spain

The role of electrostatics is studied in the adsorption of cationic proteins to zwitterionic phosphatidycholine (PC) and anionic mixed PC/ phosphatidylglycerol (PG) small unilamellar vesicles (SUVs) [1]. For model proteins the interaction is monitored vs. PG content at low ionic strength [2]. The adsorption of lysozyme-myglobin-bovine serum albumin (BSA) (isoelectric point, pI 5-11) is investigated in SUVs, along with changes of the fluorescence emission spectra of the proteins, via their adsorption on SUVs [3]. In the Gouy-Chapman formalism the activity coefficient goes with the square of charge number [4]. Deviations from the ideal model indicate asymmetric location of anionic phospholipid in the bilayer inner leaflet, in mixed zwitterionic/anionic SUVs for protein-PC/PG, in agreement with experiments-molecular dynamics simulations. Effective SUV charge stays constant. Myoglobin-, DNC-melittin- and melittin-zwitterionic associations are described by a partition model, modulated by electrostatic charging of membrane as propeller interface. Provisional conclusions follow. (1) In mixed zwitterionic/anionic vesicles the electrostatic repulsion between cationic ad proteins dominates over the electrostatic attraction between ad protein dipoles. (2) The salt effect on the protein binding model of mixed zwitterionic/anionic vesicles was analyzed. The cooperativity increases with ionic strength. The corresponding interpretation is that the electrostatic repulsion between cationic ad proteins decreases with increasing salt effect, and the electrostatic attraction between ad protein dipoles becomes dominant over the electrostatic repulsion between ad protein charges. (3) In anionic vesicles the effect of vesicle charge on protein binding shows that, with increasing anionic character of the vesicles, the protein-protein electrostatic repulsion is decreasingly important vs. the protein-vesicle attraction, and the electrostatic attraction between ad protein dipoles becomes dominant over the electrostatic repulsion between ad protein charges. (4) For lysozyme-mixed zwitterionic/anionic vesicles and myoglobin cooperativity increases with pH. With decreasing pH and decreasing cationic character of the protein, the protein-protein electrostatic repulsion is decreasingly important against the protein-SUV attraction, and the electrostatic attraction between ad protein dipoles becomes dominant over the electrostatic repulsion between ad protein charges. Furthermore the opposed effect is observed for lysozyme-zwitterionic vesicles. (5) For protein-mixed zwitterionic/anionic vesicle binding there is more dispersion in the results, which could indicate asymmetric location of anionic phospholipid.

P13
CARTESIUS: a group function based toolkit for hybrid molecular modelling
Mikhail Rudenko1*, Al Tchougreeff1,2
1Institute of Physics and Technology of RAS, Nakhimovsky ave. 36/1, 112718, Moscow, Moscow Center for Continuous Mathematical Education, 119091, Moscow, Russia; 2Aachen, Germany

Modern methods of quantum chemistry have achieved significant successes. Nevertheless, modeling of certain classes of compounds remains problematic. The main complications are both high computational complexity of these methods and the lack of a unique way to determine the ground state with correct asymptotic properties for complex systems. One of the possible solutions is development of the hybrid molecular modeling methods.

Our approach to hybrid molecular modeling is based on the ideas of the group function approximation, first suggested by McWeeny [1] and further elaborated in [2]. This approximation lays a solid basis for simultaneous usage of most appropriate methods for computation of properties of each part of a complex system and consistent interpretation of the properties of the system independently of the methods used.

The CARTESIUS toolkit implements the above mentioned scheme in both flexible and computationally efficient way. This is achieved by means of combining cutting-edge C++ techniques for computationally intensive parts of the code and the full power of Python for the control interface.

As a result we have a carefully designed software system, which possesses both the power of the quantum chemistry codes developed previously with use of the group function approximation (BF [3], EHCF [4], BSCF [5] etc) and high potential for further enhancements.

This work is partially supported by the RFBR through the grant No 07-03-01128.

References
The description of chemical structures as a collection of connected molecular fragments is a basic requirement of coarse grained simulation methods like molecular fragment dynamics. These methods use molecular fragments as their basic interacting entities ("atoms") and allow the modeling and investigation of very large chemical systems. Therefore a molecular fragments chemoinformatics is in need that supports the fragment-based representation of chemical structures as well as the elementary operations upon them. The poster outlines definitions and approaches to tackle these issues from an adequate molecular line notation up to the graphical representation of simulation boxes.

P15
A theoretical investigation of microhydration of amino acids
Catherine Michaux, J Wouters, D Jacquemin, Eric A Perpète
Unité de Chimie Physique Théorique et Structurale, Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles, 61, B-5000 Namur, Belgium

Water plays a role in stabilizing biomolecular structure and facilitating biological function. Water can generate small active clusters and macroscopic assemblies, that are able to transmit information on various scales [1]. Protonation and microhydration of proteins or DNA are of fundamental importance in biochemical processes such as proton transport, water-mediated catalysis, molecular recognition, protein folding, etc. The understanding of these hydration effects at the molecular level requires the characterization of the interactions between biomolecules and their environment. Due to its relevance in many fields, the microhydration process of nucleic acid bases or amino acids has received a widespread attention [2].

In this work, we first describe the microhydration of protonated amino acid (AA), and particularly of protonated glycine (GlyH+) [3][4], alanine (AlaH) [5] and proline (ProH) [6]. First a high-level theoretical method was setting up in order to compute the structures and properties of GlyH+-water complexes, Gly being the simplest AA and a suitable model for such a study. Then complexes with more than one water molecule as well as other amino acids (Ala and Pro) were investigated, to extend the validity domain of our computational procedure.

We then investigate a series of complexes made up of a deprotonated (anionic) AA and a single water molecule [7]. Such species have recently been identified with mass spectrometry, allowing meaningful comparisons between theoretical and experimental complexation energies. The selected systems are [Gly-H], [Ala-H], [Val-H], [Asp-H], [Gln-H], as well as the acetate anion as a model benchmark.

References:
The scoring function was parameterized by reproducing reference alignments taken from four different proteins. The performance of the new alignment algorithm will be demonstrated by pairwise alignments obtained for examples from the FlexS data set [2] and by an additional example for multiple flexible alignment.

References

Bioisosteric similarity of drugs in virtual screening
Michael C Hutter*, Markus Krier
Center of Bioinformatics, Saarland University, Building C7.1, D-66041 Saarbrücken, Germany
Journal of Cheminformatics 2010, 2(Suppl 1) P18

Choosing compounds for screening is difficult problem due the vast chemical space. The question is thus which compounds are most likely to be hits. The comparison of drugs that all target the same enzyme, however, shows reoccurring chemical modification throughout all therapeutic categories. These so-called bioisosteric replacements [1] comprise simple exchanges of terminal atoms as well as more complex structural modifications, such as ring closures or rearrangements of larger fragments. To detect and evaluate all kinds of replacements we have designed an approach that adopts the algorithmic concept used to assess the homology of amino acid sequences to chemical molecules. The mutual exchange frequencies between distinct atom types are expressed in a substitution matrix [2]. Likewise, pairwise alignment between the molecules is constructed using dynamic programming [3], with the compounds being represented as unique SMILES. To obtain the actual exchange frequencies, we refined an initial matrix based on observed chemical replacements [4] by collecting the generated alignments of 1353 drugs from 33 therapeutic categories in an automated procedure. To compute the mutual bioisosteric similarity between two molecules a specific function has been derived that makes use of the alignment [5].

To assess the suitability of this bioisosteric similarity for virtual screening, we compared the recovery of known drugs against the background of other substances. The majority of drugs possess a higher similarity within the same class than compared to substances from the ZINC [6] or the Prous Science Drugs of the Future database [7]. Likewise, drugs for the same target are usually recovered at higher values of similarities than compared to other methods based on most common substructure and fingerprint approaches. Moreover, nondrugs without any pharmaceutical function exhibit considerably lower similarities than actual drugs. Furthermore, this bioisosteric similarity can be used to express the chemical diversity within a given compound class. We found that e.g. inhibitors of the HIV Reverse Transcriptase are more divers than Angiotension-II Antagonists and Tetracycline Antibiotics.

P19
Updating existing QSAR models: selection and weighting of new data
Tomas Öberg*, T Liu
University of Kalmar, Sweden
Journal of Cheminformatics 2010, 2(Suppl 1) P19

Computational chemistry and quantitative structure-activity relationships (QSAR) are foreseen to be extensively used in the implementation of the new REACH regulation for chemicals in Europe. However, for some compound groups the data are too few in number to permit both calibration and testing of a new model. Usage and previously developed or updated models are then viable alternatives.

Perfluorocarboxylic acids (PFCAs) and fluoroteleomer alcohols (FTOHs) are two groups of environmentally relevant compounds, with unique physical and chemical properties. The subcooled liquid vapour pressure (pl) is one such property, where experimental determinations are limited and far from consistent [1]. Updating is, however, challenging when the new compounds are far outside of the original calibration domain space. But by carefully selecting and weighting only three new compounds, we have been able to update a previously developed general QSAR model [2], to cover the new domain while maintaining predictive performance for the earlier calibration and test data. The optimal weighting scheme was determined from the sample leverages and residuals in the calibration phase [3].

The performance of this re-calibrated model greatly surpassed previous modelling attempts [4], when applied to an external test set of two PFCAs and four FTOHs with pl in the range 0.2-200 Pa; with Q2Ext = 0.994 and RMSEP = 0.190 units of log Pa. The domain coverage also increased from 1% to 51%, for 426 perfluoralkylated compounds selected from the REACH registration list, the PhysProp database, and the OECD 2006 survey [5]. Selection and weighting of new calibration data can thus facilitate the extension and use of existing QSAR models. This investigation was supported by the EU FP7 project CADASTER (grant agreement no. 212668).

References

P20
DrugScorePPI for scoring protein-protein interactions: improving a knowledge-based scoring function by atomtype-based QSAR
Dennis M Krüger*, Holger Gohlke
Heinrich-Heine-University, Computational Pharmaceutical Chemistry, 40225 Düsseldorf, Germany
Journal of Cheminformatics 2010, 2(Suppl 1) P20

Protein-protein complexes are known to play key roles in many cellular processes. Therefore, knowledge of the three-dimensional structure of protein-complexes is of fundamental importance. A key goal in protein-protein docking is to identify near-native protein-complex structures. In this work, we address this problem by deriving a knowledge-based scoring function from protein-protein complex structures and further fine-tuning of the statistical potentials against experimentally determined alanine-scanning results.

Based on the formalism of the DrugScore approach1, distance-dependent pair potentials are derived from 850 crystallographically determined
protein-protein complexes. These DrugScorePPI potentials display quantitative differences compared to those of DrugScore, which was derived from protein-ligand complexes. When used as an objective function to score a non-redundant dataset of 54 targets with "unbound perturbation" solutions, DrugscorePPI was able to rank a near-native solution in the top ten in 89% and in the top five in 65% of the cases. Applied to a dataset of "unbound docking" solutions, DrugscorePPI was able to rank a near-native solution in the top ten in 100% and in the top five in 67% of the cases. Furthermore, Drugscore-PPI was used for computational alanine-scanning of a dataset of 18 targets with a total of 309 mutations to predict changes in the binding free energy upon mutations in the binding interface. Computed and experimental values showed a correlation of $R^2 = 0.34$. To improve the predictive power, a QSAR-model was built based on 24 residue-specific atom types that improves the correlation coefficient to a value of 0.53, with a root mean square deviation of 0.89 kcal/mol. A Leave-One-Out analysis yields a correlation coefficient of 0.41. This clearly demonstrates the robustness of the method. The application to an independent validation dataset of alanine-mutations was used to show the predictive power of the method and yields a correlation coefficient of 0.51. Based on these findings, DrugscorePPI was used to successfully identify hotspots in multiple protein-interfaces. These results suggest that DrugscorePPI is an adequate method to score protein-protein interactions.

References

P21
Adaptation of formal concept analysis for the systematic exploration of structure-activity and structure-selectivity relationships
Eugen Lounkine1, Jürgen Bajorath
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany
Journal of Cheminformatics 2010, 2(Suppl 1) P21

Formal Concept Analysis (FCA) is a data mining and visualization approach originating from information science. It operates on binary relationships between objects and attributes, which are reported in a formal context. Formal concepts are sets of objects that share a defined subset of attributes. FCA organizes these concepts in lattices that reflect their relationship in terms of shared objects and/or attributes and allows the identification of objects with defined sets of attributes [1]. Two adaptations of FCA that allow the systematic analysis of structure-activity and selectivity relationships are presented. Fragment Formal Concept Analysis (FrFCA) assesses the distribution of molecular fragment combinations among ligands with closely related biological targets. This allows the identification of fragment signatures that exclusively occur in compounds with a defined activity profile. FragFCA also identifies fragment combinations that are characteristic of highly potent compounds against defined targets. Fragment signatures usually represent combinations of two or three fragments and can be used to differentiate active compounds of closely related targets for different target families [2,3].

Molecular Formal Concept Analysis (MoFCA) is introduced for the systematic comparison of the selectivity of a compound against multiple targets and the extraction of compounds with complex selectivity profiles from biologically annotated databases. Selectivity is assessed based on pair-wise compound potency ratios. This allows the definition multiple selectivity queries involving the comparison of an arbitrary number of targets and compound potency values or ratios. The individual queries are applied in a sequential manner to retrieve compounds with desired selectivity against targets of interest. MoFCA operates on activity space representations of compounds and thus allows the identification of structurally diverse compounds matching a given selectivity profile [4].

References

P22
Systematic extraction of structure-activity relationship information from biological screening data
Mathias Wawer1, L. Petason, Jürgen Bajorath
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universitat Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany
Journal of Cheminformatics 2010, 2(Suppl 1) P22

The analysis of high-throughput screening data poses significant challenges to medicinal and computational chemists. The number of compounds assayed in a single screen is prohibitively large for manual data analysis and no generally applicable computational methods have thus far been developed to consistently solve the problem of how to best select hits for further chemical exploration.

Focusing on the question of how structure-activity relationship (SAR) information can be used to support this decision, we present methods for the descriptive analysis of screening data. Network representations visualize the distribution of 2D similarity relationships and potency in a data set and give an overview of local and global features of an activity landscape. Although dominated by many weakly active hits, different local SAR environments can be identified among screening hits, thus helping to focus on regions in chemical space that might show favorable SAR behavior in further exploration [1].

A more detailed analysis of the data is achieved by systematically mining the network for SAR pathways, i.e. sequences of pairwise similar compounds that connect two molecules via a gradually increasing potency gradient. The SAR pathways are calculated exhaustively for all possible compound pairs in a data set to identify those having most significant SAR information content. Often, high-scoring pathways lead to activity cliffs, i.e. pairs of similar compounds with significant differences in potency, and scaffold transitions can be observed along the pathways. Furthermore, a tree structure organizes alternative pathways that begin at the same compound but lead to different molecules and chemotypes. Similarly, SAR trees can be generated from all pathways that lead to an activity cliff in order to characterize the surrounding SAR microenvironment [2].

References

P23
High throughput in-silico screening against flexible protein receptors
Horacio Pérez-Sanchez1, Bernhard Fischer, Daria Kohl*, Holger Merz*, Wolfgang Wenzel1
1Institut für Nanotechnologie, Forschungszentrum Karlsruhe, Postfach 3640, 76021 Karlsruhe, Germany; 2Wuppertal, Germany
Journal of Cheminformatics 2010, 2(Suppl 1) P23

Based on the stochastic tunneling method (STUN) [1] we have developed FlexScreen [2], a novel strategy for high-throughput in-silico screening of large ligand databases. Each ligand of the database is docked against the receptor using an all-atom representation of both ligand and receptor. The ligands with the best evaluated affinity are selected as lead candidates for drug development. Using the thymidine kinase inhibitors as a prototypical example we documented [3] the shortcomings of rigid receptor screens in a realistic system. We demonstrate a gain in both overall binding energy and overall rank of the known substrates when two screens with a rigid and flexible (up to 15 sidechain dihedral angles) receptor are compared. We note that the STUN suffers only a comparatively small loss of efficiency when an increasing number of receptor degrees of freedom is considered. FlexScreen thus offers a viable compromise between docking flexibility and computational efficiency to perform fully automated database screens on hundreds of thousands of ligands. We also investigate enrichment rates of rigid, soft and flexible receptor models for 12 diverse receptors using libraries containing up to 13000 molecules. A flexible sidechain model with flexible dihedral angles for up to 12 aminoacids increased both binding propensity and enrichment rates: EF1 values increased by 35% on average with respect to rigid-docking (3-8 flexible sidechains). This methodology will be soon available for the Cell processor and Pipeline Pilot.
References

P24
Virtual screening by high-throughput docking using hydrogen bonding constraints for targeting a protein-protein interface in M. tuberculosis
Oliver Koch1,2*, T Jäger2, L Flohe1, PM Selzer1
1Intervet Innovation GmbH, Zur Propstei, 55270 Schwabenheim, Germany; 2MOLISA GmbH, Brenneckestraße 20, 39118 Magdeburg, Germany

The resurgence of Tuberculosis, caused primarily by Mycobacterium tuberculosis, and the appearance of drug resistant tuberculosis strains encourage the need for new drugs with alternative mode of actions [1]. An interesting drug-target is the Thioredoxin-Reductase (TrxR)/Thioredoxin (Trx) System of M. tuberculosis, which is responsible for providing reducing equivalents for many cellular processes including bacterial antioxidant defence [2].

We performed a virtual screening approach for identifying novel drugs for the treatment of Tuberculosis that used the hydrogen bonding constraint functionality of GOLD for pre-processing instead of using pharmacoaphore filtering. Two important hydrogen bonding interactions could be identified at the protein-protein interface of the TrxR-Trx complex. A high-throughput docking was applied to filter the in vitro in-house compound library (~6.5 million compounds) for possible hits that interact with these two hydrogen bonding acceptors. This reduced the number of interesting compounds to 151768 that were redocked without any constraints and more accurate docking settings. Finally, the ranking of the reduced compound set was obtained using a normalisation based on molecular weight [3] and a consensus scoring in a restrictive way using Chemscore and ASPScore.

So far, six out of the first 25 of the ranked compound list showed an activity with an IC50 value up to μM range. Especially three compounds with different scaffolds and a low molecular weight are promising candidates for further developments.

References
1. World Health Organization Tuberculosis Programme. [http://who.int/tb/]

P25
Ensemble docking revisited
Oliver Körb1, S Bowdien, T Olsson, D Frenkel, J Liebeschuetz, J Cole
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK


In recent years, the importance of considering induced fit effects in molecular docking calculations has been widely recognised in the molecular modelling community. While small-scale protein side-chain movements are now accounted for in many state-of-the-art docking strategies, the explicit modelling of large-scale protein motions such as loop movements in kinase domains is still a challenging task. For this reason ensemble-based methods have been introduced taking into account several discrete protein conformations in the conformational sampling step. Our protein-ligand docking approach GOLD [1,2] has been extended to search such conformational ensembles time-efficiently. The performance of the approach has been assessed on several protein targets using different scoring functions. A detailed analysis of pose prediction and virtual screening results in dependence of the number of protein structures considered in the conformational ensemble will be presented and limitations of the approach will be highlighted.

References

P26
Picking out polymorphs: H-bond prediction and crystal structure stability
Peter TA Galek1*, Frank H Allen, László Fábtián
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK


A methodology has been developed to predict the propensity for hydrogen bonds to form in crystal structures, treating each potential H-bond as a binary response variable, and modelling its likelihood using a set of relevant chemical descriptors [1]. Modelling is tailored to a target using chemically similar known structures, from e.g. the Cambridge Structural Database [2], making it accessible to the complete spectrum of organic structures, including solvates, hydrates and cocrystals. Recent work has developed the approach to predicting inter- and intramolecular H-bonds when either type can occur.

By way of a comparison between possible and observed H-bonds, the method has been applied to assess structural stability, which shows much promise in the domain of polymorph screening in the pharmaceutical industry. We will introduce the methodology and illustrate its application using a selection of pharmaceutical compounds, one of which will be Abbott’s well-publicised anti-HIV medication ritonavir (Norvir®). Owing to a hidden, more stable form II with much lower bioavailability, ritonavir was temporarily withdrawn from the market with significant financial impact [3]. Our method quickly suggests a real threat of polymorphism in this compound, and strongly supports the relative stability of form I over form II. For all examples, the high predictivity of the method is emphasised.

References

P27
Kernel learning for ligand-based virtual screening: discovery of a new PPARα agonist
Matthias Rupp1*, T Schroeter2, R Stein1, Evgenij Proschak1, K Hansen2, H Zettl1, O Rau1, M Schubert-Zsilavecz1, K-R Müller1, Gisbert Schneider2
1University of Frankfurt, Griesmayerstr. 70, 60323 Frankfurt am Main, Germany; 2Berlin, Germany


We demonstrate the theoretical and practical application of modern kernel-based machine learning methods to ligand-based virtual screening by successful prospective screening for novel agonists of the peroxisome proliferator-activated receptor (PPARα) [1]. PPARα is a nuclear receptor involved in lipid and glucose metabolism, and related
OrChem: an open source chemistry search engine for Oracle

Mark L Rijnbeek, Christoph Steinbeck
EMBL EBI, Wellcome Trust Genome Campus, Cambridge, UK


Registration, indexing and searching of chemical structures in relational databases is one of the core areas of cheminformatics. However, little detail has been published on the inner workings of search engines and their development is mostly closed-source. We decided to implement an open source chemical library for Oracle, the de-facto database standard in the commercial world.

We present OrChem, an extension for the Oracle 11G database that adds registration and indexing of chemical structures to support fast substructure and similarity searching. The cheminformatics functionality is provided by the Chemistry Development Kit [1,2]. OrChem enables similarity searching with response times in the order of seconds for databases with millions of compounds, depending on provided similarity cut-off. For substructure searching, OrChem can make use of multiple processor cores on today's powerful database servers to provide fast response times in equally large data sets.

OrChem is a mix of PL/SQL and Java that executes inside the database. The user interacts with OrChem with calls to PL/SQL and Java Stored Procedures. Starting with Oracle 11g there is a just-in-time (JIT) compiler for the Oracle JVM environment which makes Java run much faster inside the database than previously.

OrChem is built on top of the Chemistry Development Kit (CDK) and depends on this Java library in numerous ways. For example, compounds are represented internally as CDK molecule objects, the CDK's I/O package is used to retrieve compound data, and its subgraph isomorphism algorithms are used for substructure validation. OrChem adds its own Java layer on top of the CDK to implement fast database storage and retrieval. With the CDK loaded into Oracle, a large cheminformatics library becomes readily available to PL/SQL. With little effort developers can build database functions around the CDK and so quickly implement chemistry extensions for Oracle.

References
Successful drug discovery often requires optimization against a set of biological and physical properties. We describe de novo design studies that demonstrate successful scaffold hops between known classes of ligands for p38 MAP kinases using ligand-based and structure-based multi-parameter scoring functions coupled to the molecular invention engine Muse. The ligand-based scoring function includes pharmacophoric and steric tuplets and structural (fingerprint based) similarity. In addition various selectivity or ADME related properties (e.g. Lipinski properties, polar surface area, activity at off-targets, etc.) can be taken into account to guide the evolution of structures meeting multiple design criteria. The structure-based scoring function uses Surflex-Dock to pose and score invented structures inside the target’s active site. In addition, a number of simple molecular properties (e.g. clogP, Lipinski properties, etc.) are used as score components to focus the design on medicinally relevant chemistries. With the ability of Surflex-Dock to start the docking process with a single or multiple placed fragments, this scoring function can be applied in fragment based drug discovery to optimize attachments onto a pre-placed substructure.

References


3. [http://www.knime.org/].

P34 Progress on an open source computer-assisted structure elucidation suite (SENECA)
Stefan Kuhn , Christoph Steinbeck
Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

We suggested and developed various components for Computer-Assisted Structure Elucidation (CASE) over the years [1,2]. Our current goal is to integrate these into an easy to use and efficient platform for end users, called SENECA. This is based on Bioclipse [3], an integrated software suite for chemo- and bioinformatics providing plugins for file handling and visualisation of compounds and spectra. The SENECA feature currently comprises the following components and algorithms:

- A deterministic structure generator suitable for small chemical spaces.
- Simulated Annealing and Genetic Algorithm for random structure walks in large spaces.
- Simulation of 13C NMR spectra for ranking results.

SENECA is a Bioclipse feature, available via the update site at http://www. ebi.ac.uk/steinbeck-srv/specbioclipse/net.bioclipe.seneca-updateSite/.

References


P32 Ligand-side tautomer enumeration and scoring for structure-based drug-design
Thomas Seidel*, G Wolber
InteLigand GmbH, Marahilferstrasse 74B/11, A-1070 Vienna, Austria

Tautomeric rearrangements of a molecule lead to distinct equilibrated structural states of the same chemical compound that may differ significantly in molecular shape, surface, nature of functional groups, hydrogen-bonding pattern and other derived molecular properties [1]. Especially for the structure-based pharmacophore modeling of ligand-protein complexes [2], knowledge of the most favorable tautomeric ligand states may be crucial for the quality and correctness of the generated pharmacophore models and derived putative binding modes. We will present a ligand-side tautomer enumeration and ranking algorithm that considers both geometrical constraints imposed by the conformation of the bound ligand as well as intra- and inter-molecular energetic contributions to find the preferred tautomeric states of the ligand in the macromolecular environment of the binding-site. The presented tautomer ranking algorithm is based on scores that are derived solely from MMFF94 [3-9] energies (thus the method works for a wide range of organic molecules) and has proven to be able to top-rank known preferred tautomeric states of ligands in a series of investigated protein complexes.

References


P33 Maximum-score diversity selection for early drug discovery
Thorsten Meinl*, C Ostermann, O Nemz, A Zalani, MR Berthold University of Konstanz, 78457 Konstanz, Germany

Diversity selection is a common task in early drug discovery, be it for removing redundant molecules prior to HTS or reducing the number of molecules to synthesize from scratch. One drawback of the current approach, especially with regard to HTS, is, however, that only the structural diversity is taken into account. The fact that a molecule may be highly active or completely inactive is usually ignored. This is especially remarkable, as quite a lot of research is involved in improving virtual screening methods in order to forecast activity. We therefore present a modified version of diversity selection – which we termed Maximum-Score Diversity Selection – which additionally takes the predicted activities of the molecules into account. Not very surprisingly both objectives – maximizing activity whilst also maximizing diversity in the selected subset – conflict. As a result, we end up with a multiobjective optimization problem. We will show, that the task of diversity selection is quite complicated (it is NP-complete) and therefore heuristic approaches are needed for typical dataset sizes. A common and popular approach is using multiobjective genetic algorithms, such as NSGA-II [1], for optimizing both objectives for the selected subsets. However, we will show that usual implementations suffer from severe limitations that prevent them from finding quite a lot of possible interesting solutions. Therefore, we evaluated two other heuristic for maximum-score diversity selection. One is special heuristic (called BB2) that was motivated by the mentioned proof of NP-completeness [2]. The other is a novel heuristics called Score Erosion which was specifically developed for our actual problem. Among all three heuristics, Score Erosion is by far the fastest one while finding solutions of equal quality compared to the genetic algorithm and BB2. This will be shown on several real world datasets, both public and internal ones.

All experiments were carried out using the data analysis platform KNIME [3] therefore we will also show some example how maximum-score diversity selection can be performed inside workflow-based environments.

References


3. [http://www.knime.org/].

P31 Multi-parameter scoring functions for ligand- and structure-based de novo design
Fabian Bos1, KM Smith2, Q Liu3
1Tripos International, Martin-Kollar-Str. 17, 81829 München, Germany; 2St. Louis, USA; 3Shanghai, PR China

Score components to focus the design on medicinally relevant chemistries. Especially for the structure-based pharmacophore modeling of ligand-protein complexes [2], knowledge of the most favorable tautomeric ligand states may be crucial for the quality and correctness of the generated pharmacophore models and derived putative binding modes. With the ability of Surflex-Dock to start the docking process with a single or multiple placed fragments, this scoring function can be applied in fragment based drug discovery to optimize attachments onto a pre-placed substructure.

References


Prediction of highly-connected ‘hub’-proteins in protein interaction networks using QSAR

M Hsing1*, K Byler2, A Chernkassov2
1University of British Columbia, 2329 West Mall, Vancouver, BC Canada;
2Rochester, USA


Proteins that are most essential for functioning and viability of bacterial cell have been shown to exhibit larger number of interactions with other cell components. Thus, by identifying the most connected proteins (or hubs) in protein interaction networks (PINs), one may discover prospective drug targets that can be utilized to combat emergent and drug-resistant pathogens such as Methicillin-Resistant Staphylococcus aureus 252 (MRSA). The advantage of using such hub proteins as drug targets lies in their essentiality, non-replaceable position in the PIN and lower rate of mutation, which can help to counter bacterial resistance. However, finding or predicting such hub proteins remains a challenging task as the corresponding experiments are very costly, while traditional bioinformatics approaches generally fail in forecasting PIN data due to the general lack of agreement between the existing datasets [1].

Thus, we have decided to utilize various structural and physicochemical features of proteins, related to traditional QSAR properties for predicting highly connected proteins. Using our own in-house generated PIN for the MRSA cell we have trained a boosting tree-based classifier that uses 75 physical and chemical QSAR descriptors computed for all proteins in the interaction network [2]. The utilized parameters included molecular weight, net charge, isoelectric point, hydrophobicity, surface area, solvent accessibility, electronegativity, secondary structure composition, surface coils and flexibility among other QSAR descriptors.

The developed QSAR model has yielded a high prediction accuracy of 80% for the validation set and was used to predict additional hubs in the rest of the MRSA proteome. The predicted hubs have then been evaluated experimentally and 55% of them were confirmed as high interactors which corresponds to >5 fold dataset enrichment for potential hub-proteins provided by the developed QSAR model.

Thus, the successful development of accurate hub classifiers demonstrated that highly-connected proteins tend to share certain structural and physicochemical features that can be characterized and quantified by conventional QSAR descriptors.

It is anticipated that the developed hub classifiers will represent a useful tool for the prediction of highly-interacting proteins and can find broad application for planning and executing large-scale proteomic experiments needed for a prediction. Depending on the number of ligands that needed prediction and time available, a choice for one method can be made. However, optimization of the charged group could be the interaction with its target. Changing the charge on the functional group could improve the interaction with its associated target and result in improved binding affinity.

Different methods have been developed for the computational determination of pKₐ values. These can be based on different methods such as QSAR [2] or quantum chemistry approaches [3]. These two methods differ considerably in terms of computational resource and hence, the time needed for a prediction. Depending on the number of ligands that needed prediction and time available, a choice for one method can be made. A comparison between several pKₐ predictors (Pipeline Pilot, Moka, Epik and Jaguar) was made. All methods perform well when a diverse set of ligands is considered. However, when optimizing a series of compounds the influence of small changes to the molecule and its effect on pKₐ becomes more difficult to predict.

References

Efficient extraction of canonical spatial relationships using a recursive enumeration of k-subsets

Georg Hinselmann*, Nikolas Fechner, A Jahn, Andreas Zell
University of Tübingen, Sand 1, 72076 Tübingen, Germany

The spatial arrangement of a chemical compound plays an important role regarding the related properties or activities. A straightforward approach to encode the geometry is to enumerate pairwise spatial relationships between k substructures, like functional groups or subgraphs. This leads to a combinatorial explosion with the number of features of interest and redundant information. The goal of this work is to compute all possible k-subsets of spatial points and to extract a single canonical descriptor for each subset in sub-polynomial computation time. More precisely, the problem is to reduce the complexity of \( n_k = n(n-1)...(n-k) \) possible relationships (patterns or descriptors) for n features and k-point relationships.

We propose a two-step algorithm to solve this problem. A modified algorithm for the computation of the binomial coefficient computes the k-subsets [1] containing the possible combinations of the n relevant features. If a k-subset is completed in the inner recursion, the algorithm computes a canonical representation for it. By defining a natural order by means of the geometrical center of gravity of the k points, we extract k patterns that describe the distance to the center of gravity and type of the spatial feature \( k \in F \). Then, the algorithm returns a unique identifier for the lexicographically sorted array of patterns. If applicable \( f_{n_k} < f_{n_{k-1}} < \ldots f_{n_1} \), an additional identifier is added which has the form \( f_{n_k} \xrightarrow{d_{n_k}} f_{n_{k-1}} \xrightarrow{d_{n_{k-1}}} \ldots f_{n_1} \xrightarrow{d_{n_1}} f_{n_0} \), where \( d_k \) denotes the geometrical distance between features i, j. Else \( f_{n_k} \leq f_{n_{k-1}} \leq \ldots f_{n_1} \), this step is omitted. Therefore, this approach also considers stereochemistry. Finally, one feature is returned for each k-subset resulting in a set of \( (n, k) \) patterns describing the structure. The main result is that the number of features is reduced from \( n_k \) to \( C(n, k) \), which equals the binomial coefficient. This procedure is useful in combination with similarity approaches that use spatial relationships, like pharmacophore searches, fingerprints, or graph kernels. We experimentally validated the algorithm on numerous QSAR benchmark sets in combination with the pharmacophore kernel [2].

References
these methods do not give any information if a molecule that is predicted can be sufficiently described by the knowledge contained in the model. Thus, the estimation of the reliability of a model-based prediction is an important question in machine learning based QSAR modeling.

One approach to solve this problem is to describe the portion of the chemical space used during the training phase of a model. Any molecule included in the same subspace is then considered as a structure for which the model is regarded as valid. This concept of the description of the subspace in which a model is regarded as reliable is known as the estimation of the applicability domain of this model [1].

Most machine learning approaches for QSAR rely on a vectorial representation of the molecules. The applicability domain is expressed as a subspace of the vector space with one dimension for each descriptor used. This concept can be not directly applied to kernel-based techniques like support vector machines. These methods rely on an implicit feature space that is only defined by the applied kernel similarity and with unknown dimensions. The applicability domain of a kernel-based model therefore has to be defined by means of the kernel. Consequently, this allows to use structured similarity measures, like the Optimal Assignment Kernel [2] and its extension [3], instead of a numerical encoding. Thus, it is possible to describe the complex chemical structure of many drugs better than it would be possible using descriptors.

In this work, several approaches to define the applicability domain of a QSAR model by means of a kernel are presented and compared to each other. The approach is to extend the concept of a kernel density estimation to incorporate further information contained in a trained model. This can be achieved by using a weighted average kernel similarity of a predicted molecule to the training data set. The weights can be obtained either by exploiting the knowledge contained in the learned model or by approaches that describe the feature space structure using the kernel.

References

P39 Expanding and understanding metabolite space
Julio E Peiróncely1, Andreas Bender1, M Rojas-Cheméro2, T Reimers2, L Coulter1, T Hankemeier1
1TNO, Quality of Life, Utrechtseweg 48, Zeist, The Netherlands; 2Netherlands Metabolomics Centre, University of Leiden, LACDR, Einsteinweg 55, 2333 CC Leiden, The Netherlands
Journal of Cheminformatics 2010, 2(Suppl 1)P39

In metabolomics the identity and role of low mass molecules called metabolites that are produced in cell metabolic processes are investigated. These make them valuable indicators of the phenotype of a biological system. The ‘Metabolite Space’ is the total chemical universe of metabolites present in all compartments and in all states from any organism. The molecules exhibit common features that form what can be called ‘metabolite likeness’. Here, we focus on the human metabolite space, including both endogenous and exogenous (such as drug) metabolites. In order to analyze the ‘Metabolite Space’, we collected data from the Human Metabolome Database (HMDB) [1] which is a comprehensive database for human metabolites containing over 7000 compounds that were identified in several human biofluids and tissues. As there still remain many compounds to be identified that lay outside the boundaries of this known space, exploring this unknown region is crucial to evaluate ‘metabolite likeness’.

In order to expand ‘Metabolite Space’ in our approach we employed the Retrosynthetic Combinatorial Analysis Procedure (RECAP) [2] to generate new molecules that possess features similar to those present in metabolites, however in other (but still likely) rearrangements. We studied how discernable these new molecules are from real metabolites and, hence, whether synthetic organic chemistry reactions are indeed able to expand the known universe of metabolites. We further studied the new chemistry present in the expanded metabolite space by looking at Murcko assemblies [3], ring systems and other chemical properties.

The new metabolite space is compared to other small molecules, such as those obtained from the ZINC database, that are not metabolites. By combining all the above analyses we expect to characterize better the metabolite space, and furthermore, to predict the metabolite-likeness of a molecule and to understand its immanent properties.

References
separation of reliable and non-reliable classifications. In quantitative predictions, the standard deviation of ensemble predictions has been found as the most accurate measure distance in a recent benchmarking [3]. We propose to integrate both metrics. Rather than giving a point estimate, this approach provides us with a probability distribution of finding particular compound in one of the classes. Suggested metrics is probability

\[
\int N(a, v, x)dx
\]

where \(E\) is class domain \(a\) - ensemble’s average prediction, \(v\) – variance of ensemble’s prediction, \(N(a, v, x)\) is probability density of the Gaussian distribution. Performance of this metric and its comparison to the traditional ones are evaluated for several QSAR/QSPR classification problems. The developed approach can be freely accessed to develop and estimate applicability domain of classification models at http://qspr.eu

References

P42 Automatic pharmacophore model generation using weighted substructure assignments
Andreas Jahn*, H Planatscher, Georg Hinselmann, Nikolas Fechner, A Zell
University of Tübingen, Sand 1, 72076 Tübingen, Germany

The generation of a pharmacophore model is a challenging process, which often requires the interaction of medicinal chemists. Given a number of ligands for a specific target, the aim is to identify the pharmacophore patterns that are responsible for the biological activities of chemical compounds. A recent study of optimal assignment methods has shown that the assignment of chemical substructures is able to detect active compounds in a data set [1]. Therefore, we investigated the possibility to use this technique to identify key features of a set of active compounds.

To determine important substructures of active compounds, we integrated \(n\) weight factors, where \(n\) is the number of substructures. The substructures were defined using the pharmacophore definitions of Phase 3.0 [2]. To define the individual weights of the pharmacophore patterns, we integrated a genetic algorithm which assigns weight factors to the previously defined patterns. The experimental setup was designed as follows: Given a data set with active compounds, the most active compound was selected as query structure for the experiment. The remaining active compounds were inserted into a background data set containing inactive compounds. The genetic algorithm evolved \(n\) weights for the pharmacophore patterns of the query structure. To evaluate the fitness of an individual, we performed a single query screening with the weights of the individual. During the optimization process, the BEDROC score [3] is optimized which puts emphasis on the early recognition performance. The result of the genetic algorithm was a weight vector that assigns each pharmacophore feature the weight of the best individual.

We evaluated our approach on a subset of the Directory of Useful Decoys that is suitable for ligand-based virtual screening [1][4]. The query structure was extracted from the same complexed crystal structure used by Huang et al. [4] to determine the binding site of the protein. The presented method is able to provide valuable information about key features that are important for the biological activity of a compound. Additionally, information of the protein structure is not needed. Therefore, the method can also be used to derive a pharmacophore model if no protein structure is available (e.g. GPCRs).

References

P43 A combined combinatorial and pKa-based approach to ligand protonation states
Tim ten Brink*, Thomas E Exner
Department for Chemistry, University of Konstanz, D-78457 Konstanz, Germany

The protonation of the ligand molecule and the protein binding site has a significant influence on the results obtained by protein-ligand docking. Due to the inability of X-ray crystallography to resolve the hydrogen atom in protein and protein complex structures, the correct protonation for the protein and the ligand has to be assigned on a theoretical basis before the structures can be used. Because of the local environment inside the binding site and because of the influence of the ligand and the protein onto each other, the ligand protonation can differ from the protonation one would expect for the ligand in solution under physiological conditions. Hence for protein-ligand docking different protonation states of the ligand have to be taken into account.

Our recently introduced structure preparation tool SPORES [1] used a rule based method to generate a standard protonation for each ligand molecule and afterwards generated different protonation states in a combinatorial way by adding and removing single hydrogen atoms belonging to predefined functional groups. Docking of all these protonation states of a given ligand molecule often led to scoring problems due to the different number of hydrogen interactions formed by the different protoners of the ligand molecule. To overcome these problems two methods to filter highly charged and unstable protonation states from the docking were implemented. One based on the difference between the standard protonation of the ligand and the actual protonation state and one based on pKa values calculated with ChemAxon’s MARVIN software [2]. Here we present a new approach in which the ligand atoms considered for the combinatorial method not chosen from predefined functional groups but according to the calculated pKa values which leads to a wider variety but with a smaller overall number of protonation states.

References
2. MARVIN 5.2.0, ChemAxon 2009 [http://www.chemaxon.com].

P44 Semi-empirical derived descriptors for the modelling of properties of N-containing heterocycles
Alexander Enzian1, Horst Bögel1, Mirko Buchholz2, Ulrich Heiser2
1Martin-Luther-University Halle-Wittenberg, Institute for Organic Chemistry, Kurt-Mothes-Str. 2, 06120 Halle, Germany; 2Probiotrag AG, Dept. Medicinal Chemistry, Weinbergweg 22, 06120 Halle, Germany

Nitrogen is one of the most prominent hetero atoms found in heterocycles. The corresponding electron lone pairs of these nitrogen atoms are mainly responsible for properties like the basicity and the pkb-values of the investigated heterocycle. Nevertheless, the overall electronic state of a molecule is also directly related to observable physico-chemical properties. This fact underlines a possible connection between different investigated properties.

The electronic properties of nitrogen containing compounds were analyzed with the aim to further predict these properties from quantum-mechanical descriptors. We thereby expect a relationship between the proton affinity, the pkb-value and the strength of metal complexation.
Because basicity seems to be a fundamental property of these compounds, the work was first focused on the proton affinities of the molecules in the gas phase. We thereby consider the fact, that these values are not influenced by the solvation in liquid phases. Lone pairs in nitrogen containing heterocycles play also an important role for metal complexation or due to their nucleophilic attack in chemical reactions. Therefore 55 heterocyclic compounds were selected that belong to the compound classes of substituted pyridines, pyrazoles and imidazoles. These compound set was carefully divided into a training (37) and test set (18), and different descriptors of the electronic states were calculated by using the semi-empirical molecular orbital software MSINDO [1]. In regression analysis the following descriptors were identified and marked as important:

- Charge of the nitrogen atom,
- Energy of the lone pair electrons and
- Perturbation treatment of the interaction energy - Klopman [2].

Using this set of descriptors a model was built and validated that predicts the experimental data for proton affinity with an R² of 0.93 and a Q² of 0.91. Furthermore, it was possible to successfully develop models for the prediction of pKb-values and the structure of a metal complex formation for these compounds.

References

P45
QSAR of anti-inflammatory drugs
Veronika R Khayrullina1, H. Bögel2
1Bashkir State University, RU-450074 Ufa, Russia; 2Halle, Germany

The computer analysis of relations between molecular structures and their biological activity using fragment-based methods is very useful to draw conclusions for the understanding of drug action and for the development of more efficient non-toxic drug candidates. We used the computer system SARDB21 (Structure Activity Relationship & Design) to investigate common structural features (fragments and substituents) typical for high- and low-effective non-steroid anti-inflammatory drugs (NSAIDs) successfully. This derived information has been used for the model for prediction of anti-inflammatory effectiveness of medicines with 76% and 81% level of recognition by two methods. This information could be used for creating new highly effective NSAIDs, and for increasing effectiveness of already known components.

In a second part of this paper the interrelation between structure and efficacy for anti-inflammatory drug action is carried out using traditional QSAR with descriptors from topology and from quantum-mechanical calculations followed by regression models from modelling.

The aim of this paper is to compare both molecular approaches of molecular design of drugs.

Reference

P46
Fingerprint-based detection of acute aquatic toxicity
Luca Pireddu1, I. Michielan1, M. Floris1, P. Rodriguez-Tome1, S. Moro2
1CR54, Pula, Italy; 2University of Padova, Padova, Italy

In this work we show the effectiveness of 2D structural fingerprints in the prediction of aquatic toxicity of chemical compounds, creating a self-contained system for structure-based aquatic toxicity classification. Using the data from the U.S. Environmental Protection Agency Fat Head Minnow (EPA-FHM) dataset [1] we build a non-linear RBF SVM [2] classifier that distinguishes acutely toxic compounds from less toxic compounds, loosely according to the criterion stipulated by the E.U. Reach legislation [3]. The classifier achieves up to 86% accuracy in leave-one-out validation using 580 of the dataset’s 614 compounds. This performance is comparable with models built from the same dataset using more sophisticated molecular descriptors, such as AutoMEP and Sterimol descriptors [4]. We apply our classification model to predict the aquatic toxicity of 3M compounds in the MMsINC database [5]. Furthermore, we create a linear SVM model using the same technique and apply it to the MMsINC data, with the additional integration of the EXPLAIN system [6] which allows us to show which structural features are responsible for the model classifying a molecule as less toxic or acutely toxic.

References
Evolutionary design of selective adenosine receptor ligands
Horst van der Eecke, J Krusselbrink, A Aleman, MT Emmenich,
Andreas Bender, AP IJzerman
Department of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2333CC Leiden, The Netherlands
Journal of Cheminformatics 2010, 2(Suppl 1) P48

Evolutionary de novo design is the design of new active compounds from scratch, using evolutionary principles. It involves an iterative cycle of structure generation, evaluation, and selection of candidate structures. Selected candidates serve as input for the next generation. By repeatedly selecting only the best structures as basis for structure generation, the process evolves toward better (not necessarily best) solutions. Here, we applied an evolutionary algorithm (the Molecule Commander) for the design of selective adenosine receptor ligands. Generated compounds were all rule-of-5 compliant and had a polar surface areas between 0 and 140 Å², favorable for intestinal absorption. In addition, toxic compounds were filtered out using a categorical SVM model for prediction of mutagenicity trained on Ames-test mutagenicity data (5-fold ROC score: 0.8948). Four pharmacophores were designed, one for each human adenosine receptor subtype. The hA1 receptor pharmacophore served as objective for the evolution, while the three other pharmacophores served as negative objective in order to obtain selectivity. In order to measure similarity with known adenosine receptor ligands, rings systems and scaffolds in the generated compounds were compared with those extracted from adenosine ligands in the StARLiTe database. With each new generation, the structures displayed an increasing number of ring systems also found in adenosine ligands while the number of unique core structures (scaffolds) increased as well. Eventually, the best (ADMET and pharmacophore score) candidate structures will be proposed for synthesis and tested for activity.

Ionotropic GABA receptors: modelling and design of selective ligands
Vladimir A Palyulin*, EV Radchenko, DE Osolodkin, VI Chupakhin, NS Zefirov
Institute of Organic Chemistry, Russian Academy of Sciences, Moscow 119991, Russia
Journal of Cheminformatics 2010, 2(Suppl 1) P49

Ionotropic GABAa and GABae receptors play an important role in the operation of CNS and serve as targets for many neuroactive drugs. Using the homology modelling and molecular dynamics, the 3D models of the receptors were built and some aspects of ligand-target interactions were elucidated [1,2]. To better understand the structural factors controlling the activity and selectivity of the ligands, a series of QSAR models [3] were derived based on the Molecular Field Topology Analysis (MFTA) [4], CoMFA and Topomer CoMFA approaches. They were compared with each other as well as with the molecular modelling results.

Finally, a number of potential selective ligand structures were identified by means of the virtual screening [5] from the available chemicals databases and the generated structure libraries.

References

PoseView – molecular interaction patterns at a glance
Katrin Stierand*, Matthias Rarey
Center for Bioinformatics (ZBH), University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany
Journal of Cheminformatics 2010, 2(Suppl 1) P50

Chemists are well trained in perceiving 2D molecular sketches. On the side of computer assistance, the automated generation of such sketches becomes very difficult when it comes to multi-molecular arrangements such as protein-ligand complexes in a drug design context. Existing solutions to date suffer from drawbacks such as missing important interaction types, inappropriate levels of abstraction and layout quality. During the last few years we have developed PoseView [1,2], a tool which displays molecular complexes incorporating a simple, easy-to-perceive arrangement of the ligand and the amino acids towards which it forms interactions. Resulting in atomic resolution diagrams, PoseView operates on a fast tree re-arrangement algorithm to minimize crossing lines in the sketches. Due to a de-coupling of interaction perception and the drawing engine, PoseView can draw any interactions determined by either distance-based rules or the FlexX interaction model (which itself is user accessible). Owing to the small molecule drawing engine 2Ddraw [3], molecules are drawn in a textbook-like manner following the IUPAC regulations.

The tool has a generic file interface for other complexes than protein-ligand arrangements. It can therefore be used as well for the display of, e.g., RNA/DNA complexes with small molecules. For batch processing, an additional command line interface is available: output can be provided in various formats, amongst them gif, ps, svg and pdf. Besides the underlying interaction models, we will present new algorithmic approaches, assess usability issues and a large-scale validation study on the PDB.

References
energies. To allow large molecular databases to be screened rapidly, simple and approximative scoring functions are used as a fast filter, resulting in low hit rates. Therefore, docking hit lists are commonly rescored in a final step by more rigorous and time-consuming methods to gain a more accurate final list of ranked compounds.

Molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) or generalized Born surface area (MM-GBSA) methods are currently considered to be suitable techniques for resoring. These physically realistic approaches incorporate more sophisticated models for solvation and electrostatic interactions than most scoring functions. Hence, they can discriminate more reliably between correct and incorrect docking poses. A high-throughput resoring protocol using force field-based methods has been proposed by Brown and Muchmore [1]. They used 18 in-house uronisin-ligand crystal structures and their corresponding experimentally determined binding free energies as a test set. On the basis of this resoring protocol we tested several molecular dynamics simulation protocols in combination with different MM-PBSA, and MM-GBSA calculation procedures using Amber 10 [2] and Gromacs 4 [3]. Considering performance and accuracy, our best resoring protocol performs similarly to the one described by Brown and Muchmore [1]. It has a comparable run-time and achieves a correlation between experimental and resored values of 0.88 compared to 0.87.

However, we used uronisin-ligand complexes generated using the docking program Glide [4] instead of crystal structures. Additionally, this shows that our resoring protocol improves the correlation of 0.57 between experimental values and Glide scores significantly to 0.88, thereby achieving a more accurate list of ranked compounds. The protocol will be incorporated into BALLView [5], an open-source molecular viewer and modeling tool. Thus, it will be available free of charge and can be conveniently used to resore (docked) protein-ligand complexes.

References

PS3
The pipelined metabolite identification based on MS fragmentation
Miguel Rojas-Cherto1, PT Kasper, Peironcely Julia E, T Reijmers, RJ Vreeken, T Hankemeier
Leiden, The Netherlands

Structural characterization and identification of components of complex biological mixtures constitutes one of the central aspects of metabolomics. Metabolite identification is a challenging but essential task in studies of biological samples. Mass spectrometry, because of its high sensitivity and specificity, is widely and successfully used in analysis of biological samples. Identification of metabolites can be in principle achieved using high resolution multistage mass spectrometry (MSn) because it provides a feature rich fingerprint of the precursor structure. However, neither general methodology for the identification nor extensive databases of metabolites with multistage mass spectrometric data are available at the moment. We demonstrate in this poster the feasibility of the strategy for metabolite identification based on analysis of fragmentation trees.

High resolution multistage MS experiments were performed on LTQ-Orbitrap (Thermo) equipped with Triversa nanoMate (Advion) nanoelectrospray ion source using defined protocol. An in-house developed software, integrating among others: Chemistry Development Kit (CDK) and XCMS libraries, was used for spectral data processing. Resulting fragmentation trees were stored in a local database. Multi-stage Molecular Formula (MMF) tool, which uses a method to resolve the elemental composition of the compound and fragment ions derived from MSn data using a cyclic constraining process has been developed. The process of elemental formula assignment and fitting within elemental formula paths removes artefacts of the spectra. Background signals, spikes, contaminations, side peaks, etc., are rejected in this stage and are not included in the fragmentation tree. The resulting fragmentation trees are stored in XML format.

Repeatability, reproducibility and robustness of fragmentation tree acquisitions were tested by changing experimental conditions and varying the concentration of the metabolite of interest. An acquisition protocol was established for the reliable and reproducible acquisition of mass spectral trees. It was investigated to which extent the variation of conditions such as fragmentation energy, isolation width etc. did change the fragmentation pattern or topology of hierarchical relations between fragments.

We demonstrate how the developed analytical strategy based on analysis of fragmentation trees can be used to discriminate between metabolite isomers with the same elemental composition and an only slightly different structure, but with a significantly different biological function.

Our results provide firm basis for developing a generic, multi stage mass spectrometry based platform for efficient identification of metabolites.

PS4
SBE13, a newly identified inhibitor of inactive polo-like kinase 1
Sarah Keppner1, Evgenij Proschak2, Gisbert Schneider2, B Spänkuch1
1Goethe-University, Medical School, Department of Gynecology and Obstetrics, Theodor-Stem-Kai 7, 60590 Frankfurt, Germany;
2Goethe-University, Frankfurt, Germany.

Protein kinases are important targets for drug development. The almost identical protein folding of kinases and the common co-substrate ATP leads to the problem of inhibitor selectivity. Type II inhibitors, targeting the inactive conformation of kinases, occupy a hydrophobic pocket with less conserved surrounding amino acids [1].

Human polo-like kinase 1 (Plk1) represents a promising target for approaches to identify new therapeutic agents. Plk1 belongs to a family of highly conserved serine/threonine kinases, and is a key player in mitosis, where it modulates the spindle checkpoint at metaphase/anaphase transition. Plk1 is over-expressed in all today analyzed human tumors of different origin and serves as a negative prognostic marker in cancer patients. The newly identified inhibitor, SBE13, a vanillin derivative, targets Plk1 in its inactive conformation [2]. This leads to selectivity within the Plk family and towards Aurora A. This selectivity can be explained by docking studies of SBE13 into the binding pocket of homology models of Plk1, Plk2 and Plk3 in their inactive conformation.

SBE13 showed anti-proliferative effects in cancer cell lines of different origins with EC50 values between 5 μM and 39 μM and induced apoptosis. Increasing concentrations of SBE13 result in increasing amounts of cells in G2/M phase 13 hours after double thymidin block of HeLa cells. The kinase activity of Plk1 was inhibited with an IC50 of 200 μM. Taken together, we could show that carefully designed structure-based virtual screening is well-suited to identify selective type II kinase inhibitors targeting Plk1 as potential anti-cancer therapeutics.

References

Cite abstracts in this supplement using the relevant abstract number, e.g.: Keppner et al. SBE13, a newly identified inhibitor of inactive polo-like kinase 1. Journal of Cheminformatics 2010, 2(Suppl 1):P54.