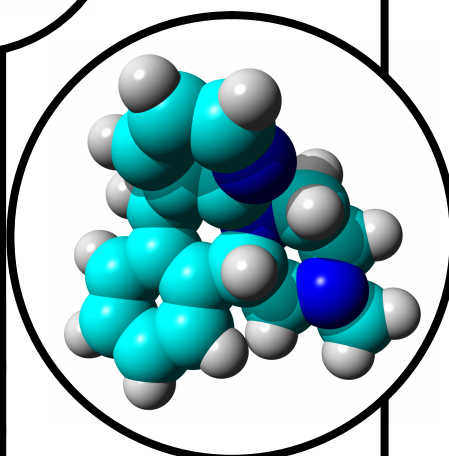


SEVENTH INTERNATIONAL
CONFERENCE ON

CHEMICAL STRUCTURES



Program &
Abstracts



June 5 - 9, 2005
Noordwijkerhout, The Netherlands

www.int-conf-chem-structures.org

Preface

Welcome to the Seventh International Conference on Chemical Structures!

With the Seventh International Conference on Chemical Structures we continue this well-established conference series that begun in 1973 as a workshop on *Computer Representation and Manipulation of Chemical Information* sponsored by the NATO Advanced Study Institute and thereafter was held under its new name every third year starting in 1987. The 2005 conference should continue the high standard of technical presentations and discussions that characterized all previous conferences. The response to the *Call for Papers* has produced an outstanding program of technical papers and posters and also attracted a sizable number of vendors and scientific institutions showing their newest software, content, and applications.

The scientific poster session has been divided into two sessions this year due to the large number of posters being presented. All posters will be exhibited during the poster sessions; however, presenters from the odd-numbered posters will be available during the Monday evening poster session (the blue team) and presenters from the even-numbered posters will be available during the Tuesday evening poster session (the red team).

The conference was chosen as the preferred venue to award the second CSA Trust Mike Lynch Award to Professor Johnny Gasteiger at the Universitaet Erlangen-Nuernberg. Professor Gasteiger will kick-off the conference by receiving the award and delivering the keynote address titled *My Love Affair with Molecules – and Reactions* on Sunday evening. Prior recipients of the CSA Trust Mike Lynch Award include Professor Peter Willett at the University of Sheffield in 2002

The Seventh International Conference on Chemical Structures kicked-off the permanent conference web site at www.int-conf-chem-structures.org. All information regarding the conference has been disseminated from the web site and this will continue for the Eighth International Conference on Chemical Structures in 2008. It will also serve as an archive for all prior ICCS conferences.

We hope that you enjoy the conference and if you ever need assistance during the week please contact one of the conference Organizing Committee or Scientific Advisory Board members listed on page 7.








Bob Snyder, Chair
Markus Wagener, Vice Chair

Contents

| | |
|--|-----|
| Organizing Committee | 7 |
| Scientific Advisory Board | 7 |
| List of Sponsors | 11 |
| List of Exhibitors | 15 |
| Exhibition Layout | 16 |
| Exhibition Hours | 16 |
| Technical Program | 19 |
| Plenary Session | 19 |
| Poster Session | 23 |
| Plenary Session Abstracts | 31 |
| Blue Poster Session Abstracts | 59 |
| Red Poster Session Abstracts | 87 |
| List of Participants | 113 |

**ORGANIZING COMMITTEE
AND
SCIENTIFIC ADVISORY BOARD**

Organizing Committee

| | | |
|---------------------------------------|--|---|
| Dr. Robert W. Snyder, USA | Chair | |
| Dr. Markus Wagener, The Netherlands | Vice Chair | |
| Dr. Lutgarde Buydens, The Netherlands | Royal Netherlands Chemical Society (KNCV) |  |
| Dr. Kimito Funatsu, Japan | Division of Chemical Information and Computer Science of the Chemical Society of Japan (CSJ) |  |
| Dr. Guenter Grethe, USA | The Chemical Structure Association Trust (CSA Trust) |  |
| Dr. Rainer Moll, Germany | Chemistry-Information-Computer Division, Gesellschaft Deutscher Chemiker (Society of German Chemists) (GDCh) |  |
| Dr. Don Parkin, United Kingdom | Chemical Information Group, the Royal Society of Chemistry (RSC) |  |
| Dr. Michele Parrinello, Switzerland | Swiss Chemical Society (SCS) |  |
| Dr. Terry Wright, USA | Division of Chemical Information (CINF), American Chemical Society (ACS) |  |

Scientific Advisory Board

| |
|--|
| Dr. Dimitris Agrafiotis, Johnson & Johnson Pharmaceutical Research & Development (2005 – 2011) |
| Dr. Kimito Funatsu, The University of Tokyo (2002 – 2008) |
| Dr. Val Gillet, University of Sheffield (2005 – 2014) |
| Dr. Michael Lajiness, Eli Lilly and Company (2002 – 2008) |
| Dr. Matthias Rarey, Universität Hamburg (2005 – 2011) |
| Dr. Robert W. Snyder, (2005 – 2011) |
| Dr. Lothar Terfloth, Universität Erlangen-Nürnberg (2005 – 2014) |
| Dr. Markus Wagener, N.V. Organon (2002 – 2008) |

LIST OF SPONSORS

List of Sponsors

Platinum Level



Gold Level



Silver Level



Student Bursaries

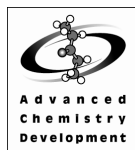


LIST OF EXHIBITORS

List of Exhibitors



Accelrys Ltd.



Advanced Chemistry
Development



Akos Consulting &
Solutions



Barnard Chemical
Information Ltd.



Bio-Rad Laboratories,
Informatics Division



Bioreason



Cambridge Crystallographic
Data Centre



CambridgeSoft



A Division of the American Chemical Society
CAS



ChemAxon
ChemAxon



Chemical Computing Group



COSMOlogic GmbH



Elsevier MDL



FIZ CHEMIE Berlin



Inte:Ligand



Molecular Networks GmbH



OpenEye Scientific
Software



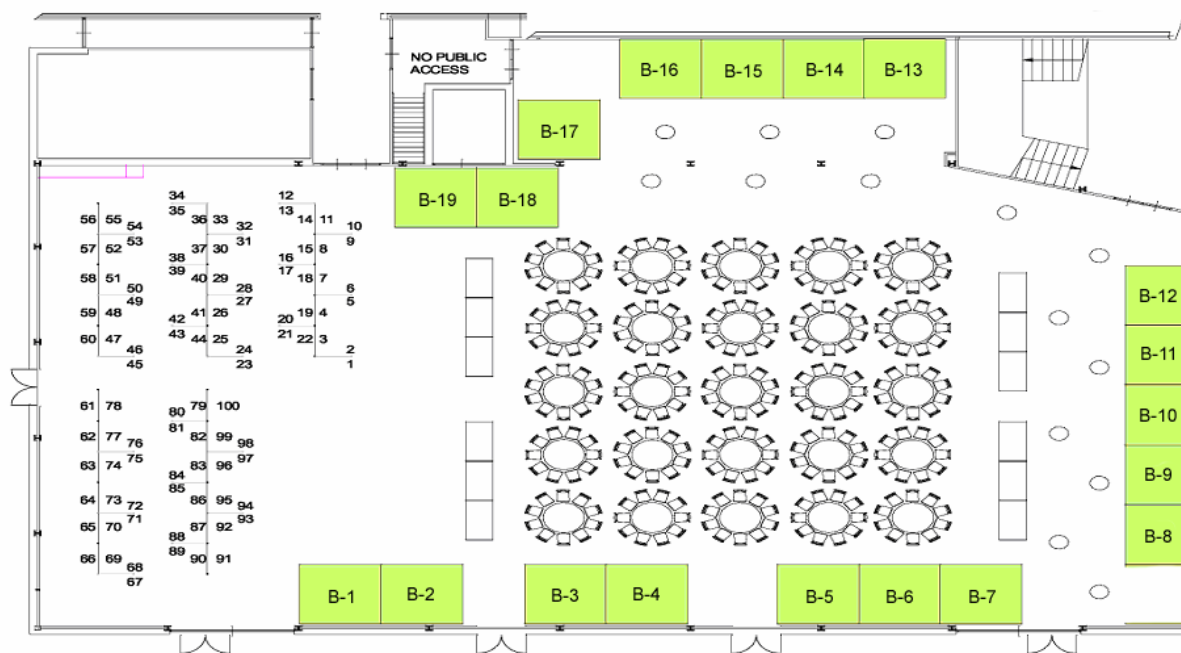
SciTeq



Tripos

Exhibition Layout

Atrium



- | | |
|-----------------------------------|--|
| 1. Tripos | 11. Barnard Chemical Information |
| 2. ChemAxon | 12. Cambridge Crystallographic Data Centre |
| 3. Chemical Computing Group | 13. SciTegic |
| 4. Akos Consulting & Solutions | 14. Accelrys Ltd. |
| 5. CambridgeSoft | 15. Inte:Ligand |
| 6. CAS | 16. COSMOlogic GmbH & Co. KG |
| 7. FIZ CHEMIE Berlin | 17. OpenEye Scientific Software |
| 8. Elsevier MDL | 18. Bio-Rad Laboratories, Informatics Division |
| 9. Advanced Chemistry Development | 19. Bioreason |
| 10. Molecular Networks GmbH | |

Exhibition Hours

| | |
|---------|---------------|
| Monday | 10:35 - 11:10 |
| | 13:00 - 14:00 |
| | 15:50 - 21:30 |
| Tuesday | 10:35 - 11:10 |
| | 13:00 - 14:00 |
| | 15:50 - 21:30 |

TECHNICAL PROGRAM

Technical Program

Plenary Session

| | |
|----------------|--|
| Sunday, June 5 | |
| 12:00-18:00 | Registration |
| 18:00 | Welcome and Introduction - Bob Snyder, ICCS Program Chair |
| 18:15 | Opening Session - Keynote Address, CSA Trust Mike Lynch Award K-1 : <i>My Love Affair with Molecules – and Reactions</i> Johnny Gasteiger, Universitaet Erlangen-Nuernberg |
| 19:00 | Welcoming Reception – Atrium , courtesy of Elsevier MDL |
| 20:00 | Rijsttafel Dinner – Atrium , courtesy of Chemical Abstracts Service |
| Monday, June 6 | |
| | Session A – Cheminformatics Michael Lajiness and Lothar Terfloth, Presiding |
| 8:15 | A-1 : <i>Similarity-Based Virtual Screening Using Data Fusion</i> Peter Willett, University of Sheffield |
| 8:45 | A-2 : <i>Quest for the Rings – A Cheminformatics Analysis to Identify Novel Bioactive Heterocyclic Systems</i> Peter Ertl, Novartis Institutes for Biomedical Research |
| 9:15 | A-3 : <i>Classification of Reactions by Type or Name</i> Guenter Grethe, Consultant |
| 9:45 | A-4 : <i>Molecular Similarity Searching Using the Conductor-Like Screening Model (COSMOSim)</i> Andreas Bender, Unilever Centre for Molecular Science Informatics |
| 10:15 | Product Review PR-1 : <i>Pharmacophore Elucidation in MOE (Molecular Operating Environment)</i> Steve Maginn, Chemical Computing Group |
| 10:25 | Product Review PR-2 : <i>Pipeline Pilot Product Update</i> Robert Brown, SciTegic |
| 10:35 | Break - Atrium |
| 11:10 | Product Review PR-3 : <i>Infotherm: Thermophysical Data of Mixtures and Pure Compounds</i> Joerg Homann, FIZ CHEMIE Berlin |
| 11:20 | Product Review PR-4 : <i>COSMO-RS Based Methods in Drug Design</i> Karin Wichmann, COSMOlogic GmbH & Co. KG |
| 11:30 | A-5 : <i>Aligning Ligand Ensembles in Multiplet Space</i> Robert D. Clark, Tripos, Inc. |
| 12:00 | A-6 : <i>Structure-Based 3D Pharmacophores: An Alternative to Docking?</i> Gerhard Wolber, Inte:Ligand Softwareentwicklungs- und Consulting GmbH |

| | |
|-----------------|---|
| 12:30 | A-7 : <i>Comparing Vector Versus Structural Coding for Predicting ADMETox Data Sets</i> Joerg Kurt Wegner, Universität Tübingen |
| 13:00 | Lunch - Atrium , courtesy of Hampden Data Services |
| 14:00 | A-8 : <i>Open Access/Open Source and the IUPAC International Chemical Identifier</i> Stephen Heller, NIST |
| 14:30 | A-9 : <i>ErG: A Two-Dimensional Pharmacophore Approach to Scaffold Hopping</i> Nikolaus Stiefl, Lilly Forschung GmbH |
| 15:00 | A-10 : <i>Improving Conformer Generation: A Sisyphean Task?</i> Matthew Stahl, OpenEye Scientific Software |
| 15:30 | Product Review PR-5 : <i>MDL Patent Chemistry Database & DiscoveryGate</i> Eva Seip, Elsevier MDL |
| 15:40 | Product Review PR-6 : <i>ChemAxon: Platform for Cheminformatics</i> Miklos Vargyas, ChemAxon Ltd. |
| 15:50 | Break - Atrium |
| 15:50 - 21:30 | Exhibits & Posters Open - Atrium |
| 16:30 - 18:30 | Poster Session Open – Atrium “Blue” authors will be present Rainer Moll & Guenter Grethe, Presiding |
| 18:30 | Reception – Atrium , courtesy of Accelrys Ltd |
| 19:30 | Buffet Dinner - Atrium |
| Tuesday, June 7 | |
| | Session B - Structure-Based Design and Virtual Screening Val Gillet, Presiding |
| 8:15 | B-1 : <i>FlexNovo: Structure-Based Searching in Chemistry Spaces</i> Jörg Degen, Zentrum für Bioinformatik |
| 8:45 | B-2 : <i>Recent Advances in De Novo Ligand Design and Optimisation</i> Krisztina Boda, University of Leeds |
| 9:15 | B-3 : <i>Combining the Power of Combinatorial Chemistry With the Efficiency of Pharmacophore-Based Docking</i> Holger Claussen, BioSolveIT GmbH |
| 9:45 | B-4 : <i>Using Molecular Fields to Derive Bound Conformations</i> Andy Vinter, Cresset BioMolecular Design Ltd |
| 10:15 | Product Review PR-7 : <i>LigandScout Review: An Application to Human Factor Xa Inhibitors</i> Monika Rella, Inte:Ligand GmbH |
| 10:25 | Product Review PR-8 : <i>The Proper Handling of Shape and Electrostatics in Ligand-based Design</i> George Vacek, OpenEye Scientific Software |

| | |
|---------------|--|
| 10:35 | Break - Atrium (Exhibits & Posters Open) |
| 11:10 | Product Review PR-9 : <i>Mogul: Rapid Retrieval of Molecular Geometry Information from a Crystallographic Database</i> Susan Robertson, Cambridge Crystallographic Data Centre (CCDC) |
| 11:20 | Product Review PR-10 : <i>KnowItAll®: An Integrated Solution for ADME/Tox, Spectroscopy, Cheminformatics, and Custom Software Development</i> Holger Ruchatz, Bio-Rad Laboratories, Inc - Informatics Division |
| 11:30 | B-5 : <i>Generation and Selection of Potential ER Ligands Using the De Novo Structure-Based Design Tool SkelGen</i> Henriette Willems, De Novo Pharmaceuticals |
| 12:00 | B-6 : <i>Knowledge-Based Design of Target-focused Libraries</i> Donovan Chin, Biogen Idec |
| 12:30 | B-7 : <i>The Nuclear Receptor Ligand Binding Domain: A Family-Based Structural Analysis</i> Simon Folkertsma; University of Nijmegen |
| 13:00 | Group Photo |
| 13:00 | Lunch - Atrium |
| 14:00 | B-8 : <i>Successful Virtual Screening for Tat-TAR RNA Interaction Inhibitors With a Fuzzy Pharmacophore Model</i> Steffen Renner, University of Frankfurt |
| 14:30 | B-9 : <i>An Effective Virtual Screening Protocol for Beta-Secretase (BACE1)</i> Tímea Polgár, Richter Gedeon Ltd |
| 15:00 | B-10 : <i>De Novo Design of PPAR-α Agonists Using ConTour Technology</i> Suresh Singh, Vitae Pharmaceuticals |
| 15:30 | Product Review PR-11 : <i>Automated Property-based Structure Design</i> Gavin Shear, Advanced Chemistry Development, Inc. |
| 15:40 | Product Review PR-12 : <i>Applications for the Modeling of Chemistry and Chemical Properties</i> Christof H. Schwab, Molecular Networks GmbH |
| 15:50 | Break - Atrium (Exhibits & Posters Open) |
| 15:50-21:30 | Exhibits & Posters Open - Atrium |
| 16:30 - 18:30 | Poster Session Open – Atrium “Red” authors will be present Don Parkin, Presiding |
| 18:30 | Reception – Atrium , courtesy of FIZ CHEMIE Berlin |
| 19:30 | Buffet Dinner - Atrium |

| Wednesday, June 8 | |
|-------------------|---|
| | Session C – Structure-Activity and Structure-Property Prediction Markus Wagener, Presiding |
| 8:15 | C-1 : <i>Modelling Cytochrome P450 Inhibition Using Large Datasets of in vitro Assays for CYP 2D6 and CYP 3A4</i> Boryeu Mao, Cerep Inc. |
| 8:45 | C-2 : <i>SPORCalc – Fingerprint Based Probabilistic Scoring of Metabolic Sites</i> Catrin Hasselgren Arnby, AstraZeneca |
| 9:15 | C-3 : <i>Relationships Between Molecular Complexity, Biologic Activity and Structural Diversity</i> Ansgar Schuffenhauer, Novartis Pharma AG |
| 9:45 | C-4 : <i>In silico Prediction of Buffer Solubility Based on Quantum-Mechanical, HQSAR- and Topology-Based Descriptors</i> Andreas Göller, Bayer Healthcare |
| 10:15 | Product Review PR-13 : <i>ClassPharmer™ Suite Automates Extraction of SAR to Maximize Antiviral Activity and Minimize Cytotoxicity</i> Vincent Vivien, Bioreason |
| 10:25 | Product Review PR-14 : <i>PASS</i> Alexander Kos, AKos GmbH |
| 10:35 | Break |
| 11:10 | Product Review PR-15 : <i>Fingerprints, Clustering and High-speed Virtual Library Analysis: BCI's Software Toolkits and Web Services</i> Geoff Downs, Barnard Chemical Information |
| 11:20 | Product Review PR-16 : <i>LITHUM Dock- A Virtual Assay System for Docking</i> Ulrike Uhrig, Tripos, Inc. |
| 11:30 | C-5 : <i>MC4PC: A Computational Tool for the Rational Evaluation of the Hazard Potential of New Pharmaceuticals and other Organic Chemicals</i> Gilles Klopman, Case Western Reserve University |
| 12:00 | C-6 : <i>Characterising Bitterness: Identification of the Key Structural Features</i> Sarah Rodgers, Unilever Research |
| 12:30 | C-7 : <i>The Molecule Evoluator: An Interactive Evolutionary Algorithm for Designing Drug-Like Molecules</i> Ad IJzerman, Leiden Center for Drug Research |
| 13:00 | Box Lunch |
| 13:00 - 23:00 | Group Excursion - once again we will visit The IJsselmeer (or Lake IJssel), where we will sail the seas, eat, drink, and be merry. Weather permitting, we will dock at one of towns bordering the IJsselmeer and create havoc with the locals. Busses will leave from the NH Leeuwenhorst Conference Hotel. You will be given a box lunch to eat during the bus ride to the boat harbor. Please bring appropriate clothing as we might see some rain on the lake and it can get a little chilly after sundown. A good pair of walking shoes is also advised. Dinner will be served during the sailing cruise and is provided courtesy of Chemical Computing Group. |

| Thursday, June 9 | |
|------------------|---|
| | Session D – Analysis of Large Data Sets Kimito Funatsu, Presiding |
| 7:30 – 8:15 | Hotel Check-out |
| 8:15 | D-1 : <i>Hit Selection From HTS Assays: Enhancing Hit Quality And Diversity</i> Iain McFadyen, Wyeth Research |
| 8:45 | D-2 : <i>Use of Multiple-Category Bayesian Modeling to Predict Side Effects</i> Robert Brown, SciTegic, Inc. |
| 9:15 | D-3 : <i>Scaffold-Hopping Using Clique Detection Applied to Reduced Graphs</i> Eleanor Gardiner, University of Sheffield |
| 9:45 | D-4 : <i>A First Look into ABCD</i> Dmitrii Rassokhin, Johnson & Johnson Pharmaceutical R&D |
| 10:15 | Product Review PR-17 : <i>SciFinder 2006: A Preview of Upcoming New SciFinder Features</i> Paul Peters, CAS |
| 10:25 | Product Review PR-18 : <i>Accord Cheminformatics Suite - Enhancements in v 6.0</i> Tim Aitken, Accelrys Ltd. |
| 10:35 | Break and Hotel Check-out |
| | Session E – Bridging the Cheminformatics-Bioinformatics Gap Matthias Rarey, Presiding |
| 11:15 | E-1 : <i>Integration of Chemical and Biological Data: The NCBI PubChem Project</i> Wolf Ihlenfeldt, National Institutes of Health |
| 11:45 | E-2 : <i>StARLite – a Chemogenomics Knowledge Base</i> Edith Chan, Inpharmatica |
| 12:15 | E-3 : <i>A Searchable Database for Comparing Protein-Ligand Binding Sites for the Discovery of Structure-Function Relationships</i> Richard Jackson, University of Leeds |
| 12:45 | Conference Closing Remarks - Markus Wagener, ICCS Vice Chair |
| 13:00 | Lunch or Box Lunch |
| 13:30 | Shuttle busses leave for Schiphol Airport |
| 14:30 | Shuttle busses leave for Schiphol Airport |

Poster Session

P-1 : *A Retrospective Docking Study of PDE4B Ligands and an Introduction into Methods of Avoiding Some Failures of Current Scoring Functions*

Chidochangu Mpamhanga; University of Sheffield, Sheffield, GB

P-2 : *Descriptors of Chemical Reactivity and Application to Mutagenicity Prediction*

Qing-You Zhang; Universidade Nova de Lisboa, Caparica, PT

P-3 : *Calculating Biases Using Artificial Intelligence in Conjunction with Data Assimilation*
Hamse Mussa; Cambridge University, Unilever Centre for Molecular Science Informatics, Cambridge, GB

P-4 : *Calculation of Interaction Energies Between DNA and Fluorescent Materials by Using Molecular Orbital Calculations*
Mitsuyo Aota; Yamaguchi University, Ube, JP

P-5 : *Storage and Processing of Chemical Information Directly from any Web Browser*
Luc Patiny; Ecole Polytechnique Fédérale de Lausanne, Lausanne, CH

P-6 : *Development of the Total System ToMoCo for 3D-QSAR and Molecular Design*
Masamoto Arakawa; University of Tokyo, Bunkyo-ku, JP

P-7 : *Incorporating the Flexibilities of Both the Ligand and the V82F/I84V Drug-Resistant Mutant HIV Protease Target During Docking: Applying the Relaxed Complex Method of Drug Design to HIV-1 Protease*
Alex Perryman; University of California at San Diego, La Jolla, CA, US

P-8 : *Incorporating Conformational Flexibility into QSAR: Validation of a Novel Alignment-Independent 4D-QSAR Technique*
Knut Baumann; University of Wuerzburg, Wuerzburg, DE

P-9 : *Making Real Molecules in Virtual Space*
Gyorgy Pirok; ChemAxon, Budapest, HU

P-10 : *Structure-Based Predictions of ¹H NMR Chemical Shifts and Coupling Constants Using Associative Neural Networks*
Yuri Binev; Universidade Nova de Lisboa, Caparica, PT

P-11 : *SAPPHIRE: Structure Aided Pharmacophore Implied Reagent Extraction – A method for in silico Screening*
Narasinga Rao; Scynexis Inc, Research Triangle Park, NC, US

P-12 : *Optimising the Effectiveness of Similarity Measures Based on Reduced Graphs*
Kristian Birchall; University of Sheffield, Sheffield, GB

P-13 : *Strategies for ACE2 Structure-Based Inhibitor Design*
Monika Rella; University of Leeds, Leeds, GB

P-14 : *Generation of a Focussed Set of GSK Compounds Biased Towards Ligand-Gated Ion Channel Ligands*
Anna Maria Capelli; GlaxoSmithKline, Verona, IT

P-15 : *Flexible Smoothed-Bounded Distance Matrix-Based Similarity Searching of the MDDR Database*
Nicholas Rhodes; University of Sheffield, Sheffield, GB

P-16 : *QSPR Study of Melting Point and Density of Imidazolium Ionic Liquids*
Gonçalo Carrera; Universidade Nova de Lisboa, Caparica, PT

P-17 : *BRUTUS: A Fully Automated Rigid-Body Superposition Tool*
Toni Ronkko; University of Kuopio, Kuopio, FI

P-18 : *New Descriptors from Energy Decomposition in Semiempirical Level*
Alexandre Carvalho; Universidade do Porto, Porto, PT

P-19 : *Modelling the Inhibition of P450 Enzymes*

Gijs Schaftenaar; Radboud University Nijmegen, Nijmegen, NL

P-20 : *Quantitative Analysis by Spectral Data Transformation in Multivalued Fingerprints and Multivariate Calibration*

Gonzalo Cerruela; University of Córdoba, Córdoba, ES

P-21 : *Treasure Island: Molecular 3D Shape-based Clustering with Neural Networks*

Paul Selzer; Novartis Institutes for Biomedical Research, Basel, CH

P-22 : *Study and Display of the Effect of Structural Similarity Approach in the Screening of Chemical Databases*

Gonzalo Cerruela; University of Córdoba, Córdoba, ES

P-23 : *Classification of Protein Kinases: Clustering Similarity Matrices Generated from Alignment and Novel Sequence-Based Descriptors*

Suresh Singh; Vitae Pharmaceuticals, Fort Washington, PA, US

P-24 : *Generation of Multiple Pharmacophore Hypotheses Using a Multiobjective Genetic Algorithm*

Simon Cottrell; University of Sheffield, Sheffield, GB

P-25 : *Distributed Search System CACTVS/SONORA: Search and Retrieval of Chemical Compounds and Associated Data from Very Large Databases*

Markus Sitzmann; National Institutes of Health, Frederick, MD, US

P-26 : *Advanced Structural Search Using ChemAxon Tools*

Ferenc Csizmadia; ChemAxon, Budapest, HU

P-27 : *Surrogate Docking: High Quality Docking at High Throughput Speeds*

Andrew Smellie; Arqule, Woburn, MA, US

P-28 : *Neural Networks for the Prediction of ¹H NMR Chemical Shifts of Sesquiterpene Lactones*

Fernando Da Costa; Universitaet Erlangen-Nuernberg, Erlangen, DE

P-29 : *ROBIA: A Reaction Prediction Program*

Ingrid Socorro; Cambridge University, Unilever Centre for Molecular Science Informatics, Cambridge, GB

P-30 : *Indexing the Chemical Semantic Web*

Nick Day; Cambridge University, Cambridge, GB

P-31 : *3D Structure-Activity Relationships of Non-Steroidal Ligands in Complex with Androgen Receptor Ligand-Binding Domain*

Annu Söderholm; Finnish IT Center for Science, Espoo, FI

P-32 : *VET: A Tool for Reaction Plausibility Checking*

Joseph Durant; Elsevier MDL, San Leandro, CA, US

P-33 : *Open Content Databases and Open Source Libraries for Chemoinformatics*

Christoph Steinbeck; Cologne University, Cologne, DE

P-34 : *Accurate Geometry Optimization Method for Molecular Mechanics*

Ödön Farkas; Eötvös Loránd University, Budapest, HU

P-35 : *Modeling the Metabolism of Xenobiotics*

Lothar Terfloth; Universitaet Erlangen-Nuernberg, Erlangen, DE

P-36 : *"Ultra-fast" Ligand-based de novo Design Using Virtual Reaction Schemes*
Uli Fechner; Goethe-Universitaet Frankfurt, Frankfurt, DE

P-37 : *BRUTUS: Rapid Optimization of Molecular Electrostatic Overlay - Evaluation of the Applicability of the Algorithm*
Anu Tervo; University of Kuopio, Kuopio, FI

P-38 : *The Molecule Evoluator: A Computer-based Tool for Drug Design*
Eric-Wubbo Lameijer, Leiden/Amsterdam Center for Drug Research, Universiteit Leiden, Leiden, NL

P-39 : *The Use of Exclusion Volume in Feature Based Alignment Pharmacophore Models: Catalyst HipHopRefine*
Samuel Toba; Accelrys, San Diego, CA, US

P-40 : *ScafReplace: Novel Tools for Scaffold Replacement*
Patrick Fricker; Center for Bioinformatics, Hamburg, DE

P-41 : *Diversity of Chemical Structure Libraries Characterized by the Distribution of Tanimoto Indices*
Kurt Varmuza; Vienna University of Technology, Vienna, AT

P-42 : *Genomic Data Analysis Using DNA Structure*
Eleanor Gardiner; University of Sheffield, Sheffield, GB

P-43 : *Detection of Toxicity Indicating Structural Patterns*
Modest von Korff; Actelion Ltd., Allschwil, CH

P-44 : *Increasing the Efficiency of Chemical Structure Storage and Retrieval in Large Relational Databases*
Sasha Gurke; Knovel Corp., Norwich, NY, US

P-45 : *Representing Structural Databases in a Self-Organising Map*
Ron Wehrens; Radboud University Nijmegen, Nijmegen, NL

P-46 : *Analysis of GRID Molecular Interaction Fields*
Sandra Handschuh; Boehringer Ingelheim Pharma GmbH, Biberach, DE

P-47 : *Drug Design Applications Based on COSMO-RS*
Karin Wichmann; COSMOlogic GmbH & Co. KG, Leverkusen, DE

P-48 : *Structural DNA Profiles*
Linda Hirons; University of Sheffield, Sheffield, GB

P-49 : *Techniques for Location-Independent Chemoinformatics Teaching and Research*
David Wild; Indiana University School of Informatics, Bloomington, IN, US

P-50 : *The Study of Bias Fusion of Chemical Similarity Searching*
John Holliday; University of Sheffield, Sheffield, GB

P-51 : *ChemXtreme: Harvesting Chemical Information From Internet Using Distributed Approach*
Muthukumarasamy Karthikeyan; National Chemical Laboratory, Pune, IN

P-52 : *A System Fusing Computational and Information Chemistry for Developing New Synthesis Routes of Compounds: An Application to the Synthesis Routes of Tropinone*
Kenzi Hori; Yamaguchi University, Ube, JP

P-53 : *On the Use of Spectra as Molecular Descriptors in QSAR Research*
Egon Willighagen; Radboud University Nijmegen, Nijmegen, NL

P-54 : *Universal Scripted Chemical Information Processing: the CACTVS Chemoinformatics Toolkit*
Wolf Ihlenfeldt; Xemistry GmbH, Lahntal, DE

P-55 : *The Development of a Machine Learning Algorithm for Ligand-Based Virtual Screening*
David Wood; University of Sheffield, Sheffield, GB

P-56 : *3D Structure Prediction and Conformational Analysis*
Gabor Imre; Eotvos Lorand University, Budapest, HU

P-57 : *Development of the Transition State Data Base*
Toru Yamaguchi; Yamaguchi University, Ube, JP

P-58 : *Uses and Potential Uses of Reasoning in Chemoinformatics*
Julian Hayward; Lhasa Limited, Leeds, GB

P-59 : *Pharmacophore Hypotheses Derived from Protein Structure and Inhibitors: Methods & Binding Site Comparisons of CYP3A4*
Litai Zhang; Bristol-Mayers Squibb, Princeton, NJ, US

P-60 : *Lead Conformers, a Thermodynamics Approach*
Adrian Kalaszi; Eotvos Lorand University, Budapest, HU

P-61 : *Automatic Classification of Chemical Reactions without Identification of Reaction Centers*
Qing-You Zhang; Universidade Nova de Lisboa, Caparica, PT

P-62 : *Finding Discriminative Substructures Using Elaborate Chemical Representation*
Jeroen Kazius; Universiteit Leiden, Leiden, NL

P-63 : *Chiral QSPR Analysis of ¹³C NMR Properties in Chiral Solvents*
Qing-You Zhang; Universidade Nova de Lisboa, Caparica, PT

P-64 : *scPDB: An Annotated Database of Three-Dimensional Structures of Binding Sites for Drug-Like Molecules*
Esther Kellenberger; University of Strasbourg, Illkirch, FR

P-65 : *Weighted Reaction Searching – Using Focused Fingerprints for Discriminated Results*
Tim Aitken; Accelrys, Cambridge, GB

P-66 : *Application of Knowledge-Based Scoring Functions for Virtual Screening*
Chrysi Konstantinou Kirtay; Cambridge University, Unilever Centre for Molecular Science Informatics, Cambridge, GB

P-67 : *Turbo Similarity Searching*
Jérôme Hert; University of Sheffield, Sheffield, GB

P-68 : *Selecting Potential Active Compounds by Matching Biological Profiles of Compounds with Known and Unknown Activities*
Alexander Kos; AKos Consulting & Solutions GmbH, Riehen, CH

P-69 : *Construction of a System Predicting Hydration Rates of Toxic Substrates in the Environmental Conditions*
Yutaka Ikenaga; Yamaguchi University, Ube, JP

P-70 : *Evaluation of the Diversity of Screening Libraries*
Mireille Krier; CNRS, Illkirch, FR

P-71 : *Structure-Based Design of Potential Novel Inhibitors of FGFR and VEGFR Tyrosine Kinase as Anti-Angiogenesis Agents*
Naparat Kammasud; Mahidol University & Centre Universitaire, Orsay Cedex, FR

P-72 : *Estimation of Environmental Compartment Half-Lives from Structural Similarity*
Ralph Kühne; UFZ Centre for Environmental Research, Leipzig, DE

P-73 : *Linking the Real and Predictive Worlds: A Conceptual Model of Chemical Information*
Chris Marshall; AstraZeneca, Macclesfield, GB

P-74 : *Water Solubility Prediction - Model Selection Based on Structural Similarity*
Ralph Kühne; UFZ Centre for Environmental Research, Leipzig, DE

P-75 : *Calculation of Physicochemical Descriptors Based on a new Structure Representation*
Jörg Maruszyk; Universitaet Erlangen-Nuernberg, Erlangen, DE

P-76 : *The Inconsistency of Medicinal Chemists in Reviewing Sets of Compounds*
Mic Lajiness; Eli Lilly & Company, Indianapolis, IN, US

P-77 : *Drug Design, Chemoinformatics and Public Web Services with Very Large Databases*
Marc Nicklaus; National Institutes of Health, Bethesda, MD, US

P-78 : *Chemical Clichés: Treasure Hidden by Obviousness?*
Eric-Wubbo Lameijer; Universiteit Leiden, Leiden, NL

P-79 : *QSAR Analysis for Infinite Dilution Activity Coefficients of Organic Compounds Using a CODESSA PRO Software*
Kaido Tämm; University of Tartu, Tartu, EE

P-80 : *¹H NMR – Based Classification of Photochemical Reactions*
Diogo Latino; Universidade Nova de Lisboa, Caparica, PT

P-81 : *A Neural Network Application in Multi-Target QSAR*
Pierre-Jean L'Heureux; Universite de Montreal, Montreal, CA

P-82 : *SOMA – Computational Molecular Discovery Environment*
Pekka Lehtovuori; CSC - Scientific Computing Ltd, Espoo, FI

P-83 : *The Quest for Bioisosteric Replacements*
Jos Lommerse; NV Organon, Oss, NL

P-84 : *Characterization and Clustering of Reagents for Combinatorial Library Design from the Products' Perspectives*
Uta Lessel; Boehringer Ingelheim Pharma GmbH, Biberach, DE

P-85 : *Similarity Searching Using Molecular Interaction Fields*
Kirstin Moffat; University of Sheffield, Sheffield, GB

PLENARY SESSION ABSTRACTS

Plenary Session Abstracts

K-1 : My Love Affair with Molecules – and Reactions

Johnny Gasteiger, Universitaet Erlangen-Nuernberg, Computer-Chemie-Centrum und Institut fuer Organische Chemie, DE

The international language of chemistry is a two-dimensional structure diagram. However, molecules are three-dimensional species much like human beings. I have always appreciated the beauty of molecules and, from the very beginning of my work in chemoinformatics, have resisted any attempt to dissect molecules into fragments. Rather, I have tried to respect the integrity of molecules and have treasured that they have a three-dimensional structure, have a skin, change shape, and have left- and right hands.

Computational approaches to the generation of 3D molecular models, to the calculation of molecular surface properties, to the generation of multiple conformations, and to the quantification of molecular chirality will be presented. It will be shown how these structure representations can be used to investigate the relationships between molecular structure and physical, chemical, and biological properties of compounds with particular emphasis on drug design.

In my PhD work I have struggled to elucidate the mechanism of some esoteric organic reactions. Since then, I have always considered the modeling of organic reactions by computer methods as a great challenge.

We have developed methods for the quantification of concepts, the organic chemist is using in discussing reaction mechanisms such as charge distribution, inductive, resonance, polarizability, and steric effect. We have used these physicochemical effects to model chemical reactivity and the course of chemical reactions, from laboratory reactions to biochemical pathways.

A-1 : Similarity-Based Virtual Screening Using Data Fusion

Peter Willett; University of Sheffield, Department of Information Studies, Sheffield, UK
Val Gillet, Jerome Hert, and Martin Whittle, University of Sheffield

Data fusion (which is referred to as consensus scoring in the ligand-docking community) is a general technique for combining the results of multiple database searches in systems for virtual screening [1]. The basic assumption of the data-fusion approach is that the use of multiple computational tools will enable a more effective prioritisation of a set of compounds for biological testing than will the use of a single such tool. In similarity-based virtual screening, data fusion has typically involved matching a bioactive target structure against the database molecules using several different similarity measures, and then merging the various rankings. For example, one could use multiple representations (e.g., a 2D fingerprint, a set of four-point pharmacophores, and a set of topological indices) or multiple similarity coefficients (e.g., the Forbes, Tanimoto and Russell-Rao coefficients). An alternative approach, and the one considered here, involves the use of a single similarity measure but multiple target structures, which we refer to as group fusion.

The idea that one can enhance retrieval effectiveness by using multiple target structures in a similarity search is not a new one (see, e.g., [2, 3]). We have recently compared several ways of combining the information from such multiple structures, using 2D fingerprints in extended simulated virtual screening searches of the MDL Drug Data report database [4]. This comparison demonstrates clearly the effectiveness of the group fusion approach: given some number (ten in our experiments) of known active target structures, match each of them against the database molecules and score each of these by the maximum of its similarities with the known actives. The similarity measure in these initial experiments was based on the Tanimoto Coefficient and Unity 2D fingerprints. More recent experiments have evaluated different types of similarity coefficient and 2D fingerprint for group fusion; these experiments suggest the general effectiveness of the Tanimoto Coefficient and Scitegic circular substructure descriptors. We have also demonstrated that group fusion is most effective when the set of actives that is being searched for is structurally heterogeneous, a situation that is difficult for conventional similarity searching and data fusion; group fusion, conversely, will add little to conventional similarity searching when the actives are structurally

homogenous [5, 6]. We are now developing a mathematical model of data fusion with the aim of being able to rationalise the effectiveness of different fusion rules.

1. Ginn, C. M. R. et al. (2000). *Perspect. Drug Discov. Design*, 20, 1-16.
2. Schuffenhauer, A. et al. (2001). *J. Chem. Inf. Comput. Sci.*, 43, 391-405.
3. Xue, L. et al. (2001). *J. Chem. Inf. Comput. Sci.*, 41, 746-753.
4. Hert, J. et al. (2004). *J. Chem. Inf. Comput. Sci.*, 44, 1177-1185.
5. Hert, J. et al. (2004). *Org. Biomol. Chem.*, 2, 3256-3266.
6. Whittle, M. et al. (2004). *J. Chem. Inf. Comput. Sci.*, 44, 1840-1848.

A-2 : Quest for the Rings – A Cheminformatics Analysis to Identify Novel Bioactive Heterocyclic Systems

Peter Ertl; Novartis Institutes for Biomedical Research, Basel, CH

Heterocyclic rings play a central role in organic synthesis, combinatorial chemistry, as well as in the search for bioactive molecules within drug design and discovery effort. Despite the enormous combinatorial possibilities the organic chemistry offers, however, the number of heterocyclic systems found in the common drugs, for example, is quite limited. Is this due to the fact that only a very small portion of ring space can exhibit biological activity, due to the synthetic inaccessibility of other heterocyclic systems, or is it simply a result of combinatorial explosion, where the whole current synthetic effort is just scratching the surface of a gigantic ring universe? The presented study tries to answer this question by analyzing heterocyclic rings with focus on their biological activity. First the ring systems found in known bioactive molecules were analyzed with the aim of identifying structural and property features necessary for activity, and which discriminate between active and inactive molecules. In the second step all "reasonable" heteroaromatic ring systems up to three fused rings were enumerated and their properties calculated. By using the knowledge derived in the first step, this large database of ring systems with calculated properties may be used to identify novel rings with a high chance of demonstrating biological activity, provide a basis for the design of novel combinatorial libraries with increased hit-rates, and generally help to fill voids in the ring universe.

A-3 : Classification of Reactions by Type or Name

Guenter Grethe; Consultant, Alameda, CA, USA

Josef Eiblmayer, Hans Kraut, and Peter Loew, Infochem

Much progress has been made over the years to facilitate reaction searching for the synthetic chemist. The complexity of reaction substructure searching was greatly simplified by classifying reactions based on reaction centers and immediate surroundings. The program, developed by InfoChem, is now implemented in most reaction databases. Despite its success, the program still requires formulating a structural query, which can be a daunting task for the occasional user.

From the beginning of synthetic organic chemistry, reactions have been described and classified according to their mechanism. Many of these reactions have been named after their discoverer or developer; they are well known and serve as useful mnemonics to the chemist. Approximately 700 of these reactions exist in the literature. Though reaction databases frequently contain this information, it is fragmented, intellectually derived and biased. It occurred to us that a web-based, hierarchical classification system taking into account reaction type and reaction name would be a valuable addition tool to reaction searching.

The hierarchy consists of the six main reaction types in organic chemistry and their subdivisions. Though built to eventually encompass all reactions our initial efforts were directed towards classifying named reactions and validating the approach. Assigning a database entry to a specific named reaction is done by substructure searches over the database. Each substructure search consisting of one or more queries assigns the corresponding keyword(s) to a reaction. For searching purposes the user has only to click on the tree displaying the individual named reactions.

We will describe the generation of the search system and its use exemplified by some searches.

A-4 : Molecular Similarity Searching Using the Conductor-Like Screening Model (COSMOsim)

Andreas Bender; University of Cambridge, Unilever Centre for Molecular Science Informatics, Cambridge, UK
 Andreas Klamt, Martin Hornig, and Karin Wichmann, COSMOlogic GmbH & Co KG
 Michael Thormann, Morphochem AG

We present a novel approach to define molecular similarity [1] and its application in virtual screening. The algorithm is based on molecular surface properties in combination with a geometric encoding scheme. The molecular surface is described by sigma values calculated using the COSMO-RS [2] (conductor-like screening model for real solvents) theoretical background. COSMO-RS is a quantum-chemical molecular description originally developed and widely validated for solubilities and partition coefficients of molecules in the liquid state. The descriptor also captures properties relevant to ligand-target binding such as hydrogen-bond donors and acceptors, positive and negative charges and lipophilic moieties. Encoding of the surface properties is performed using global sigma profiles (which encode no structural information), multiple-point pharmacophores and local sigma profiles. Various similarity / distance measures based on conventional similarity coefficients and integrals over sigma-profiles are employed for comparison.

Based on encouraging results obtained from initial applications [3], the approach was applied to a previously investigated dataset [4] derived from the MDDR which includes five classes of active compounds (5HT3 ligands, ACE inhibitors, HMG CoA reductase inhibitors, PAF antagonists and TXA2 inhibitors). Compared to other approaches, the approach presented here compares favourably with respect to the number of active compounds retrieved. In addition the descriptors employed here are able to retrieve compounds with similar properties yet different molecular scaffolds, a property commonly referred to as “scaffold hopping”. Thus the methodology possesses both a solid theoretical foundation and practical applicability. Further work will focus on the representation as well as the comparison of structures.

1. Andreas Bender and Robert C. Glen. Molecular Similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* 2004 (2) 3204 – 3218.
2. Frank Eckert and Andreas Klamt. Fast Solvent Screening via Quantum Chemistry: COSMO-RS approach. *AIChE Journal* 2002 (48) 369 – 385.
3. M. Thormann, M. Hornig, A. Klamt. Bioisosteric Similarity Search based on COSMO-RS alpha-Profiles. (in preparation for *JCIM*)
4. Hans Briem and Uta Lessel. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discov. Des.* 2000 (20) 231 – 244.

PR-1 : Pharmacophore Elucidation in MOE (Molecular Operating Environment)

Steve Maginn; Chemical Computing Group

The latest, 2005 version of the Molecular Operating Environment (MOE) system contains a new application for the derivation of pharmacophores from a database of known active compounds. This new capability builds on the Flexible Alignment, Conformation Import and Consensus Pharmacophore capabilities already available in MOE, and is receiving one of its first public showings at Noordwijkerhout.

PR-2 : Pipeline Pilot Product Update

Robert Brown; SciTegic

Pipeline Pilot is SciTegic's data pipelining product that allows the construction and rapid execution of workflows known as protocols. Individual operations within a protocol are provided by components, each of which performs a single task on the records flowing through the pipeline. The capabilities of Pipeline Pilot have been greatly expanded in recent releases through the addition of new component collections and through the availability of new components integrating 3rd party software. This product review will introduce new component collections from SciTegic for

- Report generation

- Web interface building
- Sequence analysis
- Statistical analysis
- Text analytics
- ADMET calculations.

In addition, both SciTegic and our ISV (Independent Software Vendor) partners are developing component collections that integrate 3rd party applications. We will highlight the latest developments and discuss the availability of these components.

PR-3 : Infotherm: Thermophysical Data of Mixtures and Pure Compounds

Joerg Homann; FIZ CHEMIE Berlin

The database Infotherm, currently available via Internet, comprises 167,000 tables of PVT-properties, phase equilibria, transport and surface properties, caloric properties, acoustic and optical properties of 25,000 mixtures and 6,700 pure compounds taken from journals, data collections, manuals and measurement reports some of which exclusive to Infotherm. Infotherm was relaunched in November 2004 with improved search functions in order to combine 98 properties, 25 conditions, 22 types of equilibria, 10 chemical systems, definable value ranges, substance names, formulas and CAS registry numbers by boolean operators. Besides the improved search procedure the data sources and quality assurance will be presented.

PR-4 : COSMO-RS Based Methods in Drug Design

Karin Wichmann, COSMOlogic GmbH & Co. KG

COSMOtherm is our approved program of the Conductor-like Screening Model for Realistic Solvation method (COSMO-RS), i.e. for the quantitative calculation of fluid phase thermodynamics based on quantum chemical COSMO calculations. Besides many features primarily developed for the very demanding chemical engineering mixture thermodynamics, COSMOtherm allows for predictive calculations of many properties relevant for life-science research, as water solubility and solubility in other solvents, partition coefficients between any solvents, and pKA. It has predefined models for logPOW, blood-brain partitioning, intestinal absorption, albumin binding, and allows for the definition of other models using the s-moments approach. New features of the current distribution include a graphical user interface (GUI), and improved COSMOtherm functionality and applicability, e.g. new pKA computation for bases.

COSMOfrag is a tool which performs the rapid automated fragment-wise construction of approximate s-profiles and the calculation of the resulting physicochemical properties of larger molecules, bringing the COSMO-RS technology to high-throughput performance. It is based on a database of presently approximately 40,000 COSMO files of highly diverse, small basic and larger drug-like compounds. After a highly efficient database search for the most appropriate locally similar fragments of a new compound, COSMOfrag generates COSMO meta files comprising the fragmentation description of the new compound. These meta files and the resulting fragment-based s-profiles can then be used as starting point for many COSMO-RS calculations, i.e. physicochemical, physiological or environmental property calculations, drug similarity searching or even receptor binding approaches.

COSMOsim provides a novel approach for the quantification of drug similarity, which makes use of the surface polarities *s* as defined in COSMO-RS. The histogram of these surface polarities, the s-profiles, have been proven to be the key for the calculation of all kinds of partition and adsorption coefficients, and thus of relevant ADME parameters like solubility, logBB and many others. It also carries a large part of the information required for the estimation of desolvation and binding processes responsible for the inhibition of enzyme receptors by drug molecules. Consequently, a large degree of similarity with respect to the s-profiles appears to be a necessary condition for drugs of similar physiological action. Driven by this insight, we have developed a s-profile based drug similarity measure SMS for the detection of new bioisosteric drug candidates. In several examples and in a number of real drug design projects COSMOsim already has demonstrated its statistical and pharmaceutical plausibility, its practicability for real drug research projects, and its unique independence from the chemical structure which enables scaffold hopping in a natural way.

COSMObase is a database of high quality DFT/COSMO files for about 2800 common compounds with a focus on common solvents. COSMObase can be very useful in the context of solvent and co-solvent selection and other screening applications.

A-5 : Aligning Ligand Ensembles in Multiplet Space

Robert D. Clark; Tripos, Inc., St. Louis, MO, USA

Edmond Abrahamian, Alexander Strizhev, Philippa Wolohan and Charlene Abrams, Tripos, Inc.

Existing methods of ligand alignment based on shared pharmacophoric features are deductive in nature, operate in Cartesian space and explicitly or implicitly require the selection of a single molecule as template - usually the one bearing the smallest set of features. This combination of factors sharply limits the ability of such methods to identify partial match constraints for use in 3D database searching. We have combined our recently developed fast pharmacophore multiplet technology with steric feature definitions and a novel genetic algorithm to produce a pharmacophore elucidation and alignment tool for carrying out true ensemble alignment in internal coordinate space. The program produces useful partial match search queries and allows hitlists based on such queries to be fitted to the model obtained.

A-6 : Structure-Based 3D Pharmacophores: An Alternative to Docking?

Gerhard Wolber; Inte:Ligand Softwareentwicklungs und Consultig GmbH, Vienna, AT

Thierry Langer, Inte:Ligand Softwareentwicklungs und Consultig GmbH

Chemical-feature based pharmacophore models have been established as state-of-the-art technique for virtual screening [1]. While feature-based pharmacophore recognition from a set of bio-active ligands is implemented in a number of programs [2, 3], the recognition of 3-D pharmacophores from macromolecular complex structures with bound ligands is hardly used and has not been implemented as fully automated procedure.

We present a novel program that automatically derives 3D pharmacophores from macromolecular complex data using customizable chemical feature definitions [4]. In a first step, small molecule ligands are extracted and improved from structure data using known algorithms including assignment of hybridization states and bond orders. Second, from the interactions of the interpreted ligands with relevant surrounding amino acids, pharmacophore models reflecting functional interactions like H-bonds, ionic transfer interactions or lipophilic contacts are created and projected into 3D space.

Pharmacophore objects can subsequently optionally be overlaid in order to find geometrically and chemically compatible pharmacophoric features occurring in several comparable complexes. The overlay algorithm utilizes a maximum clique detection algorithm [5] and an efficient, analytical alignment method [6]. Applications to inhibitor models of human rhinovirus serotype 16 coat protein and BCL-ABL tyrosin kinase are shown.

Finally, geometric fitting of the automatically generated pharmacophore to the bound ligands is compared with docking methods in terms of conformational recognition, flexibility and eligibility for virtual screening.

1. H. Kubinyi. In Search for New Leads, EFMC - Yearbook 2003, 14-28.
2. Catalyst, version 4.9. Accelrys Inc., San Diego, CA.
3. Molecular Operating Environment (MOE). Chemical Computing Group Inc., Montreal, Quebec, Canada.
4. G. Wolber, T. Langer. LigandScout: 3-D Pharmacophores derived from protein-bound ligands and their use as virtual screening filters J. Chem. Inf. Model., 2005, 45, 160-169
5. G. Wolber and T. Langer. CombiGen: A novel software package for the rapid generation of virtual combinatorial libraries. In Rational Approaches to drug design; Höltje, H.-D., Sippl, W., Eds.; Prous Science: 2000; pp 390-399.
6. Kabsch W. A solution for the best rotation to relate to sets of vectors. Acta Cryst. 1976, A32, 922-923.

A-7 : Comparing Vector Versus Structural Coding for Predicting ADMETox Data Sets

Joerg Kurt Wegner; Universität Tübingen, Tuebingen, DE
Holger Froehlich and Andreas Zell, Universität Tübingen

We will present an extensive Quantitative Structure Activity Relationship study comparing different coding types and hypothesis building algorithms for molecular data sets. The focus lies especially on the chemoinformatics problem for getting good hypotheses with allowing structural interpretability. We used for this analysis a LogP, LogS, Human Intestinal Absorption, Bioavailability, Blood Brain Barrier, and Toxicity data set. For the vector based coding we used a diverse set of 7000 descriptors at the beginning and applied several removal steps including several state-of-the-art feature selection approaches. For building the hypotheses we used Support Vector Machines, Decision Trees, and Nearest Neighbour induction algorithms.

For the structure based coding we used the Atom Pair coding and Maximum Common Substructure variants, for example the Highest Scoring Common Substructure search algorithm of Sheridan.

Finally, we will conclude our presentation with giving further perspectives for future research and the underlying optimization problems of the presented approaches.

A-8 : Open Access/Open Source and the IUPAC International Chemical Identifier

Stephen Heller; National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA
Stephen E. Stein and Dmitrii V. Tchekhovskoi, NIST

With the universal acceptance and use of the Internet, the ability for chemists and their colleagues in related fields to communicate more readily and at less expense has finally arrived. Open Access and Open Source are public domain, freely available projects which allow for the free exchange of information and are having and will continue to have a positive and profound effect on chemists worldwide. IUPAC has long been involved in the development of systematic procedures for naming chemical substances on the basis of their structure. The resulting rules of nomenclature, while covering a large fraction of compounds, were designed for text-based media. IUPAC has now developed an open source, public domain means of representing chemical substances in a format more suitable for digital processing and the Internet, involving the computer processing of chemical structural information (connection tables). This has led to the development of the IUPAC International Chemical Identifier, InChI. Details of InChI and related freely available Open Access information tools, such as journals will be discussed in this presentation.

A-9 : ErG: A Two-dimensional Pharmacophore Approach to Scaffold Hopping

Nikolaus Stiefl; Lilly Forschung GmbH, Hamburg, DE
Ian Watson, Eli Lilly Corporate Center
Knut Baumann, University of Wuerzburg
Andrea Zaliani, Lilly Forschung GmbH

Reducing molecules to their pharmacophoric features prior to the filtering phase is a widely accepted concept in virtual screening. Especially if the so called 'scaffold hopping', i.e. switching from one chemotype to another, is of interest, molecular abstraction of this type is frequently applied.

Here, a conversion procedure for two-dimensional structures is presented. The resulting molecular descriptor is an extension of the concept of reduced graphs [1] (ErG), which converted using radial distribution functions. That way, a highly condensed molecular descriptor of small size is obtained. Different methods of fuzzy incrementation, molecular abstraction as well as similarity measures are investigated. To assess the behaviour of the descriptor on real data sets, retrieval rates for a set of MDDR activity classes are studied and compared to DAYLIGHT fingerprints. Additionally, the ability to 'hop' between scaffolds is highlighted with a simple clustering approach. The results suggest that ErG shows retrieval rates similar to DAYLIGHT and that 'scaffold-hopping' is possible. Structural examples are given.

1. Barker, E.J.; Eleanor, J.; Gardiner, J.; Gillet, V.J.; Kitts, P.; Morris, J.; J. Chem. Inf. Comput. Sci. 2003, 43, 346-356

A-10 : Improving Conformer Generation: A Sisyphean Task?

Matthew Stahl; OpenEye Scientific Software, Santa Fe, NM, USA

The quality of answers obtained by molecular modeling is dependent on the accuracy of the models. Small molecular conformer generation is a key initial step in structure-based drug design, and accurate models can significantly improve the results of virtual screening and lead generation experiments. Improvements in conformer generation protocols, and their application in structure-based drug design will be presented.

PR-5 : MDL Patent Chemistry Database & DiscoveryGate

Eva Seip, Elsevier MDL, Frankfurt, DE

Patents are an important and under-used source of information in chemistry and life sciences research. While many text-based systems already exist for accessing patent information, structure-based searching offers a more powerful and flexible way for scientists to mine this vast pool of important information.

To provide this capability, Elsevier MDL offers the new structure-searchable MDL® Patent Chemistry Database, a factual database indexing reactions, substances and their properties from organic chemistry and life science patent publications (World and European since 1978, U.S. since 1976).

The MDL Patent Chemistry Database has been specifically designed for R&D scientists—chemists and bio-/medicinal scientists—to provide structure-based access to key patent chemistry information. The database can help researchers more effectively plan syntheses, better profile bioactivity and quickly understand the scope and relevance of patents. The database can assist researchers in designing new synthetic methods, developing drug profiles, selecting and optimizing leads and monitoring competitor activities and industry trends.

Updated every two weeks, the Patent Chemistry Database currently contains about 1.8 million reactions, about 2 million organic, inorganic, organometallic (and polymeric*) compounds and related information from approximately 360,000 patents. For more information, visit www.mdl.com.

With the Patent Chemistry Database available on the DiscoveryGate® content platform, scientists can explore patent information over a variety of integrated and complementary information sources such as Derwent Chemistry Resource or the MDL® Drug Data Report database.

The latest version of DiscoveryGate is faster, simpler and easier to use, displaying hits as soon as they are retrieved and requiring fewer clicks and windows to achieve results. See www.discoverygate.com.

* For patent applications published from December 2003 onwards

PR-6 : ChemAxon: Platform for Cheminformatics, Miklos Vargyas, ChemAxon Ltd.

Miklos Vargyas, ChemAxon Ltd., Budapest, HU

ChemAxon has been well-known in the pharmaceutical industry for its Java based web-enabled and platform independent cheminformatics components and tools, primarily for MarvinView which is a 2D structure visualizer and for MarvinSketch, a chemical drawing tool. Recently, however, ChemAxon has also been acknowledged as a cheminformatics platform provider.

The talk will move through the core technology of the JChem cheminformatics and discovery platform highlighting recent advances - in the form of connecting with posters being presented at the meeting. Mention will be made of the

results of development of JChem Base chemical database server and of our JChem Cartridge for Oracle and the integration of ChemAxon's Chemical Terms language for advanced searching.

A new product line developed in the past two years targets rational drug design. Applications include the Synthesizer and Reactor software that mimics chemical reactions in a knowledge based fashion, the Screen package coupled with JKlustor for high throughput virtual screening and library analysis.

Most recently ChemAxon made a move into the 3D world of cheminformatics. A 3D conformation generator was marketed first. The latest product to be released is MarvinSpace, an OpenGL based 3D structure visualizer component. In the future MarvinSpace will develop to a drug discovery platform. The presentation will highlight some interesting features in a live demonstration.

B-1 : FlexNovo: Structure-based Searching in Chemistry-spaces

Jörg Degen, Universität Hamburg, Zentrum für Bioinformatik, Hamburg, DE
Matthias Rarey, Universität Hamburg

At the beginning of a drug discovery project, a possible strategy for lead identification would be the structure-based search in extremely large Chemistry-spaces. Since such spaces cannot be enumerated for efficiency reasons, new tools for exploiting the combinatorial structure of the spaces are needed. This was the reason for the development of the new molecular design software FlexNovo. FlexNovo performs a structure-based search in large Chemistry-spaces following a sequential growth strategy. The Chemistry-spaces usually consist of several hundreds to thousands of chemical fragments and a corresponding set of rules, which primarily specifies how the fragments can be connected with each other.

FlexNovo is based on the FlexX[1] docking software and makes use of its incremental construction algorithm. In a first step, docking poses of single fragments under pharmacophore-type constraints are calculated and used as starting positions. Larger compounds are then subsequently generated by extending these initial fragments in a stepwise manner by adding fragments according to the specified rules. Interaction energies are calculated using a standard scoring function.

Compared to first generation de novo design software tools, FlexNovo allows the handling of large spaces and uses well-defined connection rules for the virtual synthesis of compounds. At the same time, a couple of filter criteria can be specified in order to obtain molecules which fulfill drug-like properties to a certain extent (e.g. Lipinski's Rule of Five).

FlexNovo has been used to design potential inhibitors for a couple of targets of pharmaceutical interest (e.g. cdk-2, cox-2 and estrogen receptor). This was done by using a Chemistry-space containing approx. 17000 fragments which were obtained through cleavage of retrosynthetically important (strategic) bonds in a collection of drug-like molecules. The binding energy was estimated by using the FlexX[2] as well as the ScreenScore[3] scoring function.

The determined structures were visually inspected and the proposed binding modes compared to modelled binding modes of known inhibitors. The compounds obtained show that FlexNovo is able to generate a diverse set of reasonable molecules. By comparing these to known inhibitors, similarities with respect to their binding modes are frequently observed.

1. M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.*, 261, 470 (1996).
2. H. J. Böhm, *J. Comput.-Aided Mol. Design*, 6, 593 (1992).
3. M. Stahl, M. Rarey, *J. Med. Chem.*, 44, 1035 (2001).

B-2 : Recent Advances in De Novo Ligand Design and Optimisation

Krisztina Boda; University of Leeds, School of Chemistry, Leeds, UK
Peter Johnson, S. Weaver, A.Vigh, V. Valko, University of Leeds

Systems which purport to carry out the protein structure based de novo design of ligands have now been available for over a decade. The extent to which this goal is achieved varies widely, and there are still many problems waiting to be solved in this area. The presentation will review some of these problems and discuss some of the solutions which have been developed in our laboratory. Areas covered will include: a) conformational sampling in structure generation b) methods for optimisation of existing ligands by automatic structure modification taking into account availability of starting materials c) navigation of answer sets using structural complexity analysis.

B-3 : Combining the Power of Combinatorial Chemistry With the Efficiency of Pharmacophore-based Docking

Holger Claussen; BioSolveIT GmbH, Sankt Augustin, DE
Markus Lilienthal, BioSolveIT GmbH
Hans Briem, Schering AG

A valid strategy for increasing binding affinity is to combine fragments that occupy different sub pockets. Current fragment-based experimental technologies can produce suitable fragment lists. Combined with available spacer groups following combinatorial chemistry protocols, libraries of enormous size are accessible in principle.

The docking program FlexX has a couple of additional modules containing specific functionality. FlexX-C allows for an efficient handling of combinatorial libraries by re-using partial compound placements in its incremental construction procedure. With this approach it is possible to speed up the average runtime by a factor of up to 20-30 compared to sequentially docking the enumerated library. The FlexX-Pharm module enables the user to guide the docking by using additional knowledge in the form of receptor-based pharmacophore constraints.

We revised and integrated these two modules. More powerful constraint definitions were enabled using logical expressions and SMARTS-based substructure definitions. FlexX-C, on the other hand, had to be enhanced in order to be able to maintain a key-feature of FlexX-Pharm, which is to check constraints not only after but also prior and during placement of the ligand in the context of the binding site. Correlation analysis of combinatorial and sequential docking runs gave new insights and led to an improved protocol for combinatorial library docking. Therefore the hybrid approach allows for pharmacophore-guided docking while maintaining the speed-up of FlexX-C.

In an example application we demonstrate how to define pharmacophore constraints for two specific sub pockets of a target and to dock a combinatorial library based on a set of suitable fragments for each sub pocket and a number of spacer groups. Following this approach, suitable linker groups can be efficiently detected.

B-4 : Using Molecular Fields to Derive Bound Conformations

Andy Vinter; Cresset BioMolecular Design Ltd, Letchworth, UK
Tim Cheesright, Mark Mackey, and Sally Rose, Cresset BioMolecular Design Ltd

Docking studies have been well exemplified as a useful tool for identifying hits given a protein x-ray structure or good homology model of the chosen target. Virtual screening using Molecular Fields is an alternative ligand-based similarity method which goes beyond structural comparison and can work independently of protein information. It therefore provides a logical in silico approach for finding hits against GPCR and other membrane spanning targets.

Fields model the potential binding sites on a molecule's surface. Two molecules of diverse structure can bind at the same site if they make the same interactions with the protein. We can use this observation to derive a bound conformation hypothesis for a ligand given two (or more) active compounds of diverse structure which are known to bind at the same site. This is done by running a conformational analysis on both compounds and identifying any pair of compound conformations that have the same Field pattern, and hence are both capable of making the same

binding interactions with the protein. In practice, three compounds are generally preferred to reduce the number of Field matches that are found.

We have used x-ray data from non-GPCR proteins with bound ligands to validate this hypothesis and will further report on case histories where virtual screens across commercially available databases have found active compounds with diverse structures not previously known to be active at the specified target.

PR-7 : LigandScout Review: An Application to Human Factor Xa Inhibitors

Monika Rella; Inte:Ligand GmbH

Factor Xa inhibitors are innovative anticoagulant agents that provide a better safety and efficacy profile compared to other anticoagulative drugs. We used LigandScout, a novel software package that fully automatically derives pharmacophores from protein-ligand-complexes, to identify interaction patterns of inhibitors bound to Human Factor Xa (PDB entries 1fjs, 1kns and 1eqz). The complex structures were selected regarding the criteria of high inhibitory potency (i.e. all ligands show K_i values against factor Xa in the sub-nanomolar range) and good resolution (i.e. at least 2.2Å) in order to generate selective and high quality pharmacophore models.

Common feature pharmacophores were generated using automated overlaying of several models obtained from the complex structures. The resulting chemical-feature based pharmacophore models were used for virtual screening of commercial molecular databases like the WDI 2003 database using the screening platform Catalyst [2]. The results obtained indicate that the models created using LigandScout are highly selective filters suitable for successful virtual screening experiments.

References

1. G. Wolber, T. Langer. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.*, 2005, 45, 160-169. (Version 1.0 will be available from www.inteligand.com)
2. Catalyst Version 4.7, Accelrys Inc., San Diego (CA), 2004.

PR-8 : The Proper Handling of Shape and Electrostatics in Ligand-based Design

George Vacek, OpenEye Scientific Software

ROCS and EON screen databases of molecules for shape and electrostatic similarity to a lead compound and thereby facilitate lead generation and library design for drug discovery. ROCS and EON report rigorous shape and electrostatic Tanimoto and Tversky measures between molecules, so that a molecular database can be quickly sorted according to similarity. More importantly, they provide an intuitive measure of similarity; that is, when two molecules with a high shape and electrostatic similarity are viewed, they clearly look alike and one can see that they should interact in a similar fashion. Shape and electrostatics are not only whole molecule properties but also fundamental metrics, giving them advantages over 2D descriptors and simple pharmacophore models.

Several recent publications validate the use of ROCS and EON in drug discovery. The first publication used ROCS to 'scaffold hop' to several active compounds without the same toxicity and potential intellectual property issues as their original lead compound. The new actives had been missed by the original HTS campaign. The second study, using both ROCS and EON to screen a virtual combinatorial library, showed the use of electrostatic similarity to be critical. One of the resulting leads was three-fold more active than the original query. In the third paper, shape and electrostatics were shown to be important variables in discriminating classes of active and inactive compounds for the diverse cases of COX2, progesterone, dopamine and the calcium ion channel.

Key new features in ROCS 2.1 and EON 1.1 include:

- support for CCP4 and XPLOR map files as grid queries
- new chemically aware "color" force-fields for donor, acceptor, anion, cation, rings and hydrophobic groups
- more efficient scaling to large numbers of processors

Referenced Publications

1. A Shape-Based 3-D Scaffold Hopping Method and its Application to a Bacterial Protein-Protein Interaction: Rush, T.S., Grant, J.A., Mosyak, L., Nicholls, A., *J. Med. Chem.*, 48 (2005) 1489-1495.
2. Variable Selection and Model Validation of 2D and 3D Molecular Descriptors, Nicholls, A., MacCuish, N.E., MacCuish, J.D., *J. Comp.-Aided Mol. Des.* 18 (2004) 451-474.

PR-9 : Mogul: Rapid Retrieval of Molecular Geometry Information from a Crystallographic Database

Susan Robertson, Cambridge Crystallographic Data Centre (CCDC)

The Cambridge Structural Database (CSD) currently contains over 345,000 structure determinations for small molecule organic and organometallic compounds, and thus is an excellent source of information on molecular geometries and conformational preferences. The program Mogul [1] (Molecular Geometry Library) has been developed by the CCDC to provide rapid access to this information with full statistics.

To retrieve information on a bond length, valence angle or acyclic torsion, the user inputs a molecule of interest into Mogul and selects the geometric feature to be investigated. The molecular environment of the selected fragment is used to generate atom- and bond-based "keys", which are then used to retrieve all relevant CSD hits. The use of a search tree optimises search speeds without the need for graph-based atom-by-atom matching typically used by other retrieval programs such as ConQuest [2]. Key statistics and histograms are derived and displayed along with an indication of how well the geometry of the feature in the input molecule matches observed geometries in the CSD. Mogul searches can be run automatically via an instructions file with results written to an output file, allowing easy integration with other programs. Mogul has already been integrated with CRYSTALS [3] in this way to allow validation of molecular geometries of newly refined crystal structures.

Mogul can also be used with DASH [4], software for structure solution from powder diffraction data.

1. I. J. Bruno, J. C. Cole, M. Kessler, Jie Luo, W. D. S. Motherwell, L. H. Purkis, B. Smith, R. Taylor, R. I. Cooper, S. E. Harris, A. G. Orpen, *J. Chem. Inf. Comput. Sci.*, 44(6), 2133, 2004
2. I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson and R. Taylor, *Acta Crystallogr.*, B58, 389, 2002
3. P. W. Betteridge, J. R. Carruthers, R. I. Cooper, K. Prout, D. J. Watkin, *J. Appl. Cryst.*, 36, 1487, 2003
4. W. I. F. David, K. Shankland, N. Shankland, *Chem. Commun.*, 931, 1998 [abstract]

PR-10 : KnowItAll®: An Integrated Solution for ADME/Tox, Spectroscopy, Cheminformatics, and Custom Software Development

Holger Ruchatz, Bio-Rad Laboratories, Inc - Informatics Division

Bio-Rad Laboratories, Inc. offers both desktop and enterprise-wide solutions to facilitate multiple aspects of pharmaceutical drug design and screening, industrial, and academic research including:

In Silico ADME/Tox Profiling & Consensus Modeling

Bio-Rad's KnowItAll offers a complete suite of award-winning tools for the *in silico* prediction of a potential drug's ADME/Tox profile, including over 30 predictive models, applications to validate and build predictive models (using SVM technology), and experimental ADME/Tox data.

Bio-Rad's unique environment for ADME/Tox features consensus modeling, which involves the combination of multiple, complementary models to improve the accuracy of prediction over single models. This presentation will demonstrate how drug discovery professionals can improve the accuracy of their *in silico* investigations using this technology and will review tools for global models and local models.

Spectroscopy Solutions for Multiple Spectral Techniques – Data Management & Mining

Bio-Rad offers multiple solutions for multiple spectral techniques including spectral data management for proprietary data, processing, analysis, and access over 885,000 high-quality MS, NMR, IR, and Raman spectra.

At ICCS, Bio-Rad will feature its new multi-technique spectral searching. Unlike other search software, KnowItAll searches data from ALL techniques at the same time. The system seamlessly consolidates all spectral data available to yield a single, intuitive result. By visualizing all spectral information simultaneously, KnowItAll is able to offer accuracy and efficiency beyond systems that handle only one spectral technique at a time.

Cheminformatics Solutions – Bio-Rad offers cheminformatics tools that are industry standards in laboratories worldwide with tools to draw, modify, store, search, name, and retrieve chemical structures. Using this system, complex mixtures of diastereomers can be stored and searched with elegance and ease. In addition, the system works in either a stand-alone configuration or in a state-of-the-art client-server solutions that offers the world's fastest system for searching structures, substructures, and spectral data.

Powered by KnowItAll® Program - Analytical instrument companies, pharmaceutical software and database suppliers, and other cheminformatics-related businesses can deliver fully customized solutions to their customers based on a selection of any combination of applications from the KnowItAll product toolkit. This custom application development approach provides several options to Powered by KnowItAll partners, including selection from over 20 standard applications (including structure drawing, reporting, analysis, data management, and data mining), instrumentation integration, and custom interface options. Bio-Rad will briefly introduce this program during this presentation. Bio-Rad associates would be pleased to meet with any parties interested in this program.

For more information, please visit www.knowitall.com.

B-5 : Generation and Selection of Potential ER Ligands Using the De Novo Structure-Based Design Tool, SkelGen

Henriette Willems; De Novo Pharmaceuticals, Histon, UK

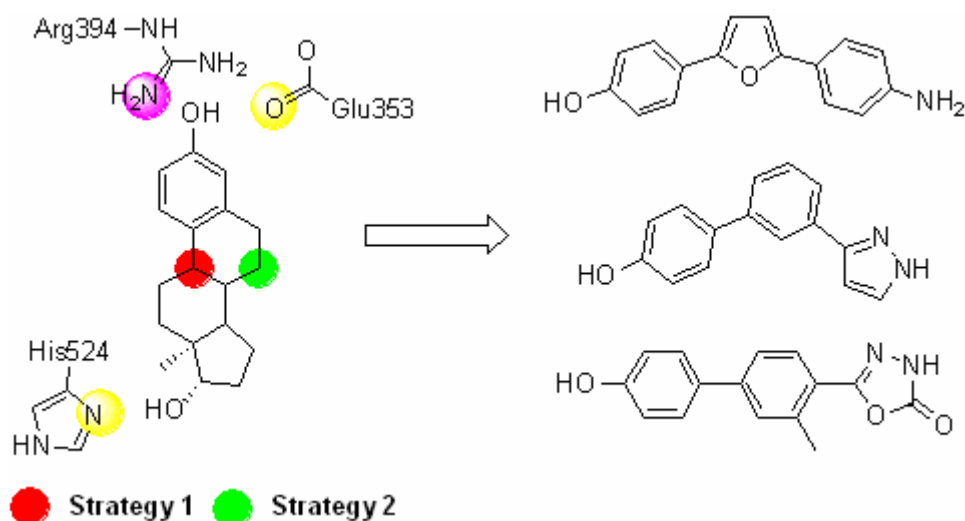
Stuart Firth-Clark, Anthony Williams, and William Harris, De Novo Pharmaceuticals

We report on the de novo structure-based design by Skelgen of novel, micromolar ligands against the Estrogen Receptor (ER).

SkelGen is an automated de novo design program that constructs ligands from fragments. SkelGen takes as input an X-ray crystal structure, a library of fragments, a set of rules on how to connect these fragments and, optionally, a pharmacophore to guide ligand generation. Here, two different pharmacophores or 'strategies' were used to guide ligand generation (see figure). Strategy 1 required hydrogen bonds to be formed with Glu-353 (donor needed), Arg-394 (acceptor needed), and His-524 (donor needed), in addition to a hydrophobic requirement in the centre of the active site. Strategy-2 required the same hydrogen bonding interactions with the protein, but the hydrophobic requirement is placed in a more buried part of the site. The default library of 1700 fragments derived from the WDI and the default connection rules were used.

SkelGen was used to design potential ligands for seven crystal structures of ER α . 1000 Ligands were generated for each combination of crystal structure and pharmacophore. Each set of 1000 was ranked using ScreenScore and the 25 best scoring ligands from each set were kept. The 350 ligands remaining were assessed on chemical diversity and synthetic accessibility, aided by the retro-synthetic analysis tool that is incorporated within SkelGen. 35 Compounds were selected for synthesis, of which 19 were produced within our chemistry time frame.

We show that we obtain micromolar activity for five of the 19 compounds (26%). Four of the active compounds are novel ER ligands. We believe this is a rare report of active ligands synthesized exactly as suggested by a de novo design program, without the use of a known active fragment.



B-6 : Knowledge-based Design of Target-focused Libraries

Donovan Chin; Biogen Idec, Cambridge, MA, USA
Claudio Chuaqui, Zhan Deng, and Juswinder Singh, Biogen Idec

In order to reduce screening costs and improve hit rates in the pharmaceutical industry there is a critical need for combinatorial chemistry to be able to design focused chemical libraries. Here we represent a new strategy for designing and filtering potentially massive combinatorial libraries by using the structure of the drug binding site to focus the chemical library. This method takes into account the 3-D structure of the active site of the target molecule and translates desirable ligand-target binding interactions into library filtering constraints. We have developed a variation of the structural interaction fingerprint (SIFt) called r-SIFt, which is tailored to handling interaction patterns from combinatorial libraries docked into protein binding sites. The R-sift incorporates the binding information of variable fragments in a combinatorial library, and we show using a test case the inflammatory kinase p38, that we can efficiently analyze and classify compounds based on their abilities to interact with the target with desired binding mode. Based on these classifications, decision tree models were generated using the molecular descriptors of the compounds as predictor variables. We have shown that these predictive models can be used as effective filters to sift through massive combinatorial libraries in order to generate smaller, target-focused subset of compounds for further investigations. Our results suggest that R-SIFT coupled with our classification models should be a valuable for structure-based focusing of combinatorial chemical libraries.

B-7 : The Nuclear Receptor Ligand Binding Domain: A Family-Based Structural Analysis

Simon Folkertsma; University of Nijmegen, Nijmegen, NL
Gert Vriend, University of Nijmegen
Paula van Noort, Ralph Brandt and Jacob de Vlieg, Organon NV
Emmanuel Bettler, BPCP

The huge amount of sequence and structural data on nuclear receptors requires automated methods to classify and analyse the role of key amino acids in the nuclear receptor ligand binding domain. By means of automated structural analysis we identified (1) frequent ligand binding residues, (2) important homo- and hetero-dimerisation residues and (3) selective cofactor binding residues. The ligand contact data shows that frequent ligand binding residues are mainly hydrophobic and form a core pocket in the nuclear receptor ligand binding domains. The identity of these frequent ligand binding residues determines the shape and the physico-chemical properties of the ligand binding pocket. Ideally, these properties are compatible with the physico-chemical properties of the ligand. Dendograms based on the most important ligand binding residues suggest novel potential cross reactivity of ligands between different subfamilies of nuclear receptors. In addition, subfamily selective ligand binding residues can be used to

guide the docking of novel ligands in these subfamilies. Finally our contact analysis revealed positions in the ligand binding domain that only interacts with antagonists and partial agonists. The contact information was translated into ligand-receptor interaction profiles that are very helpful in the design of selective compounds with a particular desired function.

B-8 : Successful Virtual Screening for Tat-TAR RNA Interaction Inhibitors with a Fuzzy Pharmacophore Model

Steffen Renner; University of Frankfurt, Frankfurt am Main, DE
Verena Ludwig, Oliver Boden, Ute Scheffer, Michael Göbel, Gisbert Schneider

We have recently introduced a fuzzy pharmacophore approach (“sophisticated quantification of interaction distributions”, SQUID) [1] for virtual screening for novel lead structures. In SQUID, an alignment of known active ligands is transformed into a set of Gaussian spheres modeling the density-distribution of potential pharmacophore points (PPPs). The result is a generalizing representation of the conservation and the spatial distribution of the PPPs over the set of aligned molecules. Finally this description is transformed into an alignment-free correlation vector (CV) representation, containing the density of pairs of PPPs. This CV can be utilized to prioritize molecules encoded with a pharmacophore-pair CV descriptor (CATS3D), [2] representing the conformations of single molecules in a screening database. The SQUID approach and CATS3D similarity searching [2] were employed to search for novel inhibitors of the Tat-TAR RNA interaction. Virtual screening of the SPECS database resulted in a small subset of 19 molecules which were experimentally tested for TAR RNA binding in a fluorescence resonance energy transfer (FRET) assay. Both methods retrieved molecules with comparable activity to the reference molecules. The best hit was found with the fuzzy pharmacophore and showed a ten-fold improvement to the references ($IC_{50} = 46\mu M$). This hit contained a different scaffold than the reference molecules. [3] We demonstrated that the fuzzy pharmacophore approach is appropriate for prospective screening.

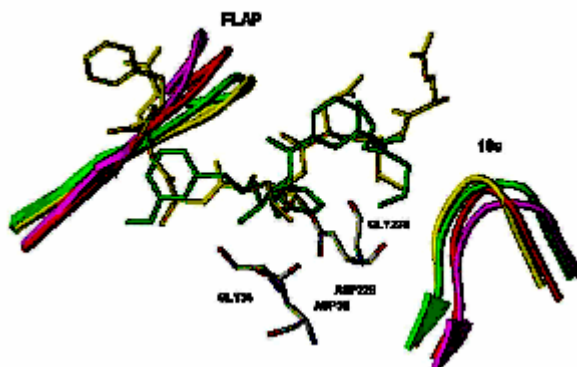
References

1. Renner, S., Schneider, G., Fuzzy Pharmacophore Models from molecular alignments for correlation vector-based virtual screening, *J. Med. Chem.*, 47 (2004), 4653-4664.
2. Fechner, U., Franke, L., Renner, S., Schneider, P., Schneider, G., Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aided Mol. Des.*, 17 (2003), 687-698.
3. Renner, S., Ludwig, V., Boden, O., Scheffer, U., Göbel, M., Schneider, G., New inhibitors of the Tat-TAR RNA interaction found with a “fuzzy” pharmacophore model. *ChemBioChem*, in press.

B-9 : An Effective Virtual Screening Protocol for β -Secretase (BACE1)

Tímea Polgár; Richter Gedeon Ltd, Budapest, HU
György M. Keserű, Richter Gedeon Ltd

A comparative virtual screen of FlexX and FlexX-Pharm considering different protonation states was performed on β -secretase. Up to now the exact protonation states of catalytic Asp residues (Asp32 and Asp228) have not been determined beyond ambiguity; various computational methods were applied and various protonation states for catalytic Asp residues were observed. Dissociation constants of titrating sites in crystal structures were computed by ZAP.



Crystal structures of BACE. Four different conformations can be seen: magenta and red: open conformations (apo-structures); green and yellow: closed conformations (ligand bound forms).

Assigning the calculated protonation states to Asp32 and Asp228 enrichment factors were significantly improved. We also show that pharmacophore constraints introduced by FlexX-Pharm maintained or improved the enrichment shown with FlexX. Hereby, effects of the binding conformations for the virtual screening output were also tackled in this particular case so we were able to show that the enrichment is not dependent here on the actual conformation.

Our virtual screening process was able to improve the enrichment given by FlexX from 14 to 40 for both apo and ligand bound conformations. Our results confirm the importance of the exact protonation states and pharmacophore constraints in virtual screening.

B-10 : De Novo Design of PPAR- α Agonists Using ConTour Technology

Suresh Singh; Vitae Pharmaceuticals, Fort Washington, PA, USA

Wenguang Zeng, Colin Tice, David Claremon, Jun Shimada, Jennifer Berbaum, Rich Harrison, Jerry McGeehan, Jean-Pierre Wery, and John Baldwin, Vitae Pharmaceuticals

We will describe a novel combinatorial de novo molecular growth technology [1] that allowed us to design a structurally distinct class of PPAR- α agonists. The underlying methodology uses a coarse-grained knowledge-based potential derived from high resolution protein-ligand structural data. A robust scoring function for predicting binding free energies of protein-ligand complexes was derived using a support vector algorithm by training on available structural data and the corresponding observed binding affinities. The iterative use of this de novo design approach rapidly allowed us to design, score, and rank thousands of novel synthetic targets with acceptable ADME properties. The select set of compounds synthesized exhibit high potency in our cell-based assay.

References:

1. Fast and accurate coarse-grained estimate of small molecule binding free energies. J. Shimada, A. V. Ishchenko, K. Jim, D. J. Lawson, P. R. Lindblom, G. Wu, J. Wery. 228th ACS National Meeting, Philadelphia, PA, August 22-26, 2004.

PR-11 : Automated Property-based Structure Design

Gavin Shear, Advanced Chemistry Development, Inc.

For the last 10 years, ACD/Labs has been dedicated to building integrated solutions that enable data transfer and connection within chemical organizations. Modern laboratories have to effectively manage multitudes of chemical structures and associated data. ACD/Labs offers chemical and analytical databasing solutions that enable chemists to manage their work effectively, and make stored chemical data available to other members worldwide. Our other key

products include renowned predictors of physicochemical properties (pKa, logD, logP, solubility, PSA, boiling point, etc.), nomenclature software, and spectral predictors, processors, and databasing tools.

Each year, we release enhanced versions of our software to provide more capabilities and superior integration between existing and new technologies. This year, we would like to present the following products:

ACD/Name: Generate Name from Structure or Structure from Name

New capabilities include:

- Support of German and French IUPAC nomenclature
- Support of the InChI protocol

ACD/ChemFolder: Database Chemical Reactions, Structures and Related Information

This universal databasing software also offers capabilities for mobile data management for Pocket PC and Palm OS, and new 2D barcode technology to encode chemical structures.

ACD/Structure Design Suite: Optimize Structures for Improved Properties

This design studio for synthetic and medicinal chemists suggests chemical modifications aiming to improve biological endpoint properties of lead compounds. Based on the relationship between the compound's structure and its physicochemical parameters, the software optimizes the desired structural fragment to offer a choice of analogs with better property of interest, providing a physicochemical basis for future synthetic directions. In our presentation we will highlight the recent applications of this approach in pharmaceutical and agrochemical environments.

PR-12 : Applications for the Modeling of Chemistry and Chemical Properties

Christof H. Schwab, Molecular Networks GmbH

Molecular Networks' core expertise is in the development of computationally fast, empirical descriptors for the modelling of electronic properties of atoms and bonds that would usually require quantum mechanical calculations.

These descriptors are the basis of software tools designed to support researchers in the pharmaceutical, biotechnology and chemical industry and have a wide range of applications: from the prediction of molecular properties used by computational chemists for discovering and optimizing lead compounds to the decision support applications that help chemists to plan a synthesis of chemical compounds.

Molecular Networks product portfolio was recently extended with the applications

- ADRIANA.Code for the calculation of molecular physicochemical properties and for computing 2D-, 3D- and surface-based vectorial molecular descriptors,
- SPINUS for the prediction of ¹H NMR chemical shifts

and will include an updated version of the current synthesis planning system WODCA in the very near future.

Molecular Networks product portfolio also includes applications for processing, manipulating, visualizing and warehousing chemical structures and reactions.

Due to Molecular Networks and SciTegic's collaboration, several applications (such as 2DCOOR, ADRIANA.Code, CONVERT and CORINA) are available as components for Pipeline Pilot.

Molecular Networks will be demonstrating its suite of software applications and decision support tools at the exhibition at Booth #10. The booth personnel will be available to answer any questions regarding Molecular Networks' technology and will be glad to help you to find the right solution for your needs.

C-1 : Modelling Cytochrome P450 Inhibition Using Large Datasets of in vitro Assays for CYP 2D6 and CYP 3A4

Boryeu Mao; Cerep Inc., Redmond, WA, USA
R.Gozalbes, F.Barbosa, J.Migeon, S.Merrick, K.Kamm, E.Wong, C.Costales, W.Shi, N.Froloff, D.Horvath, and C.Wu, Cerep SA

Cytochrome P450 enzymes (CYPs) constitute a superfamily of heme proteins involved in the metabolism of endogenous and exogenous compounds in living organisms. In particular, a number of CYP isozymes are important in the oxidative metabolism of drug molecules and other xenobiotics. Interactions with these enzymes could lead to altered pharmacokinetic parameters (half life, C_{max}, or AUC) of other drug molecules that might be co-administered. To reduce the attrition of otherwise promising lead candidates due to such toxicity liabilities, in vitro screening tests of drug candidates for inhibition of drug-metabolizing CYPs are now carried out early in drug discovery. The screening in principle can be moved to an even earlier stage, especially for large combinatorial libraries, if suitable in silico methods are available. We report the QSAR modeling of CYP inhibition using a large dataset (BioPrint®). Over 2000 marketed drugs, reference compounds and drug-like molecules have been systematically tested across a panel of in vitro assays that include the inhibition of 8 CYPs (2D6, 3A4, 2C9, 2C19, 2E1, 2B6, 1A2, and 3A5). CYPs 2D6 and 3A4 combine to metabolize the majority of all drugs and are thus targeted for in silico modeling. For CYP 2D6, a robust and validated model developed from proprietary descriptors and methods will be presented. For CYP 3A4, the inhibition data consists of assay results for four different substrates: benzyloxy coumarin, testosterone, benzoxyresorufin and midazolam. This dataset provides a complete view of drug interaction with CYP 3A4; in addition, it allows the development and the testing of a multiple-binding pharmacophore hypothesis (MPH) that was formulated as an extension of the traditional QSAR approach for dealing with the binding of a large variety of drugs to CYP 3A4. In its simplest form, the MPH approach takes advantage of the multiple substrate datasets and identifies the binding of test articles as either proximal or distal relative to that of the substrate. In addition to an improved modeling for the inhibition of CYP 3A4, the results from this modelling approach provide insights into the drug-enzyme interactions for this enzyme.

C-2 : SPORCalc –Fingerprint Based Probabilistic Scoring of Metabolic Sites

Catrin Hasselgren Arnby; AstraZeneca, Mölndal, SE
Lars Carlsson, James Smith, Robert C. Glenn, and Scott Boyer, AstraZeneca

Metabolic clearance is a very frequent problem for drug discovery projects, particularly in the early stages of lead identification and lead optimisation. One of the common strategies to gain information on the structure metabolism relationships in these phases is to run metabolite identification studies on the substances of interest. This is time consuming as it requires both pure compound and definitive metabolite identification methods. Predictive metabolism methods can be used in this context to enhance the understanding of structure metabolism relationships, however many methods are based on rules, which give many metabolites in no particular order of importance, or are based on approximations of both the structures and reactivities of the cytochromes P450. The method outlined in this study uses data mining methods to exploit actual biotransformation data that has been recorded over the past decade in the MDL Metabolite database. The database contains nearly 70K individual reactions. Atom-based fingerprints were produced of all 'substrates' in the database as were fingerprints of the reacting centres. Reacting centre fingerprints were derived from a comparison of 'substrates' and their corresponding 'products' listed in the database.

Mining of the data in the database then takes the form of submitting a new molecule and searching for matches to every atom in the new 'candidate' molecule from both the 'substrate' and the reacting centre databases. The results are presented to the user as an 'Occurrence Ratio' of matching fingerprints in both datasets. Normalisation of the occurrence ratio within a single candidate molecule enables the results of the search to be projected onto the candidate molecule in the form of rank-order-dependent colour coding for easy interpretation. Using the rank ordering as a possible measure of the likelihood of a reactions occurring to 'predict' metabolic sites in a validation set of compounds shows performance that would allow SPORCalc to be used by drug discovery teams to generate useful hypotheses regarding structure metabolism relationships.

C-3 : Relationships Between Molecular Complexity, Biologic Activity and Structural Diversity

Ansgar Schuffenhauer; Novartis Institutes of Biomedical Research, Basel, CH
Paul Selzer, Novartis Institutes of Biomedical Research

Following the theoretical model by Hann et al. [1] moderately complex structures are preferable lead compounds since they lead to specific binding events involving the complete ligand molecule. To make this concept usable in practice for library design we studied several complexity measures on the biological activity of ligand molecules. We used the historical IC50/EC50 summary data of 132 assays run at Novartis covering a diverse range of targets, among them kinases, proteases, GPCRs and protein-protein interactions and compared this to the background of "inactive" compounds which have been screened for two years, but never shown any activity in any primary screen. As complexity measures we used the number of structural features present in various molecular fingerprints and descriptors.

We found generally that with increasing number of different structural features present in molecule the activity increased and could establish a minimum number structural features in each descriptor needed for biologic activity. Especially well suitable in this context were the Similog keys[2] and circular substructure fingerprints. These are those descriptors, which also perform especially well in the identification of bioactive compounds by similarity search[3], which suggest that structural features encoded in these descriptors have a high relevance for bioactivity. Since the number of features correlates with the number of atoms present in the molecule also the number of atoms serves as a reasonable complexity measure and larger molecules have in general higher activities. However, the requirements for physicochemical properties, pharmacokinetics and synthetic feasibility constrain the maximum size of feasible lead candidates. It has therefore been suggested to prioritize lead candidates by their ligand efficiency, the biologic activity normalized by the number of atoms.[4] One can normalize the feature count metrics studied above in the same way to obtain feature density measures. We found that in our dataset the ligand efficiency increases with increasing feature density in the Similog and circular substructure keys. Molecules with especially low feature density have not only a low ligand efficiency, but are often also perceived as chemically unattractive by chemists due to their high linearity or symmetry.

Due to the relationship between feature counts and densities on hand and biologic activity on the other hand, the size bias present in almost all similarity coefficients [5] becomes especially important. Diversity selections using these coefficients can influence the overall complexity of the resulting set of molecules, and thus the potential of having a high absolute biological activity or ligand efficiency. Using sphere-exclusion based diversity selection methods together with the Tanimoto distance the average feature count distribution of the resulting selection is shifted towards lower complexity than that of the original set, especially when applying tight diversity constraints. At the same time we find in the selected set in comparison to the complete set a higher percentage of compounds with extreme feature densities on the upper and also the lower, unfavourable end of the scale. This suggests that the feature count and feature density constraints should be applied before diversity selection in order to penalize or exclude low complexity compounds.

1. Hann, Michael M.; Leach, Andrew R.; Harper, Gavin. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *Journal of Chemical Information and Computer Sciences* (2001), 41(3), 856-864.
2. Schuffenhauer, Ansgar; Floersheim, Philipp; Acklin, Pierre; Jacoby, Edgar. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *Journal of Chemical Information and Computer Sciences* (2003), 43(2), 391-405.
3. Hert, Jerome; Willett, Peter; Wilton, David J.; Acklin, Pierre; Azzaoui, Kamal; Jacoby, Edgar; Schuffenhauer, Ansgar. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Organic & Biomolecular Chemistry* (2004), 2(22), 3256-3266.
4. Hopkins Andrew L; Groom Colin R; Alex Alexander Ligand efficiency: a useful metric for lead selection. *Drug discovery today* (2004), 9(10), 430-1.
5. Holliday, John D.; Salim, Naomie; Whittle, Martin; Willett, Peter. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *Journal of Chemical Information and Computer Sciences* (2003), 43(3), 819-828.

C-4 : In silico Prediction of Buffer Solubility Based on Quantum-mechanical, HQSAR- and Topology-based Descriptors

Andreas Göller; Bayer Healthcare, Wuppertal, DE
Dr. Jörg Keldenich and Dr. Matthias Hennemann, Bayer Healthcare
Prof. Tim Clark, University of Erlangen-Nuremberg

We present an artificial neural network (ANN) model for the prediction of solubility in buffer at pH 6.5, thus mimicking the medium in human gastrol-intestinal tract. The model was derived from about 6000 in-house experimental solubilities measured consistently in the described buffer medium. Selection of the subset of significant descriptors as are VAMP AM1 quantum-chemical wave function derived, HQSAR-derived logP, and topology-based descriptors was done by means of FIRM decision tree and other mining procedures. Ten ANN were trained with always 90% of the data set as training set and 10 % of as test set. Solubility prediction is then done with all 10 ANN, resulting in a mean prediction value and the standard deviation of the ANN as quality parameter. The productive ANN gives a rcv of 0.71 and a standard deviation of 0.71 - based on the prediction of any training set compound by the net it were was not part of the training set - with 93% of the compounds having a n error of less than 1.5 log units. The performance of the ANN on validation set compounds is even better due to the prediction by 10 ANN, together with the information of how reliable the value is. By the model, we were able to assess many outliers as compounds with wrong experimental data, and the model learned implicitly about the probable charge states of compounds in buffer medium. With an approximate calculation time of 10s per molecule, the model is now applied on all compounds of the Bayer library, and can be applied on-the-fly to external compounds.

PR-13 : ClassPharmer™ Suite Automates Extraction of SAR to Maximize Antiviral Activity and Minimize Cytotoxicity

Vincent Vivien; Bioreason

Chemists involved in lead optimization use information mined from R-tables to aid in the design of new compounds. This is a labor-intensive and time-consuming process when identifying features and positions of groups that affect more than one property. Using accurate chemical classes as the starting point for extraction of statistically based SAR hypotheses can provide a computational solution for a difficult problem.

ClassPharmer Suite meets the everyday development needs of the chemist by creating an easily accessible environment that is chemistry smart and not a black box. In this presentation, we show how an individual would use the information displayed in the R-tables and provided in the hypotheses to identify properties of R-groups that can be used to optimize compounds for two competing properties: high antiviral activity and low cytotoxicity.

We used the public DTP-AIDS Antiviral Screening data with values for in vitro protection of HIV infected cells (EC50) and in vitro cytotoxicity of uninfected cells (IC50). Compound structures for molecules with a pyrimidine variation were selected and clustered using ClassPharmer™ Suite. The classification was then analyzed with two ClassPharmer™ Suite analysis modules: generation of R-tables and extraction of structure property relationships for both antiviral activity and cytotoxicity from R-group properties.

Analyses were focused on two classes: one without a ring fused to the furan (with representative structures shown below) and one with a fused ring (not shown). All the compounds in the latter were “inactive.” Several of the SAR hypotheses for the former describing the “best” groups at R6 are summarized below. In addition, the activity distribution for one of these hypotheses is shown along with example supporting compounds.

PR-14 : PASS

Alexander Kos, AKos GmbH

PASS: Prediction of Activity Spectra for Substances

The majority of known biologically active substances possess many kinds of biological activity, comprising of pharmacological effects, biochemical mechanisms of action, carcinogenicity, mutagenicity, etc. We often call this

the biological profile. It is very difficult to screen every compound in all available biological assay; and as a consequence about 30% of projects fail because serious adverse or toxic effects are discovered too late.

PASS predicts the biological activity spectra on the basis of the 2D structural formula. This provides the opportunity to select compounds with desirable effects, and without unwanted side effects in the early stage of drug discovery. PASS version 1.932.1 (January 2005) predicts 1000 kinds of biological activity with an average accuracy of 85% (leave one out cross-validation). Based on the calculated values of probability to be active and inactive (P_a and P_i respectively), one may define a flexible criteria for selecting the most promising leads with desirable level of novelty. Calculation of biological activity spectra for 100,000 compounds on a PC takes about 20 min. PASS can be effectively used to analyze large databases.

PharmaExpert helps to analyze the prediction taking into account a huge knowledgebase of activity-activity relationships. It provides the means for interactive selecting the most advantageous compounds. Structures are visualized using the Chime plug-in. Compounds can be profiled based on user-defined criteria. Selected compounds can be exported in SD-file format.

PASS CL is the command line version for inclusion in other in-house applications, or into programs like SciTegic's Pipeline Pilot.

The programs are working on PCs under Windows using MOL and SDF input formats; TXT, SDF and CSV output formats.

Contact: Dr. Alexander Kos, AKos Consulting & Solutions GmbH, Postfach 141, CH-4010 Basel, alexander.kos@akosgmbh.de, www.akosgmbh.com.

PR-15 : Fingerprints, Clustering and High-speed Virtual Library Analysis: BCI's Software Toolkits and Web Services

Geoff Downs, Barnard Chemical Information

This year, BCI is celebrating its 20th birthday. Over the past 20 years, it has established a unique reputation for the strength of its consultancy and specialist software for chemical structure representation, clustering, diversity analysis and Markush structure handling.

During the past 18 months, BCI has increased the utility of its technology by releasing the functionality contained in the original standalone programs as a range of toolkits and SOAP services. The toolkits are available for an extensive range of operating systems and can be supplied with a number of language interfaces including C/C++ and Java. The SOAP services can be called from any client supporting the SOAP protocol, including Scitegic's Pipeline Pilot.

BCI has also increased the functionality of its products, with significant advances having been made. This is particularly evident in the Markush handling of combinatorial libraries, with very fast full, overlap and substructure searching now being provided.

PR-16 : LITHIUM Dock- A Virtual Assay System for Docking,

Ulrike Uhrig, Tripos, Inc.

LITHIUM Dock allows laboratory scientists ready access to the results of a molecular modeler's docking work, improving visibility and utilization of modeling results. Through centralized capture and communication of modeler produced docking information, LITHIUM Dock improves decision support in the laboratory by allowing chemists to rapidly run candidate molecular structures against stored docking procedures.

LITHIUM Dock user benefits:

- Improved utilization of modeling resources leading to enhanced ROI on modeling infrastructure.
- Improved visibility and use of modeling results.

- Improved management of modeling knowledge.
- Improved decision support for laboratory chemists leading to enhanced efficiency.
- Simple and rapid access to expert molecular modeling for laboratory chemists.

LITHIUM Dock capabilities:

- Molecular modelers can upload and store docking procedures into the LITHIUM Dock system.
- Laboratory chemists can search, access & view stored docking procedures through LITHIUM, Tripos' desktop application for 3D chemical visualization, development, and delivery.
- Laboratory chemists run candidate molecular structures against stored docking procedures.
- Laboratory chemists can view the results of docking their own molecular structures.
- Laboratory chemists can make modifications to docked molecules in the context of the target protein and re-dock the modified structure.
- Laboratory chemists can rapidly perform iterative molecular design cycles using stored docking procedures.

C-5 : MC4PC: A Computational Tool for the Rational Evaluation of the Hazard Potential of New Pharmaceuticals and other Organic Chemicals

Gilles Klopman; Case Western Reserve University, Beachwood, OH, USA

The design of new biologically active chemicals, whether pharmaceuticals or agricultural requires a good understanding of their potential beneficial activity, their ability to reach and act upon their intended target, and lack of toxic and adverse effects.

Enormous resources go into the development of new chemicals and it is extremely desirable to find ways to assess potentially damaging properties at the earliest possible stage so as to minimize the number of expensive failures. Structure-Activity studies are important in many areas of mechanistic and exploratory chemistry. They are based on the premise that a relationship may exist between the chemical, or biological properties of a series of molecules and their chemical structure as well as some of their independently observed physical or chemical properties. The major advantage of successful Structure-Activity studies is that it permits predictions to be made for chemicals that have not yet been synthesized and thus improves productivity.

In this lecture, applications of the MCASE methodology to the evaluation of toxic and adverse effects will be described as well as its use to estimate ADME properties of new molecules. These techniques are now in use at various sites of the US Food and Drug Administration, the Japanese NIH, Health Canada and the Danish EPA, as well as at numerous Pharmaceutical and Chemical Companies worldwide for the purpose of assessing the safety of new pharmaceuticals and other chemicals.

C-6 : Characterising Bitterness: Identification of the Key Structural Features

Sarah Rodgers; Unilever Research, Vlaardingen, NH

A bitter taste is a common and undesirable attribute of many functional food ingredients. A large and diverse range of molecules is able to produce a bitter taste via approximately 26 human G protein-coupled receptors on the tongue. However, only limited information is available detailing which bitter molecules are associated with which receptors. In addition, there is no known distinction in the signal produced from the different receptors. To enable the addition of these compounds to foodstuffs their bitter taste must be managed, this can be achieved through a greater understanding of bitterness.

This paper describes the methods used to characterise bitter molecules and identify which structural fragments are involved in conferring the bitter taste. This has been achieved through an examination of the structural characteristics of a database of bitter molecules collected from the literature.

A phylogenetic-like tree (PGLT) [1] was used to summarise the fragments present in the bitter molecules. This involved generating maximal common substructures (MCSs) from clusters or sub-groups in the database, these MCSs were employed as substructural filters controlling membership of nodes in the tree. A large database of

random molecules was filtered through the tree to allow identification of those structural features that are more common to bitter molecules. A selection of the results obtained from the tree, including R-Tables, SAR rules, possible receptor groupings and key structural features, will be presented.

MOLPRINT [2], a similarity searching and classification tool, was also employed to examine the database of bitter molecules. 2D atom environment descriptors are combined with an information gain based feature selection and Naïve Bayesian classifier. The information gain component identifies those structural fragments best able to distinguish between active and inactive (bitter and random) molecules. An overview of the fragments will be presented, in addition to a comparison with those identified from the PGLT.

1. Nicolaou, C.A., Tamura, S.Y., Kelley, B.P., Bassett, S.I. and Nutt, R.F. (2002) Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees. *Journal of Chemical Information and Computer Science*, 42(5), 1069-1079.
2. Bender, A., Mussa, H.Y., Glen, R.C. and Reiling, S. (2004) Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *Journal of Chemical Information and Computer Sciences*, 44, 170-178.

C-7 : The Molecule Evuator: an Interactive Evolutionary Algorithm for Designing Drug-like Molecules

Ad IJzerman; Universiteit Leiden, Leiden Center for Drug Research, Leiden Institute of Advanced Computer Science, Leiden, NL
EW Lameijer, Th Baeck, and J Kok, Universiteit Leiden

The design of a new drug is essentially an optimization problem, in which a suitable molecule must be found in the huge chemical space. Computer science has developed many optimization methods, amongst others the evolutionary algorithms, which use Darwin-like evolution to generate new individuals by mutating and combining existing individuals and selecting the best offspring ("survival of the fittest"). While evolutionary algorithms have been applied various times in de novo design, two aspects remain problematical: having a good molecule representation to search the chemical space and, most importantly, having a good fitness function. We have developed an evolutionary algorithm-based program called "the Molecule Evuator". It can make all possible one-atom/one-bond mutations, and should therefore be able to search the entire chemical space. This will also enable the program to escape more easily from local minima than fragment-based methods. As the fitness function, we decided to use interactive evolution, i.e. letting the user judge the fitness of the molecules. This would employ the expertise of the user in (gu)estimating ease of synthesis and biological activity of the molecules. The combination of this human expertise with the capacity of the computer for fast and unprejudiced searches in chemical space could be useful in designing novel drug-like molecules. Since hundreds of mutations are possible on a medium-sized molecule, we have included several options to restrict the search to the most interesting parts of chemical space. Next to letting the user decide whether parts of the molecule can be fixed, the program can filter the compounds on physicochemical parameters, such as Lipinski's rule of five and the Polar Surface Area. Additionally, molecules with chemically undesirable substructures can be removed from the population automatically. In conclusion, we think that by capitalizing on the implicit knowledge and expertise of the medicinal chemist, the "Molecule Evuator" could be a useful new tool in computer-aided drug design.

D-1 : Hit Selection from HTS Assays: Enhancing Hit Quality and Diversity

Iain McFadyen; Wyeth Research, Cambridge, USA
Rebecca Cowling and Diane Joseph-McCarthy, Wyeth Research

The selection of hits from an HTS is a critical step in the lead discovery process. A typical 'Top X' approach simply selects enough hits to meet the throughput limit of the secondary assay from a list ranked by inhibition. Hits selected in this way are typically enriched in frequent hitters and compounds with undesirable properties, and at the same time de-enriched in diversity due to the presence of redundant members of privileged structural classes. We have developed an alternative scheme for hit selection from Confirmation HTS assays. Crucially, an activity threshold is determined by rigorous statistical analysis of the assay data, and all hits so defined are given equal initial

consideration, regardless of quantity. Structural and assay data are used to filter the hits to eliminate duplicates, highly redundant clusters, compounds with poor physical properties, uninteresting scaffolds and/or functional groups, and compounds with poor assay data. This reduces the number of hits to meet the assay throughput limit whilst simultaneously increasing their quality (as judged by physiochemical properties consistent with optimization from low affinity HTS hits to drug-like development candidates) and diversity (number of unique ring scaffolds). Each step of the workflow is customized to the unique needs of the individual project. We present results from 3 diverse projects that show significant improvements in both the diversity and quality of the selected hits. The ultimate measure of any HTS hit selection process is the successful identification of potential leads, and in each of these projects 4-8 advanced hit series have advanced.

D-2 : Use of Multiple-category Bayesian Modeling to Predict Side Effects

Robert Brown; SciTegic, Inc., San Diego, CA, USA
David Rogers, SciTegic, Inc.

The rapid growth of in the amount of biological information available from high-throughput screening studies and drug compendia such as MDL's Drug Data Reports (MDDR) and Derwent's World Drug Index would appear to provide a strong experimental base for the flagging of potential drug side effects. However, current methods, such as rational drug design, virtual docking, and QSAR modeling, do not appear to be easily applied to this task, due to slow computational speed, difficulty in automating, or difficulty in handling large, diverse data sets.

Laplacian-modified Bayesian modeling was developed to rapidly analyze high-throughput screening data using 2D molecule fingerprints. It can be extended so that a single model can predict the absolute probability of thousands of different activity classes. From this prediction, one can not only see what class a particular molecule is most likely active in, but competitive activity classes that may appear later as side-effects. The process is rapid enough to learn all of a drug compendium such as MDDR in minutes, and can then be used to suggest primary effects and side effects of hundreds of molecules per minute.

D-3 : Scaffold-Hopping Using Clique Detection Applied to Reduced Graphs

Eleanor Gardiner; University of Sheffield, Department of Information Studies, Sheffield, UK
Edward Barker, David Cosgrove, Valerie Gillet, and Paula Kitts, University of Sheffield

Similarity-based methods for virtual screening are widely used. However, conventional searching using 2D chemical fingerprints or 2D graphs may retrieve only compounds which are structurally very similar to the original target molecule. Of particular current interest then is scaffold-hopping, i.e., the ability to identify molecules that belong to different chemical series but which could form the same interactions with a receptor. Reduced graphs provide summary representations of chemical structures and therefore offer the potential to retrieve compounds that are similar in terms of their gross features rather than at the atom-bond level. Using only a fingerprint representation of such graphs, we have previously shown that actives retrieved were more diverse than those found using Daylight fingerprints [1,2].

Maximum common substructures give an intuitively reasonable view of the similarity between two molecules. However, their calculation using graph-matching techniques is too time-consuming for use in practical similarity searching in larger datasets. In this work we exploit the sparsity of the reduced graph, in graph-based similarity searching. We reinterpret the reduced graph as a fully connected graph using the bond-distance information of the original graph. We describe searches using both the RASCAL and Bron-Kerbosch clique-detection algorithms on the fully connected reduced graphs and compare the results with those obtained using both conventional chemical and reduced graph fingerprints. We show that graph-matching using fully connected reduced graphs is an effective retrieval method, and that the actives retrieved are likely to be topologically different to those retrieved using conventional 2D methods.

1. Gillet, V.J., Willett, P., Bradshaw, J. Similarity Searching Using Reduced Graphs. *Journal of Chemical Information and Computer Sciences* 43, 2003, 338-345.

2. Barker, E., Gardiner, E., Gillet, V.J., Kitts, P., Morris, J. Further Development of Reduced Graphs for Identifying Bioactive Compounds, *Journal of Chemical Information and Computer Sciences* 43, 2003, 346-356.

D-4 : A First Look into ABCD

Dmitrii Rassokhin; Johnson & Johnson Pharmaceutical Research & Development, Molecular Design & Informatics, Exton, PA, USA

We unveil ABCD (<http://www.bioitworld.com/archive/061704/discovery.html>), a modern drug discovery informatics platform for Johnson & Johnson Pharmaceutical Research & Development. ABCD is an attempt to bridge multiple continents, data systems and cultures using modern information technology, and provide scientists with tools that allow them to make better decisions. The system consists of three major components: 1) a data warehouse, which combines data from multiple chemical and pharmacological transactional databases, organized using dimensional modelling principles to support superior query performance; 2) a state-of-the-art application suite, which facilitates data upload, retrieval, mining and reporting, and 3) a workspace, which facilitates collaboration and data sharing by allowing users to share queries, templates, results and reports across project teams, campuses, and other organizational units. Chemical intelligence, performance and analytical sophistication lie at the heart of the new system, which was developed entirely in-house.

PR-17 : SciFinder 2006: A Preview of Upcoming New SciFinder Features

Paul Peters, CAS

SciFinder was launched 10 years ago as a search tool which promised to change the way scientists conduct research. Over this period of time, the SciFinder family of products has been continually enhanced with new functionality and content. SciFinder is now an enterprise-wide research service for discovery scientists at leading research-driven companies, academic institutions and government agencies throughout the world. SciFinder has grown to become an essential part of the research process. This product review will offer a preview of the upcoming release of SciFinder scheduled for later this summer.

PR-18 : Accord Cheminformatics Suite - Enhancements in v 6.0

Tim Aitken, Accelrys Ltd.

The Accord suite of Cheminformatics software ranges from individual function-specific software components, to programming toolkits, desktop applications and enterprise-wide solutions, providing both off-the-shelf applications and tools to allow custom development.

2005 has seen the launch of version 6.0 of the Accord suite, including some of the following enhancements in chemical and biological data management.

This review will focus on recent advances in the Accord Chemistry Engine underlying all Accord products.

Accord Developer Tools

Accord Chemistry SDK version 6.0 now supports creation, storage and representation of Markush schema, allowing patent information to be mined and multiple chemical queries to be submitted simultaneously.

Additionally, enhancements have been made to the Similarity & fingerprinting functionality, allowing fingerzoning & fingerlocations to be defined – this gives users a wide range of similarity search types, including molecule only, reactant vs product, reaction centre only etc. Bit density enhancements mean the fingerprints can be weighted according to the target data set allowing users to find previously unexpected chemically significant results from similarity searches on reaction databases.

The Property calculations available within Accord products have been extended, to include Hepatotoxicity & CYP450-2D6 alongside BB permeability, Aqueous Solubility, #H bond donors etc

Version 6.0 of the Accord Chemistry Cartridge allows partitioned indexes & support for parallel processors for enhanced index creation & search speeds, along with support for the Cost Based Optimiser to maximise performance of complex queries.

Accord Enterprise Informatics

AEI 6.0, the latest release of Accelrys' flagship cheminformatics enterprise solution contains a number of enhancements and improvements that assist users in managing their chemical and biological data. 6.0 now provides support for the registration and querying of chemical reactions through both DS Accord for Excel Enterprise and DS Accord Enterprise Workbench, a new querying and administration client. 6.0 also provides Markush querying, the ability to store compound characterisation data (such as MS or NMR data), support for V3000 SD files and faster searching, querying and registration over previous versions.

E-1 : Integration of Chemical and Biological Data: The NCBI PubChem Project

Wolf Ihlenfeldt; National Institutes of Health, NCBI, NLM, Bethesda, MD, USA

Evan Bolton, Jie Chen, Lewis Geer, Jane He, Siqian He, Abby Ngau, Vahan Simonyan, Paul Thiessen, Valery Tkachenko, Yanli Wang, Jian Zhang, and Steven Bryant, National Institutes of Health

The NCBI PubChem project (online at <http://pubchem.ncbi.nlm.nih.gov>) is new public information system recently launched as part of the NIH Roadmap Initiative. Within the NIH Roadmap, a number of screening centers will be funded with the mission to development novel biological screens, especially for targets of currently minor commercial interest but linked to diseases, and to run massive high-throughput experiments on these screens.

The central purpose of PubChem is to act as a repository for chemical structure data and summary assay results from these experiments. PubChem links chemical structures to the results of bioassays, and augments this information with references to literature and auxiliary data describing these assays and other results. This is a unique scope not covered by any other publicly accessible database. PubChem is tightly integrated into the cluster of biological and literature databases hosted at NCBI, such as PubMed and MMDB. Building onto the NCBI Entrez query system, which was enhanced with custom components developed for this project, interactive federated queries spanning multiple databases can be executed to discover deep relationships between compounds, assays, biological targets and literature. The PubChem interfaces provide extensive query capabilities on textual and numeric information, as well as a comprehensive set of structure-based query methodologies and assay data filter and analysis tools.

While PubChem was launched with existing legacy datasets such as the DTP, MMDB and KEGG structure sets, PubChem is designed as an open system that will in future include additional information sources. As envisioned by the interdisciplinary RoadMap Initiative, PubChem's goal is to link chemical and biological information. Using chemical structure as a key, PubChem will present to the researcher an expanding variety of information concerning the biological properties of small molecules. We hope to be able to develop PubChem into a major clearinghouse for chemical and biological structure information, acting as a hub to a multitude of more specialized information sources.

E-2 : StARLite – a Chemogenomics KnowledgeBase

Edith Chan; Inpharmatica, London, UK

Bissan Al-Lazikani, Richard Cox, Dave Michalovich, and John Overington, Inpharmatica

StARLite is a comprehensive chemogenomics knowledgebase bridging chemical and biological space. The database comprises around 300,000 bioactive tractable compounds and their related pharmacology and target information abstracted from two journals, the Journal of Medicinal Chemistry (from 1980 to present) and Bioorganic and Medicinal Chemistry Letters (from 1991 to present). The chemical 2D structures as well as their pre-calculated 2D and 3D properties are stored in a relational chemical database, using Oracle 9i and MDL technology. 3D chemical structures are stored in a Tripos Unity database enabling advanced 3D searching (e.g pharmacophore search). There are over 1.3 million activity data points, which cover functional, binding and ADMET assays.

Structures and assays are linked to their synthetic routes and assay protocols, via published references. An expertly curated target dictionary provides non-redundant molecular and functional target sequences, names, synonyms and taxonomic sources. There are over 4000 unique molecular targets searchable by sequence and various accession codes such as SWISS-PROT, TREMBL and GenBank. The database in essence provides an extensive review of drug discovery research over the last 25 years. It can be used for navigating through compound, assay, activity and target relationships, obtaining target family chemotype portfolios, and elucidating SAR, selectivity and potency profiles.

E-3 : A Searchable Database for Comparing Protein-ligand Binding Sites for the Discovery of Structure-function Relationships

Richard Jackson; University of Leeds, School of Biochemistry, Leeds, UK
Nicola D. Gold, University of Leeds

The rapid expansion of structural information for protein-ligand binding sites is potentially an important source of information in structure-based drug design and in understanding ligand cross reactivity and toxicity. We have developed a large database of ligand binding sites extracted automatically from the Macromolecular Structure Database. This has been combined with a new fast method for calculating binding site similarity based on geometric hashing (Brakoulias and Jackson, 2004). A relational database for the retrieval of site similarity and binding site superposition of the large class of nucleotide ligands (containing 5548 sites) has been created. This contains an all-against-all comparison of binding sites which is accessible to data mining. It will be expanded to include other small-molecule ligands in the future.

Similarity in the spatial arrangements of atoms between any two sites might indicate that they bind similar ligands (or parts of ligands) and thus may exhibit similar functions. The geometric hashing method provides a score of similarity, an RMSD and a superposition of equivalenced atoms for each pair of compared binding sites. The similarity score allows identification of the most similar ligand binding sites to a defined query and also allows similarity of the ligand conformation to be analyzed. The preprocessed data are stored in a World Wide Web accessible database, which is searchable with a PDB code and ligand information (such as ligand name, number and chain). The search rapidly returns a ranked list of similar ligand binding sites (above a certain similarity score cut-off). Additional annotation is also provided such as the SCOP (Structural Classification Of Proteins) codes and each hit is coloured according to its similarity to the overall family and fold of the query protein. Optimal superpositions of the binding site and ligands of interest can then be performed allowing further examination and visualisation. A multiple alignment of structurally equivalenced atoms is also provided to define binding site sequence similarity with a statistical E-value using the method of Stark and coworkers (Stark et al., 2003) in order to give a measure of confidence in the identified functional similarity.

We will give examples of its use. The method is successful in identifying close family and superfamily relationships prior to their structural classification in the SCOP or CATH databases. It has also been applied successfully to identifying more distant evolutionary relationships at the fold level and even beyond where binding site similarity and ligand conformation are conserved in the absence of sequence conservation. We are also able to detect similarity between binding sites with conserved ligand binding conformations where the ligands are chemically similar but not identical.

References

1. Brakoulias A., Jackson RM. (2004) "Towards a Structural Classification of Phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching" *Proteins; structure function and bioinformatics*, 56, 250-260.
2. Stark A, Sunyaev S and Russell RB. (2003) "A model for statistical significance of local similarities in structure" *JMB*, 326 1307-1316.

BLUE POSTER SESSION ABSTRACTS

Blue Poster Session Abstracts

P-1 : A Retrospective Docking Study of PDE4B Ligands and an Introduction into Methods of Avoiding Some Failures of Current Scoring Functions

Chidochangu Mpamhanga; University of Sheffield, Sheffield, GB
Beining Chen, Iain McLay, Daniel Ormsby, and Mika K. Lindvall, University of Sheffield

In general fast scoring functions fall short in their ability to determine the relative affinities of ligands for their receptors nevertheless this study demonstrates a successful docking and scoring methodology for PDE4B. A series of known inhibitors of PDE4B were docked into the PDE4B/ Pyrazolo[3,4-b]pyridine binding site using LigandFit a fast shape matching docking algorithm. This work was done as a preliminary study, to verify the suitability of the LigandFit/DockScore protocol for virtual screening for another project which required a method that could enrich the top 5% of a database by a factor of at least four. An RMSD comparison of the LigandFit/DockScore (in virtual screening mode) generated poses with the crystallographic poses over 19 inhibitors, whose x-ray structures were available, revealed a reasonable success rate. However, the main objective was to investigate the effectiveness of five available scoring functions (PMF, JAIN, PLP2, LigScore2 and DockScore) to enrich the top ranked fractions of nine artificial databases constructed by seeding 20 randomly selected inhibitors ($pIC_{50} > 6.5$) into 1980 inactive ligands ($pIC_{50} < 5$). PMF and JAIN showed high average enrichment factors (greater than 4) in the top 5-10% of the ranked databases. Rank-based consensus scoring was also investigated and the rational combination of 3 scoring functions resulted in more robust and generalisable scoring schemes, consensus Score DPmJ (DockScore, PMF and JAIN) and PPmJ (PLP2, PMF and JAIN) yielded particularly good results. Finally a brief analysis of the behaviour of the scoring functions across different chemo-types or chemical classes followed. This revealed the inherent bias of the docking and scoring method towards the initial crystal structure binding mode (PDE4B/Pyrazolo[3,4-b]pyridine). And this suggests a need to develop better means of avoiding the problems of using the rigid receptor in docking studies. Future work will be focused on the docking of ligands into multiple binding sites of the PDE4B. Another lesson learnt from this investigation is that scoring functions can be used to a limited extent for lead optimization.

References:

1. C. M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman. *Journal of Molecular Graphics and Modeling*, 21, 2003, 289-30.
2. D. G. Allen, D.M. Coe, C.M. Cook, M. D. Dowle, C. D. Edlin, J. N.Hamblin, M.R Johnson, P.S. Jones, R. G. Knowles, M. K. Lindvall, C. J. Mitchell, A.J. Redgrave, N. Trivedi, P. Ward, Pyrazolo[3,4-b]pyridine compounds, and their use as phosphodiesterase inhibitors. *PCT Int. Appl.* (2004), 293.
3. C. P. Mpamhanga, B. Chen, D. L. Ormsby, M. K. Lindvall, I. McLay, A Retrospective Docking and Scoring Study of PDE4 Ligands and the Consequences of Consensus Scoring. (Under review).

P-3 : Calculating Biases Using Artificial Intelligence in Conjunction with Data Assimilation

Hamse Mussa; Cambridge University, Unilever Centre for Molecular Science Informatics, Cambridge, GB
David J. Lary, NASA, Goddard Space Flight Centre
Robert C. Glen, Cambridge University, Unilever Centre for Molecular Science Informatics

In chemometrics and other parameter estimation areas, models are developed for analysing accurately linear and non-linear multivariate data. In these areas data are generally the sole constraint for fitting models which are supposed to give a mathematical representation of the underlining process(es) of the observations. In other words, the functional forms of the models are approximated solely from the data. Therefore it is crucial that multivariate analysis (MVA) methods take into account any errors in the data.

The MVA techniques deal well with random errors. however, tackling biases is a significant issue for them. Unfortunately, measurements/observations are prone to biases whose detection is often a difficult task in its own right.

In this talk, we present a novel algorithm for predicting biases in observations. The method is based on artificial intelligence algorithms in conjunction with data assimilation [1], a mathematical scheme which blends the mathematical representation of the underlining process of the observation with the observed data, such that the obtained "mixture" is the "best" estimate of the underlining process which is consistent with both the observations and the model predictions (in this case, the MVA model predictions).

The new approach will allow us not only to detect and then remove biases from observations, but it also will make it easy to calibrate instruments.

Measurements of ozone (O_3) concentrations were used to test the performance of the proposed method. Results showing the performance of the new method are presented and discussed.

References:

1. B. V. Khattatov, J. C. Gille, L. Lyjak, G.P.Brasseur, V. L. Dvortsov, A. E. Roche, and J. W. Waters, J. Geophys. Res., 104, 18,715 (1999).

P-5 : Storage and Processing of Chemical Information Directly from any Web Browser

Luc Patiny; Ecole Polytechnique Fédérale de Lausanne, Lausanne, CH
Damiane Banfi and Michal Krompiec, Ecole Polytechnique Fédérale de Lausanne

In industries and even worse at universities it is very difficult to find information, like NMR spectra or HPLC chromatograms that were acquired 5 years ago. During this presentation we will describe the system that we have developed allowing to store and process all physical characteristics, chemical structures, chromatograms and spectra in a database. At the technical point of view, this project is based on:

on the server :

- a java servlet
- a SQL database (currently MySQL, <http://www.mysql.org>)
- some shell scripts for the conversion of data coming from all the instruments

on the client :

- any web browser
- the java virtual machine (<http://www.java.com>)

We will first present the database diagram that we have design in order to keep all the information in a way that is intuitive for chemists. We will then present the various ways to import chemical information in this database, either directly from the web browser, using a home-made java application (XMLCreator) or by importing an XML file. This later will be the best choice to import straightforwardly the chemical information coming from the instruments (spectra and chromatograms). We will then describe our choice of the format that is used in order to store the information and to maintain durability in this project. For instance all the chromatograms and spectra are converted in the ubiquitous jcamp format (<http://www.jcamp.org>). Queries can be done for all the information and we can for example search for chemical structures containing a benzene ring, having a boiling point around 150°C and a NMR peak with a chemical shift around 3ppm. Finally we will show "Nemo", our jcamp visualising applet, that allows to process and analyse in a very efficient way all the spectra. One of the main advantages is that there is only one user interface for any kind of experimental data. This whole system can be used internally (on the intranet) in order to store the "private" chemical information but is also used to build a freely accessible database on the internet in which anybody can store any kind of chemical information. As a conclusion, in this project we push the javascript and java features to the limits in order to have outstanding possibilities accessible directly on any computer in the world.

P-7 : Incorporating the Flexibilities of Both the Ligand and the V82F/I84V Drug-Resistant Mutant HIV Protease Target During Docking: Applying the Relaxed Complex Method of Drug Design to HIV-1 Protease

Alex Perryman; University of California at San Diego, La Jolla, CA, US
Jung-Hsin Lin and J. Andrew McCammon, University of California at San Diego

Including the flexibilities of both the drug and the protein target can enhance the drug design process, and the inclusion of that flexibility could be critical when targeting a highly dynamic protein, such as HIV-1 protease. The results of applying our new Relaxed Complex Method of drug design (Lin, J.-H., Perryman, A.L., Schames, J.R., & McCammon, J.A. "Computational drug design accommodating receptor flexibility: the relaxed complex scheme." *J. Am. Chem. Soc.* 124(20): 5632-5633 (2002); and Lin, J.-H., Perryman, A.L., Schames, J.R., & McCammon, J.A. "The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme." *Biopolymers.* 68(1): 47-62 (2003)) to the HIV-1 protease system will be presented. Considering the huge conformational changes that HIV protease experiences, and considering both the large size and the extensive flexibility displayed by the HIV-1 protease inhibitors that are currently used clinically, applying the Relaxed Complex method to this system was a formidable challenge. The algorithmic details involved such things as converting all of the snapshots from all atom, parm99-formatted restart files into united atom, parm94-formatted pdb files and then further converting those files into AutoDock3.0.5's format in a fully automated fashion (while maintaining the same relative position of the grid points). Significant trial-and-error was then involved in optimizing the run parameters for AutoDock3.0.5's Lamarckian Genetic Algorithm (Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. *J. Comp. Chem.* 19: 1639-1662 (1998)), in order to get reproducible results in an efficient manner.

In the Relaxed Complex experiments that were performed, the completely-flexible drug JE-2147 was docked to every tenth picosecond snapshot extracted from both the 22 ns wild type HIV-1 protease MD simulation as well as from the 22 ns V82F/I84V mutant HIV-1 protease MD simulation. Those MD simulations were discussed in: Perryman A.L., Lin, J.-H., & McCammon, J.A. "HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs." *Protein Sci.* 13(4):1108-23 (2004). JE-2147 was the drug crystallized with the wild type HIV-1 protease in the 1KZK.pdb structure, which was the basis for the conventional MD simulation of the wild type HIV protease. The same set of optimized run parameters performed very robustly when docking JE-2147 against all 22 ns of both the wild type and of the V82F/I84V mutant of HIV-1 protease, even though the mutant was crystallized with Tipranavir (1D4S.pdb). When each Relaxed Complex experiment was repeated with the same set of optimized run parameters, the estimated free energy of binding that was obtained from docking against each particular snapshot had good agreement between the three independent trials that were performed on each of the two systems (see Figures 1 and 2 below).

That set of optimized run parameters is currently being utilized in Relaxed Complex experiments that involve the design and evaluation of new active site inhibitors that should hopefully be more effective against the V82F/I84V drug-resistant mutant ensemble of conformations. Structural intuition and guidance from the literature and from the control experiments were used to design a series of over 20 new, slightly different compounds, which exploit the advice that the size, the flexibility, and the asymmetry of the P2/P2' side chains should be increased when trying to design an HIV protease inhibitor that will be more effective against the drug-resistant mutants. Fourteen of those compounds are currently being screened in silico, but the results of screening the entire series will be presented.

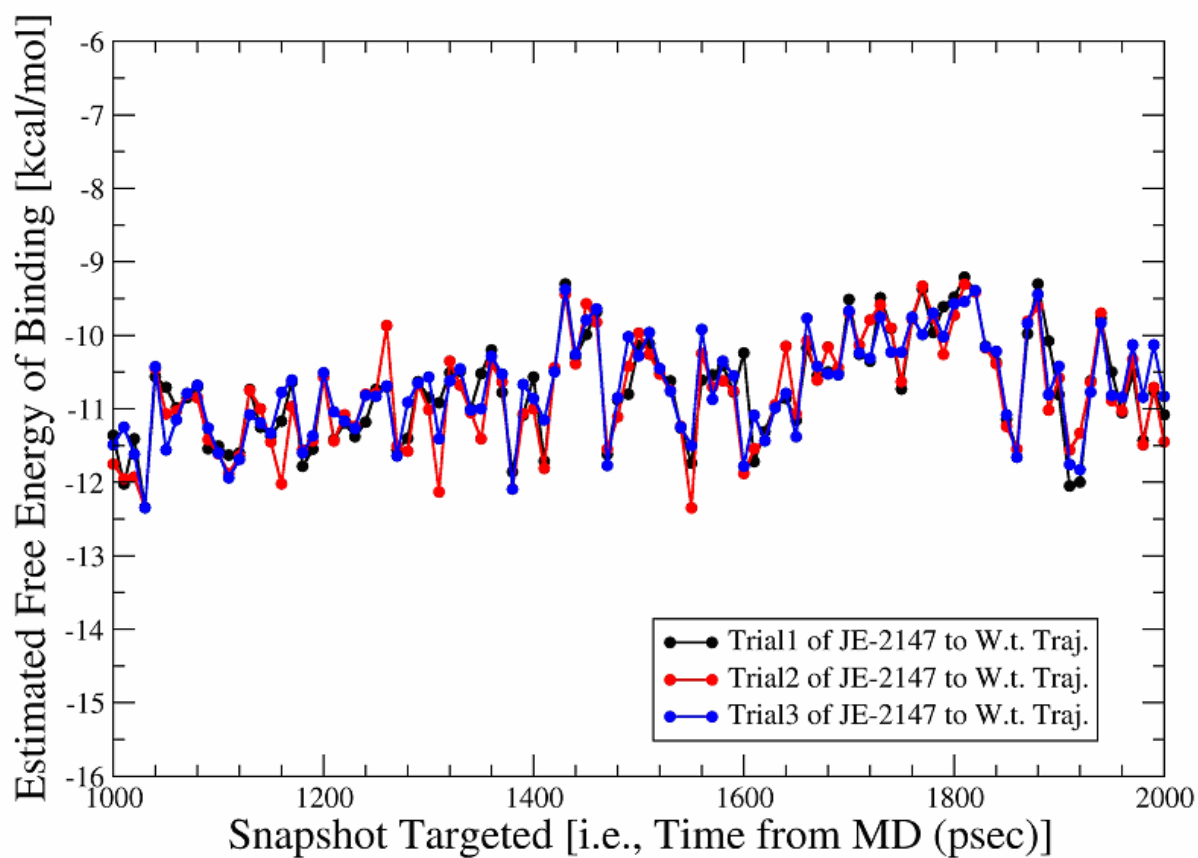


Fig. 1: The drug JE-2147 was docked to 2,200 different snapshots from the wild type HIV-1 protease MD trajectory, and the results of three independent trials were quite robust. The results of targeting 100 of those conformations (every tenth snapshot from the second ns of the 22 ns wild type trajectory) are shown above. Each circle/trial signifies the best of ten runs of docking to one particular snapshot of HIV protease, and each snapshot was targeted in 3 separate experiments.

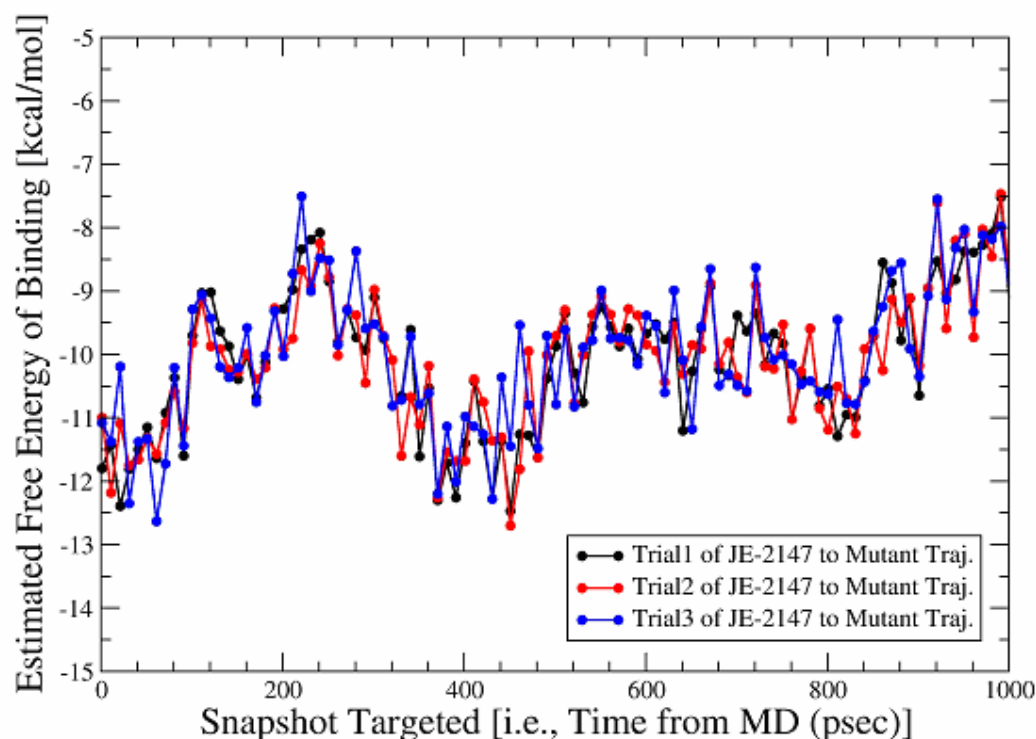


Fig. 2: The drug JE-2147 was docked in three independent trials to 2,200 snapshots of the V82F/I84V drug-resistant mutant HIV-1 protease in a reproducible manner, using the exact same optimized run parameters and procedures that were utilized when docking against the wild type's snapshots (see Fig. 1). The results of targeting 100 of those snapshots (every tenth conformation from the first ns of the 22 ns mutant trajectory) are shown above.

A.L.P. is a Howard Hughes Medical Institute Pre-doctoral Fellow. We are grateful for the generous funding provided by the Howard Hughes Medical Institute and the W.M. Keck Foundation. Additional funding was provided, in part, by grants to J.A.M. from NIH, NSF, NPACI/SDSC, NBCR, and by UCSD's new NSF Center for Theoretical Biological Physics.

P-9 : Making Real Molecules in Virtual Space

Gyorgy Pirok; ChemAxon, Budapest, HU

Nora Mate, Jenő Varga, Miklós Vargyas, Szilárd Dorant, and Ferenc Csizmadia, ChemAxon

Most virtual reaction applications require the manual intervention of experienced chemists in the enumeration phase (selection of appropriate reactants, assignment of the corresponding reaction sites, removing unlikely products). To automate the synthesis process we have moved the expertise intensive stages from the compound library design phase to the reaction library design phase. ChemAxon is building a library of the most important preparative reactions, where each reaction definition contains a generic transformation scheme and additional chemo-, regio- and stereoselectivity rules to handle specific reactants selectively.

The key component of this technology is the Java-based Reactor software able to evaluate and enumerate these "smart" reactions. Its high performance and ability to predict synthetically feasible reaction products opens new possibilities for researchers. We present some virtual synthesis and biotransformation applications, which are able to enumerate entire combichem libraries, build a diverse molecular space of synthetically feasible virtual compounds from available chemicals and predict metabolic pathways.

P-11 : SAPPHIRE: Structure Aided Pharmacophore Implied Reagent Extraction – A method for in silico Screening

Narasinga Rao; Scynexis Inc, Research Triangle Park, NC, US

In the past decade virtual screening methods have become an integral part of lead generation in drug discovery. Primarily, there have been two approaches, one based on ligand activity utilizing the key pharmacophore elements; and the other, based on the receptor structure using protein docking methods. Both approaches have proved to be invaluable tools for lead identification as well as lead optimization. However, each method has some limitations, especially, when it is necessary to incorporate both pharmacophore and receptor information into virtual screening. This work describes a composite approach that takes into account the pharmacophore information, as well as, spatial constraints of the receptor pocket for identifying potential hits. The method uses a cascade of filters that tries to select compounds based on pharmacophore descriptors, a user definable shape constraint and the receptor pocket information that could be tailored to focus hits to specific proteins of interest. The SAPPHIRE method can be particularly useful to prescreen compound libraries prior to subjecting them for high throughput in vitro screens.

P-13 : Strategies for ACE2 Structure-Based Inhibitor Design

Monika Rella; University of Leeds, Leeds, GB
Thierry Langer and Richard Jackson, University of Leeds

Angiotensin-Converting Enzyme (ACE) is an important drug target for hypertension and heart disease. Recently, a unique human ACE homologue termed ACE2 has been identified, which has been linked to hypertension, heart and kidney disease. In addition, ACE2 was shown to function as SARS-Coronavirus receptor. This surprising role and its assumed counter-regulatory function to ACE make ACE2 an interesting new cardio-renal disease target. With the recently resolved ACE2 structure in complex with an inhibitor available, a structure-based drug design project has been undertaken to identify novel potent and selective inhibitors. Computational approaches involve combinatorial library design and docking as well as pharmacophore-based virtual screening of large compound databases.

Initially a small number of fragments was selected and evaluated via docking and later used for combinatorial library design considering synthetic accessibility by mimicking a chemical reaction. Most suitable R-groups were suggested for synthesis. In a complementary approach, a protein-based pharmacophore model was created manually comprising several chemical features such as hydrogen bonding, electrostatic and hydrophobic interactions aligned in 3D resembling specific drug-receptor interactions. Selectivity of the model was ensured by initial screening for ACE inhibitors and enrichment enhanced through repeated optimisation cycles. The final model was used to search 2.5 million compounds for matching features, proving a fast and efficient alternative to docking for initial screening. Hits were further evaluated and prioritised via docking and the most promising candidates proposed for purchase and biological testing.

P-15 : Flexible Smoothed-Bounded Distance Matrix-Based Similarity Searching of the MDDR database

Nicholas Rhodes; University of Sheffield, Sheffield, GB
David Clark, Nicholas Rhodes, and Peter Willett, University of Sheffield

The work extends the autocorrelation vectors method originally described by Moreau and co-workers [1]. The method was applied first to similarity searching of 2-D structures and later to 3-D rigid similarity [2] and molecular surfaces [3]. Though recent progress has been made in the area of 3D flexible similarity [e.g. 4] few efficient methods are available. Molecular flexibility is encoded using smoothed bounded distance matrices. Eight atomic properties are employed and the vectors comprise 16 elements (the range from 0.0 to 20.8 Angstroms is divided into bins of 1.3 Angstroms). Two vectors are compared by computing (the square of) the Euclidean distance between them, the comparisons are very rapid, giving rise to short search times, even for large databases. The poster describes the application of the method to searching of MDDR with targets drawn from a number of activity classes.

1. Moreau, G.; Broto, P. The autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim.* 1980, 4, 359-360.
2. Moreau, G.; Turpin, C. Use of similarity analysis to reduce large molecular libraries to smaller sets of representative compounds. *Analysis* 1996, 24, 17-21.
3. Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* 1995, 117, 7769-7775.
4. Raymond, J. W.; Willett, P. Similarity searching in databases of flexible 3D structures using smoothed bounded distance matrices. *J. Chem. Inf. Comput. Sci.* 2003, 43, 908-916.

P-17 : BRUTUS: A Fully Automated Rigid-Body Superposition Tool

Toni Ronkko; University of Kuopio, Kuopio, FI
Anu Tervo and Antti Poso, University of Kuopio

Often, drug research projects have to be started with little knowledge on possible target. Therefore, finding new lead molecules worthy of further research is one of the first challenges such projects face. The purpose of this study was to develop a new virtual screening method for finding biologically active while structurally dissimilar lead molecules. After all, structural analogs have usually been considered by pharmaceutical industry already, and structurally dissimilar lead molecules are desirable especially in early phases of drug discovery process. The result of our work, BRUTUS, is a fully automated rigid-body superposition method for virtual screening of large molecular databases. In BRUTUS, molecular energy fields are investigated instead of molecular structures to find dissimilar while biologically active compounds. In the course of a single molecular superposition, about 12000 different alignments are evaluated before producing 1-5 most prominent alignments for further research. The amount of trial alignments allows finding less obvious matches that are more difficult to find by methods focusing solely on molecular structures. Despite of the many trial alignments, only 0.2 seconds of computer time is needed per conformation. Hence, BRUTUS is a practical virtual screening method that is useful especially in early phases of drug discovery process.

P-19 : Modelling the Inhibition of P450 Enzymes

Gijs Schaftenaar; Radboud University Nijmegen, Nijmegen, NL
P.W. van Grootel and L. Roumen, Eindhoven University of Technology

P450 enzymes are proteins that have an important role in removing xenobiotics from the human body. Other P450 enzymes are important in the biosynthesis and conversion of steroid hormones. The ability to predict whether a ligand will inhibit the activity of these enzymes will be instrumental in the design of drugs targeted at these enzymes. Quantum mechanical calculations on a model system of the active site of these proteins were performed in order to quantify the interaction between the active site and a model ligand. The mapping of the potential energy surface with respect to some key internal variables of active site - ligand complex, will produce data which can be converted to analytical forms of this interaction. This will be used in a force-field type of description of the interaction between ligand and the full protein.

P-21 : Treasure Island: Molecular 3D Shape-based Clustering with Neural Networks

Paul Selzer; Novartis Institutes for Biomedical Research, Basel, CH
Peter Ertl, Jörg Mühlbacher, and Stephen Jelfs, Novartis Institutes for Biomedical Research

Analyzing the relation between the structure of a molecule and its physicochemical and/or biological properties is a very powerful technique to identify molecular structural features of pharmacological importance. For this purpose a structural descriptor was applied that is calculated from the intramolecular atom distances in 3D space and thus describes the 3D shape of the molecule.

Transforming molecules into such molecular 3D-shape descriptors allows one to use three-dimensional structure information as input for the training of a neural network. Neural networks learn inductively about the relation between input (molecular structure) and output (physicochemical/biological property) by analyzing a set of examples, the so-called training set. After a network has been trained it is able to predict those properties for new molecules.

This versatile methodology was implemented as cheminformatics web tool on the Novartis intranet providing three main functionalities:

- Diversity checking of a molecular data set
 - Mapping sets of molecules into a neural network provides quick feedback about the coverage of chemical space and the diversity or overlap of the different data sets.
- Selection of Bioactive Molecules – “Cherry Picking”
 - Highlighting biological properties of the mapped compounds provides information about the overlap of biological and chemical space. Analyzing the close network neighborhood of active compounds indicates other compound candidates with a high probability of being active too.
- Selecting a diverse and representative subset of a large compound collection

By mapping a large compound collection into the network and taking only one representative compound from each neuron a representative subset of molecules can be created. This could be applied e.g. to reduce the size of a virtual combinatorial library with a minimum loss of chemical space coverage.

P-23 : Classification of Protein Kinases: Clustering Similarity Matrices Generated from Alignment and Novel Sequence-Based Descriptors

Suresh Singh; Vitae Pharmaceuticals, Fort Washington, PA, US
 Ansuman Bagchi, Keith Schmidt, and Robert P. Sheridan, Merck Research Laboratories
 Richard D. Hull, Axontologic Inc.

We present a classification of a set of 296 protein kinases from the Steven Hanks' protein kinase data set [1] based on sequence features. We generated classifications by clustering sequences given a cross-similarity matrix. Similarity matrices may be generated in a number of ways. The conventional method of calculating the similarity between two sequences is by alignment. For this we used BLAST2 [2] to generate the alignments and the BLOSUM62 [3] and GONNET [4] matrices to score them. We introduce a novel method for calculating similarities using sequence-based descriptors that do not require alignment. Given the descriptors, we use two different ways of calculating similarities, Dice [5] and a method based on singular-value decomposition called LaSSI [6] (Latent Semantic Structure Indexing). The clustering analysis shows that these clusters correlate well with the functional class membership defined by Steven Hanks. In addition our classification assigns Hanks defined functional class membership to most of the unassigned other protein kinase (OPK) groups. We will present a comparison of our clustering based classifications with Steven Hanks' classification.

References:

1. Hardi, G. and Hanks, S (1995). The protein kinase facts book, Vols I and II. Academic Press Inc., San Diego, CA 92101. http://pkr.sdsc.edu/html/pk_classification/pk_catalytic/pk_hanks_class.html
2. Altschul, Stephen F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nuc. Acids Res.* 25:3389-3402.
3. Benner SA, Cohen MA, Gonnet GH. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* 1994 Nov;7(11):1323-32.
4. Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89:10915-10519.
5. Hull, R.D., Fluder, E.M., Singh, S.B., Nachbar, R.B., Kearsley, S.K., & Sheridan, R.P. (2001b). Chemical Searches using Latent Semantic Structure Indexing (LaSSI). *J. Med. Chem.*, 44, 1185-1191.
6. Hull, R.D., Singh, S.B., Nachbar, R.B., Sheridan, R.P., Kearsley, S.K., & Fluder, E.M. (2001a). Latent Semantic Structure Indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.*, 44, 1177-1184.

P-25 : Distributed Search System CACTVS/SONORA: Search and Retrieval of Chemical Compounds and Associated Data from Very Large Databases

Markus Sitzmann; National Institutes of Health, Frederick, MD, US
Marc Nicklaus and Igor Filippov, National Institutes of Health
Wolf-Dietrich Ihlenfeldt, Xemistry GmbH

We present the distributed search system SONORA (Searches Optimized for Node-Operation for Rapid Answers) implemented on the basis of the chemical information system CACTVS [1]. SONORA is able to distribute any kind of searches in a chemical structure database to a set of CACTVS clients running on a computer cluster. The measured speed up for searching a database is nearly linear with the number of CPUs used. Currently, SONORA is working on our 96-CPU Beowulf-type parallel computer cluster, consisting of 48 dual AMD Athlon XP1900+ nodes. SONORA can make use of an arbitrary number of nodes and both processors of each assigned node.

As a first application based on SONORA, we are implementing a search and display service providing access to a database of over 13 million unique small-molecule structures distributed by ChemNavigator [2]. All entries of this database are searchable by various criteria similar to our "Enhanced NCI Database Browser" (<http://cactus.nci.nih.gov/ncidb2/>), e.g. molecular formula, full structure, substructure, and molecular similarity. Additionally, calculated properties (such as ADME-type properties) useful in the context of drug development are being included in all entries of the database.

1. <http://www.xemistry.com>
2. <http://www.chemnavigator.com/cnc/chemInfo/iRL.asp>

P-27 : Surrogate Docking: High Quality Docking at High Throughput Speeds

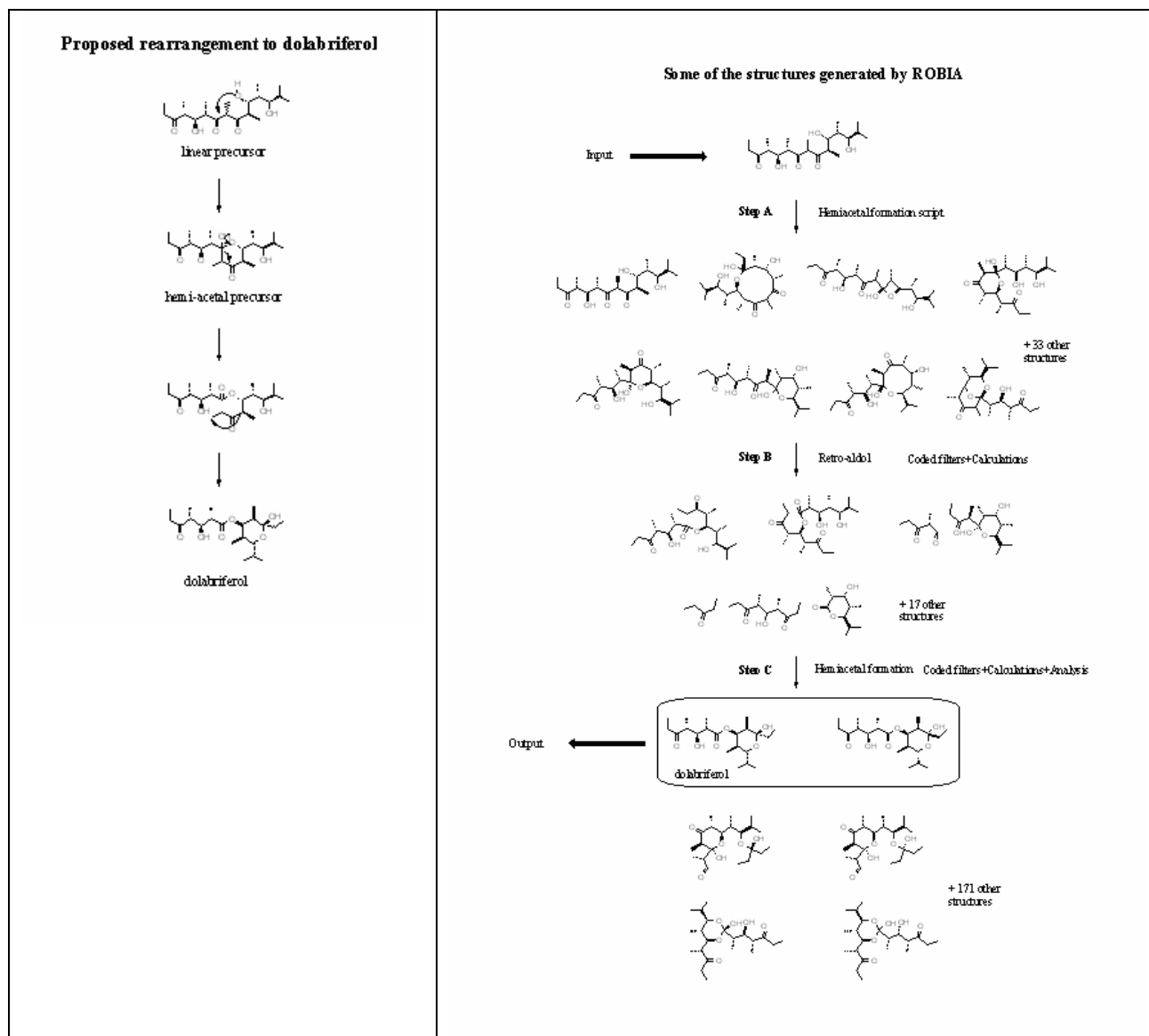
Andrew Smellie; Arqule, Woburn, MA, US
Sukjoon Yoon and Anton Filikov, Arqule

A methodology has been developed that provides a user-controlled continuum that trades off docking quality v.s. speed. By selecting a fraction of the molecules to be docked on a target, regular docking is performed and models are constructed that predict the likelihood of other molecules docking successfully. By ranking molecule sets from the scores from the model, it will be shown that most of the molecules that dock well are ranked highly, thus giving high enrichment. It will be shown that most of the good docking molecules can be obtained by docking a fraction of the database ranked by the model and that these molecules contain a high proportion of active compounds. Details will be shown that show the effectiveness of different descriptor sets and different model building methods. Examples will be described from docking studies of molecules from the NCI database on CDK2 and with known inhibitors of the Estrogen Receptor

P-29 : ROBIA: A Reaction Prediction Program

Ingrid Socorro; Cambridge University, Unilever Centre for Molecular Science Informatics, Cambridge, GB
Jonathan M. Goodman, Cambridge University, Unilever Centre for Molecular Science Informatics
Keith T. Taylor, Elsevier MDL

We are developing a computer program, ROBIA, with the purpose of predicting and analysing organic reactivity. This interactive computer program predicts the products of organic reactions from the starting materials and the reaction conditions, based on the selected transformations within its database. This mechanistic approach generates a large number of products, from which the most important are selected using filters and molecular modeling calculations. The procedure has been applied successfully to the biosynthesis of dolabriferol [1] as shown in the example below.



1. Ciavatta, M. L.; Gavagnin M.; Puliti R.; Cimino G. *Tetrahedron*, 1996, 52, 12831.

P-31 : 3D Structure-Activity Relationships of Non-Steroidal Ligands in Complex with Androgen Receptor Ligand-Binding Domain

Annu Söderholm; Finnish IT Center for Science, Espoo, FI
Lehtovuori Pekka and Nyrönen Tommi, Finnish IT Center for Science

Androgen receptor (AR) is a member of the nuclear receptor superfamily and functions as a ligand-dependent transcription factor in the regulation of AR targeted gene expression. The transcriptional activation of ARs is regulated through agonist and antagonist binding to the ligand-binding domain (LBD), resulting in conformational changes and subsequent recruitment of co-regulators. The agonistic mechanism and the agonist-bound structure of AR LBD are known, unlike the antagonist-bound structure and the antagonistic mechanism.

In this study [1], we applied the Comparative Molecular Similarity Indices Analysis (CoMSIA) method to gain insight into the physicochemical properties contributing to the binding affinity of a set of non-steroidal AR ligands. We combined molecular docking to the 3D-QSAR analysis to identify their preferred binding modes within the AR ligand-binding pocket (LBP) and to generate the ligand alignment for the 3D QSAR analysis.

The data for 70 AR ligands containing 67 non-steroids were obtained from the literature [2-6]. This panel represents a diverse set of compounds in terms of structure and function. The ligands were divided into a training set of 61 compounds and a test set of 9 compounds. Model validation was carried out by leave-one-out (LOO) and random groups (RG) cross-validation methods.

The CoMSIA model derived from the hydrophobic and hydrogen bond acceptor fields using five PLS components is stable and statistically significant as indicated by internal validation ($Q^2_{LOO}=0.656$, $SDE_{PLOO}=0.576$; $Q^2_{RG10}=0.612$, $SDE_{PRG10}=0.612$; $R^2=0.911$, $SEE=0.293$). The external validation using the test set indicate a model with good predictive power ($pred-R^2=0.800$, $SEE=0.367$).

The interpretation of the model is compatible with the protein environment. The superposition is thus likely to represent the biologically active conformations of the non-steroidal ligands, and the results provide information on how the ligands bind and interact with the AR LBD.

1. Söderholm AA, Lehtovuori PT, Nyrönen TH, J. Med. Chem. DOI: 10.1021/jm0495879.
2. Dalton JT, Mukherjee A, Zhu Z, Kirkovsky L, Miller DD, Biochem. Biophys. Res. Commun. 1998, 244, 1-4.
3. Kirkovsky L, Mukherjee A, Yin D, Dalton JT, Miller DD, J. Med. Chem. 2000, 43, 581-590.
4. Van Dort ME, Robins DM, Wayburn BJ, J. Med. Chem. 2000, 43, 3344-3347.
5. Van Dort ME, Jung YW, Bioorg. Med. Chem. Lett. 2001, 11, 1045-1047.
6. Yin D, He Y, Perera MA, et al., Mol. Pharmacol. 2003, 63, 211-223.

P-33 : Open Content Databases and Open Source Libraries for Chemoinformatics

Christoph Steinbeck; Cologne University, Cologne, DE
Stefan Kuhn and Christian Hoppe, Cologne University
Egon Willighagen, Radboud University Nijmegen

Traditionally, scientists share information and knowledge to enable others to build upon their results. While software development in early computational chemistry adhered to this paradigm, there were many areas of academic software and database development in chemistry which adopted a closed source/closed content culture. In contrast to bioinformatics, for example, chemistry thus lacks those valuable open data collections that allow scientists all over the world to perform research based on community-collected data.

With this talk we want to emphasize the importance of building open data repositories in chemistry using open source software. We will exemplify how this can be done based on two projects run by our group.

The Chemistry Development Kit (CDK) is an open-source Java library for Structural Chemo- and Bioinformatics [1]. Its architecture and capabilities as well as the development as an open-source project by a team of international collaborators from academic and industrial institutions will be described. The CDK provides methods for many common tasks in molecular informatics, including 2D and 3D rendering of chemical structures, I/O routines, SMILES parsing and generation, ring searches, isomorphism checking, structure diagram generation, QSAR descriptor calculation, and much more. The CDK forms the basis of a number of applications [2-4], such as the open web database for organic structures and their NMR data, NMRShiftDB [5], which is available to the public at <http://www.nmrshiftdb.org>. NMRShiftDB, with now more than 13.000 organic compounds and their NMR spectra, will serve as an example for an effort to create a community-built open content, open submission database, which grows by contributions from the user community. We will describe how the quality of data entered by the users is ensured by combining automated controls with a peer review by registered human reviewers. NMRShiftDB allows for (sub-) structure, (sub-) spectra and textual searches. It can further perform ^{13}C -NMR spectrum predictions based on the HOSE code method and the database material.

Finally, a number of other Open Source projects developed by us and collaborating groups will be introduced to the audience in a short overview.

1. Steinbeck, C., Han, Y. Q., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E., *Journal of Chemical Information and Computer Sciences*, 2003, 43, 493.
2. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D., *Nucleic Acids Research*, 2004, 32, D431.
3. Steinbeck, C., *Journal of Chemical Information & Computer Sciences*, 2001, 41, 1500.
4. Krause, S., Willighagen, E., and Steinbeck, C., *Molecules*, 2000, 5, 93.
5. Steinbeck, C., Kuhn, S., and Krause, S., *Journal of Chemical Information & Computer Sciences*, 2003, 43, 1733

P-35 : Modeling the Metabolism of Xenobiotics

Lothar Terfloth; Universitaet Erlangen-Nuernberg, Erlangen, DE

Inappropriate pharmacokinetic properties are often responsible for the attrition of a new drug in a late phase of its development process. For this reason, the prediction of an acceptable ADME-Tox (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profile for a new compound at an early stage is a key step in the drug discovery process. Xenobiotics are oxidized, reduced, or hydrolyzed in the first phase of the metabolism. Cytochrome P450 is involved in more than 90 percent of the oxidation reactions. The two major isoforms of cytochrome P450 which contribute to these oxidation reactions are 3A4 with about 50 percent and 2D6 with about 25 percent. In the second phase, a conjugation reaction such as a glucuronidation, acetylation, methylation, or glutathione conjugation follows. Here, we direct a special focus to the investigation of the metabolism of xenobiotics by cytochrome P450.

X-ray crystal structures for the membrane-bound, human Cytochrome P450 enzymes are available just recently. Therefore, we applied methods from ligand-based drug design to pursue our studies. A knowledge-based approach using neural networks will be presented [1]. For the classification of substrates of Cytochrome P450 3A4 and 2D6 the program SONNIA [2] (Self-Organizing Neural Network for Information Analysis) has been used.

Furthermore, we report on the development of a reaction database related to the metabolism of xenobiotics by human Cytochrome P450 enzymes.

1. J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design* (Wiley-VCH, Weinheim, ed. 2, 1999).
2. SONNIA, available at <http://www.mol-net.de/software/sonnia/index.html>

P-37 : BRUTUS: Rapid Optimization of Molecular Electrostatic Overlay - Evaluation of the Applicability of the Algorithm

Anu Tervo; University of Kuopio, Kuopio, FI
Toni Rönkkö and Antti Poso, University of Kuopio
Tommi H. Nyrönen, Finnish IT Centre for Science

BRUTUS is a fast molecular field-based superposition algorithm developed for chemical similarity searching. The properties of chemical compounds (e.g., possible biological activity to a certain target protein) are initially dependent on the distribution of their electrons surrounding the atomic nuclei. Therefore, the usage of molecular field information that is modeling the charge distribution of the compounds can be considered as appropriate in virtual screening of chemical databases, especially when structurally dissimilar compounds with similar properties are of particular interest.

Here we present the evaluation results of BRUTUS. It was utilized in chemical similarity searching on the basis of steric and electrostatic molecular fields of human immunodeficiency virus protease (HIV-1 PR) and cyclooxygenase-2 (COX-2) inhibitors. The search results of BRUTUS were structurally diverse, and comparable in

magnitude to the reference results obtained using Unity fingerprints with Tanimoto coefficient as a measurement of similarity degree. The results suggest BRUTUS as a fast molecular field-based algorithm that can be successfully used in field-based similarity searching of large molecular databases.

P-39 : The Use of Exclusion Volume in Feature Based Alignment Pharmacophore Models: Catalyst HipHopRefine

Samuel Toba; Accelrys, San Diego, CA, US
Al Maynard, Jon Sutter, and Marvin Waldman, Accelrys

This presentation provides an overview of the Catalyst HipHop and HipHopRefine pharmacophore generation and refinement algorithms. HipHop performs feature-based alignment of a collection of compounds and generates pharmacophore models. HipHop is used to match features, such as surface-accessible hydrophobes, surface-accessible hydrogen bond donors/acceptors, and charged/ionizable groups, against a set of active candidate molecules. HipHop does not incorporate any penalty for incompatible sterics, and the HipHopRefine algorithm has been designed as a post processing technique, suited to HipHop pharmacophores, which targets the addition of excluded volumes as just such a penalty. Details of the detection, ranking and selection of locations for excluded volume addition to pharmacophores based on a set of active and inactive molecules is given, along with details of improved enrichments and elimination of false positives in database searches compared with the original pharmacophores.

P-41 : Diversity of Chemical Structure Libraries Characterized by the Distribution of Tanimoto Indices

Kurt Varmuza; Vienna University of Technology, Vienna, AT
Heinz Scsibraný, Vienna University of Technology

Chemical structures of organic compounds are represented by binary 2D-substructure descriptors. Similarity between two structures is characterized by the Tanimoto index calculated from 1365 substructures. Software SubMat has been developed for an easy and automatic calculation of binary substructure descriptors for a set of molecular structures and a set of substructures. SubMat generates a text file with a line for each molecular structure that contains a string of 0's and 1's for absence or presence of the substructures. Computing time for 1000 molecular structures and 200 substructures is typically one second (Pentium IV, 2.6 GHz). SubMat can be optionally executed by calling it from another program. In this case a command file is used to transfer file names and parameters to SubMat. During execution so called semaphore files are used to communicate with the calling program. The contents of spectral databases (IR, MS), with up to about 100,000 structures, have been characterized by the Tanimoto indices of randomly selected structure pairs. The distribution of typically up to one million Tanimoto indices describes the structural diversity of a database. Shape and parameters of such distributions are discussed.

P-43 : Detection of Toxicity Indicating Structural Patterns

Modest von Korff; Actelion Ltd., Allschwil, CH
Thomas Sander, Actelion Ltd.

Since several years the early stages of drug discovery are driven by multiobjective optimization. Besides a ligand's binding affinity to the target protein its ADMET features came increasingly into the researchers' focus. Where reliable high-throughput assays to evaluate ADMET properties are expensive or missing there is a high demand for predictive in-silico techniques. Thus, computational methods delivering indicators of a compounds' toxicity potential are of high interest to medicinal chemists.

Based on the assumption that structurally similar compounds are also similar concerning their toxicity profile, we started a topological exploration of the RTECS database, which covers various toxicity classes. The IDDB database was used as reference for non-toxic compounds. The clustering and classifications methods applied were Naive Bayesian Clustering, k Next Neighbor Classification and Support Vector Machines. To find the optimum molecule

representation we analyzed the behavior of three different descriptors. We underlying descriptors comprised one fragment based chemical fingerprint, a topological walk based chemical fingerprint and a topological pharmacophore end point descriptor. Any toxicity model based on one of these classification algorithms and one of these descriptors were trained and tested on independent datasets.

Furthermore, we created a toxicity alerting system for compounds that are outside of the known chemical space of toxic compounds. Considering that a compound's toxicity is often caused by a certain chemical substructure we shredded the RTECS database yielding 100.000s of substructure fragments. By introducing query features we retained the original substitution patterns within these fragments. The occurrence of each fragment was counted within all molecules that expose a certain class of toxicity and then normalized by the fragment's natural occurrence found in the reference database. Fragments with both a statistically relevant overall frequency and significantly higher occurrence in the toxic group of compounds were listed as potentially risky fragments. For predicting the toxicity of an unknown molecule we run a substructure search of all risky fragments, which, depending on the result, may give us an indication of potential toxic behavior.

To visualize the chemical space of all RTECS and IDDB molecules available we trained a self-organizing map (SOM) with the compounds of both databases. The map topology was quadratic, toroidal and contained 10000 neurons. Finally we mapped the compounds of various toxicity classes onto the SOM, and colored those compounds accordingly. The result shows the distribution of toxic compounds in the chemical descriptor space. Areas with a high density of toxic compounds indicate regions in the chemical space to be avoided. Mapping external compounds onto the 'toxicity tinged' SOM provides for another toxicity measure purely based on the toxicity profile of the nearest neighbors and their similarity distances to the test compound.

Summarizing, we found that a Support Vector Machine in combination with the Fragment Based Fingerprint is well suited to classify compounds into various toxicity classes. Also SOMs perform excellent in separating toxic from nontoxic substances. While these methods reliably detect toxicity risks for compounds being similar to already known toxic compounds, our fragment based approach also covers compounds being dissimilar to anything known.

P-45 : Representing Structural Databases in a Self-Organising Map

Ron Wehrens; Radboud University Nijmegen, Nijmegen, NL

Rene de Gelder, Willem Melssen, and Lutgarde Buydens, Radboud University Nijmegen

We present a way to visualise large numbers of crystal structures, as represented by their simulated powder diffraction patterns, in a Kohonen feature map. Essential is the application of a recently introduced similarity criterion, the weighted cross-correlation. It will be shown that good results are obtained even if the network is trained with a small subset of the complete database. This makes it possible to construct the map, using common hardware, in a few hours. This two-dimensional visualisation has a number of important applications, such as fast and easy screening of a database, the selection of a representative set, and providing an overview of the contents of the database in terms of structural diversity of specific chemical classes of compounds.

P-47 : Drug Design Applications Based on COSMO-RS

Karin Wichmann; COSMOlogic GmbH & Co. KG, Leverkusen, DE

Andreas Klamt, COSMOlogic GmbH & Co. KG

Screening charge (σ) surfaces and screening charge distributions (σ -profiles) of molecules derived from quantum chemical COSMO (Conductor-like Screening Model) calculations offer a broadly applicable description of molecular interactions in liquid phases, which has already been used successfully in many applications in chemical engineering and in ADME property prediction. This approach is perpendicular to the force field based modelling methods commonly used in drug design and thus it has the potential to complement the established methods.

Although σ -profiles don't encode any structural information, they can be used to investigate receptor-ligand interactions: If the σ -profile of a receptor is known, the receptor can be considered as a pseudo-liquid and the chemical potential of a ligand molecule in such a pseudo-liquid receptor (PLR) can be calculated. Since the calculation of the chemical potential of the ligand in water is a standard task for COSMO-RS, the partition

coefficient of the ligand between the PLR and an aqueous phase can easily be calculated. This partition coefficient can be used as a measure for the overall polarity suitability of drug and receptor and may be a useful number in the selection of promising drug candidates. Since the receptor σ -profile has to be calculated only once, large numbers of drug candidates can be screened. For the COSMO treatment of enzymes a procedure based on linear scaling semi-empirical calculations has been developed. AM1-COSMO calculations and subsequent BP/SVP single point calculations were performed for 161 tri-peptides. On the basis of bond order and atom types of the neighboring atoms, 39 asymmetric bond types were found, and dipole and quadrupole corrections were fitted. Thus, the original rms deviation between the AM1 and BP/SVP electrostatic potentials was reduced from 0.0112 e/Å to 0.0058 e/Å.

First applications of this novel approach focus on factor Xa and kinase receptor-inhibitor interactions and will be presented.

P-49 : Techniques for Location-Independent Chemoinformatics Teaching and Research

David Wild; Indiana University School of Informatics, Bloomington, IN, US
Gary Wiggins, Indiana University School of Informatics

We have recently developed a chemoinformatics teaching curriculum and research base at Indiana University that allows participation at multiple geographic locations and institutions, with an aim to engage students, lecturers and researchers in multiple locations and disciplines and to bolster the connectivity of chemoinformatics to the wider research community. In this presentation we will discuss the programs we have created, evaluate their effectiveness, and detail the distance learning technologies that we have found most effective.

P-51 : ChemXtreme: Harvesting Chemical Information From Internet Using Distributed Approach

Muthukumarasamy Karthikeyan; National Chemical Laboratory, Pune, IN
S Krishnan, National Chemical Laboratory

Internet is a resource of large amount of unstructured chemical information. There is a wealth of scientific information available with experimental data, which require data mining and analysis tool. The java based software entitled "ChemXtreme" developed for harvesting chemical information from Internet using distributed approach is presented. In the present investigation, a novel and secured method of harvesting chemical information from public resources using distributed systems. The technology used for this purpose utilizes the "searching the search engine" strategy, where the URLs returned from search engines are analyzed for required patterns 'word by word' for chemical information and transformed automatically into structured format compatible for database operations. The query data from server is encoded, encrypted and compressed and sent to all the 'participating' active clients in the network with internet connectivity. The data received from the clients after web search and analysis is decompressed, verified and added to the database for data mining and further analysis. Given list of CAS-RN or Chemical Name (preferably common name) of the substances with selective keywords as patterns returns more accurate and useful data from URLs from the search engines. With 2MBPS connectivity speed the program is able to search (2-3 sec per query) and analyse (6-7 sec per URL) per client. The more detailed results on some case studies involving search of about common 100,000 CAS-RN with MSDS context in distributed environment will be presented. The biological activity data retrieved for list of chemical substance can be directly used for QSAR or QSTR analysis in combination with some of the descriptor generator and statistical tools.

P-53 : On the Use of Spectra as Molecular Descriptors in QSAR Research

Egon Willighagen; Radboud University Nijmegen, Nijmegen, NL
R. Wehrens and L.M.C. Buydens, Radboud University Nijmegen

Several papers have been published that describe the use of infrared and NMR spectra as molecular descriptors in quantitative structure activity relationship (QSAR) research. This research further explores the use of spectra and

applied to several data sets with biochemical and physical properties. Results are compared against models based on conventional descriptors, such as topological, geometrical and electronic descriptors. Several modern modelling methods, like Partial Least Squares (PLS), Support Vector Regression (SVR) and Classification and Regression Trees (CART), are used and compared.

P-55 : The Development of a Machine Learning Algorithm for Ligand-Based Virtual Screening

David Wood; University of Sheffield, Sheffield, GB
Peter Willett and Beining Chen, University of Sheffield
Xiaoqing Lewell, GlaxoSmithKline

Machine learning techniques for ligand-based virtual screening (VS) generally involve inputting an algorithm with a training set of examples of active and inactive compounds in a descriptor format. The algorithm then uses the information in the training set to develop a model of activity which can be used to classify compounds with unknown activity. Such algorithms can therefore be used to rapidly identify promising compounds from large virtual libraries or supplier's catalogues.

Kernel discrimination methods are a class of machine learning algorithms that classify unknown items by comparing population density distributions of each of the training set classes over the descriptor space. Binary Kernel Discrimination (BKD) is an example of a kernel-based classification technique that is input with multivariate binary data: a commonly used method of representing chemical compounds. BKD has only fairly recently been applied to VS [1]. Several comparative studies have indicated that it outperforms many commonly used ligand-based VS techniques [2], and has comparable performance to Support Vector Machines [3]. Further development of this technique could yield a useful tool for drug discovery.

A series of experiments were performed to develop BKD for VS. The performance of a range of commonly used fingerprint descriptors, when used in conjunction with BKD, was compared over 11 different activity classes. The fingerprints included BCI, Daylight, Unity and SciTegic's Extended Connectivity fingerprints. The SciTegic fingerprint system was found consistently to demonstrate the best performance over many of the activity classes and so is the 2D fingerprint descriptor of choice for BKD. A further experiment demonstrated that the performance of the system is improved by increasing the number of inactive compounds in the training set, although the improvement rapidly tails off once sufficient inactive compounds have been included. Finally, alternative methods of optimising the model of activity in the algorithm's training stage were considered and an improved variation to the original method is proposed. This poster will report in detail the experiments that have been conducted to develop the BKD method for VS.

1. Harper, G., et al., Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comp. Sci.*, 2001. 41: p. 1295-1300.
2. Wilton, D.J., et al., Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci.*, 2003. 43: p. 469-474.
3. Wilton, D., Willet, P., Delaney, J., Lawson, K., & Mullier, G, The Use of Binary Kernel Discrimination and Support Vector Machines for Virtual Screening in Pesticide-Discovery Programmes. (In preparation, 2004).

P-57 : Development of the Transition State Data Base

Toru Yamaguchi; Yamaguchi University, Ube, JP
Kenzi Hori, Yamaguchi University

We have been developing a data base including information of transition states (TS) such as TS structures, activation energies and so on for various reactions. It is called the Transition States Data Base (TSDB) which assists to develop synthesis routes of compounds by using information of TSs together with the help of the KOSP program, a synthesis routes designing system developed by Funatsu and his coworkers. The TSDB is a system consisting of six programs as follows.

1. TSLB is the main library storing information of transition states.
2. TS_Search assists to search transition states of reactions.
3. TSDB_View manages data in TSLB.
4. FIND_Tsinfo searches information in TSLB related to synthesis routes from KOSP. This data base is constructed with ChemFinder of Cambridge Soft.
5. Reaction_View is used for viewing molecular structures, animations of molecular vibrations and geometry transformations along the intrinsic reaction coordinates. We are now using the JMol program in this purpose.
6. Auto_PTOD calculates TS structures at the B3LYP/6-31G* level of theory from initial geometries of PM3 calculations stored in the TSLB.

In the present study, we will represent the TSDB in detail and demonstrate how the data base works for developing synthesis routes of compounds.

P-59 : Pharmacophore Hypotheses Derived from Protein Structure and Inhibitors: Methods & Binding Site Comparisons of CYP3A4

Litai Zhang; Bristol-Mayers Squibb, Princeton, NJ, US

Cytochromes P450 (CYP 450) are the major enzymes involved in the oxidative metabolism of various drugs and other xenobiotics. A large number of pharmacologically active compounds are often rejected as new drug candidates due to either unsuitable pharmacokinetics or their interference with the metabolism of existing therapeutic agents. In many cases this is because the compounds are either low K_m substrates or potent inhibitors of one or more CYP 450 isozymes, such as CYP 3A4, 2C9, 2D6, 2C19. The CYP 3A4 ranks among the most important drug metabolizing CYP isoforms present in human liver. Numerous inhibitory drug interactions of high clinical significance involving CYP 3A4 substrates have been described[1]. Historical precedents within BMS have indicated that CYP 3A4 has been particularly problematic.

Pharmacophore hypotheses derived from protein structure and inhibitors to elucidate the active site for the CYP 3A4 can be used in the triage of HTS & actively direct synthesis away from CYP 3A4 issues.

P-61 : Automatic Classification of Chemical Reactions without Identification of Reaction Centers

Qing-You Zhang; Universidade Nova de Lisboa, Caparica, PT
Joao Aires-de-Sousa, Universidade Nova de Lisboa

Automatic classification of chemical reactions is of high importance for the analysis of reaction databases, reaction retrieval, reaction prediction, or synthesis planning. After a period of almost inactivity, this topic is re-emerging particularly due to the current interest in metabolic reactions.

Well-documented methods for reaction classification have been based on a) physicochemical properties of bonds or atoms at the reaction center; [1,2] b) variation of physicochemical properties of one or more atoms attached to the reaction center; [3] c) degree of overlap of weighted fragment sets; [4] d) numerical codes of the neighborhoods of the reaction center. [5] These approaches require atom mapping and identification of the bonds broken or made in the reaction (the reaction center). Some also require ranking of the bonds involved in the reaction, and a scheme to compare reactions with a different number of bonds involved.

In the course of our QSAR studies with molecular maps of bonds (MOLMAPS), it was recognized that the difference between the map of the products bonds and the map of the reactants bonds could be interpreted as a map of the reaction. A MOLMAP is based on a Kohonen self-organizing map previously trained with bonds represented by a series of physicochemical properties calculated by PETRA. [6] The MOLMAP of a molecule is the pattern of neurons that are activated by the bonds existing in that molecule. Bonds far apart from the reaction center are unchanged during the reaction and activate the same position (neuron) of the MOLMAP both in the reactant map and in the product map. Therefore, the difference map (MOLMAP of the reaction) gets a zero value at that neuron. The

pattern of neurons in the MOLMAP of the reaction with non-zero values relates to the bonds of the reactants that break or change, and to the bonds of the products that are made or changed in the reaction. The former lead to negative values, while the latter lead to positive values.

Following this approach, 543 photochemical reactions were encoded that involve two reactants and one product. They were manually assigned to eight classes – seven types of reactions (203 reactions) and a class with the remaining 340 reactions. The data set was divided into a training set consisting of 429 reactions and a test set with 114 reactions.

Classification of the reactions by machine learning on the basis of MOLMAPS of size 12x12 was investigated with an unsupervised method (Kohonen self-organizing map) and a supervised method (random forest). Kohonen SOMs achieved 97% of correct classifications for the training set and 87% for the test set. Random forests obtained 99% of correct predictions for the training set and 93% for the test set.

This work demonstrates that the difference between MOLMAP descriptors of products and reactants can be used to represent and classify reactions without assignment of reaction centers.

ACKNOWLEDGMENTS. The authors thank InfoChem GmbH (Munich, Germany) for sharing the data set of photochemical reactions. Z. Q. Y. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for a post-doctoral grant under the POCTI program (SFRH / BPD / 14476 / 2003).

REFERENCES

1. L. Chen; J. Gasteiger; J. Am. Chem. Soc. 1997, 119, 4033-4042.
2. O. Sacher. Classification of organic reactions by neural networks for the application in reaction prediction and synthesis design. Ph. D. Thesis, University of Erlangen-Nuremberg. http://www2.chemie.uni-erlangen.de/services/dissonline/data/dissertation/Oliver_Sacher/html/
3. H. Satoh; O. Sacher; T. Nakata; L. Chen; J. Gasteiger; K. Funatsu; J. Chem. Inf. Comput. Sci. 1998, 38, 210-219.
4. T. E. Moock; D. L. Grier; W. D. Hounshell; G. Grethe; K. Cronin; J. G. Nourse; J. Theodosiou; Tetrahedron Comput. Methodol. 1988, 1, 117-128.
5. <http://www.infochem.de/eng/downloads/Classify.pdf>
6. <http://www2.chemie.uni-erlangen.de/software/petra>

P-63 : Chiral QSPR Analysis of ¹³C NMR Properties in Chiral Solvents

Qing-You Zhang; Universidade Nova de Lisboa, Caparica, PT
Joao Aires-de-Sousa, Universidade Nova de Lisboa

The NMR chemical shifts of two opposite enantiomers are not necessarily the same in a chiral environment such as a chiral solvent. Similarly, a single enantiomer may exhibit different chemical shifts if the NMR spectrum is taken in two enantiomeric solvents. Kishi and co-workers [1-3] observed different ¹³C NMR chemical shifts for chiral alcohols in (S,S)- or (R,R)-BMBA-p-Me chiral solvents. A database was produced with those differences for the atoms adjacent to the chiral hydroxymethine center in 24 chiral alcohols. The prediction of such differences, and comparison with experimental values, can assist in the assignment of the absolute configuration.

In this poster we show how counterpropagation neural networks (CPG NNs) were trained and applied to estimate the difference between the chemical shifts of a given carbon atom in the two enantiomeric solvents. The neural networks receive as input a representation of the atom (an atomic chirality code) and give as output the difference of its chemical shift in the (R,R)-BMBA-p-Me and (S,S)-BMBA-p-Me solvent.

We have previously developed molecular chirality codes that represent the chirality of the whole molecule.[4,5] Introduction of an atomic chirality code is here necessary to account for local chirality around an atom, which is relevant for modeling atomic properties such as the NMR chemical shift. An atomic code means that every atom rather than the whole molecule has its own chirality code. For the generation of the molecular chirality codes, all the sets of 4 atoms are analyzed. The atomic version of the chirality code is computed in the same way, but only sets of

4 atoms that include a specific atom in the molecule are processed – to produce the atomic chirality code for that specific atom.

We used a dataset of 94 atoms and the corresponding chemical shifts differences. The dataset was partitioned into a 74-objects training set and a 20-objects test set. Parameters of the atomic chirality codes were optimized for the training set, and CPG NNs were trained to predict the sign of the differences. Correct predictions could be achieved for all the cases of the test set. With the same procedure, the quantitative prediction of the chemical shifts differences was investigated, and r^2 of 0.839 and 0.936 between the predicted and the experimental values were obtained for the training and test set respectively.

The results show that atomic chirality codes describe the chirality of an atom's environment in a way that can be correlated with a physical property - its NMR behavior in a chiral solvent. This work is also a contribution to the assignment of absolute configuration from NMR data, particularly for its implementation in automatic systems.[6]

ACKNOWLEDGMENTS. Z. Q. Y. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for a post-doctoral grant under the POCTI program (SFRH / BPD / 14476 / 2003).

REFERENCES:

1. Kobayashi, Y.; Hayashi, N.; Kishi, Y. Toward the creation of NMR databases in chiral solvents: bidentate chiral NMR solvents for assignment of the absolute configuration of acyclic secondary alcohols. *Org. Lett.* 2002, 4, 411-414.
2. Kobayashi, Y.; Hayashi, N.; Kishi, Y. Application of chiral bidentate NMR solvents for assignment of the absolute configuration of alcohols: scope and limitation. *Tetrahedron Lett.* 2003, 44, 7489-7491.
3. Kobayashi, Y.; Czechtizky, W.; Kishi, Y. Complete stereochemistry of tetrafibricin. *Org. Lett.* 2003, 5, 93-96.
4. Aires-de-Sousa, J.; Gasteiger, J. Prediction of enantiomeric selectivity in chromatography. application of conformation-dependent and conformation-independent descriptors of molecular chirality. *J. Molec. Graph. Model.* 2002, 20, 373-388.
5. Aires-de-Sousa, J.; Gasteiger, J.; Gutman, I.; Vidovič, D. Chirality codes and molecular structure. *J. Chem. Inf. Comput. Sci.* 2004, 44, 831-836.
6. Zhang, Q. Y.; Carrera, G.; Gomes, M. J. S.; Aires-de-Sousa, J. Automatic assignment of absolute configuration from 1D NMR data. *J. Org. Chem.*, 2005, in press.

P-65 : Weighted Reaction Searching – Using Focused Fingerprints for Discriminated Results

Tim Aitken; Accelrys, Cambridge, GB
Ian Buchan, Accelrys

Similarity searching has long been used within the Drug Discovery industry for retrieving chemically similar hits from compound databases, clustering and carrying out diversity studies. Several algorithms for fingerprint comparisons are commonly used and a wide variety of fingerprinting methods have been implemented in commercial and non-commercial database search systems. With the widespread use of Reaction databases, however, the traditional molecule-oriented approaches have not proven to be as useful and search methods tend to focus on substructure, exact structure and text searches.

A new approach is demonstrated using a hash fingerprinting algorithm weighted towards reaction centre transformation and discriminating reactants from products within a reaction database. This encapsulation of molecular, reaction and reaction centre information within a 'zoned' fingerprint can return unexpected hits from a database search, which traditional searches would fail to return.

P-67 : Turbo Similarity Searching

Jérôme Hert; University of Sheffield, Sheffield, GB
 Peter Willett and David J. Wilton, University of Sheffield
 Kamal Azzaoui, Edgar Jacoby, and Ansgar Schuffenhauer, Novartis Institute for Biomedical Research

Previous work has shown that fusing the outputs of similarity searches carried out using different isoactive reference compounds produces a more effective ranking than one based on just a single reference compound. Turbo similarity searching applies this strategy using a reference molecule and its nearest neighbours. The similar property principle implies that these neighbour compounds are likely to have a similar bioactivity profile; accordingly it may be worth including them in a fusion procedure. The effectiveness of this method is investigated by means of simulated virtual screening experiments using the MDL Drug Data Report Database. Extensive searches are carried out for eleven diverse activity classes and consistently demonstrate the superiority of turbo similarity searching over conventional similarity search. This method hence represent a simple way of enhancing similarity-based virtual screening methods.

P-69 : Construction of a System Predicting Hydration Rates of Toxic Substrates in the Environmental Conditions

Yutaka Ikenaga; Yamaguchi University, Ube, JP
 Kenzi Hori, Yamaguchi University

Recently, we are required to investigate decomposition of toxic substances emitted to nature in order to avoid environment disruptions. However, it is impossible to measure rates of decomposition reactions for all the compounds since there are enormous numbers of toxic compounds. Theoretical calculations can play important role in predicting whether or not a toxic compound easily decomposes to others in the environmental conditions. It is because calculated activation energies (E_a) are closely related to the rates of decomposition of substrates. The E_a values should be correlated with experimental ones and used for predicting substrate to be decompose in the environmental conditions, i.e., a reaction with E_a more than, for example, 50 kcal mol⁻¹ does not proceed in rivers, lakes or seas. In order to confirm this concept, we adopted esters which are forced to measure decomposition rates by a law in Japan. There are many esters which are toxic and widely used in industrial field. For this purpose, the mechanism of ester hydrolysis in the acidic conditions was investigated at the B3LYP/6-31G* level of theory. The DFT calculations did not locate any tetrahedral intermediates which many text books adopted as key intermediates. We offered an alternative mechanism which was an activation barrier of 31.0 kcal mol⁻¹. We are also measuring E_a 's for several esters and to correlate them with calculated ones. We will use this correlation to construct a system predicting easiness of decomposition of esters in the environmental conditions.

P-71 : Structure-Based Design of Potential Novel Inhibitors of FGFR and VEGFR Tyrosine Kinase as Anti-Angiogenesis Agents

Naparat Kammassud; Mahidol University & Centre Universitaire, Orsay Cedex, FR
 Isabelle André and David S.Grierson, Centre Universitaire
 Opa Vajragupta, Mahidol University

Tumor angiogenesis is often the consequence of an angiogenic imbalance in which pro-angiogenic factors predominate over anti-angiogenic factors. Many growth factors and cytokines mediate cellular signaling through the activation of tyrosine kinases (TKs). In the area of cancer, receptor tyrosine kinases (RTKs) play an important role in the process of tumor growth, metastasis development, and angiogenesis. Thus, the search for small molecules modulating the biological activity of such enzymes is of special relevance to develop new potential therapeutic agents. Recently two indolinones developed by SUGEN Inc., SU4984 and SU5402 (Figure 1), were reported to exhibit moderate activity against the Fibroblast Growth Factor Receptor-1 tyrosine kinase (IC₅₀s of 10-20 nM), only SU5402 displayed selective activity .1μ

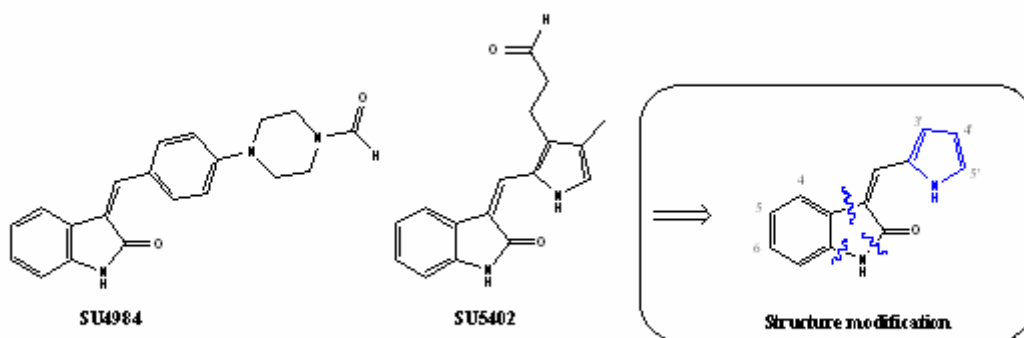


Figure 1: Indolinone inhibitors of Fibroblast Growth Factor Receptor-1 tyrosine kinase.

These inhibitors were co-crystallized with FGFR-1 tyrosine kinase [1]. Based on this structural information, SU5402 was used as a starting point for the conception of new families of inhibitors that can potentially display enhanced potency and selectivity.

Some earlier reports [2] have suggested that modifications at position 6 of SU5402 derivatives could increase the potency as well as the selectivity against VEGFR-2 and FGFR-1 tyrosine kinases. We report here the synthesis of new 6-substituted oxindole derivatives that will allow us to further explore the Structure-Activity Relationship (SAR) at this position.

In addition, we have used the information available on these derivatives to help identify alternate scaffolds by means of molecular modeling. In our strategy, we have generated focused “virtual” chemical libraries around the oxindole scaffold, varying e.g. the nature of the substituents at different chemically accessible positions (4-6, or 3'-5') of the indolinone core, replacing the pyrrole ring by a new ring, or breaking the indolinone ring (Figure 1). The virtual chemical libraries generated were then screened “in silico” using the FGFR-1 tyrosine kinase crystallographic structure in order to identify new small molecules that could potentially inhibit FGFR-1 and other highly homologous RTKs such as VEGFR-2. These studies were used as a guide in our chemistry efforts.

1. M. Mohammadi et al., *Science* 276 (1997), 955-960.
2. L. Sun et al., *J. Med. Chem.* 42 (1999) 5120-5130.

This work was partially supported by the Royal Golden Jubilee Ph.D. Program, Thailand.

P-73 : Linking the Real and Predictive Worlds: A Conceptual Model of Chemical Information

Chris Marshall; AstraZeneca, Macclesfield, GB

The purpose of prediction is to provide insight into the likely properties and behaviours of real world materials. But real world materials aren't always as accurately characterised as we would like. As part of the merger of Astra and Zeneca we have developed a conceptual model of chemical information based on the idea of independent but related container, sample and compound entities. The model is part of a project extending to cover the whole of the pharmaceutical Discovery process - the Discovery Information Model. In this paper I will show how these entities not only help to manage compound registration but also set the framework for bringing together virtual and real properties. By understanding the features of the entities and appropriately assigning them we are able to distinguish and relate experimental and predicted data in a way which takes into account our evolving understanding of a sample's chemical content.

P-75 : Calculation of Physicochemical Descriptors Based on a new Structure Representation

Jörg Marusczyk; Universitaet Erlangen-Nuernberg, Erlangen, DE
Thomas Kleinöder, Achim Herwig, and Johann Gasteiger, Universitaet Erlangen-Nuernberg

The handling of chemical structures, their input from and output to physical media, and the calculation of molecular descriptors for *in silico* predictions are the core of Chemoinformatics. Traditionally, software systems developed and used in the field of Chemoinformatics process and store chemical structures as connection tables. Such a representation, based on valence bond theory, comprises certain problems: a compound may be written in different resonance structures which all denote the same molecule. Nitrobenzene, for example, can be found in databases in the two different ionic forms and even with a pentavalent nitrogen. On the other hand, a connection table is always the same for different electronic states of a molecule, which have distinct physicochemical properties, e.g., there is no way to specify a carbene in the singlet state.

In order to overcome some of these problems, we developed a structure representation based on the ideas of the Hückel molecular orbital theory, namely the sigma/pi separation [1]. The sigma framework of a molecule consists of two-centers-two-electron sigma systems as in a traditional connection table. Pi electrons can constitute larger electron systems spanning over more than two atoms and can also include lone pairs and radical electrons. This representation scheme is called RAMSES (Representation Architecture for Molecular Structures as Electron Systems) and is part of a comprehensive C++ toolkit library called MOSES (MOlecular Structure Encoding System) that was recently developed in our group.

Based on the RAMSES representation described above, we developed a new model for the calculation of atomic partial charges that makes use of the sigma/pi separation. The sigma charge distribution is quantified by a modified version of the well established Partial Equalization of Orbital Electronegativities (PEOE) [2]. For the pi charge distribution a modified Hückel calculation is used. Both calculation schemes were calibrated with charges from natural population analysis [3] of DFT wave functions for both uncharged and charged compounds. For a wide range of organic compounds a very good correlation can be found. Further, we extend the model for the charges to calculate the resonance energies of charged molecules. In future, we plan to use the combination of atomic partial charges and resonance energies in the field of reaction prediction and evaluation.

1. S. Bauerschmidt, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* 1997, 37, 705-714.
2. J. Gasteiger, M. Marsili, *Tetrahedron* 1980, 36, 3219-3228.
3. A. E. Reed, F. Weinhold, *J. Chem. Phys.* 1983, 78, 4066-4073.

P-77 : Drug Design, Chemoinformatics and Public Web Services with Very Large Databases

Marc Nicklaus; National Institutes of Health, Bethesda, MD, US
Markus Sitzmann and Igor V. Filippov, National Institutes of Health
Wolf-Dietrich Ihlenfeldt, Xemistry GmbH

We report on the newest versions of the tools and resources used in the drug design and *in silico* screening work of the CADD Group at LMC, CCR, NCI. Many of these resources are implemented in the form of web services, and most of these are made freely available to the public. We present web-based search interfaces for databases with millions of compounds using a search engine operating in distributed mode across a Linux cluster. Many of these databases including multi-million collections of commercial screening samples, as well as data sets from various U.S. Government agencies, are being made publicly available. We present new automated tools for generating such web services as well as new calculable CACTVS hash code-based identifiers useful for rapid compound identification and database overlap analyses. We also briefly mention other chemoinformatics type services and tools available on our server at <http://cactus.nci.nih.gov>.

P-79 : QSAR Analysis for Infinite Dilution Activity Coefficients of Organic Compounds Using a CODESSA PRO Software

Kaido Tämm; University of Tartu, Tartu, EE
Peeter Burk, University of Tartu

A quantitative structure activity relationship (QSAR) study of the infinite dilution activity coefficients for a set of 38 organic compounds in ionic liquids, such as 1-methyl-3-ethylimidazolium bis((trifluoromethyl)sulfonyl)imide ([emim][N(Tf)₂]), 1,2-dimethyl-3-ethylimidazolium bis((trifluoromethyl)sulfonyl)imide ([em2im][N(Tf)₂]), and 4-methyl-N-butylpyridinium tetrafluoroborate ([bmpy][BF₄]) provided a general three-parameter QSAR models. QSAR study was carried out using the CODESSA PRO program. Three orthogonal theoretical molecular descriptors satisfactorily correlate with the activity coefficients. The descriptors, such as the complementary information content, the fractional partial negative surface area and the count of hydrogen donor sites directly describe the dilution mechanism in ionic liquids.

P-81 : A Neural Network Application in Multi-Target QSAR

Pierre-Jean L'Heureux; Universite de Montreal, Montreal, CA
Olivier Delalleau, Dumitru Erhan, Yoshua Bengio, and Shi Yi Yue, Universite de Montreal

Building a QSAR model of a new target for which few screening data is available is a daunting task. Hopefully, the new target may be part of a bigger family, for which we have plenty of screening data. We introduce a neural network architecture based on collaborative filtering that can use family information to produce predictive model of an undersampled target. We show it's performance on compound prioritization for an HTS campaign.

P-83 : The Quest for Bioisosteric Replacements

Jos Lommerse; NV Organon, Oss, NL
Markus Wagener, NV Organon

It is a major challenge to convert a compound resulting from lead finding activities into a successful drug. Whereas the initial lead compound may already bind with high affinity to the biological target, it will usually have some undesirable characteristics regarding oral bioavailability, metabolic stability, selectivity and/or toxicity.

One strategy to address these issues and to convert a lead compound into a development candidate is based on the concept of bioisosterism [1]: Structurally related compounds that both elicit the same biological activity are considered as bioisosters. A bioisosteric replacement transforms an active compound into another compound by exchanging a group of atoms with another, broadly similar group of atoms. The resulting new compound still has the original biological activity while improving on the undesirable characteristics.

We report a method that suggests potential bioisosteric replacements based on a topological pharmacophore description of the fragments. Based on that description, databases of R-groups, linkers and cores are searched for the most promising replacements. In order to focus the search on an improved ADMET profile, a number of search constraints (e.g. lipophilicity, flexibility, acidity/basicity) can be imposed. The method has been implemented as IBIS (Intranet BioIsoster Search) at Organon.

The topological pharmacophore description has been validated using the BioSter database [2,3], a database that collects examples of bioisosteric compounds from the literature. Several thousand pairs of bioisosteric fragments have been extracted from that database using unbiased criteria. Comparison of the true pairs of bioisosteric R-groups, linkers and cores from the BioSter database with random pairs confirmed the validity of the approach.

Several examples will be given which show the type of suggestions achievable with IBIS underlining the usefulness of the approach.

1. Patani GA, LaVoie EJ. Bioisosterism: A Rational Approach. *Chem. Rev.* 1996; 96: 3147-76.
2. Ujvary I. BIOSTER- A database of Structurally Analogous Compounds. *Pestic. Sci.* 1997; 51: 92-5.

- The BIOSTER database is available from Accelrys Inc. at <http://www.accelrys.com/>.

Keywords: bioisosters, topological pharmacophore fingerprint, descriptor validation

P-85 : Similarity Searching Using Molecular Interaction Fields

Kirstin Moffat; University of Sheffield, Sheffield, GB
Val Gillet, University of Sheffield
Gianpaolo Bravi and Andrew Leach, GlaxoSmithKline

Similarity searching is a valuable tool for aiding in the identification of possible lead compounds in drug discovery. Methods based on 3D field descriptors can be of particular use, since, unlike 2D methods, they do not consider molecular frameworks, but instead consider characteristics such as the overall shape, electrostatics and hydrophobicity of the molecules. These methods can therefore give a more representative view of what is “seen” by the receptor to which a compound will bind.

Field-based similarity searching methods have generally been based on atom-centred fields, for example, the Field Based Similarity Searcher (FBSS) program calculates similarity based on steric, electrostatic or hydrophobic fields or any combination thereof (Thorner et. al. (1996), Wild & Willett (1996)). The fields are represented by atom-centred gaussians and a genetic algorithm (GA) is used to find an alignment that maximises the similarity between two molecules using the Carbo coefficient applied to the gaussian representations.

The aim of this work is to calculate the similarity between two molecules based on their molecular interaction fields (MIFs) rather than atom-centred fields. The rationale for the approach is that two molecules may exhibit similar interactions with a receptor even when the atoms that give rise to the interactions are in different locations in the active site. Molecular interaction fields are calculated using the GRID program (Goodford, 1985) which places a molecule within a 3D grid and determines the interaction energy at each grid point between the molecule and a probe atom. The interaction energy at each grid point, xyz, is based on a number of components including the Lennard-Jones potential (E_{ij}), the electrostatic potential (E_{el}) and the hydrogen bonding potential (E_{hb}):

$$E_{xyz} = \sum E_{ij} + \sum E_{el} + \sum E_{hb}$$

Various probe atoms are available and more than one probe can be used at a time. The similarity between two molecules is then calculated from the grid representations using a methodology similar to that used in FBSS. The first step is to derive gaussian approximations of the MIFs. Points of minimum energy are identified in the grid and gaussians are centred at each of the minima. Each gaussian takes the form:

$$E_i = h \exp\left(-\frac{dis^2}{2\sigma^2}\right)$$

where h is the height of the gaussian, σ is the rate of decay of the gaussian and dis is the distance between the centre of the gaussian and the point, i , at which the interaction energy is being calculated. A simplex method is then used to optimise the parameters of the gaussians. Finally, a genetic algorithm is used to find the alignment of two molecules that maximises the similarity between the gaussian approximations.

Many different possibilities exist for deriving the gaussian approximations, for example, the number of gaussians and the number of grid points used in the optimisation can be varied, and the grid resolution and the number and type of probes used to calculate the fields can also be varied. The effectiveness of the various representations at approximating the MIFs has been determined by calculating a derived grid from the Gaussian approximations and comparing it with the original grid. Finally, the effectiveness of the method in similarity searching has been determined by comparison with established similarity methods via enrichment plots.

- Goodford (1985). “A Computational-Procedure For Determining Energetically Favorable Binding-Sites On Biologically Important Macromolecules”. *Journal of Medicinal Chemistry*, 28, 849-857.

2. Thorner, D.A., Wild, D.J., Willett, P., Wright, P.M. (1996). "Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials". *Journal of Chemical Information and Computer Science*, 36, 900-908.
3. Wild, D.J., Willett, P. (1996). "Similarity Searching in Files of Three-Dimensional Chemical Structures. Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm". *Journal of Chemical Information and Computer Science*, 36, 159-167.

RED POSTER SESSION ABSTRACTS

Red Poster Session Abstracts

P-2 : Descriptors of Chemical Reactivity and Application to Mutagenicity Prediction

Qing-You Zhang; Universidade Nova de Lisboa, Caparica, PT
Joao Aires-de-Sousa, Universidade Nova de Lisboa

Mutagenicity is strongly related to chemical reactivity, namely to the ability of a compound to be metabolically activated and to react with DNA. [1] Chemical reactivity depends on the properties of chemical bonds, which determine how bonds break and rearrange in the presence of certain reactants, catalysts and conditions.

In this communication we will show our studies with descriptors of molecular reactivity (physicochemical properties of bonds) for the prediction of mutagenicity in *Salmonella* (Ames assay). Those empirical descriptors are easily calculated from the molecular structure and can be quickly generated for large data sets of compounds.

In order to use the information concerning several properties of bonds for an entire molecule, and at the same time to keep its representation within a reasonable fixed length, all the bonds of a molecule are mapped into a fixed-length 2D self-organizing map.

A self-organizing map (SOM) is trained beforehand with a diversity of bonds from different structures (each bond described by seven bond properties calculated by PETRA [2]). Then all the bonds of one molecule are submitted to the trained SOM, and the pattern of activated neurons is interpreted as a map of the reactivity features of that molecule (MOLMAP) – a fingerprint of the bonds available in that structure.

MOLMAP descriptors were generated for 548 compounds, and were complemented with 17 general molecular descriptors such as the molecular weight, maximum charge, or ring strain energy. On their basis, a random forest established a predictive model for mutagenicity. Learning in a random forest results from training an ensemble of classification trees. [3] Each tree is grown with a random subset of descriptors and a random subset of objects. The final prediction is obtained by majority voting. Random forests additionally associate a probability to every prediction, and report the importance of each descriptor in the global model.

We used data from the Berkeley Carcinogenic Potency Database [4] consisting of SMILES strings and the corresponding outcome of the Ames test. [5] After excluding inorganic and organometallic compounds, salts, duplicates, and structures not accepted by PETRA 3.11, [2] the remaining 548 structures were partitioned into a training and a test set with 445 and 103 objects respectively. Correct predictions were achieved for 81-84% of the independent test set, and an internal cross-validation error of 22% was obtained for the training set (out of bag estimation). These results compare well with the experimental interlaboratorial reproducibility error of ca. 15% usually associated with the Ames assay. [6]

Inspection of the results reveals that the MOLMAP descriptors do not simply correspond to a code of structural fragments. The model has some ability to base predictions for unknown functional groups on the detection of reactivity sites.

REFERENCES:

1. For a revision on QSAR for predicting mutagenicity see: G. Patlewicz; R. Rodford; J. D. Walker. *Environ. Toxicol. Chem.* 2003, 22, 1885-1893.
2. <http://www2.chemie.uni-erlangen.de/software/petra>
3. V. Svetnik; A. Liaw; C. Tong; J. C. Culberson; R. P. Sheridan; B. P. Feuston. *J. Chem. Inf. Comput. Sci.* 2003, 43, 1947-1958.
4. <http://potency.berkeley.edu>
5. Downloaded from <http://www.epa.gov/nheerl/dsstox>. Version 15Oct03.
6. J. Kazius; R. McGuire; R. Bursi. *J. Med. Chem.* 2005, 48 (1), 312 -320.

P-4 : Calculation of Interaction Energies Between DNA and Fluorescent Materials by Using Molecular Orbital Calculations

Mitsuyo Aota; Yamaguchi University, Ube, JP
Kenzi Hori, Yamaguchi University

The interaction between DNA and a fluorescent material has been investigated for a long time. These studies usually use molecular dynamic (MD) or Monte Carlo (MC) simulations which adopt empirical force fields. As the programs for the simulations such as Amber, CHARMM, BOSS and so on have good parameters for DNA, they succeeded in producing dynamic features of complexes of DNA and fluorescent materials (FMs). However, it is necessary to always make parameters for each FM when new FMs are designed for a specific DNA sequence. We have to perform enormous trials calculating interaction energies for different combinations of DNA and FM. These calculations are required to make potential parameters which have to produce reasonable interaction energies for the complexes although this should be very difficult. When we use molecular orbital (MO) calculations for this purpose, no parameterization is required. In the present study, we adopted MO calculations to investigate interactions between FMs and eight DNA hexamers with sequences such as AAAAAA, TTTTTT, AAATTT, TTAAA, ATATAT, TATATA, GGGGGG and CCCCCC. The specific interaction between DNA and FMs, Hoechst33342, Hoechst33258, DB183, DB210, Netropsin were investigated. Docking of FMs into the minor groove of DNA was carried out using the BioMed CACHE program. This program was also used for optimizing geometries of the DNA complexes at the PM3 level of theory. Interaction energies at the RHF/6-31G//PM3 level of theory were also calculated and compared with those of the PM3 calculations.

P-6 : Development of the Total System ToMoCo for 3D-QSAR and Molecular Design

Masamoto Arakawa; University of Tokyo, Bunkyo-ku, JP
Kimito Funatsu, University of Tokyo

In our laboratory, methodologies for quantitative structure-activity relationships and molecular design are investigated and several related softwares have been developed. And recently, we started development of the total system ToMoCo by integrating these softwares. By using this system, various analyses in common user interface become possible and the result of analysis can be easily interpreted with computer graphics. The ToMoCo includes some useful functions such as molecular alignment method using Hopfield Neural Network, QSAR by CoMFA, region selection by Genetic Algorithm (GA) in CoMFA, automatic drug-like structure generation under restriction of QSAR model. In this conference, we will introduce these functions of the ToMoCo and some applications.

P-8 : Incorporating Conformational Flexibility into QSAR: Validation of a Novel Alignment-Independent 4D-QSAR Technique

Knut Baumann; University of Wuerzburg, Wuerzburg, DE
J. Scheiber, University of Wuerzburg
N. Stiefl, Eli Lilly

A novel molecular descriptor called xMaP (extended MaP descriptor) is introduced and validated. The descriptor is the 4D extension of the previously published alignment-independent MaP descriptor (Mapping Property distributions onto the molecular surface) [1]. In addition to MaP, xMaP is to a great extent independent of the chosen starting conformation of the encoded molecules. This is achieved by using ensembles of conformers which are generated with molecular dynamics simulations or by conformational searches. This step of the procedure is similar to Hopfinger's 4D-QSAR [2].

A five step procedure is used to compute the xMaP descriptor. First, the conformer ensemble for each molecule is generated. Next, for each of the conformers the molecular surface is computed. Third, molecular properties are projected onto this surface. Afterwards, the properties are assigned one of the following property categories: H-bond acceptor/donor, hydrophilic, weakly/strongly hydrophobic. Fourth, areas of identical properties are merged to surface patches. Finally, the distribution of the patches representing surface area size and surface properties are

converted into an alignment-independent descriptor which is based on potential 2 point pharmacophores. The latter step uses the information of the entire conformer ensemble.

To systematically study the influence of several important operational parameters, the novel descriptor was applied to several data sets. The results were compared to the original Map procedure [1] and to 4D-QSAR [2]. It turns out that xMaP is more robust than MaP. In addition to that it is an alignment independent descriptor as opposed to Hopfinger's 4D-QSAR. The results expressed as average over many test set predictions (R2Test) are quite satisfactory and range between 0.5 and 0.7. Although a huge amount of structural information is encoded, the novel descriptor remains interpretable. Data processing for the interpretation step is challenging and various strategies for this purpose will be presented.

1. N. Stiefl, K. Baumann, J. Med. Chem. 2003, 46, 1390-1407.
2. A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, C. Duraiswami. J. Am. Chem. Soc. 1997, 119, 10509-24.

P-10 : Structure-Based Predictions of ¹H NMR Chemical Shifts and Coupling Constants Using Associative Neural Networks

Yuri Binev; Bulgarian Academy of Sciences/Universidade Nova de Lisboa, Caparica, PT
João Aires-de-Sousa, Universidade Nova de Lisboa

Fast and accurate predictions of ¹H NMR spectra of organic compounds are highly desired particularly for automatic structure elucidation or validation. The large amount of compounds prepared in parallel syntheses need to be analysed and the structures of the products need to be verified. ¹H NMR plays an important role in this endeavour and the simulation of spectra to compare with the experimental spectra is of high interest.

The SPINUS program has been developed for the prediction of ¹H-NMR chemical shifts from the molecular structure. It is based on ensembles of Feed-Forward Neural Networks (FFNN), which were trained using a series of empirical proton descriptors (physicochemical, geometrical and topological). [1] The FFNNs were incorporated into Associative Neural Networks (ASNN). [2] An ASNN corrects a prediction obtained by the FFNNs with the observed errors for the k nearest neighbours in an additional memory. The additional memory consists of a list of protons and the corresponding experimental chemical shifts. The search for the k nearest neighbours is performed in the output space, i.e. they are the k protons with the most similar profile of outputs (each output comes from a different FFNN of the ensemble).

In this poster we evaluate a procedure to estimate coupling constants with the ASNN previously trained for chemical shifts. Now a memory of coupled protons and the corresponding coupling constants is built. The output profiles from the ASNNs are used for the prediction of coupling constants. To obtain a prediction for the coupling constant between two protons, the output profile is obtained for both protons, and the memory of coupling constants is searched to find the most similar pair of coupled protons. The prediction is based on the experimental values found. The web-based tool for predicting ¹H NMR chemical shifts and coupling constants and for simulating spectra will be presented.

ACKNOWLEDGEMENTS. Y.B. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for financial support under a postdoctoral grant (SFRH/BPD/7162/2001).

REFERENCES

1. Y. Binev, J. Aires-de-Sousa, "Structure-Based Predictions of ¹H NMR Chemical Shifts Using Feed-Forward Neural Networks", J. Chem. Inf. Comp. Sci. 2004, 44, 940-945.
2. Y. Binev, M. Corvo, J. Aires-de-Sousa, "The Impact of Available Experimental Data on the Prediction of ¹H NMR Chemical Shifts by Neural Networks", J. Chem. Inf. Comp. Sci. 2004, 44, 946-949.

P-12 : Optimising the Effectiveness of Similarity Measures Based on Reduced Graphs

Kristian Birchall; University of Sheffield, Sheffield, GB
Val Gillet, Stephen Pickett, and Gavin Harper, University of Sheffield

Similarity searching is widely used in an attempt to identify molecules that exhibit biological activity similar to a query molecule. Traditional approaches to similarity searching that are based on 2D descriptors tend to identify compounds that are structural analogues of the query. However, functional similarity is not limited to structural similarity, and, consequently, there is considerable interest in developing descriptors to identify compounds that share biological activity yet belong to different chemical series. Reduced Graphs (RGs) are one such descriptor. RGs summarise a molecular graph by grouping atoms into nodes based on properties that are likely to be important for bioactivity (H-bond donors/acceptors, aromatic rings etc.). Thus, they emphasise the properties of molecules in a type of topological pharmacophore. In previous work, RGs have been mapped to fingerprints and similarity has been quantified using the Tanimoto coefficient applied to the fingerprints (Gillet et al. 2003, Barker et al. 2003). The RGs were found to increase the diversity of actives retrieved in similarity searches compared to using more conventional descriptors such as Daylight fingerprints. Their performance has been improved further by combining fingerprint similarity with edit-distance similarity, in a method devised by Harper et al (2004).

The edit-distance approach to quantifying the similarity between two RGs involves extracting linear paths of nodes from the RGs and finding the minimum cost required to transform the paths from one molecule to those derived from the other, as determined using dynamic programming. There are three basic types of transformation; insertion, deletion and substitution of nodes and each transformation can be assigned a different cost according to the perceived severity of the operation with regards to the specific node types involved. For example, transforming an acyclic joint donor/acceptor node to an acyclic donor node may be given a low cost to signify the functional similarity between the two nodes, whereas transforming the same node to an aromatic ring node may be given a higher cost to reflect the lower functional similarity of the nodes. The edit-distance similarity measure thus copes naturally with insertions, deletions and substitutions when comparing two RGs, something which the fingerprint methods are not able to achieve.

In the published edit-distance method, Harper et al (2004) assigned the penalties based on intuition. Here, we describe the use of a Genetic Algorithm to evolve penalties that maximise the enrichments found in similarity searches. Results are presented for several different activity classes taken from the MDDR and they show significant improvement over the original penalties. The derived penalties also offer insights into the relative importance of features in active molecules and could provide suggestions for possible replacements of groups in the design of novel compounds.

1. Gillet, V.J., Willett, P., Bradshaw, J. (2003) Similarity Searching Using Reduced Graphs. *Journal of Chemical Information and Computer Sciences* 43, 2003, 338-345.
2. Barker, E., Gardiner, E., Gillet, V.J., Kitts, P., Morris, J. (2003) Further Development of Reduced Graphs for Identifying Bioactive Compounds, *Journal of Chemical Information and Computer Sciences* 43, 346-356.
3. Harper, G. Bravi, G.S., Pickett, S.D., Hussain, J., Green, D.V.S. (2004) The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data, *Journal of Chemical Information and Computer Sciences*, 44, 2145-2156.

P-14 : Generation of a Focussed Set of GSK Compounds Biased Towards Ligand-Gated Ion Channel Ligands

Anna Maria Capelli; GlaxoSmithKline, Verona, IT
Aldo Feriani, Giovanna Tedesco, and Alfonso Pozzan, GlaxoSmithKline

Several "datamining" methodologies have been recently developed to bias compound selection and library design for generic therapeutics targets (i.e. antimicrobials, anticancer agents etc.) in order to improve the effectiveness of high-throughput screening in the discovery of novel leads [1]. Among them, substructural analysis has been reported as a methodology that allows the identification of structure-activity relationships of large and disparate data sets, characterised by qualitative and quantitative activity [2]. A "datamining" methodology based on substructural

analysis and standard 1024 Daylight fingerprint as descriptors, successfully applied previously both to antibacterials [3] and 7TM ligands [4] was applied to a set of known antagonists of a sub-family of ligand-gated ion channels comprising nAChRs, 5-HT₃, GABAA and GlyR receptors [5]. The derived scoring function was used to generate a focussed set that was screened for alpha7 nAChR, resulting in the identification of novel and chemically tractable alpha7 ligands. Finally, the same scoring function was applied retrospectively to other in house sets screened for the same target in the same assay. Results and performance of the method are presented in details.

1. a) Gillet V.J., Willet P., Bradshaw J., *J. Chem. Inform. Com. Sci.*, 1998, 38, 165-179; b) Ajay, Walters W.P., Murcko M.A., *J. Med. Chem.*, 1998, 41, 3314-3324; c) Sadowski J., Kubinyi H., *J. Med. Chem.*, 1998, 41, 3325-3329; d) Ghose A.K., Viswanadhan V.N., Wendelowski J.J., *J. Comb. Chem.*, 1999, 1, 55-67; e) Harper G., Bradshaw J., Gittins J.C., Green D.V., Leach A.R. *J. Chem. Inf. Comput. Sci.* 2001, 41, 1295-1300
2. a) Hert J., Willet P., Wilton D.J., *J. Chem. Inf. Comput. Sci.* 2004, 44, 1177-85. b) Ormerod A. et al. *Quant. Struct.-Act. Relat.* 1989, 8, 115.; c) Ormerod A. et al. *Quant. Struct.-Act. Relat.*, 1990, 9, 302.
3. Feriani A., Pozzan A., Capelli A. and Tedesco G., XVIth International Symposium on Medicinal Chemistry, September 18-22, 2000, Bologna, Italy, P19.
4. Tedesco G., Feriani A., Pozzan A., Capelli A.M., EuroMUG2002, September 24-26 2002, Cambridge, UK
5. Le Novere N., Changeux, J.-P., *Nucleic Acids Research*, 1999, 27(1), 340-2. See also <http://www.pasteur.fr/units/neubiomol/LGIC.html>

P-16 : QSPR Study of Melting Point and Density of Imidazolium Ionic Liquids

Gonçalo Carrera; Universidade Nova de Lisboa, Caparica, PT
 Carlos M. Afonso, Universidade Técnica de Lisboa
 João Aires-de-Sousa, Universidade Nova de Lisboa

Ionic liquids (IL) are salts with melting points near the room temperature. Their negligible vapour pressures allow for their potential use as environmentally friendly substitutes of organic volatile solvents [1]. Judicious choice of anion and cation permits to obtain IL's with physical and chemical properties fitted to a specific problem [2]. The first decisive property is the melting point.

Others [3,4] and we [5] have reported QSPR analysis of the melting point of ionic liquids. These works have considered datasets of bromide salts. Here we present QSPR models of density and melting points, which are based on both cationic and anionic descriptors accounting for the diversity of both the cation and the anion of the salts. Random Forests [6] were used for regressions using a pool of near 300 descriptors.

For modeling the melting point we used a dataset of 235 imidazolium salts with mp between -88 and 370 Å°C, and including six different anions - BF₄⁻, Cl⁻, Br⁻, PF₆⁻, CF₃SO₃⁻, and NTf₂⁻. The dataset was divided into a training set with 155 objects and a test set with 80 objects. For the QSPR study of density, a dataset of 106 imidazolium salts was collected from the literature, with density ranging from 0.96 to 2.80, and covering five families of anions - BF₄⁻, PF₆⁻, NTf₂⁻, CF₃SO₃⁻, CF₃CO₂⁻ and CH₃CH(OH)CO₂⁻. This dataset was partitioned into a 73-objects training set and a 33-objects test set.

Three types of cationic descriptors were used based on 3D molecular structures generated by CORINA [7]: radial distribution function vector, surface spatial autocorrelation function vector and a set of empirical general molecular descriptors such as the molecular weight, maximum charge, or polarizabilities. Several descriptors were defined for the anion: binary descriptors each encoding a specific anion, descriptors based on the molecular weight of the anion, and descriptors based on the miscibility of an anionic family in different solvents. The miscibility descriptors assume that miscibility is mainly determined by the anion of the IL [8]; these descriptors are calculated from the proportion of salts belonging to a specific anionic family that are miscible in a certain solvent.

For the melting point, good correlations were obtained between the experimental and the predicted values for the test set ($r^2 = 0.80$, RMSE = 34 Å°C). Excellent predictions were obtained for the density ($r^2 = 0.98$ for the test set).

ACKNOWLEDGEMENTS

G.C. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for financial support under a PhD grant (SFRH/BD/18354/2004).

REFERENCES

1. J. S. Wilkes; *Green Chemistry*; 2002; 4; 73.
2. R. Sheldon; *Chem. Commun.*; 2001; 2399.
3. A. R. Katritzky, A. Lomaka, R. Petrukhin, R. Jain, M. Karelson, A. E. Visser, R. D. Rogers; *J. Chem. Inf. Comput. Sci.*; 2002; 42; 71.
4. D. M. Eike, J. F. Brennecke, E. J. Maginn; *Green Chemistry*; 2003; 5; 323.
5. G. Carrera, J. Aires-de-Sousa; *Green Chemistry*; 2005; 7; 20.
6. L. Breiman.; *Machine Learning*; 2001; 45; 5.
7. J. Gasteiger, C. Rudolph, J. Sadowski; *Tetrahedron Comput. Methodol.*; 1992; 3; 537.
8. C. Chiappe, D. Pieraccini; *Journal of Physical Organic Chemistry*; (early view).

P-18 : New Descriptors from Energy Decomposition in Semiempirical Level

Alexandre Carvalho; Universidade do Porto, Porto, PT
 André Melo, Universidade do Porto

In this work, we used the partition method introduced by Carvalho and Melo [1] which enable the decomposition stabilization energies of molecular association processes into physical meaningful components (conformational rearrangement, non-bonding, bonding and polarization plus charge transfer). This partition scheme has been developed within a semi-empirical formalism, which enables a complete separability of the above-mentioned components. We have study the complex between Cucurbita Maxima trypsin inhibitor (CMTI-I) and glycerol. Every residue was considered a fragment. This computational procedure enables to evaluate the range of the perturbation originated by the association process and evaluate the energetic contribution from each residue. The results obtained enable us to conclude that the present decomposition scheme can be used for understanding the cohesive phenomena and produces a new set of descriptors.

1. Alexandre R. F. Carvalho, André Melo, "Energy partitioning in association processes", *Int. J. Quantum Chem.* (2005).

P-20 : Quantitative Analysis by Spectral Data Transformation in Multivalued Fingerprints and Multivariate Calibration

Gonzalo Cerruela; University of Córdoba, Córdoba, ES
 Manuel Urbano Cuadrado, María Dolores Luque de Castro, and Miguel Ángel Gómez-Nieto, University of Córdoba

Spectroscopic techniques employing multichannel detection have become a powerful tool for the characterization of materials. A number of qualitative and quantitative approaches based on the collection of the spectrum (a large data set) in a short time and the subsequent multivariate treatment have been developed aimed at substituting time-consuming and expensive methods. The research carried out in this study is an attempt on improving multivariate calibration based on a new chemometric technique for the quantitative property prediction of samples through preprocessing of spectral data and their transformation in multivalued fingerprint. The transformation of a spectrum in a multivalued fingerprint involves the following steps:

1. Normalization of the spectral data by standard, logarithmic and maximum methods in order to transform the absorbance matrix into a new data set within the [0,1] range.
2. Selection of n-1 threshold values taking into account the maximum-minimum range, where n is the number of cases per variable.
3. Assignment of a given case to each variable if the normalized value surpasses the threshold value.
4. Different multivalued transformations of the input spectra (i.e. binary, ternary, quaternary and quintal valued transformation) have been studied and the results compared with each other.

The work presented here deals with the study of statistic parameters of PLSR equations built with discrete spectra as data matrix aimed at enlarging spectral differences. The parameters employed for this study were the Determination

Coefficient (R^2), Standard Error in Cross Validation (SECV), bias and slope. The results were compared with those obtained by authors without the proposed preprocessing (In Comparison and Joint Use of Near Infrared Spectroscopy and Fourier Transform Mid Infrared Spectroscopy for the Determination of Wine Parameters. *Talanta*. Accepted for publication. Available on-line).

Data employed corresponded to the 3000-800 cm^{-1} absorbance spectra of 136 samples of wine. Each spectrum consisted of 1142 predictor variables. The properties studied were total acidity and the content of reducing sugars, which were determined by titration to obtain the reference values. Model fitting was carried out by cross validation - six series in which the training and validation sets were composed by 116 and 20 samples, respectively, in such way that all the samples were used for validation.

The maximum normalization showed the best statistic parameters. Regarding the dimension, the ternary spectra drove to the best determinations. All the statistic parameters were improved using the proposed preprocessing with the exception of R^2 for total acidity. Bias was the most improved parameter, close to zero for the two properties. The software employed for spectra normalization and building the discrete spectra was developed by the authors in the C programming language. The Unscrambler 7.8 (Camo Process AS, Oslo, Norway) was used for developing PLSR equations.

P-22 : Study and Display of the Effect of Structural Similarity Approach in the Screening of Chemical Databases

Gonzalo Cerruela; University of Córdoba, Córdoba, ES

Manuel Urbano Cuadrado, Miguel Ángel Gómez-Nieto, and Irene Luque Ruiz, University of Córdoba

Similar structures have similar properties is a fact well known and widely accepted by the scientific community. Molecular similarity can be assessed in conceptually different ways including a variety of algorithms, metrics and high-level description of molecular structure, properties and conformation. Usually, molecules are represented by means of molecular graphs and then a structural similarity measure can be obtained. This measure is generally based on considering the set of common subgraphs (MCS (Maximum Common Edge Subgraphs)) and to apply some of the well-known similarity metrics (Tanimoto, Cosine, Simpson, etc.) which consider the sizes (nodes and edges) of the molecular graphs that are compared and the size of the common subgraphs.

The use of the measure of structural similarity has been applied to the prediction of properties, clustering, and screening of chemical databases. However, when the approach of structural similarity is used in the screening processes, it is observed that the size of the molecules and the utilized index of similarity affect considerably the number of recovered molecules for a given threshold of similarity.

In this work we present a study of the behavior of some of the structural similarity metrics as a function of the characteristics of the molecules and the approach or algorithm used in the calculation of the structural similarity. We carry out an analysis of the calculation of the structural similarity based on the MCS (Maximum Common Subgraph) with regard to the MCEs (generally utilized) and we observe the dependence that exists among the different similarity indexes as a function of parameters as: graph size (number of nodes and edges), size of the common subgraph, relationship among the sizes of the MCS and MCEs, etc.

The results obtained show information on the thresholds of similarity as a function of the different similarity indexes and the structural characteristics of the set of recovered molecules that should be used in the processes of screening of chemical databases. The relationships obtained allow us to establish the maximum and minimum threshold values for different similarity indexes in the screening process with the purpose of recovering molecules that: a) only contain a complete substructure equal to the search criteria, b) contain substructures that exist partially (as connected or non-connected to each other), and c) to delimit the size or relationship between the MCS and MCEs among the search criteria structure and the recovered molecules.

P-24 : Generation of Multiple Pharmacophore Hypotheses Using a Multiobjective Genetic Algorithm

Simon Cottrell; University of Sheffield, Sheffield, GB
 Valerie J. Gillet, University of Sheffield
 Robin Taylor, Cambridge Crystallographic Data Centre

Pharmacophore elucidation involves identifying a three-dimensional arrangement of features that is common to a set of ligands with the same biological activity. This normally involves superimposing the ligands such that functional groups relevant to biological activity are overlaid. Since most drug-like molecules are conformationally flexible, the conformational space of each ligand must be searched.

Existing methods for pharmacophore elucidation generally fall into one of two categories. Firstly, there are methods that attempt to exhaustively generate all pharmacophore hypotheses from pre-generated sets of conformers. Secondly, there are methods that return a single solution that optimises a scoring function, which is usually an arbitrary combination of several objectives. These methods usually search the conformational space dynamically during the overlay, so as to find the set of conformations that produces the best overlay. The former methods generally return a large number of solutions which take considerable effort to analyse and are highly dependent on the method used to generate the conformers, whilst the single solution returned by the latter methods implies an unrealistic degree of certainty in the result.

This work has involved applying a multiobjective genetic algorithm (MOGA) to the pharmacophore elucidation problem [1]. The MOGA evaluates solutions quantitatively; however, it does not score solutions using a single fitness function but considers each objective independently, according to the principles of Pareto dominance. Three objectives are considered in evaluating hypotheses, namely the closeness of the alignment of features in the different ligands, the volume overlap of the ligands and the internal energy of the ligands. The program generates several different hypotheses which represent different, but equally valid compromises between the objectives.

An important aim of this work has been to generate a set of solutions that are diverse from a biochemical point of view. Ensuring a diverse range of different compromises between the three objectives has proved to be a necessary but not sufficient condition for achieving this. Considerable effort has therefore been directed towards explicitly taking into account chemical diversity within the MOGA population. A recent development of the program has been to allow the identification of pharmacophore features that are common to some, but not all of the ligands.

The results of the MOGA are illustrated using datasets for binding sites of pharmaceutical interest. In each case, the MOGA generates a manageable number of different hypotheses. Thus, it takes a realistic view of the uncertainty which is inherent when the binding site structure is unknown, but still allows a quantitative comparison of the hypotheses that it generates.

1. Cottrell, S.J.; Gillet, V.J.; Taylor, R.; Wilton, D.J. Generation of Multiple Pharmacophore Hypotheses. *Journal of Computer-Aided Molecular Design*, in press.

P-26 : Advanced Structural Search Using ChemAxon Tools

Ferenc Csizmadia; ChemAxon, Budapest, HU
 Gyorgy Pirok, Szilard Dorant, Miklos Vargyas, Peter Kovacs, Nora Mate, and Szabolcs Csepregi, ChemAxon

Structural search techniques are invaluable tools in all cheminformatics systems including but not limited to rational drug design, compound registration systems and laboratory information management systems.

JChem, one of ChemAxon's major suites of programs, provides a very rich set of features related to structural search. These features are demonstrated by examples. Covered topics are: substructure, exact, superstructure, MCS (maximum common substructure) and similarity search.

Reaction and R-group search (including R-logic) are also available, which are complemented by a rich set of query features. SMARTS and query features of the MDL formats are supported. An example of a fast MCS-based

clustering is also presented. Finally the recently developed descriptive Chemical Terms Language is demonstrated by powerful structural searches.

P-28 : Neural Networks for the Prediction of ¹H NMR Chemical Shifts of Sesquiterpene Lactones

Fernando Da Costa; Universitaet Erlangen-Nuernberg, Erlangen, DE

Y. Binev and J. Aires-de-Sousa, Universitaet Erlangen-Nuernberg

J. Gasteiger, Universitaet Erlangen-Nuernberg

The sesquiterpene lactones (STLs) comprise a large group of natural products, which are basically found in plant species of the family Asteraceae. They have ecological importance, show several biological activities and are taxonomic markers of this family [1]. ¹H NMR spectroscopy plays a crucial role in the structure elucidation of STLs in all the studies concerning this class of compounds. This work describes the estimation of ¹H NMR chemical shifts of STLs using the SPINUS program [2], which is based on associative neural networks (ASNN). This system incorporates an ensemble of backpropagation neural networks (previously trained with a general set of structures and the corresponding chemical shifts) and an additional user-defined memory [3,4]. Several physicochemical, geometric and topological descriptors were used to represent the hydrogen atoms of the compounds. In a previous work, 392 ¹H NMR experimental chemical shifts of 20 STLs were used as additional memory to predict the chemical shifts of two different STLs. The results showed a high level of accuracy [5].

In the present work, the prediction of 1,902 chemical shifts from 100 structures of STLs was made using the same additional memory described above. The methylenic sp³ diastereotopic protons belonging to rigid substructures, as well as sp² protons of C-C double bonds of 3D structures, can be distinguished by the geometrical descriptors. The inclusion of the user-defined additional memory led to a considerable improvement in the accuracy of the predictions. When no memory was used, the mean absolute error (MAE) for the predictions of the ¹H NMR chemical shifts of the 100 STLs was 0.27 ppm. Using only the user-defined additional memory (392 ¹H NMR chemical shifts of 20 STLs), the MAE was improved to 0.23 ppm. When the user-defined memory was combined with a large general memory, a MAE of 0.22 ppm was achieved.

REFERENCES:

1. F.C. Seaman. Bot. Rev. 48:123-551, 1982.
2. A web interface of SPINUS is freely accessible at <http://www.dq.fct.unl.pt/spinus> and <http://www2.chemie.uni-erlangen.de/services/spinus>
3. Y. Binev, J. Aires-de-Sousa. J. Chem. Inf. Comput. Sci. 44:940-945, 2004.
4. Y. Binev, M. Corvo, J. Aires-de-Sousa. J. Chem. Inf. Comput. Sci. 44:946-949, 2004.
5. F.B. Da Costa, Y. Binev, J. Gasteiger, J. Aires-de-Sousa. Tetrah. Lett. 45:6931-6935, 2004.

ACKNOWLEDGMENTS:

F.B.C. acknowledges Alexander von Humboldt-Stiftung. Y.B. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for a post-doctoral grant under the POCTI program (SFRH/BPD/7162/2001). J.A.S. thanks Deutscher Akademischer Austauschdienst (DAAD) for travel grants.

P-30 : Indexing the Chemical Semantic Web

Nick Day; Cambridge University, Cambridge, GB

Peter Murray-Rust, Cambridge University

The semantic web is a vision where information can be retrieved and analysed robotically using agreed metadata and indexes. Molecular information is ideally suited for this and we report the use of the recently developed InChI (International Chemical Identifier) (to be released very shortly). InChI is an IUPAC project whose aim is to create a non-proprietary unique identifier for chemical structures to enable easier linking of diverse electronic data

compilations. We have recently shown that the InChI is a powerful and precise tool for indexing chemical structures both in databases and on the web.

We have shown that the structure and information held in an InChI confer several advantages over current approaches. It provides layers describing degree of certainty in our knowledge of chemical identity which can allow variable precision in the queries. Our analysis of web indexes show that it provides a robust query in today's search engines. Documents combining Chemical Markup Language with InChIs have both high recall and high precision making them the natural choice for publishing and hence building the Chemical Semantic Web. We have implemented a variety of web services for generating and using InChI.

We thank Steve Stein, Steve Heller, Dmitrii Tchekhovskoi (NIST) and Alan McNaught (IUPAC) for help and advice on InChI.

P-32 : VET: A Tool for Reaction Plausibility Checking

Joseph Durant; Elsevier MDL, San Leandro, CA, US
James G. Nourse, Elsevier MDL

Construction of reaction databases can involve the extraction of chemical reactions from textual descriptions. As part of this process chemical names are converted to structures, reaction roles are assigned to components, and "obvious" facts are supplied. Errors can occur in each of these activities, leading to errors in the extracted reactions. Trapping such errors in "real world" reaction databases is a complex, but important, task.

One problem is that the study of chemical reactivity lacks a simple and systematic way to partition possible reactions into plausible and implausible sets. Instead, one is presented with a wealth of rules and examples which make construction of an expert system for reaction planning a challenging endeavor.

Further complicating the task are the properties of real world data. For example, it is common to represent a reaction as an unbalanced reaction, where one or more products (the "uninteresting ones") are not represented. In this way the reaction representation can highlight the important features of the reactions. However, the increased clarity resulting from this flexibility in reaction representation introduces ambiguity into the interpretation of the reaction, and complicate the evaluation of its plausibility.

In order to support creation of new reaction databases we have created a program, VET, to evaluate the plausibility of chemical reactions to be included in the database. This method focuses on trapping common classes of input and representation errors and minimizing the occurrence of incorrect reactions allowed into the database. VET is presently in use at Elsevier MDL as part of the Patent Chemistry database creation workflow.

We will discuss the various strategies developed, as well as their relative strengths and weaknesses.

P-34 : Accurate Geometry Optimization Method for Molecular Mechanics

Ödön Farkas; Eötvös Loránd University, Budapest, HU

The present method [1] allows the efficient utilization of the full power of Newton optimization techniques in molecular mechanics. The examples show the weakness of the currently used optimization methods as they usually provide much higher energy local minima, started from the same structure, due to the lack of their accuracy. The proposed algorithm with applying multiple optimization criteria, which are regularly used in quantum chemistry [2], can result final RMS forces less than 10^{-6} kcal/mol/Å, a previously impossible goal for biomolecules. The accurate geometry optimization is also essential as part of QM/MM, ONIOM [3,4] procedures.

1. Farkas, Ö "Fast and robust geometry optimization algorithm for large systems", CESTC 2004, Tihany, Hungary
2. Gaussian 98, Revision A.7, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, Å-. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi,

- B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, Gaussian, Inc., Pittsburgh PA, 1998.
3. Torrent, M.; Vreven, T.; Musaev, D. G.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B. "Effects of the protein environment on the structure and energetics of active sites of metalloenzymes. ONIOM study of methane monooxygenase and ribonucleotide reductase", *Journal of the American Chemical Society* 2002, 124, 192-193.
 4. Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. "Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints", *Journal of Computational Chemistry* 2003, 24, 760-769.

P-36 : "Ultra-fast" Ligand-based de novo Design Using Virtual Reaction Schemes

Uli Fechner; Goethe-Universität Frankfurt, Frankfurt, DE
Gisbert Schneider, Goethe-Universität Frankfurt

We developed a software tool that is capable of performing a virtual retro-synthesis of compounds following the RECAP [1]. The employed set of eleven common organic reactions was specified in the SMIRKS language. The virtual retro-synthesis was carried out on a dataset of approximately 5000 drug molecules [2]. The SMILES strings of the resultant fragments were labeled to allow for the storage of the position where the reaction took place and the type of the reaction. These fragments were then used as building blocks in our ligand-based de novo design program Flux (Fragment-based ligand builder reactions) which is grounded on the TOPAS method [3]. The same set of virtual reaction schemes guided the assembly of candidate compounds thereby leading to an increased chance of designing structures that are synthetically accessible. Molecular similarity between a known active compound for a particular biological target (template) and the candidate compounds served as a scoring function. The molecular similarity was calculated using Daylight Fingerprints and the Ghose & Crippen substructure fingerprints as descriptors and the Tanimoto Index and the Euclidian Distance as similarity indices. The ligand-based scoring function facilitates the application of our design approach where the three-dimensional structure of the biological target is not available, as is the case, for example, with the large group of G-protein coupled receptors. An evolutionary algorithm with a specifically tailored mutation operator was accountable for navigation through the fitness landscape. Both the program for virtual retro-synthesis and our de novo design software extensively rely on the Daylight Toolkit (<http://www.daylight.com/>).

We evaluated our method with two retrospective design studies: Gleevec, a Abelson tyrosine kinase inhibitor, and a Factor Xa inhibitor synthesized with the four-component UGI reaction served as molecular templates. In case of Gleevec our program was able to re-assemble the template structure with our set of building blocks and virtual reaction schemes. For both molecular templates Flux proposed several candidate compounds with interesting chemical moieties. Visual inspection supported our hypothesis that the likelihood of chemical accessibility indeed is increased with our design approach.

References:

1. Lewell, X.O., Budd, D.B., Watson, S.P. & Hann, M.M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry, *J. Chem. Inf. Comput. Sci.* 1998, 38, 511-522.
2. Schneider, P. & Schneider, G. Collection of Bioactive Reference Compounds for Focused Library Design, *QSAR Comb. Sci.* 2003, 22, 713-718.
3. Schneider, G., Lee, M.-L., Stahl, M. & Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.* 2000, 14, 487-494.

P-38 : The Molecule Evuator: A Computer-based Tool for Drug Design

Eric-Wubbo Lameijer, Leiden/Amsterdam Center for Drug Research, Universiteit Leiden, Leiden, NL
Thomas Bäck and Joost Kok, Leiden Institute for Advanced Computer Sciences, Universiteit Leiden
Ad P. IJzerman, Leiden/Amsterdam Center for Drug Research, Universiteit Leiden

Computers are nowadays very common in the laboratory. However, their tasks are usually restricted to instrument control or generic functions such as word processing. Chemistry-specific applications are much less frequently used, mainly to draw molecules or to find compounds in databases.

However, it should be possible to get more out of the memory and processing speed of our machines. We have created a computer program (Molecule EvuatorTM) that does not only allow the user to manipulate molecules but can also manipulate molecules itself. The evolutionary algorithm incorporated in the software allows it to create populations of molecules, either new or derived from one or more lead structures. It can also mutate and cross these molecules into novel derivatives. The chemist, meanwhile, can select those molecules which might be synthesized or be biologically active/interesting. These molecules will then be used as "parents" for the next generation of derivatives.

To help the chemist direct the evolution, we have also implemented various other options, such as allowing the user to select parts of the molecule that have to remain constant, calculating physicochemical properties and using chemical filters to automatically discard compounds with undesirable chemical substructures.

A program such as this, which combines the speed and calculation abilities of the computer with the knowledge and imagination of the individual chemist, will improve upon the purely passive "drawing programs". We therefore hope that this interactivity between man and computer will lead to interesting novel ideas for drug-like compounds.

P-40 : ScafReplace: Novel Tools for Scaffold Replacement

Patrick Fricker; Center for Bioinformatics, Hamburg, DE
Tanja Schulz-Gasch, Martin Stahl and Matthias Rarey, Center for Bioinformatics

A frequently occurring task during lead finding and optimization is to replace a central element (linker or ring system) of a compound series. Based on a geometric arrangement of 2-3 exit vectors and additional pharmacophoric features, the task is to find molecular fragments fulfilling these constraints. Here we present a novel approach for this task based on the concepts of 3D-shredding and geometric rank searching.

In a first step, a database of fragments suited for scaffold replacements is derived from a database of 3D structures like the Cambridge Crystallographic Structure Database (CSD). A set of disconnection rules is used to split the molecular structures into molecular fragments by "cutting" the molecules at strategic acyclic bonds.

In order to avoid strained conformations for the suggested replacements, the 3D structural information of the molecules is retained. We avoid the explicit creation of fragments. Instead, we consider all connected fragments that result from any possible combination of cuts within the original compound database. In this way, the need to join fragments decreases and therefore the conformational information in the database is fully exploited.

To discard unwanted fragments on the topological level, filters are used to mark particular combinations of cuts as unwanted. These filters include fragment size, distances between cuts, number of cuts and certain substructure patterns.

Based on this new idea of 3D-shredding, a search method was developed. The method in its first version does not try to combine fragments, but searches for single fragments fulfilling "feature points" from a query. A query consists of two exit-vectors and an arbitrary number of potentially directed features, for example hydrogen-bond donors and acceptors or hydrophobic interactions.

By design, the combinatorial search algorithm finds matching fragments ordered by deviation from query features. This leads to the property that the user does not have to give tolerance ranges for the query features but can specify the number of results to be returned or can get hits incrementally.

We applied the 3D-shredding and rank search algorithm to a subset of 90000 structures from the CSD. The 3D-shredding routine results in 120000 potential fragments with more than 2 cuts. The construction of the database typically takes 2 minutes, a query of the type mentioned above can be processed on this database within 6 seconds for creating the first 100 hits.

P-42 : Genomic Data Analysis Using DNA Structure

Eleanor Gardiner; University of Sheffield, Sheffield, GB
Christopher Hunter and Peter Willett, University of Sheffield

Only 1-2% of the DNA of the human genome codes for proteins. Much of the remainder may be 'junk', but comparative genomics suggests that a significant amount must serve some purpose. For example, recent comparisons between the mouse and human genomes found that 5% of the genome is conserved between the two species. This means that in addition to the protein-coding regions about 3% of the genome is likely to be under evolutionary selection for some as yet unknown function. Recent comparative studies have revealed sets of Conserved Non-Genic sequences (CNGs) and sets of Ultra Conserved Elements (UCEs). CNGs are hundreds of bases long and are much more highly conserved than protein-coding genes: more than a quarter of the CNGs on human chromosome 21 have been found in at least ten other species. UCEs are also hundreds of bases long and are absolutely conserved, between human and mouse, without gaps. Over half of the UCEs identified have no overlap with any exonic sequences. The extraordinarily high degree of conservation of these sequences strongly suggests their significance and emphasises the need to develop new approaches to understanding the function of non-coding sequences. One hypothesis is that their function could be related to protein binding, as part of a system of DNA repair, gene expression, replication, packaging or scaffold attachment. One factor that governs protein-DNA interactions and is expressed over length scales of hundreds of bases is structural properties of the DNA. Thus the identification of common DNA structural motifs could allow annotation of the functional properties of non-coding parts of the genome and provide insights into likely protein partners.

We have previously compiled a database of the structural properties of all 32,896 unique DNA octamer sequences, including information on stability, the minimum energy conformation and flexibility. We have used Fourier techniques to analyse the UCEs and CNGs in terms of their octamer structural properties, in order to reveal long-range structural correlations which may indicate possible functions for some of these sequences.

P-44 : Increasing the Efficiency of Chemical Structure Storage and Retrieval in Large Relational Databases

Sasha Gurke; Knovel Corp., Norwich, NY, US
Sergei Trepalin, Institute of Physiologically Active Compounds

Chemical structure search was integrated with full text and fielded text and numeric searches in a large relational database providing consistent web-based access to a variety of technical content, from e-books to property databases.

In the exact chemical structure search, exact molecular weight was used as an index. Molecular weight unambiguously defines molecular formula of an organic molecule with molecular weight less than 1,000. In addition, two molecular topology sensitive indices were calculated. The 12-byte numbers thus obtained were sorted in descending order and searched using bisection algorithm. Tautomeric structures were converted to canonical format.

In substructure search, 256 screens were used for the initial record filtration. The sorting of the atoms and bonds of molecular fragments was same as that of the whole structures. A molecular formula search (the most efficient for fragments containing non-C atoms) was then run to eliminate irrelevant structures. Thereafter, the atoms and bonds of the fragment were matched sequentially to structures from the pre-filtered records. The results of each attempted match were recorded in a pair of Boolean matrices with the dimensions number of query atoms (bonds) x number of structure atoms (bonds). The matrices were populated with TRUE elements for a match and FALSE elements for a non-match. As bonds were populated, a reference was made to the atoms matrix to ensure that TRUE bonds in fact join atoms of the correct type. If the matrices contain at least one TRUE element for each atom and bond of the

fragment, a back-tracking algorithm was used to determine if it maps to the structure. This algorithm was applied starting with a maximally coordinated non-C atom, providing it was present. This is a critical function in terms of its performance. Although the time required increases with the number of nodes (atoms and bonds) in the fragment (exponential order), the time required to manipulate the Boolean matrices is proportional to their size (polynomial order) and this is a considerable advantage when back-tracking to verify a possible match.

MS SQL server was used to handle the database. Exact structure search algorithm was the standard SELECT command in an SQL expression. Substructure search was created as a stored procedure, COM object being used for subgraph isomorphism search.

P-46 : Analysis of GRID Molecular Interaction Fields

Sandra Handschuh; Boehringer Ingelheim Pharma GmbH, Biberach, DE

Anne Techau Jørgensen, Kerstin Höhfeld, and Thomas Fox, Boehringer Ingelheim Pharma GmbH

The calculation of molecular interaction fields using the program GRID[1] is an important technique in structure-based drug design. These fields can be used to identify favourable interaction hot spots, and thus support the manual design and optimisation of ligands for a given target. The molecular interaction fields also provide the basis for a range of methods, amongst others they have been used to identify regions important to ligand selectivity between various targets (GRID/CPCA) and for the calculation of *in silico* ADME descriptors (Volsurf).

The CCDC/Astex data set [2], a large publicly available data set of protein-ligand complexes, was used for a systematic analysis of the GRID-generated molecular interaction fields. Following protein classification based on the CATH system [3] comparisons were performed within and between protein classes.

All complexes were characterized by the molecular interaction fields for 10 GRID probes. Then the MIFs were described using a range of parameters, such as lowest energy found within the active site, and the percentage of grid points below specified energy levels. Furthermore, the presence of MIF hot spots around relevant functional groups was probed. Several aspects of the analysis are presented.

1. Goodford, P. J. J. *Med. Chem.* 1985, 28, 849-857.
2. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein-ligand interaction. *Proteins* 2002;49:457-471.
3. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM CATHA Hierarchic Classification of Protein Domain Structures. *Structure* 1997;5:1093-1108.

P-48 : Structural DNA Profiles

Linda Hirons; University of Sheffield, Sheffield, GB

E J Gardiner, C A Hunter, and P Willett, University of Sheffield

A DNA sequence's function is commonly predicted by measuring its nucleotide similarity to known functional sets. However the use of structural properties to identify patterns within families is justified by the discovery that many very different sequences have similar structural properties. This means that by looking at the information hidden within the structure, similarities between DNA sequences will be found that would otherwise be unrecognised.

A database containing structural properties of all 32,896 unique DNA octamer sequences has previously been constructed. The calculated descriptors include the step parameters that collectively describe the energy minima conformations, the force constants and partition coefficients that describe the flexibility of an octamer and several ground state properties such as the RMSD, which measures the straightness of an octamer's path.

The development of tools that use the contents of the octamer database to identify structural DNA activity fingerprints would be of great value in predicting unknown DNA functions. This is illustrated here by the generation of structural profiles that examine patterns common to a set of pre-aligned promoter sequences. A promoter sequence being one to which RNA-polymerase II binds before travelling downstream to transcribe a gene into messenger RNA.

P-50 : The Study of Bias Fusion of Chemical Similarity Searching

John Holliday; University of Sheffield, Sheffield, GB
Jenny Chen, University of Sheffield
John Bradshaw, Daylight Chemical Information Systems Inc.

This study has carried out experiments on bias fusion of chemical similarity searching based on four coefficients. The four coefficients were identified from set of 13, in a previous study using Naïve Bayes Classifiers, as having the highest retrieval rates in a selection of 20 active size ranges. The study indicated that Forbes and Simple Matching are the best at retrieving smaller size of actives, Tanimoto is the best at retrieving the medium size of actives and Russell-Rao is the best at retrieving the larger size of active. The purpose of this study was to find out whether retrieval performance could be improved by altering the weightings of these four coefficients in the fusion process.

A systematic approach was used to explore all possible combinations of the weights. Previous studies indicated that the choice of coefficients is class-dependent. Therefore, ten classes have been trained using the systematic approach resulting in one best weighted combination for each class. A similar methodology was also applied using the modal fingerprint of each training set rather than the full set of fingerprints.

The best combinations of weightings resulted from both pure systematic and modal fingerprints based systematic approaches were then tested. The measure of retrieval used was the number of the retrieved actives in the top 500 nearest neighbours. These were compared with results using the 13 single coefficients.

The results show that equal-weighting fusion has an average of improvement rate of 19% over Tanimoto, bias fusion using the results of pure systematic approach has an average of improvement rate of 25% and bias fusion using the results of modal fingerprints based systematic approach has an average of improvement rate of 32%.

In a separate experiment, a genetic algorithm has been used to generate class-dependent formulae for similarity searches. The purpose of this study is to find out whether an effective formula can be determined for each active class.

P-52 : A System Fusing Computational and Information Chemistry for Developing New Synthesis Routes of Compounds: An Application to the Synthesis Routes of Tropinone

Kenzi Hori; Yamaguchi University, Ube, JP

Synthesis route design systems have been practically used for more than ten years to create new synthesis routes of compounds. The number of routes diverges for multi-step syntheses as the systems usually offer several routes for each step. It is very difficult for experimental chemists to determine which is the best route for the target compound in the created routes. Quantum mechanical calculations including searches of transition states (TSs) are very effective to clarify possibility of synthesis routes, i.e., if there is the TS for a route, it is possible to synthesize the target by using the route and vice versa. Therefore, we should be able to find useful synthesis routes without experiments by use of the method fusing computational chemistry and information chemistry. The former analyzes reaction mechanisms of synthesis routes which the latter creates. However, there are few studies concerning with the promising method. We have been developing a data based named the transition state data base (TSDB) which makes it possible to effectively use the synthesis route system for developing new synthesis routes of compounds. The present study describes how to use the data base for developing new synthesis routes. As an example, we will show the results for synthesis routes of tropinone from the KOSP program using the DFT calculations at the B3LYP/6-31G* level of theory.

P-54 : Universal Scripted Chemical Information Processing: the CACTVS Chemoinformatics Toolkit

Wolf Ihlenfeldt; Xemistry GmbH, Lahntal, DE

While there are many established systems for the handling and manipulation of chemistry-specific data in standardized and structured environments, the researcher often encounters non-trivial problems when ad-hoc needs in the fields of data preparation for import or export, data filtering, or structure and reaction manipulation develop.

The CACTVS Chemoinformatics Toolkit was designed to provide solutions for above scenarios by means of extensive VHLL-scripting functionality with extensible, chemistry-specific high-level objects. Many typical problems encountered in the chemical information processing environment can be solved with just a few lines of script code. We will display a sample of prototypical solutions, with a special focus on Web-related applications.

P-56 : 3D Structure Prediction and Conformational Analysis

Gabor Imre; Eotvos Lorand University, Budapest, HU
Odon Farkas, Eotvos Lorand University

Numerous theoretical methods in the field of computational chemistry falls back on the availability of 3D structural information about compounds. Determining molecular structures without human interaction is an essential component of several techniques, like QSAR, 3D pharmacophore analysis, reaction prediction, etc. Current computational tools used for structure determination including force-fields and quantum chemical methods, even require a complete set of initial 3D coordinates. The efficiency of 3D structure based HTS tools also can be enhanced by employing conformational analysis to yield multiple valid structures.

Our approach utilises a composition of several methods ranging from pure rule based (as classified in [3]) multi-dimensional distance geometry method [1] to data based stored substructure lookup features in a flexible software framework. The actual implementation is a highly portable JAVA software (available at [2]), which fits a broad scale of applications: it is used in small web drawing applets as well as standalone database processing component.

The coordinate determination process can be best characterized by the "divide and conquer" approach: the structure is composed of fragments, which are joined together. From the available fragment conformers the conformers of the joined structures can be generated during the fusing step. The fragment conformers are generated either through further fragmentation or with an elemental structure/conformer prediction method, consequently the conformational analysis is an inherent part of the building process (in contrast with methods which proceed from 3D initial structures, like [4]). The novelty of our approach lies in the diversity of the utilised elemental methods and the arisen scalability options.

References

1. G. Imre, G. Veress, A. Volford and O Farkas, "Molecules from the Minkowski Space: An approach to building 3D molecular structures", *J. Mol. Struct. (Theochem)*, 666-667, 51-59 (2003)
2. <http://www.chemaxon.com/marvin>
3. J. Sadowski and J. Gasteiger, "From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders", *Chem. Rev.*, 93, 2567-2581 (1993)
4. J. Weiser, M. C. Holthausen, L. Fitjer, "HUNTER: A Conformational Search Program for Acyclic to Polycyclic Molecules with Special Emphasis on Stereochemistry", *J. Comput. Chem.*, 18, 1265-1281 (1997)

P-58 : Uses and Potential Uses of Reasoning in Chemoinformatics

Julian Hayward; Lhasa Limited, Leeds, GB
Philip Judson, Lhasa Limited

There are situations in chemoinformatics where the need for reliable, sometimes numerical, information to support algorithms for the quantification or ranking of output causes problems. For example, algorithms may require a “yes” or “no” answer to the question “Is this atom aromatic”; systems for finding structures similar to a query may depend on the processing of numerical measures of similarity in order to rank output. In connection with our work on the prediction of the toxicity of chemicals and the metabolism of xenobiotic chemicals we have developed reasoning methods tolerant of uncertainty which could be of much wider use in chemoinformatics.

P-60 : Lead Conformers, a Thermodynamics Approach

Adrian Kalaszi; Eotvos Lorand University, Budapest, HU
Odon Farkas, Eotvos Lorand University

Finding drug like compounds is a challenging process in drug discovery. The 3D structure of the binding site of the target protein is often unknown and dealing with the flexibility of the ligand molecules is still problematic. However, flexible ligand molecules can be considered as nanoscale scanning devices adapting to the 3D structure of the active site[1]. The special thermodynamic properties of the binding of flexible molecules, as derived here, show that the probability of the binding conformations in solution determines the likelihood of binding. The binding activities, which correlate with binding probability, can be evaluated experimentally, while the probability of conformations in solution can be obtained via molecular dynamics simulations. If both data for a set of flexible molecules are available, a model to the spatial arrangement of the pharmacophores can be constructed.

1. A. Kalaszi, O. Farkas *Journal of Molecular Structure (Theochem)* 666 667 (2003) 645 649

P-62 : Finding Discriminative Substructures Using Elaborate Chemical Representation

Jeroen Kazius; Universiteit Leiden, Leiden, NL
Siegfried Nijssen, Joost Kok, Thomas Bäck, and Ad IJzerman, Universiteit Leiden

In pharmaceutical research, knowledge of molecular substructures relevant to physico-chemical and biological properties can aid in synthesis decision, library design for high throughput screening (HTS), hit prioritisation, lead optimisation and prioritisation of pharmacological or toxicological assays. Therefore, data mining algorithms to determine substructures discriminative of these properties are ever more important. Often however, limited chemical information is considered, such as linear sequences of atom and bond types. Furthermore, the increasingly large amount of available data (e.g. from HTS) requires significant efficiency of such algorithms.

We employed a means of chemical representation in which single atoms can be represented by atomic hierarchies. Any matching SMARTS expression[1] can be used to represent an atom while further expressions can be appended as extra nodes with additional chemical information. For example, a wildcard such as a hydrogen donor label can be supplemented with specifiers for atom type, charge, ring size, number of hydrogens, etc. Molecules are consequently represented as elaborate graphs.

We have developed a novel graph-based data mining system, called GASTON. GASTON makes use of the fact that sequences, free trees and graphs are contained in each other and it efficiently splits up the substructure finding process by finding all sequences, free trees and graphs, respectively. For extra speed, constraints can be applied to, for instance, the maximum size of a substructure, the minimum number of molecules that it needs to detect and/or its maximum p-value for a binary biological or physico-chemical classification. The combination of this representation method and a substructure finding algorithm enables the automated detection of discriminative substructures, which can consist of very general and very specific components.

A final selection step is required in order to extract a small set of discriminative, nonredundant and informative substructures from a dataset of compounds with binary classifications. Therefore, a simple p-value-based greedy selection method was written and employed. As a practical example, several results of the application of the described methods on large datasets for mutagenicity and aqueous solubility are discussed. The outputted substructures do not only provide insight into relevant moieties, they can also be used for predictive purposes. As such, these substructures can directly serve as either structural alerts for mutagenicity / insolubility or as chemical solutions for increasing solubility / decreasing mutagenicity. These promising findings confirm the utility of employing the discussed methods of chemical representation, substructure finding and descriptor selection.

1. Daylight Chemical Information, Inc., Santa Fe, NM, at www.daylight.com/dayhtml/doc/theory/theory.smarts.html

P-64 : scPDB: An Annotated Database of Three-Dimensional Structures of Binding Sites for Drug-Like Molecules

Esther Kellenberger; University of Strasbourg, Illkirch, FR
Guillaume Bret, Pascal Muller, and Didier Rognan, University of Strasbourg

The scPDB is a collection of about 7000 three-dimensional structures of putative binding sites found in the Protein Data Bank (PDB). Binding sites were extracted from all high resolution crystal structures in which a complex between a protein cavity and a drug-like ligand was detected. Ligands consist of small molecules like nucleotides (3 mer), peptides (5 mer), endogeneous ligands and drugs, but not water, metal ions or unwanted molecules (e.g., non specific ligands, solvents, detergents). The binding site is formed by all protein residues (including amino acids, cofactors and important metal ions) with at least one atom within 6.5 Å of a ligand atom. The scPDB was carefully annotated. Information from PDB entries and corresponding SWISSPROT files were merged in order to assign to every binding sites the following features: protein name, function, source, domain and mutations, ligand name and structure.

The scPDB was designed for docking purposes; the virtual screening of the binding sites database against a given ligand can predict the most likely target for the molecule and also suggest a selectivity profile [1]. It may also be used to analyse the similarity between cavities and to derive rules that describe the relationship between ligand pharmacophoric points and protein site properties.

The scPDB is periodically updated. It is accessible on the web at <http://bioinfo-pharma.u-strasbg.fr/scpdb/>

1. Paul, N., Bret, G., Kellenberger, E., Muller, P., Rognan, D. (2004) Recovering the true targets of selective ligands by virtual screening of the Protein Data Bank. *Proteins*, 54, 671-680.

P-66 : Application of Knowledge-Based Scoring Functions for Virtual Screening

Chrysi Konstantinou Kirtay; Cambridge University, Unilever Centre for Molecular Science Informatics, Cambridge, GB
J.B.O. Mitchell, Cambridge University, Unilever Centre for Molecular Science Informatics
J.A. Lumley, Arrow Therapeutics Ltd

Flexible docking algorithms, applied to virtual compound libraries, are able to predict protein ligand complexes with reasonable accuracy and speed. However the major weakness lies in the functions used for predicting the binding affinity between the receptor and ligand, also known as scoring functions. Scoring functions can be applied during the docking process as fitness functions for the optimization of ligand orientation and conformation, as well as in the post docking comparison of molecules for the estimation of their binding affinity to a specific protein target [1].

We have implemented BLEEP [2, 3], a knowledge-based scoring function based on high resolution ($\leq 2\text{\AA}$) structural data from the Protein Data Bank (PDB), in order to determine how well a candidate docked structure resembles a typical real protein-ligand complex. Possible protein-ligand interactions are assessed using their atom-atom distance

distributions which are converted into pseudo-energy functions by the implementation of Boltzmann hypothesis. BLEEP generates good ($R_s \approx 0.6$) correlations with experimental binding energies for diverse sets of non-covalent complexes and substantially better correlations ($R_s \approx 0.75$) for a series of related complexes [4].

In a recent study we generated an updated version of the BLEEP statistical potential, using a dataset of 196 complexes, performing similarly to the existing BLEEP. An algorithm was implemented to allow the automatic calculation of bond orders, and hence of the appropriate numbers of hydrogen atoms present. A potential specific to strongly and weakly bound complexes was generated. However there was no further improvement to the prediction of binding affinities. In addition, we also investigated the range of binding energies found as a function of either ligand molecular weight or number of heavy atoms and we derived some simple functions describing this behaviour. Our research is currently focusing on the application of BLEEP to a number of virtual screening case examples utilising different combinations of docking/scoring functions (GOLD 2.2 & Silver, FlexX).

1. Stahl M and Rarey M. Detailed Analysis of Scoring Functions for Virtual Screening. *Journal of Medicinal Chemistry* 2001;44:1035-1042.
2. Mitchell JBO, Laskowski RA, Alex A, Forster MJ, and Thornton JM. BLEEP - A potential of Mean Force Describing Protein-Ligand Interactions II. Calculation of Binding Energies and Comparison with Experimental Data. *Journal of Computational Chemistry* 1999;20:1177-1185.
3. Mitchell JBO, Laskowski RA, Alex A and Thornton JM. BLEEP - A potential of Mean Force Describing Protein-Ligand Interactions. I. Generating the Potential. *Journal of Computational Chemistry* 1999;20:1165-1176.
4. Marsden PM, Puvanendrapillai D, Mitchell JBO and Glen RC. Predicting protein ligand binding affinities: a low scoring game ? *Organic Biomolecular Chemistry* 2004;2:3267-3273.

P-68 : Selecting Potential Active Compounds by Matching Biological Profiles of Compounds with Known and Unknown Activities

Alexander Kos; AKos Consulting & Solutions GmbH, Riehen, CH
 Dusan Toman, DIMENSION 5, Ltd.
 Vladimir V. Poroikov, Institute Biomedical Chemistry of Rus. Acad. Med. Sci.
 Ulrich Jordis, Technische Universität Wien
 Timo Knuutila, Visipoint Oy

In silico methods appear to have more merit in the early phase of compound selection than high throughput screening experiments. Whereas the first screening experiments give physical binding parameters for compounds, one does not learn anything about possible side effects or about the possible metabolic reactions of a compound. PASS (Prediction of Activity Spectra of Substances) is one of the first programs able to develop reliable 1000 biological parameters for compounds from its 2D structure. The biological parameter is the measurement as percentage that a compound has a certain biological activity, like being an acetylcholine-esterase inhibitor (AChEI). Every drug has many actions, known as side effects. A biological profile is the set of activities that a drug should have, and a set of activities that a drug should not have, i.e. being carcinogenic. Using PharmaExpert we can develop a biological profile for a desired class of compounds by analysing the statistics of the PASS parameters. Using clustering software we match the biological profiles of compounds with known and unknown activities. Furthermore, a graphic representation tells us which functional groups or atoms of a compound are used in deriving a prediction. PASS Metabolite predicts metabolic reactions. PASS, PharmaExpert and PASS Metabolite are very effective, but only some of a large number of useful in silico methods. PASS CL is a command line version that can be incorporated in ones own application, or components are available for programs like Pipeline Pilot.

P-70 : Evaluation of the Diversity of Screening Libraries

Mireille Krier; CNRS, Illkirch, FR
 Guillaume Bret and Didier Rognan, CNRS

High-throughput screening nowadays requires compound libraries in which the maximal chemical diversity is reached with the minimal number of molecules. Medicinal chemists have traditionally realized assessment of

diversity and subsequent compound acquisition although a recent study suggests that experts are usually inconsistent in reviewing large datasets[1]. In order to analyze the chemical diversity of commercially available screening collections, we have developed a general workflow aimed at (1) identifying drug-like compounds, (2) cluster them by common substructures (scaffolds) using a phylogenetic-like tree growing algorithm[2], (3) measure the scaffold diversity encoded by each screening collection independently of its size, and finally (4) merge all common substructures in a non-redundant scaffold library that can easily be browsed by structural and topological queries. Starting from 2.4 compounds described by 12 commercial sources, three categories of libraries could be identified: combi-chem libraries (low scaffold diversity, large size), screening libraries (medium diversity, medium size) and diverse libraries (high diversity, low size). The chemical space enclosed in the scaffold library can be easily searched to prioritize scaffold-focused libraries.

1. Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J Med Chem* 2004, 47, 4891-4896.
2. Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J Chem Inf Comput Sci* 2002, 42, 1069-1079.

P-72 : Estimation of Environmental Compartment Half-Lives from Structural Similarity

Ralph Kühne; UFZ Centre for Environmental Research, Leipzig, DE
Ralf-Uwe Ebert and Gerrit Schüürmann, UFZ Centre for Environmental Research

Environmental fate modelling requires knowledge of the total degradation rates in environmental compartments. However, the availability of respective data is rather limited. Compartment half-lives from literature have been compiled in a database. A k nearest neighbours approach is applied to obtain a prediction for a new compound from this database.

The similarity measure to select the nearest neighbours is an approach based on atom centred fragments. First, the molecule skeleton is separated in its individual atoms. Then, atom centred fragments are built from the atoms by consideration of the neighbour atoms and several atom properties as e.g. aromaticity. Last, the selection of the most similar compounds in the database for a test chemical is achieved by comparison of the atom centred fragments.

The results of this method are compared to a few existing methods for degradation rates of individual processes, and the reliability is tested by statistical means. The impact of the prediction uncertainty to environmental fate modelling is examined.

P-74 : Water Solubility Prediction - Model Selection Based on Structural Similarity

Ralph Kühne; UFZ Centre for Environmental Research, Leipzig, DE
Ralf-Uwe Ebert and Gerrit Schüürmann, UFZ Centre for Environmental Research

To predict water solubility for organic compounds from chemical structure, a number of quite accurate estimation methods have been published. These models differ in their performance for individual compounds and compound classes. To select the optimum method for a particular chemical, a k nearest neighbours approach is suggested. For a large training set of validated experimental water solubility data, seven estimation methods have been applied. Concerning the respective method errors for the individual compounds, a scoring system for these models has been stored in a database. To select a model for prediction, the nearest neighbours are looked for in this database by a similarity approach based on atom centered fragments. The model with the best average score for these compounds is selected.

The efficiency of this approach is demonstrated by an external test set, compared to the individual models, and proved by statistical means.

P-76 : The Inconsistency of Medicinal Chemists in Reviewing Sets of Compounds

Mic Lajiness; Eli Lilly & Company, Indianapolis, IN, US

Medicinal chemists are frequently asked to review lists of compounds to assess their drug or lead-like nature, and to evaluate the suitability of lead compounds based on their attractiveness; and/or synthetic. Presumably, one medicinal chemist's opinion is as good as any others -- but is it? In an attempt to answer this question an experiment was performed in conjunction with a compound acquisition program conducted at Pharmacia. Historically, this involved a review of many thousands of compounds by medicinal chemists who eliminate anything deemed undesirable for any reason. In the present experiment, 22,000 compounds requiring review by medicinal chemists were broken down into eleven lists of approximately 2,000 compounds each. Unbeknownst to the medicinal chemists a subset of 250 compounds, previously rejected by a very experienced senior medicinal chemist, was added to each of the lists. Most of the thirteen medicinal chemists who participated in this process reviewed two lists, although some only reviewed a single list and one reviewed three lists. Those compounds that were deemed unacceptable were recorded and tabulated in various ways to assess the consistency of the reviews. It was found that medicinal chemists were not very consistent in the compounds they rejected as being undesirable. This has important implications for pharmaceutical project teams where individual medicinal chemists review lists of primary screening hits to identify those compounds suitable for follow-up. Once a compound is removed from a list it and other structurally-similar compounds are effectively removed from further consideration. This can also have an impact on computational chemists who are developing models for assessing the desirability or attractiveness of different classes of compounds for lead discovery.

P-78 : Chemical Clichés: Treasure Hidden by Obviousness?

Eric-Wubbo Lameijer; Universiteit Leiden, Leiden, NL

Thomas Bäck, Joost Kok, Ad IJzerman, and Sara van de Geer, Universiteit Leiden

Molecules can be considered to be collections of atoms connected by bonds. The arrangement of atoms and bonds is however far from arbitrary. Synthetic and medicinal chemists commonly think in "groups", which are molecular fragments common to many molecules. Examples are the phenyl group, the methyl group and the keto group.

While most chemists know dozens of these groups by heart and hundreds by recognition, we decided to get a much more elaborate set of fragments by mining chemical databases. This has the advantages that the number of fragments found will exceed the number of fragments known to any individual chemist, and that the occurrence and therefore the relative importance of the different fragments can be determined more quantitatively.

In this project we mined the public database of the American National Cancer Institute, containing about 250,000 compounds. The molecules were split into fragments by breaking the bonds connecting the rings and the non-ring parts of the molecule. An example is shown in figure 1.

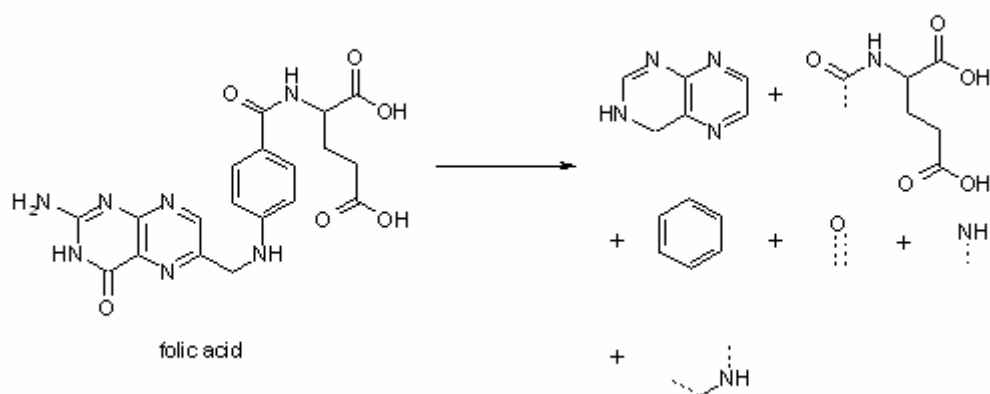


Figure 1: Decomposing folic acid into fragments

This splitting process resulted in 13574 different ring fragments and 10661 different non-ring fragments, which were sorted by occurrence and stored in separate databases. Frequencies of the different fragments varied widely, ranging from many thousands (phenyl) to one-time occurrences of the rarest groups.

We subsequently performed a co-occurrence analysis to see whether the distribution of the fragments in the molecules is mainly random or whether there are groups which co-occur uncommonly often, real “chemical clichés”. We found that there were many combinations which occurred much more often than expected by chance. Most of these were known biologically active compounds or chemical substructures which were easy to combine synthetically.

These lists of clichés may extend the choice a medicinal chemist has when modifying a compound. The groups at the top of the list can be used for the initial stages of lead optimization, to give a greater chance that the derivatives will be easy to synthesize. Alternatively, the investigator can consciously avoid clichés by looking lower on the list for suitable groups which are rarer, but for which synthetic information is available nevertheless. So knowledge of clichés could lead to both synthesizable and novel compounds.

P-80 : ¹H NMR - Based Classification of Photochemical Reactions

Diogo Latino; Universidade Nova de Lisboa, Caparica, PT

Filomena F. M. Freitas, Fernando M. S. Silva Fernandes, and João Aires-de-Sousa, Universidade Nova de Lisboa

Automatic analysis of changes in the ¹H NMR spectrum of a mixture, and their interpretation in terms of chemical reactions taking place, has a diversity of possible applications. For example the changes in the ¹H NMR spectrum of a stored chemical can be interpreted in terms of the chemical reactions responsible for degradation. Or the alterations in the spectrum of a biofluid can be related to changes in metabolic reactions.

The SPINUS program previously developed (1-3) for the estimation of ¹H NMR chemical shifts from the molecular structure allows linking a database of chemical reactions to the corresponding ¹H NMR data.

Clearly ¹H NMR spectroscopy has its own limitations, particularly for reactions with a small number of hydrogen atoms in the neighbourhood of the reaction centre. Having this in mind, the use of ¹H NMR has considerable advantages comparing to NMR of other nuclei, for example in terms of speed and amount of required sample.

Here we demonstrate the classification of photochemical reactions by Kohonen self-organizing maps (SOMs) taking as input the difference between the ¹H NMR spectra of the products and the reactants. We used a data set of 189 chemical reactions, each reaction having been manually assigned to one of 7 classes - [3+2] photocycloaddition of azirines to C=C, [2+2] photocycloaddition of triazinones to C=C, [3+2] photocycloaddition of pyridazines to C=C, [4+2] photocycloaddition of C=C to C=C (photo-Diels-Alder reactions), [2+2] photocycloaddition of C=C to C=O, [2+2] photocycloaddition of C=C to C=C, and [2+2] photocycloaddition of C=C to C=S. The chemical shifts of reactants and products were generated by SPINUS and were fuzzified to obtain a crude representation of the spectrum. All the signals arising from all the reactants of one reaction were taken together (a simulated spectrum of the mixture of reactants) and the same would hold for products (although in this work we only considered reactions yielding a single molecule as the product). The simulated spectrum of the reactants is subtracted from the spectrum of the products and the difference spectrum is taken as the representation of the chemical reaction- the input to the neural networks.

Kohonen neural networks were trained with a set of 147 reactions and then tested with the remaining 42 reactions. The test set was randomly selected to cover the whole space of reactions. A reasonable clustering of the reactions by reaction type was observed. An ensemble of five networks was considered, predictions being obtained by majority voting and associated with a prediction score. Correct predictions could be achieved for more than 90% of the training set and for 75-90% of the test set. The prediction score gave a robust indication of the reliability of the prediction.

The results support our proposal of linking reaction and NMR data for automatic reaction classification.

ACKNOWLEDGEMENTS

D.A.R.S. Latino acknowledges Fundação para a Ciência e Tecnologia for financial support under a PhD grant (SFRH/BD/18347). The authors thank InfoChem GmbH (Munich, Germany) for sharing the dataset of photochemical reactions.

REFERENCES

1. Y. Binev, J. Aires-de-Sousa, "Structure-Based Predictions of ¹H NMR Chemical Shifts Using Feed-Forward Neural Networks", *J. Chem. Inf. Comp. Sci.*, 2004, 44(3), 940-945.
2. Y. Binev, M. Corvo, J. Aires-de-Sousa, "The Impact of Available Experimental Data on the Prediction of ¹H NMR Chemical Shifts by Neural Networks", *J. Chem. Inf. Comp. Sci.*, 2004, 44(3), 946-949.
3. SPINUS can be accessed at: <http://www.dq.fct.unl.pt/spinus> or <http://www2.chemie.uni-erlangen.de/services/spinus>

P-82 : SOMA - Computational Molecular Discovery Environment

Pekka Lehtovuori; CSC - Scientific Computing Ltd, Espoo, FI
Tommi Nyrönen, CSC - Scientific Computing Ltd

We are developing a computational molecular discovery environment for the Finnish universities and research institutions. The goal is to build an integrated software environment and enhance the usability of the software in the supercomputing grid of CSC - the Finnish Information Technology Center for Science. CSC's services are based on a versatile supercomputing environment, fast data communications connections and on expertise in different scientific disciplines and information technology. CSC offers a large selection of computational tools (e.g. bioinformatics, chemoinformatics, databases, quantum chemistry, QSAR, molecular dynamics).

Transferring data from one program to another is a nerve-wracking step, simply because programs are seldom designed to work together. SOMA consists of scientific applications, linked together in a www-interface to form a user-friendly computing and data management environment.

The environment currently consists of 1. extended markup language (XML) description of the scientific programs (program-XML), 2. template scripts for the configuration files of programs and possible batch job system, 3. program piper, which moves the results from one program to the input of the next one and reports the state of each step to the job status database, 4. tools for internal data format conversion 5. tools, which build up a www-interface based on the program-xml and 6. tools, which presents the results in a web-browser.

The key function of the SOMA environment is to improve the data flow between programs in a supercomputing environment. The users can search and build small molecules *in silico* and use structural, physicochemical and biological information on the target macromolecule with less technical obstacles than before.

P-84 : Characterization and Clustering of Reagents for Combinatorial Library Design from the Products' Perspectives

Uta Lessel; Boehringer Ingelheim Pharma GmbH, Biberach, DE

Reagent-biased product based design was published by Pearlman and Smith in 1999. The method is available with Diverse Solutions and works well e.g. to select a set of reagents leading to a combinatorial library with products broadly distributed in a BCUT property space. If the BCUT property space is divided in cells the products of the designed library occupy an optimized number of cells. But a new design cycle has to be started, if one or more of the selected reagents are exchanged. Manual exchange of reagents by the combi chemists usually leads to a clear loss of cell occupation shown by the final library. For this purpose a new method for reagent selection was created based on the original idea of Pearlman.

Each reagent is characterized by a fingerprint encoding the cells which are occupied by its products. In the next step the reagents can be clustered according to their fingerprints. This way reagents group together that lead to similar

products. Additionally, reagents can be prioritized e.g. according to the number of cells occupied by their products, by the number of products obeying Lipinski's rules, etc.

In the presentation the similarities of reagents based on the cell occupancy of their products will be discussed in more detail and it will be illustrated by some design examples how the method can be implemented in Combinatorial Library Design.

LIST OF PARTICIPANTS

List of Participants

Mr. Tim Aitken
Accelrys
334 Science Park
Cambridge CB4 0WN
UK

Dr. Jerome Amaudrut
Laboratoires Fournier
50 rue de Dijon
21121 DAIX
France

Dr. Giuliana Angonoa-Doehnert
BASF Aktiengesellschaft
GVW/I - C006
67056 Ludwigshafen
Germany

Dr. Soheila Anzali
Merck KGaA
Glob Tech BCI
Frankfurter Str. 250
64271 Darmstadt
Germany

Miss Mitsuyo Aota
Yamaguchi University
Graduate School of Science and Engineering
2-16-1, Tokiwadai
Ube, 755-8611
Japan

Dr. Masamoto Arakawa
The University of Tokyo
Hongo 7-3-1, Bunkyo-ku
Tokyo, 113-8656
Japan

Dr. Janet Ash
The Roundel
Cranbrook TN17 2EP
UK

Dr. J. Christian Baber
Neurocrine Biosciences
12790 El Camino Real
San Diego, CA 92130
USA

Dr. David Bardsley
CambridgeSoft Corporation
8 Signet Court, Swanns Road
Cambridge CB5 8 LA
UK

Dr. Knut Baumann
University of Wuerzburg
Am Hubland
D-97074 Wuerzburg
Germany

Mr. Andreas Bender
University of Cambridge
Unilever Centre for Molecular Science Informatics
Lensfield Road
Cambridge CB2 1EW
UK

Dr. Yuri Binev
Universidade Nova de Lisboa
Department of Chemistry
campus FCT-UNL
Caparica 2829-516
Portugal

Mr. Kristian Birchall
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. Krisztina Boda
University of Leeds
ICAMS School of Chemistry
Leeds LS2 9JT
UK

Jost Bohlen
FIZ CHEMIE Berlin
Franklinstr. 11
10587 Berlin
Germany

Dr. Hans Briem
Schering AG
Computational Chemistry
Muellerstraße 178
13342 Berlin
Germany

Dr Natahn Brown
Novartis Pharma AG
Novartis Institutes for BioMedical Research
WSJ-350.U1.13
CH-4002 Basel
Switzerland

Dr. Robert Brown
SciTegic Inc
9665 Chesapeake Dr
Suite 401
San Diego, CA 92123
USA

Dr. Ian Bruno
Cambridge Crystallographic Data Centre
12 Union Rd
Cambridge CB2 1EZ
UK

Dr. Anna Maria Capelli
GSK
v. Fleming, 4
37135 Verona
Italy

Mr. Goncalo Carrera
Universidade Nova de Lisboa
Department of Chemistry
campus FCT-UNL
Caparica 2829-516
Portugal

Dr. Alexandre Carvalho
Universidade do Porto
REQUIMTE/Departamento de Química
Rua de Vilar 210 10ºA
Porto 4050 - 625
Portugal

Dr. Gonzalo Cerruela García
Córdoba University
Dept. Computer Science
Campus de Rabanales C2 Building
Cordoba 14071
Spain

Dr. Edith Chan
Inpharmatica
60 Charlotte Streey
London W1T 2NU
UK

Dr. Donovan Chin
Biogen Idec Inc.
14 Cambridge Center
Bio8
Cambridge, MA 2420
USA

Ms. Linda Clark
Thomson Scientific
14 Great Queen St
London WC2B 5DF
UK

Dr. Robert Clark
Tripos, Inc.
1699 S. Hanley Rd.
St. Louis, MO 63144
USA

Dr. Holger Claußen
BioSolveIT GmbH
An der Ziegelei 75
53757 Sankt Augustin
Germany

Mr. Simon Cottrell
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. Ferenc Csizmadia
ChemAxon Ltd
Máramaros köz 3/a
1037 Budapest
Hungary

Dr Elena Cubero
Laboratorios Dr. Esteve, S.A.
Av. Mare de Deu de Montserrat, 221
Barcelona 8041
Spain

Prof. Fernando Da Costa
Universitaet Erlangen
Computer-Chemie-Centrum
Naegelsbachstrasse 25
91052 Erlangen
Germany

Dr. Cyril Daveu
Sanofi Aventis
1 Avenue Pierre Brossolette
91385 Chilly Mazarin
France

Dr. Joseph Durant
Elsevier MDL
14600 Catalina Street
San Leandro, CA 94577
USA

Mr. Nick Day
University of Cambridge
Unilever Centre for Molecular Informatics
70 Coleridge Road
Cambridge CB1 3PJ
UK

Dr. Michael Engels
Gruenthal GmbH
Zieglerstrasse
52099 Aachen
Germany

Dr. Hans de Bie
Chemcad
Veldsingel 49
NL-6581 TD Malden
Netherlands

Mr. Dumitru Erhan
Universite de Montreal
167 Gordon St
Verdun, H4G2R2
Canada

Dr. Jacob de Vlieg
NV Organon
Molecular Design & Informatics
P.O. Box 20
5340 BH Oss
Netherlands

Dr. Peter Ertl
Novartis Institutes for BioMedical Research
WKL-125.14.20
CH-4002 Basel
Switzerland

Mr. Jörg Degen
Universität Hamburg
Zentrum für Bioinformatik
Bundesstraße 43
20146 Hamburg
Germany

Dr. Ödön Farkas
Eotvos Lorand University
Department of Organic Chemistry
1/A Pazmany Peter setany
H-1117 Budapest
Hungary

Dr. Geoff Downs
Barnard Chemical Information Ltd
30 Kiveton Lane
Todwick
Sheffield S26 1HL
UK

Mr. Uli Fechner
University of Frankfurt
Marie-Curie-Strasse 11
60439 Frankfurt am Main
Germany

Dr. Sigmar Dressler
Novartis Pharma AG
WSJ-350.P10
CH-4002 Basel
Switzerland

Mr. Simon Folkertsma
CMBI Radboud University
Toernooiveld 1
6525 ED Nijmegen
Netherlands

Mr. Matt Drew
Array Biopharma
3200 Walnut St
Boulder, CO 80304
USA

Dr. Xavier Fradera
Organon Laboratories
Newhouse ML1 5SH
UK

Mr. Patrick Fricker
Center for Bioinformatics
Bundesstrasse 43
20146 Hamburg
Germany

Dr. Arnaud Gohier
Institut de Recherches Servier
125, chemin de Ronde
78290 Croissy-sur-Seine
France

Prof. Kimoto Funatsu
University of Tokyo
Department of Chemical System Engineering
7-3-1 Hongo, Bunkyo-ku
Tokyo, 113-8656
Japan

Dr. Andreas Göller
Bayer Healthcare AG
PH-R Chemical Research
Aprather Weg 18a
42096 Wuppertal
Germany

Dr. Eleanor Gardiner
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. Guenter Grethe
InfoChem
352 Channing Way
Alameda, CA 94502
USA

Prof. Johnny Gasteiger
University of Erlangen
Computer-Chemie-Centrum
Nägelsbachstraße 25
91052 Erlangen
Germany

Dr. Judith Guenther
Schering AG
Computational Chemistry
Muellerstraße 178
13342 Berlin
Germany

Dr. Peter Gedeck
Novartis Institutes for BioMedical Research
Wimblehurst Road
Horsham RH12 5AB
UK

Dr. Antonio Guerreiro
Sanofi Aventis
Chemical Sciences / Drug Design
13 Quai Jules Guesde
94400 Vitry Sur Seine
France

Dr. Valerie Gillet
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Mr. Sasha Gurke
Knovel Corp.
13 Eaton Avenue
Norwich, NY 13815
USA

Dr. Sanne Glad
Nuevolution A/S
Rønnegade 8
DK-2100 Copenhagen
Denmark

Dr. Sandra Handschuh
Boehringer Ingelheim Pharma GmbH & Co. KG
Birkendorferstr. 65
88397 Biberach a.d. Riss
Germany

Dr. Mauro Gobbini
Prassis Istituto Di Ricerche Sigma-Tau S.P.A.
Via Forlanini, 3
20019 Settimo Milanese
Italy

Dr. Catrin Hasselgren Arnby
Astrazeneca
Computational Toxicology, Safety Assessment
SC3, Pepparedsleden 1
SE-43183 Molndal
Sweden

Dr. Julian Hayward
Lhasa Limited
School of Chemistry, University of Leeds
Woodhouse Lane
Leeds LS2 9JT
UK

Dr. Ulrich Heigl
Elsevier MDL
Gewerbestrasse 12
4123 Allschwil
Switzerland

Dr. Stephen R. Heller
NIST
MS-838
100 Bureau Drive
Gaithersburg, MD 20899
USA

Dr. Harold Helson
CambridgeSoft, Inc.
100 CambridgePark Dr
Cambridge, MA 2140
USA

Dr. Begoña Hernandez
Almirall Prodesfarma, S.A.
Cardener 68
Barcelona 8024
Spain

Mr. Jerome Hert
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. Achim Herwig
Molecular Networks GmbH
Nägelsbachstraße 25
91052 Erlangen
Germany

Dr. Martin Hicks
Beilstein-Institut
Trakehnerstr. 7-9
60487 Frankfurt
Germany

Miss Linda Hirons
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. John Holliday
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Joerg Homann
FIZ CHEMIE Berlin
Franklinstr. 11
10587 Berlin
Germany

Prof. Kneji Hori
Yamaguchi University
Faculty of Engineering
1/16/2001 Tokiwadai
Ube, 755-8611
Japan

Dr Wolf Ihlenfeldt
National Library of Medicine
Building 38A - Lister Hill Ctr, Room 6S614H
9000 Rockville Pike
Bethesda, Maryland 20894
USA

Prof. Ad IJzerman
Leiden University
Leiden/Amsterdam Center for Drug Research
PO Box 9502
2300RA Leiden
Netherlands

Mr. Yutaka Ikenaga
Yamaguchi University
Graduate school of science and engineering
2/16/2001 Tokiwadai
Ube, 755-8611
Japan

Mr. Gabor Imre
ELTE Dept. of Organic Chemistry
BME Dept. of Aut. and Appl. Info
1/A Pazmany Pet. set.,
H-1117 Budapest
Hungary

Dr. Richard Jackson
University of Leeds
School of Biochemistry & Microbiology
Leeds LS2 9JT
UK

Dr. Esther Kellenberger
University of Strasbourg
Faculté de Pharmacie
74 route du Rhin, BP 24
F-67100 Illkirch
France

Dr. Anne Techau Joergensen
Boehringer Ingelheim Pharma GmbH & Co. KG
Birkendorferstr. 65
88397 Biberach a.d. Riss
Germany

Dr. Matthew Kellett
Thomson Scientific
3501 Market Street
Philadelphia, PA 19070
USA

Dr. Theresa Johnson
Pfizer Inc.
620 Memorial Drive
Cambridge, MA 01970
USA

Mrs. Chrysi Kirtay
University of Cambridge
Unilever Centre for Molecular Science Informatics
Cambridge CB2 1EW
UK

Mr. Adrian Kalaszi
Eötvös University
Dept. of Organic Chemistry
1/A Pázmány Péter Setényi
H-1117 Budapest
Hungary

Dr. Achim Kless
Gruenthal GmbH
Zieglerstrasse
52099 Aachen
Germany

Miss Naparat Kammasud
Institut Curie
Centre Universitaire Bat 110
91405 Orsay
France

Prof. Gilles Klopman
Multicase Inc.
23811 Chagrin Blvd., # 305
Beachwood, OH 44122
USA

Mr. Sinan Karaboga
Laboratoires Fournier
50 rue de Dijon
21121 Daix
France

Dr. Jean-Pierre Kocher
Molecular Networks GmbH
Naegelsbachstr. 25
91052 Erlangen
Germany

Dr. Muthukumarasamy Karthikeyan
National Chemical Laboratory
Digital Information Resource Center
Pune 411008
India

Dr. Maria Kontoyianni
Procter & Gamble Pharmaceuticals, Inc.
8700 Mason Montgomery Rd.
Cincinnati, OH 45227
USA

Mr. Jeroen Kazius
Leiden University
Leiden/Amsterdam Center for Drug Research
PO Box 9502
2300 RA Leiden
Netherlands

Dr. Alexander Kos
AKos GmbH
Picassoplatz 4
Postfach 141
CH-4010 Basel
Switzerland

Dr. Hans Kraut
InfoChem GmbH
Landsberger Strasse 408
81241 Munich
Germany

Miss Mireille Krier
University of Strasbourg
Faculté de Pharmacie
74 route du Rhin, BP 60024
F-67100 Illkirch
France

Dr. Ralph Kühne
UFZ Centre for Environmental Research
Permoserstr. 15
4318 Leipzig
Germany

Dr. Mic Lajiness
Eli Lilly and Company
Lilly Corporate Center
DC:1523
Indianapolis, IN 46285
USA

Mr. Eric-Wubbo Lameijer
Leiden University
Einsteinweg 55
2333 CC Leiden
Netherlands

Mr. Diogo Latino
Universidade Nova de Lisboa
Department of Chemistry
campus FCT-UNL
Caparica 2829-516
Portugal

Dr. Pekka Lehtovuori
CSC - Scientific Computing Ltd.
P.O.Box 405
FI-02101 Espoo
Finland

Dr. Uta Lessel
Boehringer Ingelheim Pharma GmbH & Co. KG
Department of Lead Discovery
Birkendorfer Str. 65
88397 Biberach
Germany

Dr. Pierre-Jean L'Heureux
Universite de Montreal
Dept. IRO
C.P. 6128, Succ.Centre-Ville
Montreal, H3C 3J7
Canada

Dr. Marguerita Lim-Wilby
Accelrys
10188 Telesis Ct
San Diego, CA 92121
USA

Dr. Peter Loew
InfoChem GmbH
Landsberger Strasse 408
81241 Munich
Germany

Dr. Andreas Löffler
Tripos GmbH
Martin-Kollar-Str. 17
81829 Munich
Germany

Dr. Jos Lommerse
NV Organon
Molecular Design & Informatics
P.O. Box 20
5340 BH Oss
Netherlands

Mrs. Karen Lucas
Chemical Abstracts Service
2540 Olentangy River Road
Columbus, Oh 43210
USA

Mr. Daisuke Machii
Kyowa Hakko Kogyo Co., Ltd.
1188 Shimotogari
Nagaizumi-cho Sunto-gun, 411-8731
Japan

Dr. Stephen Maginn
Chemical Computing Group
St Johns Innovation Centre, Cowley Road
Cambridge CB4 0WS
UK

Dr. Boryeu Mao
Cerep Inc
15318 NE 95th St
Redmond, WA 98052
USA

Dr. Chris Marshall
Astrazeneca
Mereside, Alderley Park
Macclesfield SK10 4TG
UK

Mr. Jörg Maruszyk
Universität Erlangen-Nürnberg
Computer-Chemie-Centrum
Nägelsbachstr. 25
91052 Erlangen
Germany

Miss Yvonne McCormick
Chemical Abstracts Service
Westhill Croft
Butterton
Leek ST13 7TD
UK

Dr. Iain McFadyen
Wyeth Research
200 Cambridge Park Drive
Cambridge, MA 2144
USA

Dr. Claire Minoletti
Sanofi Aventis
Chemical Sciences / Drug Design
13 Quai Jules Guesde
94400 Vitry Sur Seine
France

Miss Kirstin Moffat
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. Rainer Moll
CASAF Computerchemie GmbH
Schwindstr. 6
D-04425 Taucha
Germany

Mr. Chido Mpamhanga
University of Sheffield
Department of Chemistry (room e45c)
Dainton Building, Brook Hill
Sheffield s3 7hf
UK

Dr. Ingo Muegge
Boehringer Ingelheim Pharmaceuticals, Inc.
900 Ridgebury Road
P.O. Box 368
Ridgefield, CT 06877-0368
USA

Dr. Hamse Mussa
Cambridge University
Unilever Centre
Lensfield Rd
Cambridge CB2 1EW
UK

Dr. Heike Nau
MDL Information Systems
Theodor-Heuss-Allee 108
60486 Frankfurt
Germany

Mr. Peter Nichols
Hampden Data Services Ltd.
Highstone House
165 High Street
Barnet EN5 5SU
UK

Dr. Marc Nicklaus
NCI-Frederick
Bldg. 376, Rm. 207
376 Boyles Street
Frederick, MD 21228
USA

Dr. Frank Oellien
Intervet Innovation GmbH
BioChemInformatics
Zur Propstei
D-55271 Schwabenheim
Germany

Dr. Don Parkin
CCLRC
Chemical Database Service
Warrington WA4 4AD
UK

Mr. Yogendra Patel
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. Luc Patiny
EPFL
BCH 5121
1015 Lausanne
Switzerland

Mr. Alex Perryman
University of California, San Diego
Howard Hughes Medical Institute
9500 Gilman Drive; Mail code 0365
La Jolla, CA 92093-0365
USA

Mr. Paul Peters
CAS
SIIL
Krophollerkade 15
2552 ZS Den Haag
Netherlands

Dr. Matthias Pförtner
Molecular Networks GmbH
Naegelsbachstr. 25
91052 Erlangen
Germany

Dr. György Pirok
ChemAxon Ltd
Máramaros köz 3/a
1037 Budapest
Hungary

Ms. Timea Polgar
Richter Gedeon Ltd
Gyömrői út 19-21
1103 Budapest
Hungary

Dr. Anders Poulsen
SBIO Pte. Ltd.
1 Science Park Road
#05-09 The Capricorn
117528 Singapore
Singapore

Dr. Eric Raimbaud
Institut de Recherches Servier
125, chemin de Ronde
78290 Croissy-sur-Seine
France

Dr. B. N. Narasinga Rao
Scynexis, Inc
3501C Tricenter Blvd
Durham, NC 27713
USA

Prof. Matthias Rarey
University of Hamburg
Center for Bioinformatics
Bundesstrasse 43
20146 Hamburg
Germany

Dr. Dmitriy Rassokhin
Johnson & Johnson PRD
665 Stockton Drive
Exton, PA 19341
USA

Ms. Monika Rella
University of Leeds
School of Biochemistry and Molecular Biology
Garstang Building
Leeds LS2 9JT
UK

Mr. Steffen Renner
University of Frankfurt
Marie-Curie-Strasse 9
60323 Frankfurt am Main
Germany

Dr. Jóhannes Reynisson
Chemistry department
15 Cotswold Road
Sutton sm2 5ng
UK

Mr. Nicholas Rhodes
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Dr. Uwe Richter
JeriniAG
Invalidenstrasse 130
10115 Berlin
Germany

Dr. Peter Rusch
CINF
162 Holland Court
Mountain View, CA 94040
USA

Dr. Susan Robertson
Cambridge Crystallographic Data Centre
12 Union Rd
Cambridge CB2 1EZ
UK

Dr. Thomas Sander
Actelion Pharmaceuticals Ltd.
Gewerbestr. 16
4123 Allschwil
Switzerland

Dr. Sarah Rodgers
Unilever Research
Olivier van Noortlaan 120
3130 AC Vlaardingen
Netherlands

Dr. Gijs Schaftenaar
Radboud University of Nijmegen
CMBI
Toernooiveld 1
6525 ED Nijmegen
Netherlands

Dr. Lucia Rodriguez-Monge
Cambridge Crystallographic Data Centre
12 Union Rd
Cambridge CB2 1EZ
UK

Martin Schmidt
FIZ CHEMIE Berlin
Franklinstr. 11
10587 Berlin
Germany

Dr. Didier Rognan
CNRS UMR 7081
Bioinformatics Group
74, route du Rhin
67400 Ilkirch
France

Dr. Ansgar Schuffenhauer
Novartis Institutes of Biomedical Research
Novartis Campus
WSJ-250.3.11
CH-4002 Basel
Switzerland

Mr. Toni Ronkko
University of Kuopio
P.O. Box 1627
FI-70211 KUOPIO
Finland

Dr. Christof H. Schwab
Molecular Networks GmbH
Naegelsbachstr. 25
91052 Erlangen
Germany

Mr. Luc Roumen
Eindhoven University of Technology
Den Dolech 2
5612 AZ Eindhoven
Netherlands

Dr. Rob Scoffin
CambridgeSoft Corporation
8 Signet Court
Swanns Road
Cambridge CB5 8 LA
UK

Dr. Holger Ruchatz
Bio-Rad Laboratories Ltd., Informatics Division
Bio-Rad House, Maylands Avenue
Hemel Hempstead
Herts HP2 7TD
UK

Dr. Eva Seip
Elsevier MDL
Theodor-Heuss-Allee 108
60486 Frankfurt
Germany

Dr. Paul Selzer
Novartis Institutes for Biomedical Research
WKL-125.14.19
4002 Basel
Switzerland

Ms. Annu Söderholm
CSC - Scientific Computing Ltd.
P.O.Box 405
FI-02101 Espoo
Finland

Mr. Gavin Shear
Advanced Chemistry Development, Inc.
110 Yonge Street
14th Floor
Toronto, M5C 1T4
Canada

Dr. Matthew Stahl
OpenEye Scientific Software
3600 Cerrillos Rd
Suite 1107
Santa Fe, NM 87507
USA

Dr. Kirk Simmons
DuPont Stine-Haskell Research Center
S300/201G
P.O. Box 30
Newark, DE 19714
USA

Dr. Christoph Steinbeck
Cologne University
Bioinformatics Center (CUBIC)
Zuelpicher Str. 47
50674 Koeln
Germany

Dr. Suresh Singh
Vitae Pharmaceuticals
502 W. Office Center Drive
Fort Washington, PA 19034
USA

Dr. Nikolaus Stiefl
Lilly Forschung GmbH
Essener Bogen 7
22419 Hamburg
Germany

Dr. Markus Sitzmann
National Cancer Institute
Laboratory of Medicinal Chemistry
376 Boyles Street
Frederick, MD 21702
USA

Dr. Guenter Stiegler
BASF Aktiengesellschaft
GVW/I - C006
67056 Ludwigshafen
Germany

Dr. Andrew Smellie
Arqule Inc.
19 Presidential Way
Woburn, MA 1801
USA

Dr. Kaido Tämm
University of Tartu
2 Jakobi str
51014 Tartu
Estonia

Dr. Robert Snyder
580 Wilderness Peak Drive NW
Issaquah, WA 98027
USA

Mr. Meinolf Tegethoff
SciTegic
Südstr. 7
D 53909 Zülpich
Germany

Miss Ingrid Socorro Gutierrez
University of Cambridge
Unilever Centre for Molecular Science Informatics
Lensfield Road
Cambridge CB2 1EW
UK

Dr. Lothar Terfloth
University of Erlangen-Nuremberg
Computer-Chemie-Centrum
Nägelsbachstraße 25
91052 Erlangen
Germany

Ms. Anu Tervo
CSC - Scientific Computing
P.O.Box 405
2101 Espoo
Finland

Dr. Miklós Vargyas
ChemAxon Ltd
Máramaros köz 3/a
1037 Budapest
Hungary

Dr. Johanna Timmerman
Accelrys
20, rue Jean Rostand
91898 ORSAY
France

Prof. Kurt Varmuza
Vienna University of Technology
Laboratory for Chemometrics
Getreidemarkt 9/166
Vienna, A-1060
Austria

Dr. Samuel Toba
Accelrys
10188 Telesis Ct, Suite 100
San Diego, CA 92121
USA

Dr. Andy Vinter
Cresset BioMolecular Discovery Limited
Suite 503, Spirella Building, Bridge Road
Letchworth SG6 4ET
UK

Dr. Ulrike Uhrig
Tripos GmbH
Martin-Kollar-Str. 17
81829 Munich
Germany

Vincent Vivien
Bioreason
40 rue de Saint Sylvestre sur Lot
68660 Liepvre
France

Dr. George Vacek
OpenEye Scientific Software
3600 Cerrillos Rd
Suite 1107
Santa Fe, NM 87507
USA

Dr. Modest von Korff
Actelion Pharmaceuticals Ltd.
Gewerbestr. 16
4123 Allschwil
Switzerland

Dr. Opa Vajragupta
Faculty of Pharmacy
447 Sri-Ayudhya Road
Bangkok, 10400
Thailand

Dr. Markus Wagener
NV Organon
Molecular Design & Informatics
P.O. Box 20
5340 BH Oss
Netherlands

Miss Sofie Van Damme
University of Ghent
Krijgslaan 281 S3
9000 Ghent
Belgium

Dr. Wendy Warr
Wendy Warr & Associates
6 Berwick Court
Holmes Chapel
Cheshire CW4 7HZ
UK

Mr. Pieter van Grootel
Eindhoven University of Technology
Den Dolech 2
5612 AZ Eindhoven
Netherlands

Mr. Joerg Wegner
Universität Tübingen
Sand 1
72076 Tuebingen
Germany

Dr. Ron Wehrens
Radboud University Nijmegen
Analytical Chemistry
Toernooiveld 1
6525 ED Nijmegen
Netherlands

Mr. Alec Westley
Accelrys
334 Cambridge Science Park
Cambridge CB4 0WN
UK

Dr. Karin Wichmann
COSMOlogic GmbH & Co. KG
Burscheider Str. 515
51381 Leverkusen
Germany

Dr. David Wild
Indiana University School of Informatics
2480 Kimberly Road
Ann Arbor, MI 48104
USA

Dr. Henriette Willems
De Novo Pharmaceuticals
Compass House, Vision Park
Chivers Way, Histon
Cambridge CB4 9ZR
UK

Prof. Peter Willett
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello Street
Sheffield S1 4DP
UK

Mr. Egon Willighagen
Radboud University Nijmegen
Toernooiveld 1
NL-6525 ED Nijmegen
Netherlands

Dr. Gerhard Wolber
Inte:Ligand GmbH
Mariahilferstrasse 74B/11
1070 Vienna
Austria

Mr. David Wood
University of Sheffield
Room 323
Regents Court
Sheffield S10
UK

Dr. Matthew Wright
Barnard Chemical Information Ltd
32 Cookridge Drive
Leeds LS16 7LT
UK

Dr. Terry Wright
Elsevier MDL
14600 Catalina Street
San Leandro, CA 94530
USA

Mr. Hans-Peter Wrona-Metzinger
Schering AG
Computational Chemistry
Muellerstraße 178
13342 Berlin
Germany

Mr. Toru Yamaguchi
Yamaguchi University
Graduate School of Science and Engineering
16-02-2001, Tokiwadai
Ube, 755-8611
Japan

Dr. Litai Zhang
Bristol-Mayers Squibb
P. O. Box 4000
Princeton, NJ 8543
USA

Dr. Qingyou Zhang
Universidade Nova de Lisboa
Department of Chemistry
campus FCT-UNL
Caparica 2829-516
Portugal

Sponsoring Societies

- Division of Chemical Information (CINF), American Chemical Society (ACS)
- Chemistry-Information-Computer Division, Gesellschaft Deutscher Chemiker (Society of German Chemists) (GDCh)
- Division of Chemical Information and Computer Science of the Chemical Society of Japan (CSJ)
- Chemical Information Group, the Royal Society of Chemistry (RSC)
- Royal Netherlands Chemical Society (KNCV)
- The Chemical Structure Association Trust (CSA Trust)
- Swiss Chemical Society (SCS)

