Making “Real” Molecules in Virtual Space

György Pirok,*[,†] Nóra Máté,[†] Jeno Varga,[‡] József Szegezdi,[†] Miklós Vargyas,[†] Szilárd Dóránt,[†] and Ferenc Csizmadia[†]

ChemAxon Ltd., Máramaros köz 3/a, 1037 Budapest, Hungary, and Hungarian Academy of Science, Chemical Research Center, Institute of Biomolecular Chemistry, Budapest, H-1525 Hungary

Received September 3, 2005

Predicting “realistic” compounds of given chemical reactions with virtual synthesis tools usually requires the manual intervention of experienced chemists in the enumeration phase for the selection of appropriate reactants, assignment of the corresponding reaction sites, and removal of the unlikely products. To automate the virtual synthesis process, we have moved the expertise intensive parts from the compound library design phase to the reaction library design phase. ChemAxon is building an in silico reaction library containing important preparative transformations, where each reaction definition contains a generic transformation scheme and additional rules to handle the various starting compounds according to the corresponding chemo-, regio-, and stereoselectivity issues. Having well designed reaction definitions in hand, our software tool is able to generate synthetically feasible compound libraries with minimal effort in the enumeration phase.

INTRODUCTION

The first computer programs trying to predict the outcome of chemical reactions appeared in the 1960s—in a relatively early age of computational science—and were initialized by E. J. Corey’s pioneering work on retrosynthetic analysis[1−3] providing a systematic approach to synthesis planning.

The famous and mature LHASA[5−7] program (Logic and Heuristics Applied to Synthetic Analysis) is still actively being developed. Many other synthesis design or reaction prediction applications have been born since then including EROS and WODCA (Workbench for the Organization of Data for Chemical Applications) from the Gasteiger group,[8−11] SYNCHEM (Gelernter[12,13]), and SYNGEN (Hendrickson[14,15]). Either model-based or data-driven, virtual synthesis programs are still not used widely by preparative chemists.

One possible reason is the complexity of the area. Except for a few cases, the outcome of a reaction depends on many factors (temperature, pH, solvent, impurities, stirring, etc.), and it is sometimes hard to repeat even one’s own experiments. The mathematical foundation of chemical reactions is too complicated to solve, and there is no standard method to formulate the valuable knowledge of experienced synthetic chemists for computers. The design of a rational synthesis route usually needs an experienced and intuitive mind rather than a systematic one, and that is why it is sometimes considered as art, not just science.

GOALS

The importance of computer generated and analyzed libraries has increased recently, but current virtual synthesis based library enumerations usually require the control of an experienced preparative chemist for manually selecting reactive starting compounds, excluding molecules giving side reactions, assigning the corresponding reaction sites, and determining the expected main products. This method is not just labor intensive, but the chemical and synthetic quality of the resulting virtual library greatly depends on the experience and skills of the operating chemist.

We build a generic reaction library where the synthetic knowledge is encoded in the reactions themselves. The main advantage of moving the most “knowledge intensive” parts from the library enumeration phases to the reaction design phase is that it speeds up the entire enumeration process. Expertise is needed for specifying the reusable reactions, but a good reaction library can be used for creating chemically feasible products in various synthesis tasks with predictable synthetic quality. We do not expect perfect prediction of the outcome of chemical reactions, but the system should be able to process compounds according to the expectations of the chemists. Customizability is a very important issue for the harmonization of the virtual reactions with the experimental results.

CHEMICAL TERMS, THE LANGUAGE OF CHEMISTRY RULES

Since a simple reaction scheme can hardly describe most of the chemical prerequisites of a successful preparative reaction, we developed a language for describing the related chemical knowledge base. The syntax of the Chemical Terms[16] language is designed to be parsable by computer software programs as well as understandable by chemists. Its function list contains arithmetic and logic operators, substructure matching and similarity functions, and lots of property calculations (pK_a, log$P$, log$D$, partial charge distribution, Hückel localization energy, etc.). The available functions are integrated via an open plugin system providing a flexible entry point for permanent enhancements. Apart from reaction rules, user defined chemical expressions can be also used in other cheminformatics areas, such as pharmacophore screening, chemical searching, evolutionary drug design, or QSAR. The Chemical Terms language is an invaluable tool for filtering druglike, leadlike, or bioavailable compounds (Figure 1).

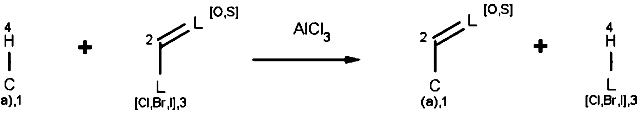
* Corresponding author e-mail: pirok@chemaxon.hu.
[†] ChemAxon Ltd.
[‡] Hungarian Academy of Science.

10.1021/ci050373p CCC: $33.50 © 2006 American Chemical Society
Published on Web 01/13/2006

```
(mass() <= 500) &&
(logP() <= 5) &&
(donorCount() <= 5) &&
(acceptorCount() <= 10);
```

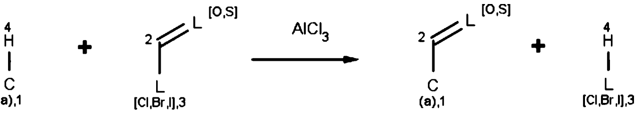
Figure 1. Lipinski's rule of five[17] in chemical terms.

Scheme 1. Generic Transformation Scheme of the Virtual Friedel−Crafts Acylation Reaction (drawn with MarvinSketch[21])[a]



[a] C(a) = aromatic carbon; L[Cl,Br,I] = chlorine, bromine, or iodine; L[O,S] = oxygen or sulfur atom.

Scheme 2. Refining the Description of Reactive Functional Groups in the Reactivity Rule



REACTIVITY: `charge(ratom(1), "aromaticsystem") < -0.2`

MODELING THE FRIEDEL-CRAFTS ACYLATION REACTION

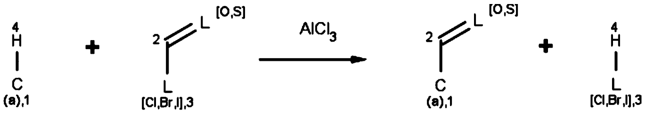
The way we encode synthetic knowledge in reaction definitions can be demonstrated by the example of the Friedel−Crafts acylation.[18−20] It is an electrophilic aromatic substitution reaction for the acylation of aromatic rings with the corresponding acid halides in the presence of aluminum chloride or another Lewis acid.

The Generic Reaction Scheme. The transformation of the functional groups during the reaction is described by a generic reaction scheme, which keeps the equation simple and applicable to various reactants. The hydrogen of an aromatic carbon atom is substituted by an acyl group. The corresponding atoms on the two sides of the reaction arrow are identified by map numbers. The reaction processing software understands some different mapping styles and has an automapper function as well. In the example below (Scheme 1), atoms having changing bonds are mapped. Stereospecific transformations such as inversion, retention, or double-bond alteration can also be marked on the reaction scheme.

The Reactivity Rule. If a reaction equation is generic and simple, it cannot describe the reactive sites well enough. The structural neighborhood and other required physicochemical properties of the reactive functional groups are specified in a Chemical Terms field called reactivity (Scheme 2). In a reaction context, any reactants or products can be referred from a Chemical Terms expression, and atoms can be easily identified by their corresponding map numbers. In our example case, Friedel−Crafts acylation does not occur if the aromatic system is deactivated (estimated from the sum of the partial charge values of the atoms in the aromatic ring system).

The Selectivity and Tolerance Rules. Reactive functional groups can be recognized now, but what should one do when a starting compound contains more than one reactive site? Main product(s) can be specified with the help of the selectivity and tolerance rules (Scheme 3). The expression in the selectivity rule is used to determine the most reactive region. In the case of our Friedel−Crafts acylation example, a Hückel calculation of the localization energy in the transition state helps to emulate the directing rules of the electrophilic aromatic substitution reactions.

Scheme 3. Determining the Main Product(s) with the Selectivity and Tolerance Rules

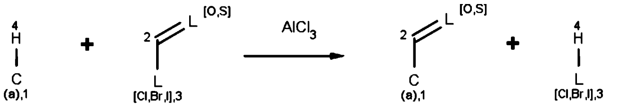


REACTIVITY: `charge(ratom(1), "aromaticsystem") < -0.2`

SELECTIVITY: `-energyE(ratom(1))`

TOLERANCE: `0.02`

Scheme 4. Complete Definition of the Virtual Friedel−Crafts Acylation Reaction Containing an Exclude Rule To Avoid Reactions with "Problematic" Compounds



REACTIVITY: `charge(ratom(1), "aromaticsystem") < -0.2`

SELECTIVITY: `-energyE(ratom(1))`

TOLERANCE: `0.02`

EXCLUDE:

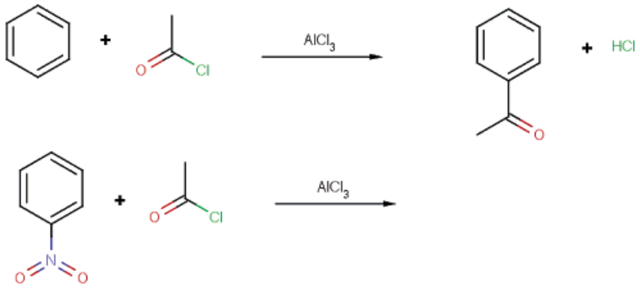
```
match(reactant(1), "[Cl,Br,I]C(=[O,S])C=C") ||
match(reactant(0), "[H][O,S]C=[O,S]") ||
match(reactant(0), "[P][H]") ||
(max(pka(reactant(0), filter(reactant(0),
"match('[O,S;H1]')"), "acidic")) > 14.5) ||
(max(pka(reactant(0), filter(reactant(0),
"match('[#7:1][H]', 1)"), "basic")) > 0)
```

The selectivity expression is calculated for all possible reaction sites. The reaction site having the greatest selectivity value leads to the main product. However, sometimes there can be more than one main product. If the selectivity difference of two sites is lower than the tolerance value, then both lead to main products.

The tolerance value of the Friedel−Crafts reaction allows the recognition when ortho and para regioisomers are both main products (the steric effect of the substituents is not considered in this example).

The Exclude Rule. The last field of a virtual reaction definition is for excluding molecules giving side reactions or destroying the catalyst (Scheme 4). We would like to avoid generating simple Friedel−Crafts products from acryloyl halides (reactant(1)), since it can result indanone derivatives owing to ring closure.[22] Aromatic reaction components (reactant(0)) having carboxylic groups and their thio-analogues can destroy the Lewis acid catalyst, so we exclude them. Furthermore, the aromatic components should not contain nucleophilic groups such as PH[23] or OH, SH compounds with a higher pK_a than 14.5, or NH compounds with a higher pK_b than 0, because they could easily be acylated on their nucleophilic groups instead of the aromatic ring. The pK_a calculation based rule excludes alcohols, thiols, amines, and anilines but permits the Friedel−Crafts acylation of phenols and many nitrogen heterocycles (see Chemical Terms documentation[16] for function details).

Scheme 5. Benzene Is Acylated, but the Deactivated Nitrobenzene Ring Is Not (Determined by the Reactivity Rule)



Finding out the appropriate rules for the reaction in question is often straightforward in practice. When a chemist is able to tell why the reaction goes a certain way and what are influencing chemical properties, his experience can be formulated in reaction rules (for example, Markovnikov and Zaitsev rules and Baeyer−Villiger oxidation). Sometimes a reaction rule is the result of a continuous refinement containing some intuitive steps (like in the case of the Friedel−Crafts reaction), and the values are set by a trial and error method. Please note that the actual values depend on the prediction software, and in some cases their applicability might be sensitive to the quality of the corresponding calculations.

In case of reactions where no reactivity and selectivity factors are known from the literature, the situation is more difficult. A computational method for creating the rules by the analysis of reaction databases would be very useful, but this is out of our focus at the moment. Our goal is only to process reactions according to the defined expectations of the chemists.

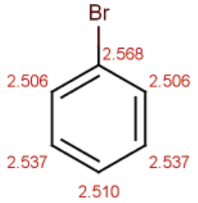


Figure 2. The calculated electrophilic localization energies of bromobenzene. They are the smallest in the ortho positions, but para is also within the tolerance range.

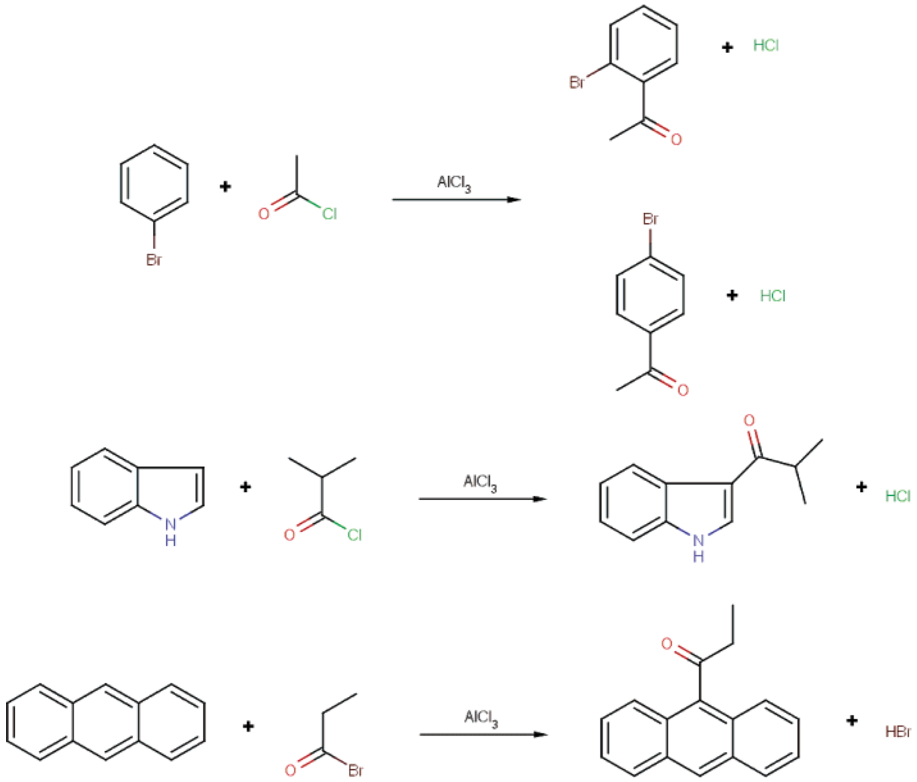
REACTOR, THE ENUMERATION ENGINE

We developed a Java-based software tool called Reactor[24,25] to process reactions according to simple reaction schemes and also those including rules defined in the above way. If a reaction is well designed, no further preparation is needed, large compound libraries can be enumerated, and Reactor will process reactive starting compounds taking care of the corresponding chemo-, regio-, and stereoselectivity issues. The effect of the rules on the virtual Friedel−Crafts acylation is shown in the next few examples.

Benzene is acylated, but the deactivated nitrobenzene ring is not (Scheme 5), which is determined by the reactivity rule (Scheme 2). Since acyl groups are also deactivating, the Friedel−Crafts acylation is well controlled, and no further acylation occurs on the same ring.

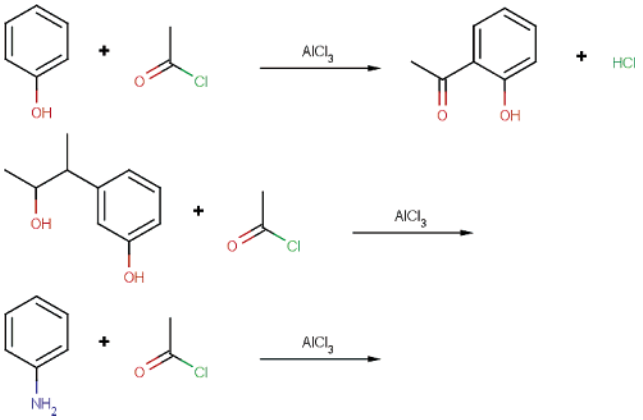
Although, halogen substituents also deactivate the aromatic ring, their effect is weaker than that of nitro or acyl groups. Bromobenzene can undergo the Friedel−Crafts acylation reaction, and the bromine ligand directs the acyl group to ortho and para positions (Scheme 6). According to the selectivity and tolerance rules (Scheme 3), the electrophilic localization energies are calculated for each possible reaction site, and those with the smallest ones lead to the main product

Scheme 6. Halogens Are Ortho/Para Directing, Meta Isomer Is Not Produced from Bromobenzene[a]



[a] The main products from heteroaromatic and fused ring systems can also be predicted by the selectivity and tolerance rules.

Scheme 7. Exclude Rule Disclosed Reagents Containing Nucleophilic Functional Groups Causing Side Reactions



(Figure 2). The selectivity rule worked out for the Friedel–Crafts acylation reaction can be applied to other aromatic electrophilic substitutions as well (bromination, nitration) to determine the main products. We plan to enhance the regioselectivity prediction by including a new function to consider the steric effects.

The exclude rule (Scheme 4) uses the pK_a calculation to avoid generating false Friedel–Crafts products. Phenols react, but compounds containing alcoholic groups are excluded. Indoles and amides are processed, while reactants having amino groups are not (Scheme 7).

Reactor has runtime options to ignore the rules defined in a reaction allowing the examination of their effects or the enumeration of all possible reaction products. Reactor supports the reagent combination, and huge combinatorial libraries can be built in a single batch process. We processed a reaction of several hundred isocyanates and isothiocyanates with almost 3000 nucleophiles such as alcohols, thiols, and amines (Figure 3). More than 1.2 million products were enumerated in 1.5 h (P4, 1.8 GHz, 256 MB RAM).

We have built Synthesizer, an application of the Reactor engine which is able to handle complex reaction graphs. The reaction and compound dispatching modules make Synthesizer a versatile tool. The most straightforward application is the multistep combinatorial synthesis. We followed the steps of a "real" combinatorial synthesis[26,27] with Synthesizer to produce a half million member library (Figure 4) in a database in 2.5 h (P4 1.8 GHz, 256 MB).

Additional rules can be assigned to each synthesis step to define custom conditions depending on the current synthetic goals. It is possible to exclude compounds containing certain functional groups or having unpreferred physicochemical

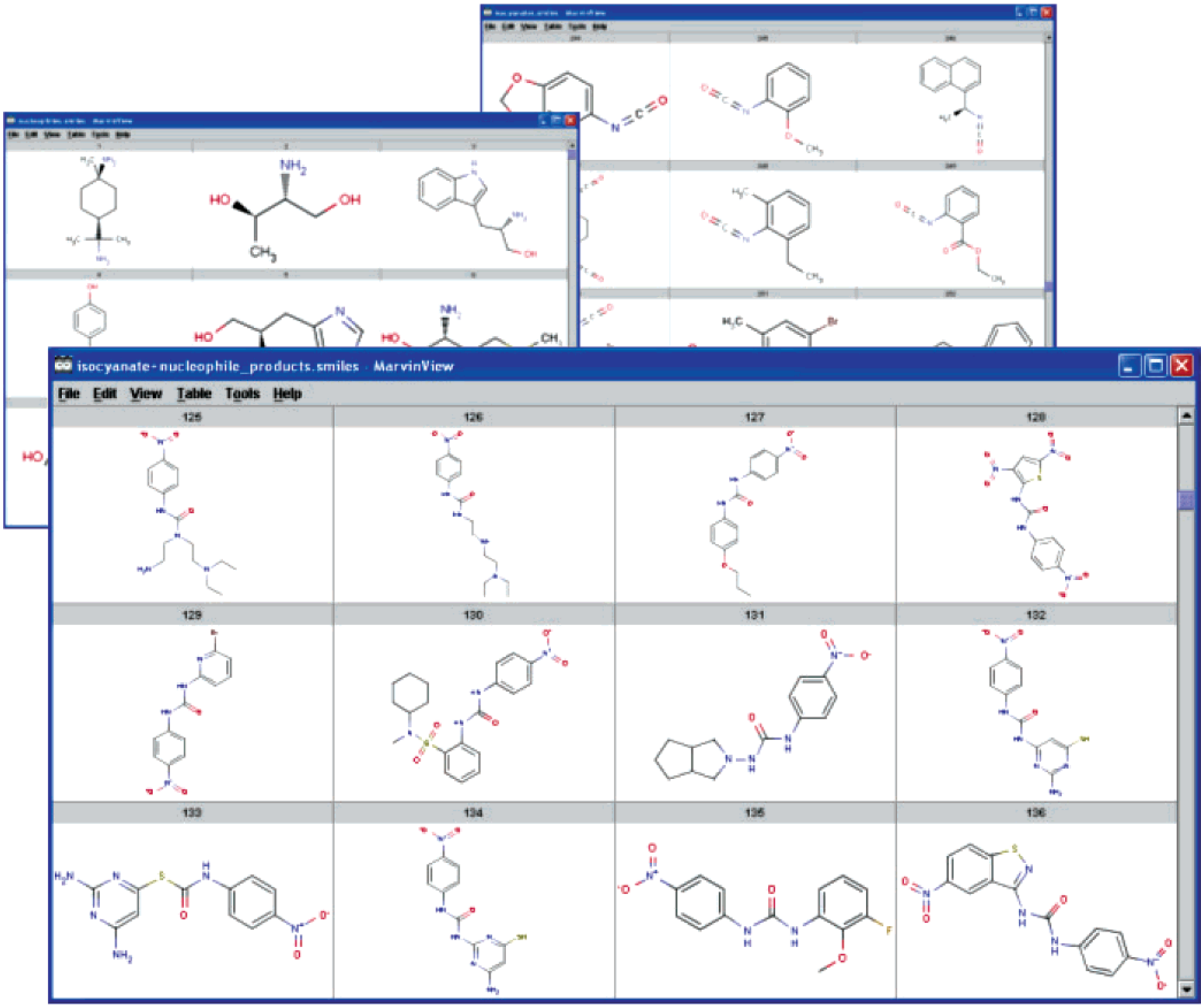
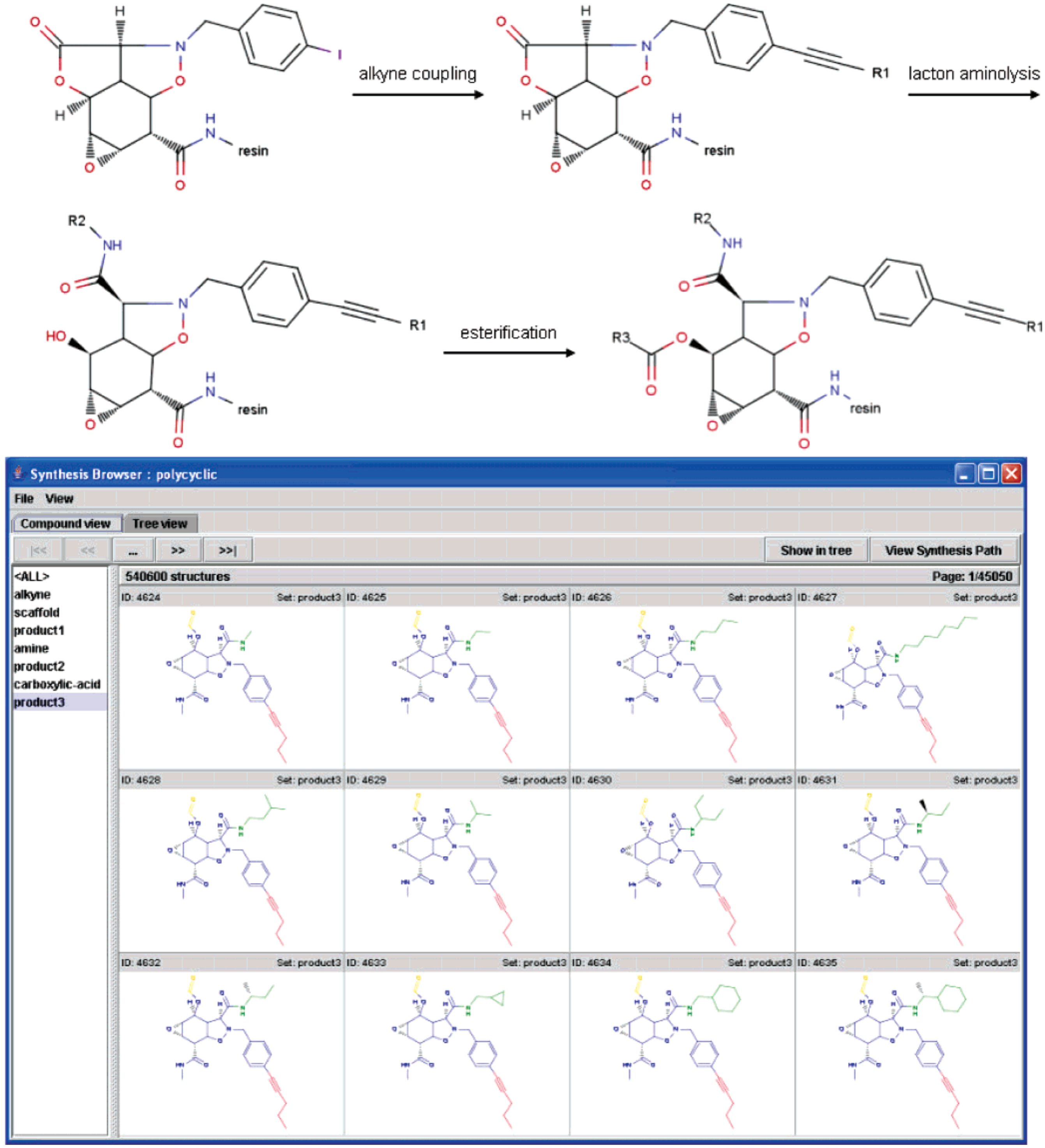


Figure 3. The reactants and the products of an in silico combinatorial reaction.



<ALL>
alkyne
scaffold
product1
amine
product2
carboxylic-acid
product3

540600 structures Page: 1/45050

ID: 4624 Set: product3
ID: 4625 Set: product3
ID: 4626 Set: product3
ID: 4627 Set: product3
ID: 4628 Set: product3
ID: 4629 Set: product3
ID: 4630 Set: product3
ID: 4631 Set: product3
ID: 4632 Set: product3
ID: 4633 Set: product3
ID: 4634 Set: product3
ID: 4635 Set: product3

Figure 4. Three step virtual synthesis of a combinatorial library.

properties even if they could be synthesized. Manual reagent or product selection is still not necessary. This way, the custom requirements do not influence the reaction definitions, so the reactions remain chemically intact and can be reused in further library enumerations.

Combinatorial synthesis is not the only use of the Reactor technology. We developed other reaction dispatching modules, one of which can be used for modeling the metabolic biodegradation of small molecules.[28] Another interesting application can build a large and diverse compound space of molecules which could be synthesized from available chemicals. These applications will be introduced in future publications.

FUTURE DEVELOPMENTS

We will extend the ChemAxon Reaction Library to include more preparative reactions useful for small molecule synthesis (about 70 reactions are currently available).

To improve the quality of the reactions, we plan to develop an automatic validation system, which will compare the processed virtual reactions to those published in reaction databases. This way, we will be able to recognize problematic cases and also to assign a credibility value to each reaction.

We plan to support multistep reaction schemes containing transition states and intermediates enabling more accurate predictions.

All reactions contain additional information about the basic literature references, general recipes for the preparation, and lots of technical details valuable for scoring (yield range, reaction time, difficulty, reproducibility, generality, environmental hazard, etc.) which can form the basis of a future retrosynthesis application.

CONCLUSIONS

We introduced the basic elements of our virtual synthesis technology: the Chemical Terms language, the ChemAxon Reaction Library, and the Reactor engine. These components, the corresponding API,[29] and some applications are currently available for developers to create their own custom solutions.

IMPLEMENTATION

All software applications mentioned above are parts of ChemAxon's JChem software suite[30] (100% Java). Physicochemical calculations are provided by ChemAxon's prediction plugins.[31] Structures and reactions were designed with MarvinSketch.[21] Hardware and software requirements are as follows: any system running Java Runtime Environment 1.4 or above.

REFERENCES AND NOTES

(1) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14*, 19−37.

(2) Corey, E. J. Computer-assisted Analysis of Complex Synthetic Problems. *Quart. Rev. Chem. Soc.* **1971**, *25*, 455−482.

(3) Corey, E. J.; Cheng, X. M. *The Logic of Chemical Synthesis*; John Wiley and Sons: New York, 1989.

(4) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178−192.

(5) Corey, E. J.; Cramer, R. D.; Howe, W. J. Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates. *J. Am. Chem. Soc.* **1972**, *94*, 440−459.

(6) Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-Assisted Analysis of Organic Synthesis. *Science* **1985**, *228*, 408−418.

(7) http://lhasa.harvard.edu.

(8) Gasteiger, J.; Ihlenfeldt, W. D.; Fick, R.; Rose, J. R. Similarity Concepts for the Planning of Organic Reactions and Syntheses. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 700−712.

(9) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. A collection of computer methods for the synthesis design and reaction prediction. *Recl. Trav. Chim. Pays-Bas.* **1992**, *111*, 270−290.

(10) Ihlenfeldt, W. D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613−2633.

(11) http://www2.chemie.uni-erlangen.de/software/wodca.

(12) Gelernter, H. K.; Sanders, A. F.; Larsen, D. L.; Agarwal, K. K.; Boivie, R. H.; Spritzer, G. A.; Searleman, J. E. Empirical explorations of SYNCHEM. *Science* **1977**, *19*, 1041−1049.

(13) Krebsbach, D.; Gelernter, H. K.; Sieburth, S. McN. Distributed Heuristic Synthesis Search. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 595−604.

(14) Hendrickson, J. B.; Toczko, A. G. Systematic synthesis design: the SYNGEN program. *Pure Appl. Chem.* **1989**, *61*, 589−592.

(15) http://syngen2.chem.brandeis.edu/syngen.html.

(16) http://www.chemaxon.com/jchem/doc/user/EvaluatorTables.html.

(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.

(18) Friedel, C.; Crafts, J. M. A new general method for the synthesis of hydrocarbons, acetones, etc. *Compt. Rend.* **1877**, *84*, 1450.

(19) Price, C. C. The Alkylation of Aromatic Compounds by the Friedel−Crafts Method. *Org. React.* **1946**, *3*, 1−82.

(20) Gore, P. The Friedel−Crafts Acylation Reaction and its Application to Polycyclic Aromatic Hydrocarbons. *Chem. Rev.* **1955**, *55*, 229−281.

(21) http://www.chemaxon.com/marvin/.

(22) Hart, R. T.; Tebbe, R. F. Acylation-Alkylation Studies. *J. Am. Chem. Soc.* **1950**, *72*, 1950, 3286.

(23) Alberts H.; Künzel W.; Schuler W. Acylated arsine and phosphine derivatives and isoarsiles. *Chem. Ber.* **1952**, *85*, 239−248.

(24) Bode, J. W. Computer Software Reviews − Reactor. *J. Am. Chem. Soc.* **2004**, *126*, 15317.

(25) http://www.chemaxon.com/jchem/doc/user/Reactor.html.

(26) Schreiber, S. L.; Tan, D. S.; Foley, M A.; Shair, M. D. Stereoselective Synthesis of over Two Million Compounds Having Structural Features Both Reminiscent of Natural Products and Compatible with Miniaturized Cell-Based Assays. *J.Am. Chem. Soc.* **1998**, *120*, 8565−8566.

(27) Schreiber, S. L.; Tan, D. S.; Foley, M. A.; Stockwell, B. R.; Shair, M. D. Synthesis and Preliminary Evaluation of a Library of Polycyclic Small Molecules for Use in Chemical Genetic Assays. *J. Am. Chem. Soc*. **1999**, *121*, 9073−9087.

(28) http://umbbd.ahc.umn.edu.

(29) http://www.chemaxon.com/jchem/doc/api/.

(30) http://www.chemaxon.com.

(31) http://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html.

CI050373P