Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups

Peter Ertl*

Novartis Pharma AG, Molecular Simulation Group, WKL-125.14.20, CH-4002 Basel, Switzerland

Received August 14, 2002

A large set of more than 3 million molecules was processed to find all the organic substituents contained in the set and to identify the most common ones. During the analysis, 849 574 unique substituents were found. Extrapolated to the number of known organic molecules, this result suggests that about 3.1 million substituents are known. Based on these findings the size of virtual organic chemistry space accessible using currently known synthetic methods is estimated to be between 10^{20} and 10^{24} molecules. The extracted substituents were characterized by calculated electronic, hydrophobic, steric, and hydrogen bonding properties as well as by the drug-likeness index. Various possible applications of such a large database of drug-like substituents characterized by calculated properties are discussed and illustrated by reference to a Web-based tool for automatic identification of bioisosteric groups.

INTRODUCTION

The concept of organic substituents and their influence on molecular properties is one of the pillars of modern organic chemistry as well as a basis of QSAR analysis. Prominent examples from this area are the pioneering contribution of Hammett with regard to the effect of substituents on reactivity (including the introduction of substituent σ constants characterizing the electron donating and accepting power of substituents), the reaction mechanism theory of Ingold (including the definition of mesomeric and inductive effects of substituents), the QSAR concept of Hansch (including the definition of substituent hydrophobicity π constants and the application of substituent constants in QSAR[1]), and the ideas of Craig and Topliss concerning the influence of substituent properties on biological activity applied in early drug design.[2,3] The whole combinatorial chemistry is also based on the concept of substituents, acting in this case under the name building blocks.

The present study focuses on organic substituents from the point of view of cheminformatics and tries to answer questions about the total number of substituents in known organic chemistry space and the implications of this number for the size of virtual organic chemistry space. The characterization of substituents by calculated properties is also discussed, including a procedure for calculating substituent drug-likeness based on a comparison of the distribution of substituents in a large database of drugs versus a large database of nondrugs. Finally an example of the application of a large database of drug-like substituents with calculated properties is presented—a Web-based tool for automatic identification of bioisosteric groups.

* Corresponding author phone: ++41 61 6967413; fax: ++41 61 6964069; e-mail: peter.ertl@pharma.novartis.com.

IDENTIFICATION OF MOST COMMON SUBSTITUENTS, THE NUMBER OF KNOWN ORGANIC SUBSTITUENTS

To estimate the number of substituents and identify the most common organic substituents, we processed a large in-house database of commercially available molecules containing 3 043 941 entries. Structures stored as SMILES strings were processed, and all substituents up to 12 non-hydrogen atoms were extracted. Any group of atoms connected by a single chemically activated (breakable) nonring bond to the rest of the molecule was considered a "substituent". Charged groups were neutralized when possible. Only substituents containing "organic" atoms (including Si, P, and Se) were collected. The whole molecular manipulation and processing was done using an in-house molecular development kit written in Java.

Analysis of the molecular database yielded 849 574 unique substituents. The substituents with the highest frequency in the database are shown in Figure 1. The size of the database processed is large enough for generalization, so we can claim that substituents displayed in this figure are the most common organic substituents.

Out of all 849 574 substituents found, however, only a very small fraction, namely 50, may be considered to be truly "common", i.e., present in more than 1% of all molecules in the database, while 438 substituents are present in more than 0.1% of all molecules. About 64% of all substituents found are singletons (i.e., they are present only in a single molecule in the whole database processed). A similar frequency distribution of fragments in molecules from the Derwent World Drug Index was also obtained in a study by Lewell et al.[4]

Figure 2 shows the dependence of the total number of substituents identified on the number of molecules processed

10.1021/ci0255782 CCC: $25.00
Published on Web 12/06/2002

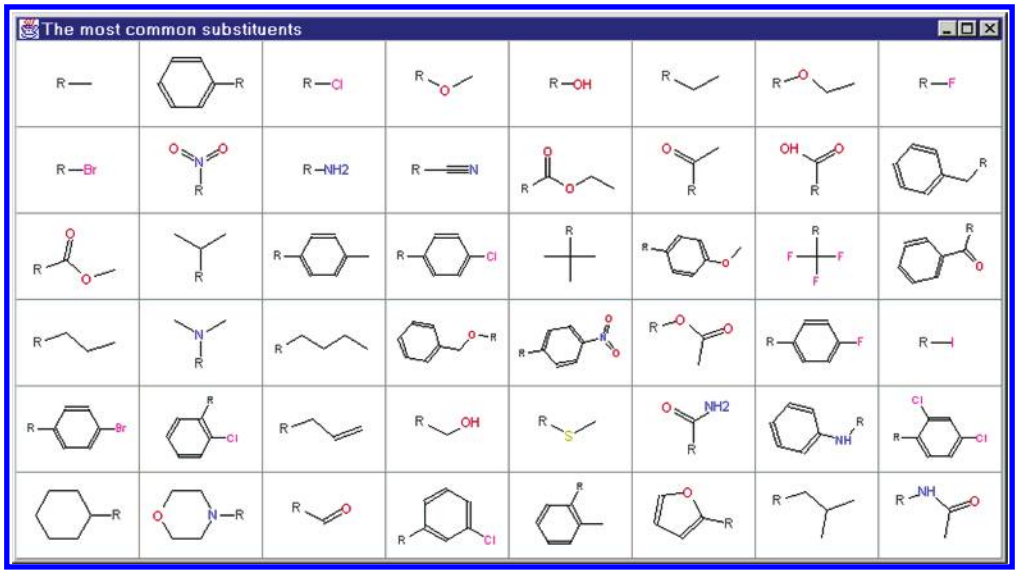

Figure 1. The most common substituents identified after processing a database of 3 043 941 molecules.

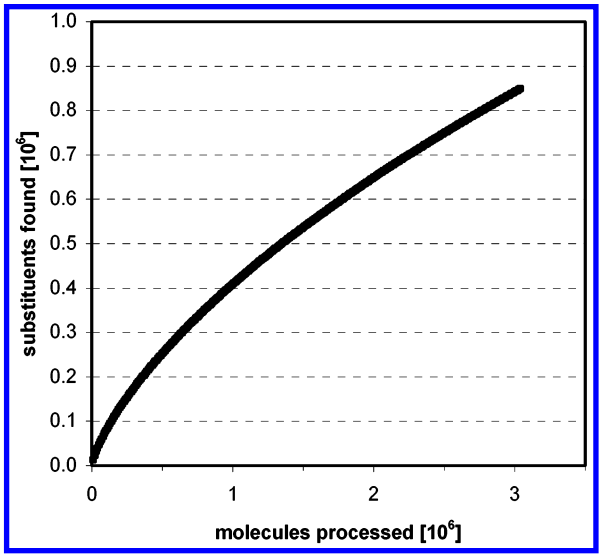

Figure 2. Dependence of the number of unique substituents found on the number of molecules processed.

(this ratio was recorded after processing of every 10 000 molecules). The form of this function was a slight surprise to us. Before the study, we expected a curve that would reach a plateau or limit after a certain number of molecules had been processed. But this was not the case: a steady growth was seen in the number of substituents found. This was even more clear when the two axes were transformed to a logarithmic scale (not shown). In this case, the dependence is clearly linear with a correlation coefficient of $r^2 = 0.998$ for the 304 points recorded. This proves that the diversity of organic chemistry is enormous, and even a relatively small unit consisting of 12 organic atoms allows so many combinations that current organic chemistry is still very far from reaching its limits. Similar conclusions about practically unlimited growth have also been made in studies on the number of types of organic reactions in a large reaction database.[5]

Using the information shown in Figure 2, one can extrapolate to the total size of known organic molecules (about 19 million). According to this analysis, the total number of organic substituents (up to 12 atoms) known in organic chemistry is about 3.1 million. This number of course depends on the size of the substituent: if substituents of up to 15 atoms were considered, the result would be 7.8 million.

SIZE OF VIRTUAL ORGANIC CHEMISTRY SPACE

Using the information derived in the previous section, one can speculate about the total size of virtual organic chemistry space which is accessible using currently known synthetic methods. To estimate this number, we used a very simple computational experiment, namely to determine how many molecules of the general formula R_1-X-R_2 may be constructed. R_1 and R_2 in this formula are substituents, and X is a scaffold with two attachment points. We analyzed our database of 3 million molecules using the same approach as described in the previous section to identify also the number of double-bonded scaffolds. When using 12-atomic building blocks, the number of structures which may be constructed[6] by using the above-mentioned model is 5.2×10^{19}; when using 15-atomic units, this number increases[6] to 8.0×10^{20}. When using a slightly more complex formula R_1-X(-R_2)-R_3 and 9-atomic fragments (to limit the maximum size of products to 36 atoms), then the number of molecules which may be constructed[6] by this approach is 6.7×10^{23}.

One may argue that many substituents used in this analysis are themselves composed of smaller building blocks/fragments. We therefore decomposed our set of 849 574 substituents in a recursive manner by using the same set of fragmentation rules as described previously, until no further fragmentation was possible. This analysis provided 62 975 "elementary" fragments, the most frequent of which are shown in Figure 3. About 35% of these fragments are monovalent, 34% bivalent, 18% trivalent, and 13% have more than 3 connection points. When constructing molecules by combining these "elementary" fragments, one can easily get more than 10^{100} structures, but most of them would be nonrealistic (not stable, or synthetically inaccessible, at least by currently known synthetic methods).

We therefore estimate the number of molecules which may be prepared by currently known synthetic methods to be somewhere between 10^{20} and 10^{24}. We are aware that there are several deficiencies in this simplified approach (for

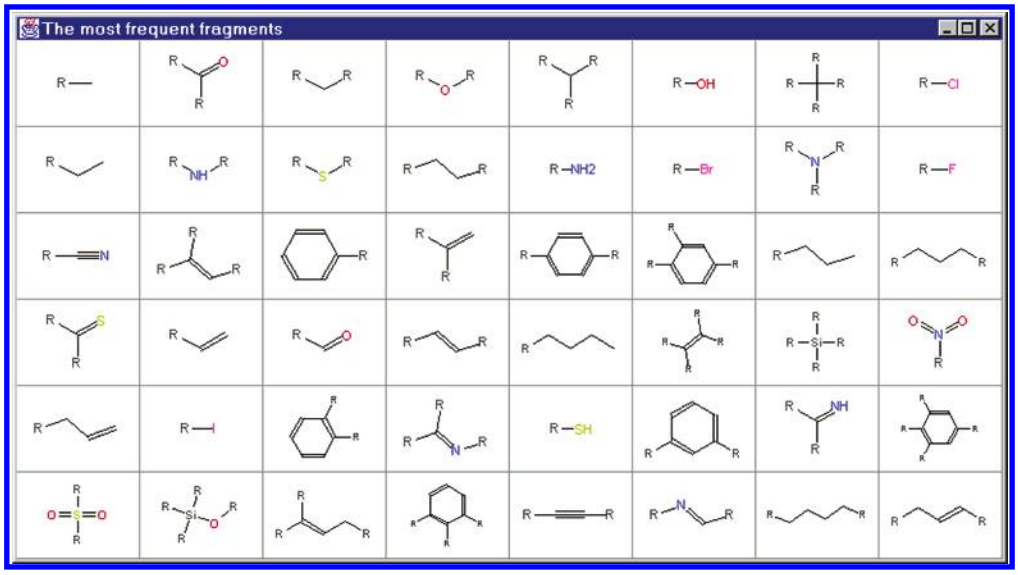

Figure 3. The most common fragments obtained by decomposition of 849 574 substituents.

example neglect of symmetry of the X scaffolds). Nevertheless, we believe the results presented here may be a useful approximation of the overall size of virtual organic chemistry space.

CALCULATION OF SUBSTITUENT PROPERTIES

An analysis of the substituents described hitherto produced a large database of about 850 000 substituents. To use this database in practical applications, the substituents need to be classified in some way. For applications in drug design, a classification based on substituent physicochemical properties is more useful than a simple classification based on structural similarity. In this study, we used substituent size, hydrophobicity, electronic properties (donating/accepting power), and hydrogen bonding characteristics. Steric substituent properties were characterized simply by topological size (number of non-hydrogen atoms) and maximum topological length. Substituent hydrophobicity (π constant) was calculated using the Ghose-Crippen atom contribution method.[7] The interdependence between calculated and experimental π constants for 256 common substituents (Figure 4) is quite acceptable. For the calculation of substituent electronic properties (Hammett σ constant) we used methodology based on quantum chemical parameters developed in-house. Since the method has been described in detail in the original paper,[8] the description here is confined to its basic features. The method is based on attaching a small model system (probe) to the substituent. Atomic charges on the probe are then calculated using a semiempirical quantum chemical method. The charges are of course influenced by the attracting or donating power of the substituent. In a training process these calculated charges were correlated with experimental σ constants for a set of substituents. Various probes, semiempirical Hamiltonians, and calculation protocols were tested to find a model with the best predictive power. Correlation between experimental and calculated σ constants for 368 substituents is shown in Figure 5. The correlation in this case is influenced by a low reliability of experimental σ constants for substituents that are not so common.[8] And as a final parameter, hydrogen bonding substituent properties were generated by a simple approach taking into account the topological distance of the hydrogen bond donor and acceptor centers from the substituent root.

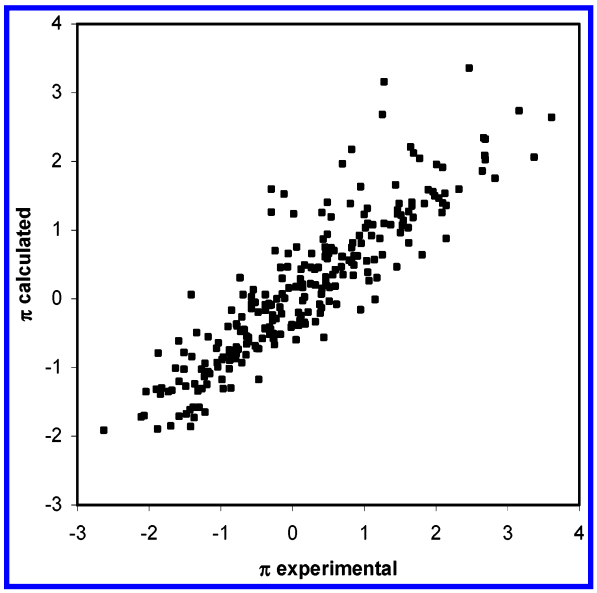

Figure 4. Correlation between calculated and experimental hydrophobicity π constants for 256 substituents.

A two-dimensional view (hydrophobicity versus electronic properties) of this four-dimensional substituent property space for more than 850 000 substituents is shown in Figure 6. This graph is actually nothing more than a well-known Craig plot[2] with the number of points, however, slightly larger than usual.

MOST COMMON SUBSTITUENTS IN DRUGS AND CALCULATION OF SUBSTITUENT DRUG-LIKENESS

To use the substituent database in drug design, not only substituent properties are important but also a characteristic known as drug-likeness or drugability. This may be defined as an ability of the substituent to generally increase the probability that its parent molecule is a drug. To identify the most common substituents in drugs, we performed an analysis of molecules in the World Drug Index database.[9] The most common substituents identified by this procedure are shown in Figure 7. A similar analysis of substituents for

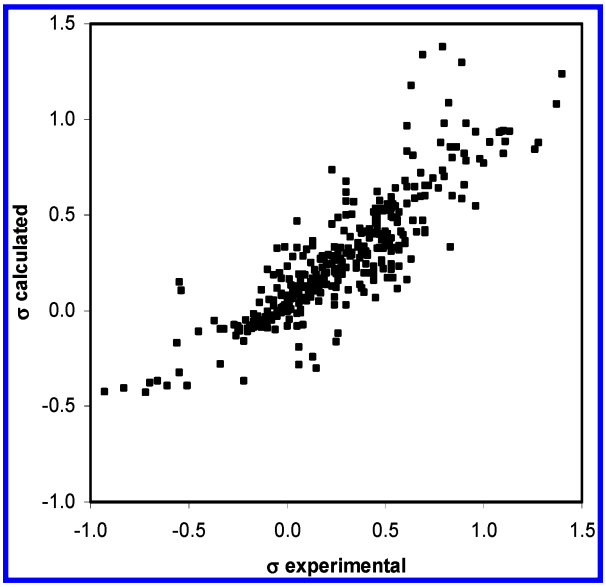

Figure 5. Correlation between calculated and experimental Hammett σ constants for 368 substituents.

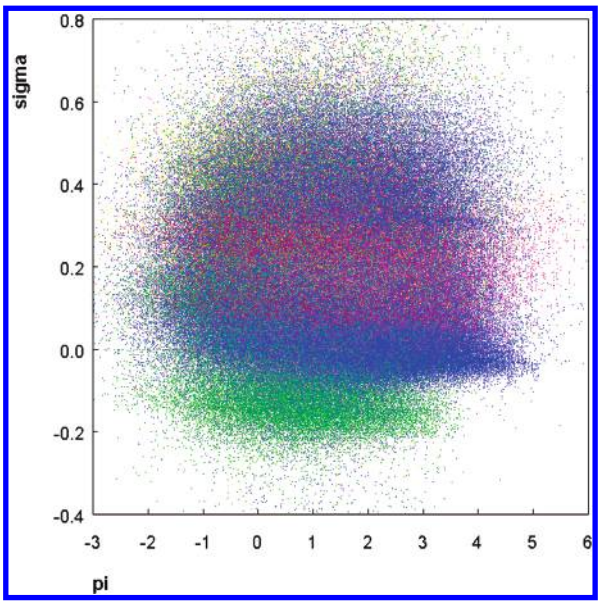

Figure 6. Distribution of calculated hydrophobic (π constant) and electronic (σ constants) properties for about 850 000 substituents (points representing aliphatic carbon substituents are blue, aromatic carbon ones violet, nitrogen green, oxygen red, and sulfur yellow).

5090 molecules from the Comprehensive Medicinal Chemistry database was published recently[10] and despite a slight difference in the definition of "sidechains" used in this study and definition of substituents in our analysis, the results of these two studies are very similar.

A comparison of most common substituents found in drugs (Figure 7) and most common substituents identified in average organic molecules (Figure 1) reveals quite a marked overlap between these two sets. For example, out of 500 most common substituents in both sets 243 are common. We also analyzed a large database of molecules with toxic effects, and the most common substituents identified here (not shown) were again similar to those presented in Figures 1 and 7. These functionalities are simply the most common organic substituents and will be found at the top after analyzing any large collection of molecules. The list of substituents presented in Figure 7 is therefore not the information needed when looking for substituents with drug-like properties. To identify substituents which beneficially influence the drug-likeness of their parent molecules one needs to identify substituents with *relatively* higher distribution in drugs than in nondrugs. A similar approach of identifying the substructure motifs present more frequently within a specific therapeutic class than in the whole World Drug Index has been described by Lewell et al.[4]

To find substituents with the highest drug likeness, we compared the distribution of substituents in two large databases of molecules: the database of nondrugs (or rather the reference set of average organic molecules) used in our earlier studies (about 3 000 000 molecules) and a subset of the World Drug Index.[9] The World Drug Index database was preprocessed by removing various "helper" molecules (imaging, radioprotecting, or dental agents, etc) as well as organometallic structures to leave 36 482 molecules which may be termed "true drugs".

Substituent drug-likeness index was calculated according to the following formula

$$f_i = \log(nact_i/ninact_i * ninact_{total}/nact_{total})$$

where $nact_i$ is the number of drugs which contain substituent i, $ninact_i$ is the number of nondrugs which contain substituent i, $nact_{total}$ is the total number of drugs, and $ninact_{total}$ is the total number of nondrugs processed.

In several in-house studies we compared the performance of various indices for separating drugs and nondrugs, and the simple index described above provided the best results. This index ranked very well also in the analysis of Ormerod et al.[11] Statistical significance of results was checked by contingency analysis at the 95% level. This ensured that rare substituents with possibly random distribution were not considered in the analysis.

Examples of substituents with high drug-likeness are shown in Figure 8. These substituents (or their derivatives in the proper reactive state) may be used as reagents in drug optimization processes or as building blocks in the design of drug-like combinatorial libraries. At the same time, our analysis also provided a list of substituents with a high negative drug-likeness index, examples of which are shown in Figure 9. The statistical analysis, of course, does not provide any reasons for these groups to be in the "bad" list. They have been identified solely based on the fact that they are present with significantly lower frequency in drugs than in average organic molecules. Therefore groups in this list should be checked manually to identify substructures with potential toxic effects, too reactive groups or metabolic weakpoints. Such groups should be avoided in synthesis planning, because products containing them would very probably be molecules with poor drugability. Such "negative lists" may also be used as exclusion criteria in virtual screening.

A WEB-TOOL FOR AUTOMATIC IDENTIFICATION OF BIOISOSTERIC SUBSTITUENTS

A large database of drug-like substituents/building blocks characterized by calculated properties may be used with advantage in various drug design applications, ranging from "classical" use in QSAR studies or molecular design by hand, through selection of building blocks for synthesis of com-

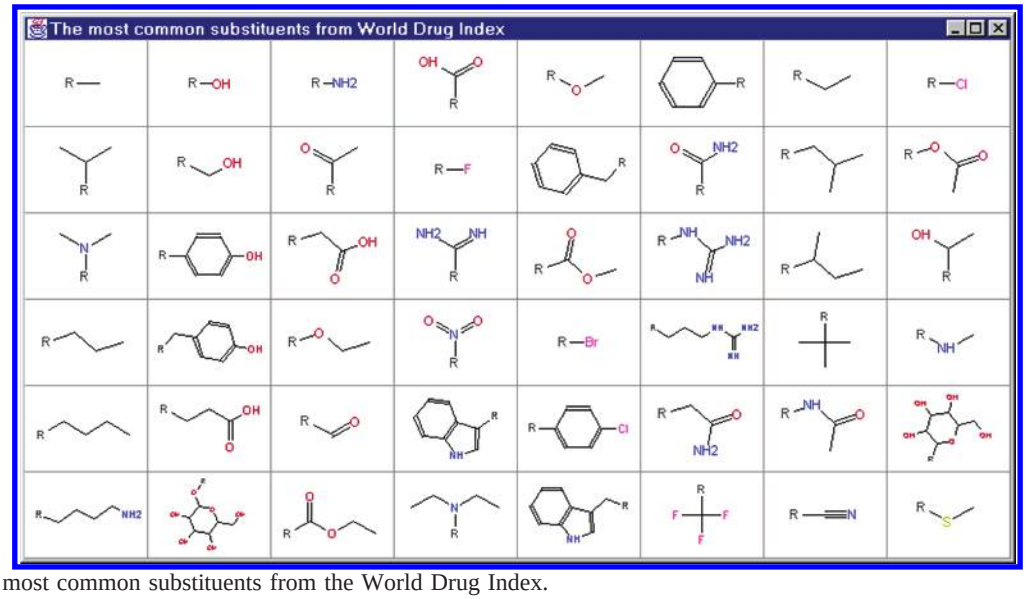

Figure 7. The most common substituents from the World Drug Index.

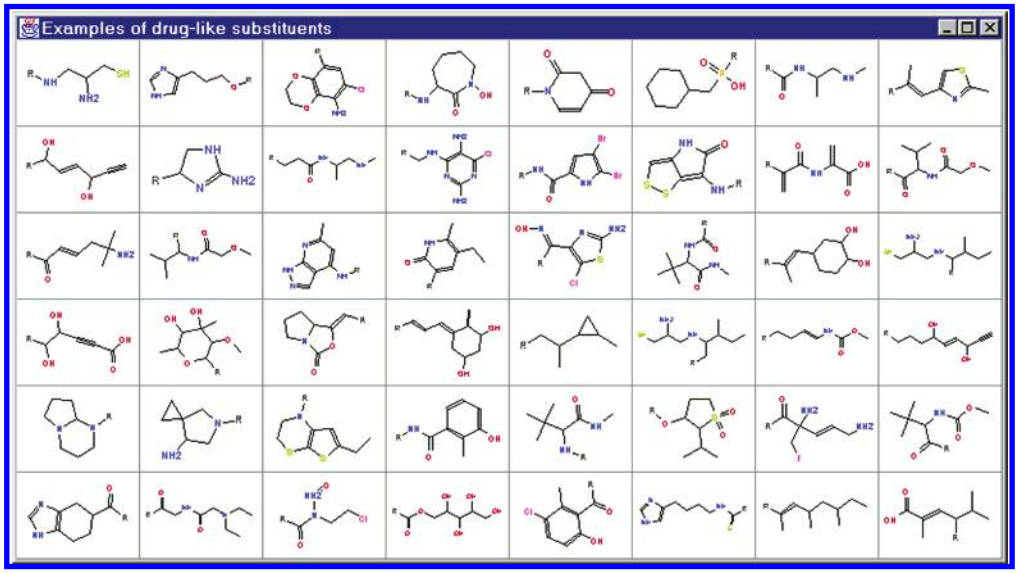

Figure 8. Examples of substituents with the highest drug-likeness index.

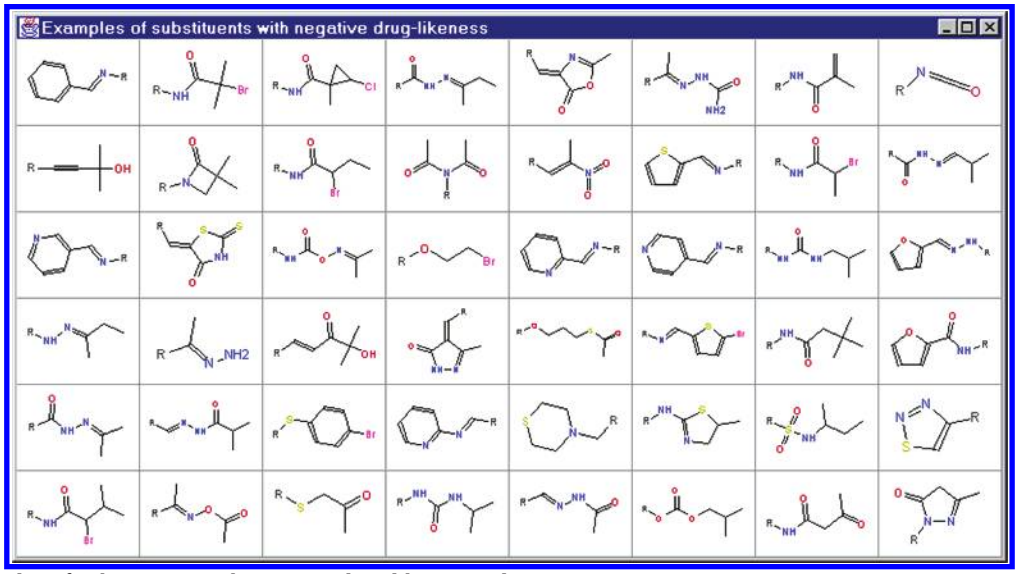

Figure 9. Examples of substituents with negative drug-likeness index.

binatorial libraries with desired characteristics, to automatic design of molecules with optimal properties by evolutionary algorithm. As an example of such applications, a Web tool for automatic identification of bioisosteric functional groups,

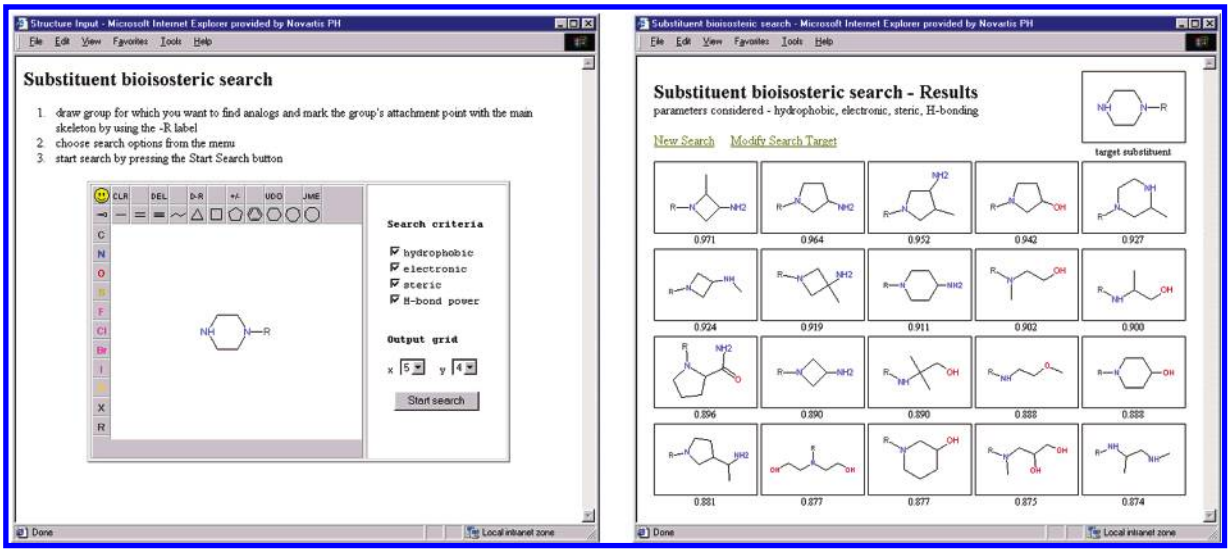

Figure 10. Web-based system for identification of bioisosteric substituents.

which is installed on the Novartis cheminformatics intranet,[12−14] is presented here.

Bioisosteric replacement is one of the standard techniques used in medicinal chemistry.[15] It can be defined as the replacement of a functional group in a bioactive molecule by another functionality of similar size and physicochemical properties. Bioisosteric transformations are used in the process of drug optimization to improve the properties of drug candidates, such as bioactivity, selectivity, and transport characteristics, and to remove unwanted side effects such as toxicity or an excessively rapid metabolism. Bioisosteric replacement is also used to design molecules more easily to synthesize or avoid patented structural features. Classical textbook examples of bioisosterically equivalent groups are phenyl and thiophenyl or carboxylic group and tetrazole. To find bioisosteric analogues of more complex groups, however, is not so easy. It requires a lot of experience, and even then the identification of a group with an optimal balance of steric, hydrophobic, electronic, and hydrogen-bonding properties, all of which influence ligand−receptor interactions, usually calls for a demanding procedure of trial and error.

That is the reason that a tool for automatic identification of substituent and linkers bioisosteric with the given target was developed. The tool presented here is an updated and modernized version of a Web application briefly presented earlier.[16] The heart of the system is a database of about 10 000 drug-like substituents characterized by calculated hydrophobic, electronic, steric, and hydrogen-bonding properties. When performing a search, the substituent for which analogues are to be identified must be entered into the system. This may be done easily with help of a Java molecular editor applet[17] incorporated directly into the Web page (Figure 10). After search options are chosen, the search is started. Substituents having properties compatible with those of the target are identified in the substituent database and displayed. Such lists of bioisosteric groups may serve as an "ideas generator" for bench chemists, helping them to design new bioactive molecules. The advantage of this approach is that it facilitates the identification not only of traditional analogues but also of nonclassical bioisosteric analogues, i.e., groups which are structurally different but are compatible in terms of their physicochemical properties.

ACKNOWLEDGMENT

I thank my colleagues Paul Selzer and Bernhard Rohde for many stimulating scientific discussions concerning all areas of cheminformatics and Sigmar Dressler for making the database of 3 million molecules available.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(2) Craig, P. N. Interdependence between Physical Parameters and Selection of Substituent Groups for Correlation Studies. *J. Med. Chem.* **1971**, *14*, 680−684.

(3) Topliss, J. G. Utilization of Operational Schemes for Analogue Synthesis in Drug Design. *J. Med. Chem.* **1972**, *15*, 1006−1011.

(4) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. N. 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.

(5) Rohde, B. Reaction type informetrics of chemical reaction databases: How "large" is chemistry? In *Further Advances in Chemical Information*; Royal Soc. Chem.: 1994; pp 109−127.

(6) The result 5.2×10^{19} was obtained simply by multiplying 5.4×10^6 (number of bivalent scaffolds up to 12 atoms) and two times 3.1×10^6 (number of substituent up to 12 atoms), similarly also for 15 atomic groups ($8.0 \times 10^{20} = 13.2 \times 10^6 * 7.8 \times 10^6 * 7.8 \times 10^6$) and for R1X(R2)R3 molecules ($6.7 \times 10^{23} = 3.1 \times 10^6 * 0.6 \times 10^6 * 0.6 \times 10^6 * 0.6 \times 10^6$).

(7) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(8) Ertl, P. Simple Quantum Chemical Parameters as an Alternative to the Hammett Sigma Constants in QSAR Studies. *Quant. Struct.-Act. Relat.* **1997**, *16*, 377−382.

(9) Derwent World Drug Index WDI 2001.2, published by Derwent Information.

(10) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.

(11) Ormerod, A.; Willett, P.; Bawden, D. Comparison of Fragment Weighting Schemes for Substructural Analysis. *Quant. Struct.-Act. Relat.* **1989**, *8*, 115−129.

(12) Ertl, P.; Jacob, O. WWW-based Chemical Information System. *J. Mol. Struct. (THEOCHEM)* **1997**, *419*, 113−120.
(13) Ertl, P. QSAR Analysis through the World-Wide Web. *Chimia* **1998**, *52*, 673−677.
(14) Ertl, P.; Miltz, W.; Rohde, B.; Selzer, P. Web-based Cheminformatics for Bench Chemists. *Drug Discovery World* **Fall 2000**, 45−50.
(15) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *95*, 3147−3176.
(16) Ertl, P. World Wide Web-based System for the Calculation of Substituent Parameters and Substituent Similarity Searches. *J. Mol. Graph. Mod.* **1998**, *16*, 11−13.
(17) JME molecular editor applet allowing creation or editing of molecules may be obtained directly from the author; see also the JME HomePage at www.molinspiration.com/jme/.

CI0255782