Journal of Computer-Aided Molecular Design (2005) 19: 705–713
DOI 10.1007/s10822-005-9014-2

Why relevant chemical information cannot be exchanged without disclosing structures

Dmitry Filimonov* & Vladimir Poroikov
Institute of Biomedical Chemistry of Rus. Acad. Med. Sci., Pogodinskaya Str., 10, 119121, Moscow, Russia

Received 9 June 2005; accepted 6 August 2005

Key words: biological activity spectra, computer prediction, PASS, relevant chemical information, reverse engineering, safety exchange, structure disclosing

Summary

Both society and industry are interested in increasing the safety of pharmaceuticals. Potentially dangerous compounds could be filtered out at early stages of R&D by computer prediction of biological activity and ADMET characteristics. Accuracy of such predictions strongly depends on the quality & quantity of information contained in a training set. Suggestion that some relevant chemical information can be added to such training sets without disclosing chemical structures was generated at the recent ACS Symposium. We presented arguments that such safety exchange of relevant chemical information is impossible. Any relevant information about chemical structures can be used for search of either a particular compound itself or its close analogues. Risk of identifying such structures is enough to prevent pharma industry from relevant chemical information exchange.

Introduction

Chemical information is equal to almost 50% of all available information, and about half of chemical data are related to biology & medicine. Currently, approximately 25 million organic chemical structures are presented in CAS databases [1], 8 million can be found in Beilstein databases [2], 14 million unique chemical structures, which are available as samples, are searchable by ChemFinder [3]. However, all these correspond to only a small fraction of general chemical space: combinations from only 30 C, N, O, and S atoms already give about 10^{60} molecules [4].

A lot of chemical information related to biological activity has not ever been published. Possessed by the pharmaceutical industry, this information is stored in "in house" databases, and only a small part of this information is disclosed to public by patents and publications. Such proprietary information constitutes essential assets of companies providing them with significant advantages at highly competitive pharmaceutical market.

After the Thalidomide's tragedy happened in 1959–1962, many pharmaceuticals have been withdrawn from the market due to their adverse effects and toxicity. Baycol and Vioxx are just two recent examples of cases that badly influenced on the reputation of companies and decreased the cost of their shares at the stock market. Even more important that, according to the statistics, about 100,000 annual deaths of patients in US are currently associated with adverse effects of drugs [5].

Therefore, it is obvious that both society and industry are interested in increasing the safety of

*To whom correspondence should be addressed. Fax: +7-095-245-0857; E-mail: Dmitry.Filimonov@ibmc.msk.ru

pharmaceuticals. One way to achieve this goal is to filter out potentially dangerous chemical compounds at the early stage of R&D by computer prediction of biological activity [6, 7], and ADMET characteristics [8]. However, the accuracy and reliability of computational estimates strongly depend on the quality & quantity of information, which is used as the basis for computer-aided prediction. Since the amount of information about biologically active compounds in public domain is limited, the question arises: Could it be possible to improve the accuracy of computational models by adding the information available within companies to the public data? As nobody is naive enough to suggest that every company readily discloses all its proprietary structural data to the public, the topic of recent Symposium "Safe exchange of chemical information: can relevant chemical information be exchanged without disclosing chemical structures?" [9] is really vital. Two answers were given at the Symposium, positive and negative one. Our arguments, why relevant chemical information cannot be exchanged without disclosing of structures, are presented below.

Results and discussion

General reflections

First of all, what is relevant chemical information without chemical structure? "Relevant" means "having a bearing on or connection with the subject at issue". Since structure is one of the main issues in modern chemistry, relevant information detached from the chemical structure looks like the Cheshire cat's smile.

Secondly, everyday practice of analytical chemistry clearly provides the evidence that structure of molecule could be reconstructed on the basis of data on its properties. Otherwise, no information about structure of millions of compounds could be available today from databases like CAS, Beilstein, etc.

Certainly, there are 100s of descriptors associated with chemical structures that can be used to build predictive models. However, information about such descriptors could serve as the basis for reverse engineering or identification of compound's class (see below).

Reverse engineering or identification of compound's class?

The basic hypothesis of SAR/QSAR/QSPR is based on the suggestion that molecular property can be presented as a function of molecular structure: $Property = Function(Structure)$. The inverse problem (reverse engineering) requires a solution of another equation: $Structure = Function(Property)$. Actually, the main purpose of reverse engineering is design of compounds with the required properties.

The most general representation of both functions is the structure–property relationship or, in other words, the set of tuples $\{<Structure, Property>\}$. So, any chemical database includes the partial functions $\{<Structure, Property_1>\}$, $\{<Structure, Property_2>\}$, ... Since the relevant information is presented by the values of descriptors (that could be exchanged), the set of these descriptors can be used as a fingerprint, to search for a particular molecule itself or class of molecules with a particular property. Let us consider some examples that illustrate such possibility.

Experiments with MDDR database and PASS training set

First of all, we tried to compare two sub-sets of compounds with molecular weight less than 1500 D. The first sub-set that is called SET1 includes 31,644 principal compounds from the MDDR database [10]; the second one that is called SET2 includes 41,602 compounds from the training set of computer program PASS [6, 7]. Both SET1 and SET2 are relatively small in size comparing to the large databases provided by CAS, ChemFinder, etc. Distribution of molecular weights for compounds from SET1 and SET2 is shown on Figure 1. Median values are 422 D and 390.5 D for SET1 and SET2 respectively. Smaller median value for SET2 could be probably explained by removing counter ions in compounds from PASS training set. Neighborhoods of median values are the most populated in both sub-sets.

The data presented in Table 1 illustrates that even molecular weight could be successfully used as a parameter to search for a particular compound in databases. All compounds in two

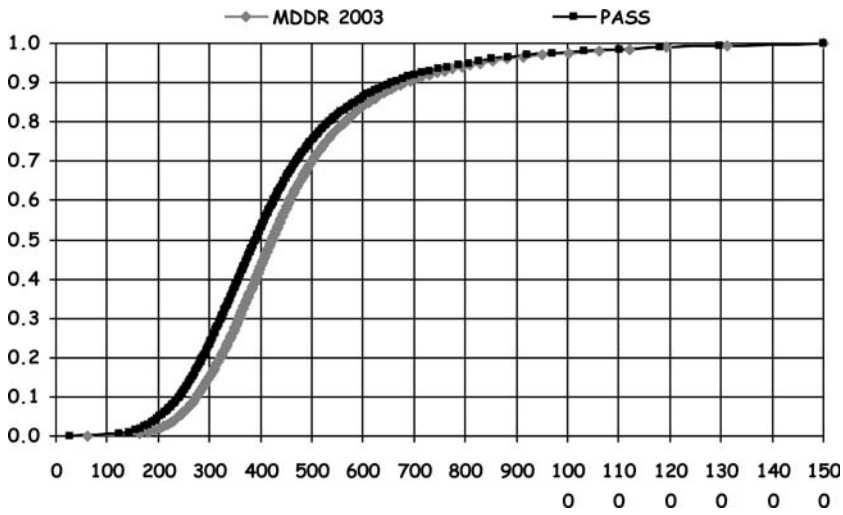


Figure 1. Molecular weight distributions for compounds from SET1 and SET2.

sub-sets with the same molecular weight have also the same molecular formula. These compounds likely have the same structural formula but might differ as stereoisomers.

Table 1. Molecular formula and molecular weight (MW) of compounds in SET1 and SET2 found in neighborhood of MW median value for principal compounds from MDDR database.

$C_{23}H_{22}N_2O_6$	422.44159
$C_{23}H_{22}N_2O_6$	422.44159
$C_{22}H_{25}F_3N_2O_3$	422.45135
$C_{25}H_{23}FO_5$	422.45746
$C_{23}H_{23}FN_4O_3$	422.46316
$C_{23}H_{23}FN_4O_3$	422.46316
$C_{25}H_{23}N_2NaO_3$	422.46346
$C_{24}H_{19}FO_4S$	422.47903
$C_{25}H_{26}O_6$	422.48237
$C_{24}H_{26}N_2O_5$	422.48522
$C_{20}H_{30}N_4O_6$	422.48530
$C_{27}H_{22}N_2O_3$	422.48799
$C_{23}H_{26}N_4O_4$	422.48807
$C_{23}H_{26}N_4O_4$	422.48807
$C_{23}H_{26}N_4O_4$	422.48807
$C_{26}H_{22}N_4O_2$	422.49084
$C_{22}H_{26}N_6O_3$	422.49092
$C_{22}H_{26}N_6O_3$	422.49092
$C_{25}H_{22}N_6O$	422.49369
$C_{25}H_{22}N_6O$	422.49369
$C_{17}H_{23}N_2NaO_5S_2$	422.50106
$C_{26}H_{27}FO_4$	422.50109

Complexity of chemical structures comparing to the complexity of scientific text

Is this just an occasional result or reverse engineering of chemical structures is not so difficult in general? To estimate the complexity of chemical structure, we used a compression procedure in accordance with the Shennon's theorem of coding. Structure samples were exported for both SET1 and SET2 from ISIS/Base as SDfiles, and were further converted into SMILES format with ConSystant software. The SMILES format is probably one of the most compact representation of chemical structures. At the next step we compressed the text files with chemical structures represented by the SMILES format with free software 7zip, which allows optimization of tuning parameters to obtain the maximal compression ratio. The results represented a number of bits per molecule for compounds with different molecular weights are presented in Figure 2.

The complexity of chemical structure varies in range 60–150 bit/molecule for molecular weights variation in range 250–675 D. Median equals to 100 bit/molecule at 400 D. It must be stressed that this value is the upper estimation, and therefore the real chemical structure complexity is less.

For comparison we performed such procedure with the text file included all abstracts of the Symposium [9]. As a result we obtained 2.4 bit/letter or

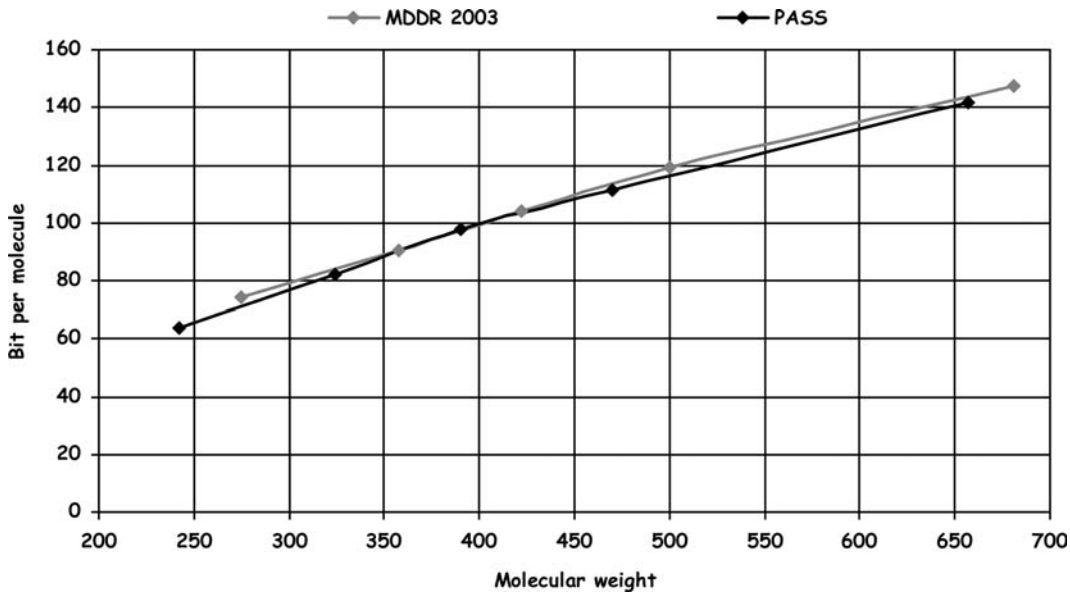


Figure 2. Complexity of chemical structure for compounds from SET1 and SET2.

about 16 bit/word. So, chemical structure of typical drug-like compound has complexity, which is equivalent to those of usual scientific text of 6 words (40 letters). Of course, the number 2^{100} is rather great, but the complexity order of ~100 bit is not a problem that could not be solved by modern cryptographic analysis. Based on these estimations, one may conclude that even direct reverse engineering with the use of complete enumeration (''brute force method'') is possible.

Of course, our experiments only demonstrated the principal possibility of reverse engineering. They could not guarantee that in any particular case the structures will be reconstructed unambiguously. However, while the positive answer to the question posed in the title of the ACS Symposium should be proved for a general case, the negative answer does not require such evidences. Demonstration that significant risk of structure disclosing exists is already a sufficient argument for pharma industry, to avoid such ''safety'' exchange of chemical information.

How many molecular descriptors represent a relevant information about chemical structure?

The next question is: How many descriptors are necessary to get a relevant information about chemical structure? We considered this problem by case study of Multilevel Neighborhood of Atoms (MNA) descriptors [11] used in PASS [6, 7]. A number of MNA descriptors per one molecule in SET1 scattering close to lognormal distribution with an average value equals to 30 MNA/molecule (Figure 3).

We also calculated an average numbers of structures, which includes 1, 2, ... common with MNA descriptors of a particular molecule. The results are presented in Figure 4.

It is clear, that on average 10 of 30 randomly chosen MNA descriptors is enough to find one unique structure in the set included more than 10,000,000 structures. So, even part of MNA descriptors, which represent a compound, is enough to identify this compound in the database. Certainly, we suggest that this compound is included into the database used for the search, but even if it is not so, the close analogs of the compound will be probably found. But this is in the most cases enough to identify the compound's class, to generate a plausible hypothesis about its activity/property, and finally to create a ''me-too-drug''.

Experiments with open NCI database

We have investigated a possibility of reverse engineering or identification of the appropriate compound's class on the basis of some other types

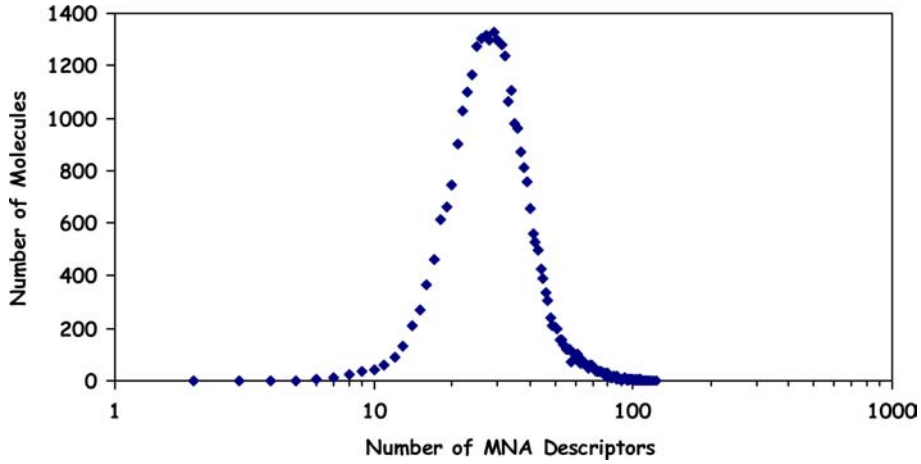


Figure 3. Distribution of MNA descriptors number for one molecule in SET1.

of descriptors available in open NCI database [12]. Since this database is freely available via Internet, it is widely used for validation of various database mining methods [13–15].

Molecular weight and log*P* represent rather simple kinds of descriptors that are widely used in QSAR/QSPR studies. Both descriptors are available in the NCI database in a searchable mode using toolkit CACTVS [12]. We have calculated how many chemical compounds from NCI database correspond to different ranges of molecular weight and log*P* (calculated by KOW method). The results are presented in Table 2.

As one may conclude from the results presented in Table 2, only four chemical structures belong to MW & log*P* range 400–401 & 3.65–3.75 or 400–400.5 & 3.65–3.75. This result clearly demonstrates that using only combination of two very simple molecular descriptors it is possible to identify a few compounds that correspond to the values of descriptors. In the majority of cases this means the disclosure of structure itself or at least their chemical class.

Biological activity spectra components used as a query

Computer program PASS (version 1.913.2) predicts 986 kinds of biological activity on the basis of

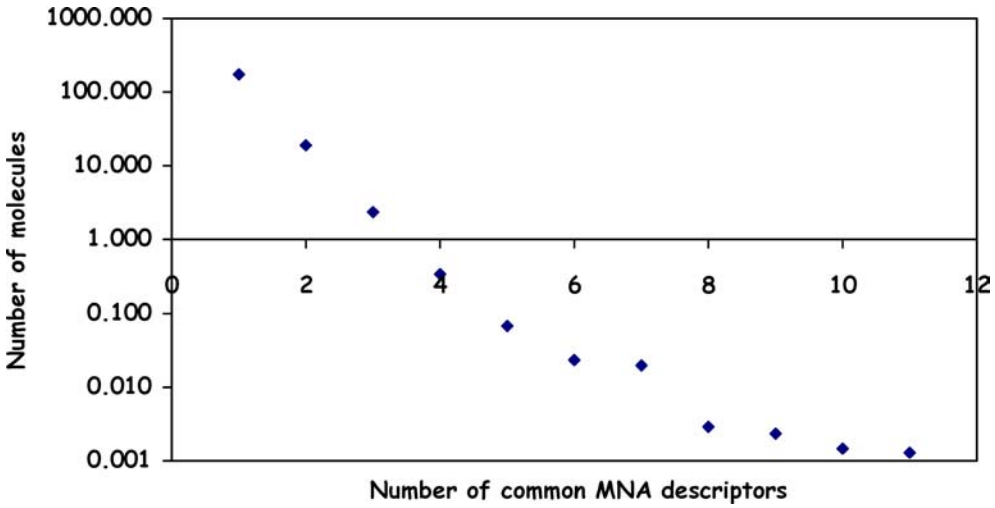


Figure 4. Average numbers of structures, which includes 1, 2, ... common with particular molecule MNA descriptors.

Table 2. Number of chemical structures in NCI database corresponding to the particular ranges of molecular weight and log*P* values.

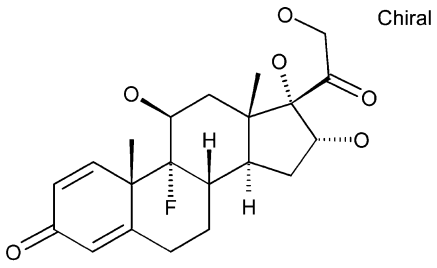
MW	*N*	Log*P*	*N*	MW & log*P*	*N*
400–402	903	3.67–3.73	2019	400–402 & 3.6–3.8	21
400–401	536	3.68–3.72	1338	400–402 & 3.65–3.75	10
400–400.5	338	3.69–3.71	644	400–402 & 3.67–3.73	8
				400–401 & 3.4–4.0	31
				400–401 & 3.6–3.8	12
				400–401 & 3.65–3.75	**4**
				400–400.5 & 3.4–4.0	21
				400–400.5 & 3.6–3.8	7
				400–400.5 & 3.65–3.75	**4**

Here: MW is the range of molecular weights; log*P* is the range of logarithm of *n*-octanol/water distribution coefficients calculated by KOW method; *N* is the number of chemical structures fallen into the appropriate range of the descriptors.

compound's structural formula with reasonable accuracy (~85% in leave one out cross-validation) [6, 7]. Example of such prediction for one compound taken by chance from Prestwick database [16] is given in Figure 5. The result of prediction is presented as the list of activities with appropriate P_a and P_i, sorted in descending order of the difference $(P_a - P_i) > 0$. P_a is the probability of belonging to the class of "actives", and P_i is the probability of belonging to the class of "inactives". Only activities for which the predicted probability $P_a > 0.5$ are given in Figure 5.

As one may see from Figure 5, this compound is presented in PASS training set but during the prediction the compound with all its known biological activities (Antiallergic; Antiinflammatory; Antipruritic; Antipruritic, allergic; Antipruritic, non-allergic; Antipsoriatic; Arachidonic acid antagonist; Dermatologic; Glucocorticoid agonist; Immunosuppressant; Steroid-like) have been excluded, to provide more objective results of prediction. The structure contains 50 different MNA descriptors. 23 of 986 kinds of biological activity are predicted with $P_a > 0.5$. The majority of known kinds of biological activity are successfully predicted. Only one activity "Arachidonic acid antagonist" is not predicted.

PASS represents the properties of molecules in biological space in contrast to many other descriptors, which reflect the structural properties of molecules. PASS parameters can be used for clustering of compounds according to their



> <ACTIVITY_PREDICTION>
50 Substructure Descriptors; 0 new.
Exclude structure with activities:
Antiallergic
Antiinflammatory
Antipruritic
Antipruritic, allergic
Antipruritic, non-allergic
Antipsoriatic
Arachidonic acid antagonist
Dermatologic
Glucocorticoid agonist
Immunosuppressant
Steroid-like
23 of 986 Possible Activities at Pa > 0.500

Pa	Pi	for Activity:
0.980	0.004	Antiinflammatory
0.960	0.005	Antiallergic
0.958	0.003	Antipruritic
0.946	0.003	Antipruritic, allergic
0.931	0.003	Arachidonic acid antagonist
0.921	0.003	Steroid-like
0.882	0.002	Antipruritic, non-allergic
0.852	0.002	Antiinflammatory steroid
0.842	0.005	Antiinflammatory, ophthalmic
0.766	0.001	Glucocorticoid agonist
0.745	0.007	Immunosuppressant
0.766	0.105	Antiseborrheic
0.663	0.015	Dermatologic
0.576	0.002	Corticosteroid-like
0.567	0.014	Rhinitis treatment
0.554	0.002	Lipocortins synthesis agonist
0.549	0.030	Ophthalmic drug
0.530	0.015	Dopamine release stimulant
0.538	0.027	Antipsoriatic
0.526	0.032	Gonadotropin antagonist
0.486	0.021	Gestagen antagonist
0.536	0.100	Inflammatory Bowel disease treatment
0.546	0.129	Metallic radical formation agonist

Figure 5. Structure and results of biological activity spectra prediction for Triamcinolone (No 438 in Prestwick database).

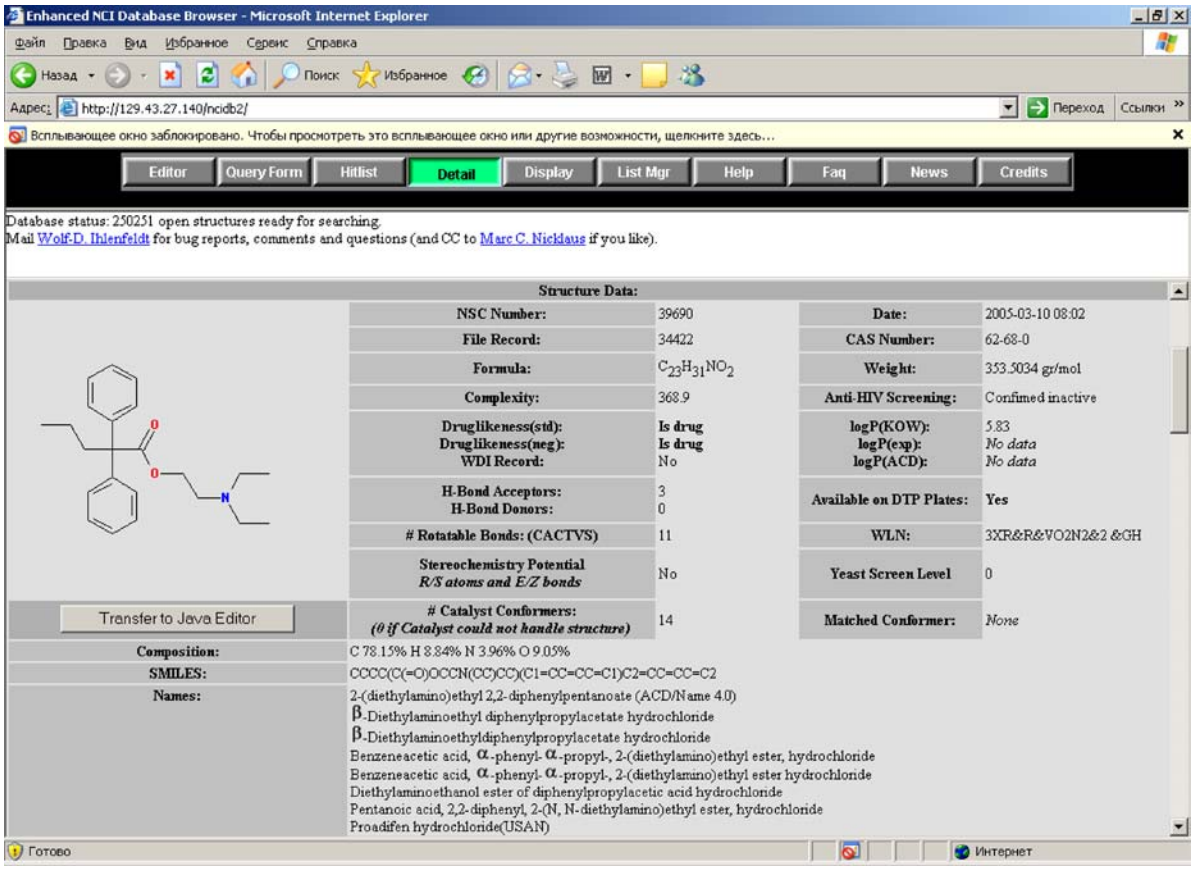


Figure 6. Coincidence of randomly chosen structure from Prestwick database with the structures from NCI database.

biological properties, not according to their structural similarity.

We tried to apply PASS predictions for three molecules chosen randomly from Prestwick database as a search queries to analyze NCI database. One molecule was Triamcinolone discussed above, two others were Proadifen Hydrochloride and Oxybutinine Chloride (numbers 124 and 621 in Prestwick database). Biological activity spectra were predicted by PASS 1.913.2. Top four activities with P_a range 0.7–1.0 were used as a query to Enhanced CACTVS Browser. For example, the query for Triamcinolone has the following form: (Antiinflammatory and P_a~0.7–1.0) and (Antiallergic and P_a~0.7–1.0) and (Antipruritic and P_a~0.7–1.0) and (Antipruritic, allergic and P_a~0.7–1.0).

As a result, we found in NCI database one coincidence with the query for Proadifen Hydrochloride (Figure 6) and several structures similar to the structures used as a query (e.g., see Figure 7 for Oxybutinine Chloride).

It should be emphasized that PASS predictions used as a query were obtained using PASS version 1.913.2, whereas PASS prediction stored in the NCI database were obtained with PASS version 1.41 [17]. PASS version 1.41 was able to predict only 565 kinds of biological activities [12], while the current version of PASS predicts 986 kinds of biological activity. Also, Prestwick database is not too close to NCI database. The first one contains about 1000 approved drugs [16], but the second one contains about 250,000 chemical compounds that were selected for study as potential antineoplastic and anti-HIV leads.

However, despite of these differences for three structures randomly selected from Prestwick database we found either the same structure or its close analog in NCI database, using only four kinds of biological activity predicted by PASS with the

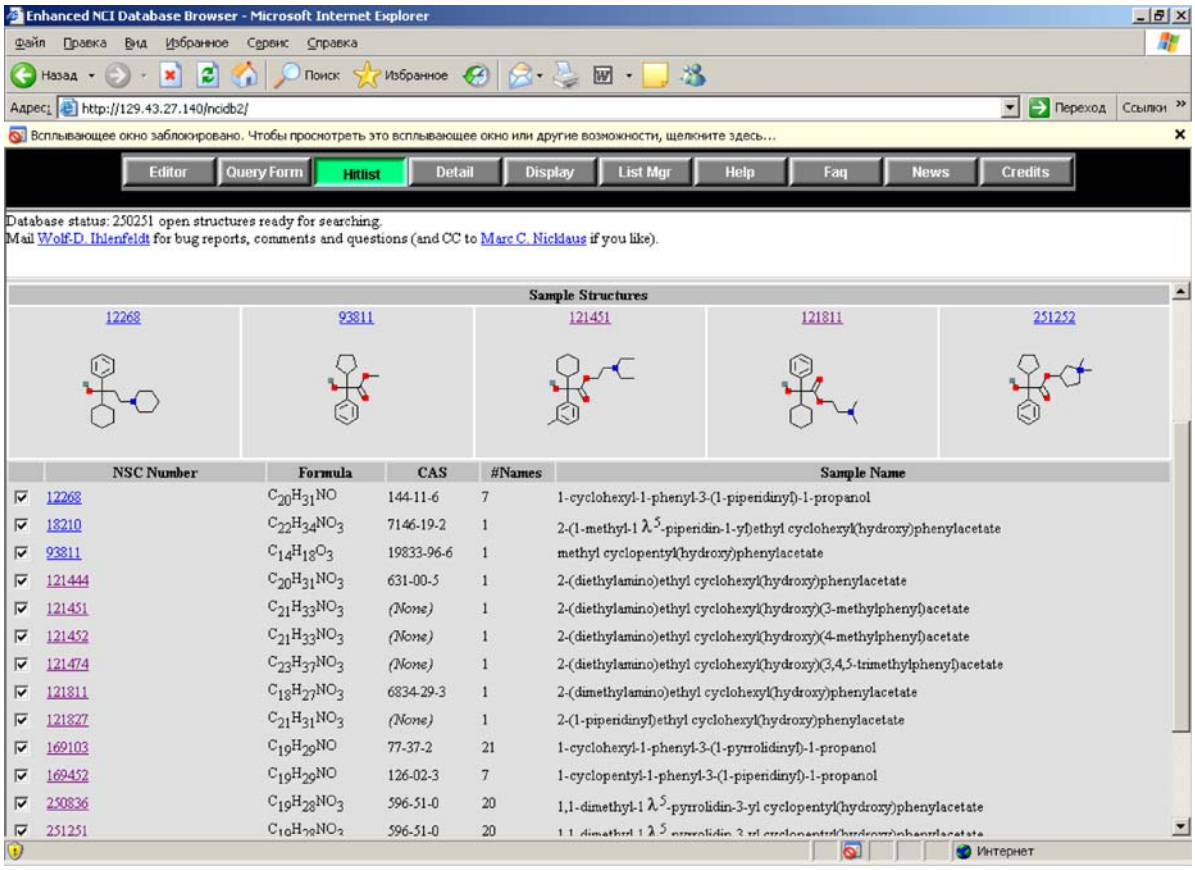


Figure 7. Similarity of randomly chosen structure from Prestwick database with the structures from NCI database.

highest probability. This experiment clearly demonstrates that using molecular descriptors in biological space it is possible to identify at least a compound's class, and therefore to disclose the structures of interest.

Conclusions

Based on the data discussed above it is obvious that a significant risk of structure disclosing exists when relevant chemical information (descriptors etc.) becomes publically available. Using the information about relevant descriptors as a query, it is possible to find either compound itself in the existing databases or at least identify their chemical class that will be in many cases enough to recognize what are target compounds.

Even if particular compounds classes are absent in the available databases like MDDR, NCI, Beilstein, ChemFinder or CAS, chemical structure generators can be applied that might provide more appropriate virtual structures under the restrictions of known descriptors' values.

In general, according to Bruce Schneider the problem of information security cannot be solved forever [18]. Struggle for security of information is permanent process, and nobody could be sure that he already won. Security of information vs. security threats is always a tradeoff between time and costs of the first and the second issue.

Keeping in mind that people from pharmaceutical industry are more than just careful concerning the confidentiality of its research and development, it is not realistic to expect that they would be ready to present any relevant information about compounds even if a very small risk of structure disclosing exists. Only increasing requirements of society to the drugs safety and strong legal measures could provide reasonable stimuli

for pharmaceutical industry to provide access to information that might help to improve significantly methods for filtering off potentially dangerous compounds.

References

1. http://www.cas.org.
2. http://www.mdl.com/products/knowledge/crossfire_beilstein/.
3. http://www.chemfinder.com.
4. Bohacek, R.S., McMartin, C. and Guida, W.C., Med. Res. Rev., 16 (1996) 3.
5. Pirmohamed, M. and Park, B.K., Trends Pharm. Sci., 22 (2001) 298.
6. Poroikov, V.V. and Filimonov, D.A., J. Comput. Aided Mol. Des., 16 (2002) 819.
7. Poroikov, V. and Filimonov, D. In Christoph Helma (Ed.), Predictive Toxicology, Taylor & Francis, 2005, pp. 459–478.
8. Van de Waterbeemd, H. and De Groot, M., Nat. Rev. Drug. Discov., 2 (2003) 192.
9. Safe exchange of chemical information: can relevant chemical information be exchanged without disclosing chemical structures. Symposium in the framework of 229th National Spring ACS Meeting, San Diego, CA (March 13–17, 2005).
10. http://www.mdl.com.
11. Filimonov, D., Poroikov, V., Borodina Yu. and Gloriozova, T., J. Chem. Inf. Comput. Sci., 39 (1999) 666.
12. http://cactus.nci.nih.gov.
13. Sadowski, J., J. Comput. Aided. Mol. Des., 11 (1997) 53.
14. Baurin, N., Mozziconacci, J.-C., Arnoult, E., Chavatte, P., Marot, C. and Morin-Allory, L., J. Chem. Inf. Comput. Sci., 44 (1997) 276.
15. Fang, X., Shao, L., Zhang, H. and Wang, S., J. Chem. Inf. Comput. Sci., 44 (1997) 249.
16. http://www.prestwickchemical.com.
17. Poroikov, V.V., Filimonov, D.A., Ihlenfeldt, W.-D., Gloriozova, T.A., Lagunin, A.A., Borodina, Yu.V., Stepanchikova, A.V. and Nicklaus, M.C., J. Chem. Inform. Comput. Sci., 43 (2003) 228.
18. Schneier, B. Secrets and Lies: Digital Security in a Networked World. John Wiley & Sons, 2000, p. 432.