Building a Chemical Space Based on Fragment Descriptors

Igor Baskin[1] and Alexandre Varnek[*,2]

[1]*Department of Chemistry, Moscow State University, Moscow 119992, Russia*

[2]*Laboratoire d'Infochimie, UMR 7177 CNRS, Universite´ Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000, France*

Abstract: This article reviews the application of fragment descriptors at different stages of virtual screening: filtering, similarity search, and direct activity assessment using QSAR/QSPR models. Several case studies are considered. It is demonstrated that the power of fragment descriptors stems from their universality, very high computational efficiency, simplicity of interpretation and versatility.

Keywords: Fragmental approach, fragment descriptors, QSAR, QSPR, filtering, similarity, virtual screening, *in silico* design.

1. INTRODUCTION

Chemogenomics aims to discover active and/or selective ligands for biologically related targets by conducting screening, ideally, of all possible compounds against all possible targets, or at least, in practice, available libraries of compounds against main target families [1]. One can hardly imagine to screen experimentally the chemical universe containing from 10^{12} to 10^{180} druglike compounds [2] against biological target universe. Nowadays, the number of experimentally screened compounds does not exceed several millions per biological target, whereas a single inexpensive computational study allows one to screen the libraries up to 10^{12} molecules and this number tends to grow up with the evolution of hardware and related software tools. Therefore, this is not surprising that the virtual, or *in silico*, screening approaches play a key role in chemogenomics.

Virtual screening is usually defined as a process in which large libraries of compounds are automatically evaluated using computational techniques [3]. Its goal is to discover putative hits in large databases of chemical compounds (usually ligands for biological targets) and remove molecules predicted to be toxic or those possessing unfavorable pharmacodynamic or pharmacokinetic properties. Generally, two types of virtual screening are known: structure-based and ligand-based. The former explicitly uses 3D structure of a biological target at the stage of hit detection, whereas the latter uses only information about structure of small molecules and their properties (activities). Structure-based virtual screening (docking, 3D pharmacophores) has been described in series of review articles, see references [4-6] and citations therein.

In this paper mostly ligand-based virtual screening involving fragment descriptors is considered. Fragment descriptors, represent selected substructures (fragments) of 2D molecular graphs and their occurrences in molecules; they constitute one of the most important types of molecular descriptors [7]). Their main advantage is related to simplicity of their calculation, storage and interpretation (see review articles [8-12]). Substructural fragment are information-based descriptors [13] which tend to code the information stored in molecular structures. This contrasts with knowledge-based (or semiempirical) descriptors issued from the consideration of the mechanism of action. Selected descriptors form a "chemical space" in which each molecule us represented as a vector. Due to their versatility, fragment descriptors could be efficiently used to create a chemical space which separates active and non-active compounds.

Historically, molecular fragments were used in first additive schemes developed in 1950-ies to estimate physicochemical properties of organic compounds by Tatevskii [14,15], Bernstein [16], Laidler [17], Benson and Buss [18] and others. The Free-Wilson method [19], one of the first QSAR approaches invented in 1960-ies, is based on the assumption of the additivity of contributions of structural fragments to the biological activity of the whole molecule. Later on, fragment descriptors were successfully used in expert systems able to classify chemical compounds as active or inactive with respect to certain type of biological activity. Hiller [20,21], Golender and Rosenblit [22,23], Piruzyan, Avidon *et al.* [24,25], Cramer [26], Brugger, Stuper and Jurs [27,28], and Hodes *et al.* [29] pioneered in this field.

An important class of fragmental descriptors, so-called *screens* (structural *keys*, *fingerprints*), has been developed in seventies [30-34]. As a rule, they represent bit strings which can effectively be stored and processed by computers. Although their primary role is to provide efficient substructure searching capabilities in large chemical databases, they are efficiently used for similarity searching [35,36], to cluster large data sets [37,38], to assess chemical diversity [39], as well as to conduct SAR [40] and QSAR [41] studies. Nowadays, application of modern machine-learning techniques significantly improves predictive performance of structure-property models based on fragment descriptors.

This paper briefly reviews the application of fragment descriptors in virtual screening of large libraries of organic compounds focusing mostly on its three approaches: (i) filtering, (ii) similarity search, and (iii) direct activity/property assessment using QSAR/QSPR models.

*Address correspondence to this author at the Laboratoire d'Infochimie, UMR 7177 CNRS, Universite´ Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000, France; E-mail: varnek@chimie.u-strasbg.fr

1386-2073/08 $55.00+.00

2. TYPES OF FRAGMENT DESCRIPTORS

Due to their enormous diversity, one could hardly review all types of 2D fragment descriptors used for structural search in chemical database or in SAR/QSAR studies. Here, we focus on some of them which are the most efficiently used in virtual screening and *in silico* design of organic compounds.

Generally, molecular fragments could be classified with respect to their topology (atom-based, chains, cycles, polycycles, etc), information content of vertices in molecular graphs (atoms, groups of atoms, pharmacophores, descriptor centers) and the level of abstraction when some information concerning atom and bond types is omitted.

Purely structural fragments are used as descriptors in ACD/Labs [42], NASAWIN [43], ISIDA [12] and some other programs. These are 2D subgraphs in which all atoms and/or bonds are represented explicitly and no information about their properties is used. Their typical example is sequences of atoms and/or bonds of variable length, branch fragments, saturated and aromatic cycles (polycycles) and atom-centered fragments (ACF). The latter consist of a single central atom surrounded by one or several shells of atoms with the same topological distance from the central one. The ACF were invented by Tatevskii [14] and Benson and Buss [18] in 1950-ies as elements of additive schemes for predicting physicochemical properties of organic compounds. In earlier seventies, Adamson [44] investigated the distribution of one shell ACF in some chemical databases with respect to their possible application as screens. Hodes reinvented one shell ACF as descriptors in SAR studies under the name *augmented atoms* [29], and also suggested *ganglia augmented atoms* 45 representing two shells ACF with generalized second-shell atoms. Later on, one shell ACF were implemented by Baskin *et al.*. in the NASAWIN [43] software and by Solov'ev and Varnek in ISIDA [12] package (Fig. **1**). Atom-centered fragments with arbitrary number of shells were implemented by Filimonov and Poroikov in the PASS [46] program under the name *multilevel neighborhoods of atoms* [47], by Xing and Glen under the name *tree structured fingerprints* [48], by Bender *et al.*. as *atom environments* [49,50] and *circular fingerprints* [51-53], and by Faulon under the name *molecular signatures* [54-56].

It has been found that characterizing atoms only by element types is too specific for similarity searching and therefore does not provide sufficient flexibility needed for large-scaled virtual screening. For that reason, numerous studies were devoted to increase an informational content of fragment descriptors by adding some useful empirical information and/or by representing a part of molecular graph implicitly. The simplest representatives of those descriptors were *atom pairs and topological multiplets* based on the notion of *descriptor center* representing an atom or a group of atoms which could serve as centers of intermolecular interactions. Usually, descriptor centers include heteroatoms, unsaturated bonds and aromatic cycles. An *atom pair* is defined as a pair of atoms (**AT**) or descriptor centers separated by a fixed topological distance: $\mathbf{AT}_i\text{-}\mathbf{AT}_j\text{-}\boldsymbol{Dist}$, where $Dist_{ij}$ is the shortest path (the number of bonds) between AT_i and AT_j. Analogously, a *topological multiplet* is defined as a multiplet (usually triplet) of descriptor centers and topological distances between each pair of them. In most of cases, these descriptors are used in binary form in order to indicate the presence or absence of the corresponding features in studied chemical structures.

The atom pairs were first suggested for SAR studies by Avidon under the name *SSFN* (*Substructure Superposition Fragment Notation*) [25,57]. Then they were independently reinvented by Carhart and co-authors [58] for similarity and trend vector analysis. In contrast to SSFN, Carhart's atom pairs are not necessarily composed only of descriptor centers, but account for the information about element type, the number of bonded non-hydrogen neighbors and the number of π electrons. Nowadays, Carhart's atom pairs are rather popular for conducting virtual screening. *Topological Fuzzy Bipolar Pharmacophore Autocorrelograms* (*TFBPA*) [59] by Horvath are based on atom pairs, in which real atoms are replaced by pharmacophore sites (hydrophobic, aromatic, hydrogen bond acceptor, hydrogen bond donor, cation, anion), while $Dist_{ij}$ corresponds to different ranges of topological distances between pharmacophores. These descriptors were successfully applied in virtual screening against a panel of 42 biological targets using similarity search based on several fuzzy and non-fuzzy metrics [60], performing only slightly less well than their 3D counterparts [59]. *Fuzzy Pharmacophore Triplets* (*FPT*) by Horvath [61] is an extention of *FBPF* [60] for three sites pharmacophores. An important innovation in the *FPT* concerns accounting for proteolytic equilibrium as a function of pH [61]. Due to this feature, even small structural modifications leading to a pK_a shift, may have a profound effect on the fuzzy pharmocophore triples. As a result, these descriptors efficiently discriminate structurally similar compounds exhibiting significantly different activities [61].

Some other topological triplets should be mentioned. Thus, *Similog pharmacophoric keys* by Jacoby [62] represent

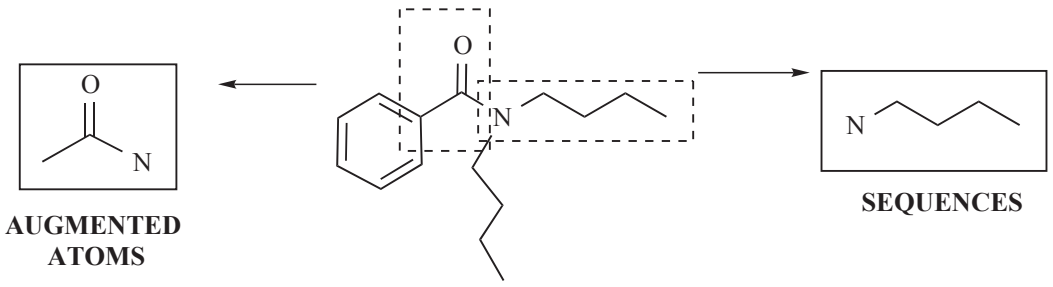


Fig. (1). Decomposition of a chemical structure into fragments. Examples of *sequences* and *augmented atoms* used as descriptors in the ISIDA program [12].

triplets of binary coded types of atoms (pharmacophoric centers) and topological distances between them. Atomic types are generalized by four features (represented as four bits per atom): potential hydrogen bond donor or acceptor; bulkiness and electropositivity. The *topological pharmacophore-point triangles* implemented in the MOE software [63] represent triplets of MOE atom types separated by binned topological distances. Structure-property models obtained by support vector machine method with these descriptors have been successfully used for virtual screening of COX-2 inhibitors [64] and D_3 dopamine receptor ligands [65].

Topological torsions by Nilakantan *et al.* [66] is a sequence of four consecutively bonded atoms $\boldsymbol{AT_i}$-$\boldsymbol{AT_j}$-$\boldsymbol{AT_k}$-$\boldsymbol{AT_l}$, where each atom is characterized by a number of parameters similarly to atoms in Carhart's pairs. In order to enhance efficiency of virtual screening, Kearsley *et al.* [67] suggested to assign atoms in the Carhart's atom pairs and Nilakantan's topological torsions to one of seven classes: cations, anions, neutral hydrogen bond donors, neutral hydrogen bond acceptors, polar atoms, hydrophobic atoms and other.

In contrast to QSPR studies based mostly on the use of complete (containing all atoms) or hydrogen-suppressed molecular graphs, handling biological activity at the qualitative level, often demands more abstractions. Namely, it is rather convenient to approximate chemical structures by *reduced graphs*, in which each vertex is an atom or a group of atoms (descriptor or pharmacophoric center), whereas each edge is a topological distance $Dist_{ij}$. Such biology-oriented representation of chemical structures was suggested by Avidon *et al.* as descriptor center connection graphs [25]. Gillet, Willett and Bradshaw have proposed the GWB-reduced graphs which use the hierarchical organization of vertex labels. This allows one to control the level of their generalization which may explain their high efficiency in similarity searching.

An alternative scheme of reducing molecular graph proposed by Bemis and Murcko [68,69] involves four-level hierarchical scheme of molecular structure simplification: (1) full molecular structure with all atoms; (2) structure without hydrogen atoms; (3) *scaffolds*, i.e. structures without substituents (which are deleted recursively by means of eliminating the "leaves" of molecular graph), (4) *molecular frameworks*, i.e. scaffolds, in which all heteroatoms are substituted by carbon atoms, while all multiple bonds are replaced by single bonds. This presentation of molecular graph was found very useful for diversity analysis of large databases [68,69].

3. APPLICATION OF FRAGMENT DESCRIPTORS IN VIRTUAL SCREENING AND IN SILICO DESIGN

In this chapter, the application of fragment descriptors at different stages of virtual screening is considered.

3.1. Filtering

Filtering is a rule-based approach aimed to perform fast assessment of usefulness molecules in the given context. In drug design area, the filtering is used to eliminate compounds with unfavorable pharmacodynamic or pharmacokinetic properties as well as toxic compounds. Pharmacodynamics considers binding drug-like organic molecules (ligands) to chosen biological target. Since the efficiency of ligand-target interactions depends on spatial complementarity of their binding sites, the filtering is usually performed with 3D-pharmacophores, representing "optimal" spatial arrangements of steric and electronic features of ligands [70,71]. Pharmacokinetics is mostly related to absorption, distribution, metabolism and excretion (ADME) related properties: octanol-water partition coefficients (*log P*), solubility in water (*log S*), blood-brain coefficient (*log BB*), partition coefficient between different tissues, skin penetration coefficient, etc.

Fragment descriptors are widely used for early ADME/Tox prediction both explicitly and implicitly. The easiest way to filter large databases concerns detecting undesirable molecular fragments (*structural alerts*). Appropriate lists of structural alerts are published for toxicity [72], mutagenicity [73], and carcinogenicity [74]. Klopman *et al.* were the first to recognize the potency of using fragmental descriptors for this purpose [75-77]. Their programs CASE [75], MultiCASE [78,79], as well as more recent MCASE QSAR expert systems [80], proved to be effective tools to assess mutagenicity [76,79,80] and carcinogenicity [77,79] of organic compounds. In these programs, sets of biophores (analogs of structural alerts) were identified and used for activity predictions. A number of more sophisticated fragment-based expert systems of toxicity assessment - DEREK [81], TopKat [82] and Rex [83] – have been developed. DEREK is a knowledge-based system operating with human-coded or automatically generated [84] rules about toxicophores. Fragments in the DEREK knowledge base are defined by means of linear notation language PATRAN which codes the information about atom, bonds and stereochemistry. TopKat uses a large predefined set of fragment descriptors, whereas Rex implements a special kind of atom-pairs descriptors (*links)*. To read more information about fragment-based computational assessment of toxicity, including mutagenicity and carcinogenicity, see review [85] and references therein.

The most popular filter used in drug design area is based on the Lipinsky "rule of five" [86], which takes into account the molecular weight, the number of hydrogen bond donors and acceptors, along with the octanol-water partition coefficient *logP*, to assess the bioavailability of oral drugs. Similar rules of "drug-likeness" or "lead-likeness" were later proposed by by Oprea [87], Veber [88] and Hann [89]. Formally, fragment descriptors are not explicitly involved there. However, many computational approaches to assess *logP* are fragment-based [42,90,91]; whereas H-donors and acceptor sites are simplest molecular fragments.

3.2. Similarity Search

The similarity-based virtual screening is based on an assumption that all compounds in a chemical database, which are similar to a query compound, could also have similar biological activity. Although this hypothesis is not always valid (see discussion in [92]), quite often the set of retrieved compounds is enriched by actives [93].

To achieve high efficacy of similarity-based screening of databases containing millions compounds, molecular structures are usually represented by *screens* (structural keys) or fixed-size or variable-size *fingerprints*. Screen and fingerprints can contain both 2D- and 3D-information. However,

the 2D-fingerprints, which are a kind of binary fragment descriptors, dominate in this area. Fragment-based structural keys, like MDL keys [40], are sufficiently good for handling small and medium-sized chemical databases, whereas processing of large databases is performed with fingerprints having much higher information density. Fragment-based Daylight [94], BCI [95] and UNITY 2D [96] fingerprints are the most known examples.

The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto (or Jaccard) coefficient *T* [97]. Two structures are usually considered similar if $T>0.85$ [93] (for Daylight fingerprints [94]). Using this threshold, Taylor estimated a probability to retrieve actives as 0.012-0.50 [98], whereas according to Delaney this number raises to 0.40-0.60 [99] (using Daylight fingerprints [94]). These computer experiments confirm usefulness of the similarity approach as an instrument of virtual screening.

Schneider *et al.* have developed a special technique for performing virtual screening referred to as CATS (Chemically Advanced Template Search) [100]. In its framework chemical structures are described by means of correlation vectors, each component of which is equal to occurrence of certain atom pair in a chemical structure divided by the total number of non-hydrogen atoms. Each atom in an atom pair is attributed to one of five classes: hydrogen-bond donor, hydrogen-bond acceptor, positively charged, negatively charged, and lipophilic. Topological distances of up to 10 bonds are considered in the atom-pair specification. The similarity is assessed in [100] using Euclidean distance between the corresponding correlation the MERLIN system from Daylight [94] for retrieving thrombin inhibitors in a virtual screening experiments.

Hull *et al*. have developed the *Latent Semantic Structure Indexing* (LaSSI) approach to perform similarity search in low-dimensional chemical space [101, 102]. To reduce the dimension of initial chemical space, the singular value decomposition method is applied for the descriptor-molecule matrix. Ranking molecules by similarity to a query molecule was performed in the reduced space using the cosine similarity measure [103], whereas the Carhart's atom pairs [58] and the Nilakantan's topological torsions [66] were used as descriptors. The authors claim that this approach "has several advantages over analogous ranking in the original descriptor space: matching latent structures is more robust than matching discrete descriptors, choosing the number of singular values provides a rational way to vary the 'fuzziness' of the search" [101].

The issue of "fuzzification" of similarity search was addressed by Horvath *et al.* [59-61]. The first fuzzy similarity metric suggested in work [59] relies on partial similarity scores calculated with respect to the inter-atomic distances distributions for each pharmacophore pair. In this case the "fuzziness" enables to compare pairs of pharmacophores with different topological or 3D distances. Similar results [60] were achieved using fuzzy and weighted modified Dice similarity metric [103]. Fuzzy pharmacophore triplets FPT (see section 2) can be gradually mapped onto related basis triplets, thus minimizing binary classification artifacts [61]. In new similarity scoring index introduced in reference [61], the simultaneous absence of a pharmacophore triplet in two molecules is taken into account. However, this is a less-constraining indicator of similarity than simultaneous presence of triplets.

Most of similarity search approaches require only a single reference structure. However, in practice several compounds with the same type of biological activity are often available. This motivated Hert *et al.* [104] to develop the *data fusion method* which allows one to screen a database using all available reference structures. Then, the similarity scores are combined for all retrieved structures using selected fusion rules. Searches conducted on the MDL Drug Data Report database using fragment-based UNITY 2D [96], BCI [95], and Daylight [94] fingerprints have proved the effectiveness of this approach.

The main drawback of the conventional similarity search concerns an inability to use experimental information on biological activity to adjust similarity measures. This results in inability to discriminate between relevant and non-relevant fragment descriptors being used for computing similarity measures. To tackle this problem, Cramer *et al.* [26] developed *substructural analysis* in which each fragment (represented as a bit in a fingerprint) is weighted by taking into account its occurrence in active and in inactive compounds. Later on, many similar approaches have been described in the literature [105].

One more way to conduct a similarity-based virtual screening is to retrieve the structures containing a user-defined set of "pharmacophoric" features. In *Dynamic Mapping of Consensus positions* (DMC) algorithm [106] those features are selected by finding common positions in bit strings for all active compounds. The *potency-scaled DMC* algorithm (POT-DMC) [107] is a modification of DMC in which compounds activities are taken into account. The latter two methods may be considered as intermediate between conventional similarity search and probabilistic SAR approaches.

Batista, Godden and Bajorath developed the MolBlaster method [108], in which molecular similarity is assessed by *Differential Shannon Entropy* [109] computed from populations of randomly generated fragments. For the range $0.64 < T < 0.99$, this similarity measure provides with the same ranking as the Tanimoto index *T*. However for the smaller values of *T* the entropy-based index is a more sensitive, since it distinguishes between pairs of molecules having almost identical *T*. To adapt this methodology for large-scale virtual screening, the *Proportional Shannon Entropy* (PSE) metrics was introduced [110]. A key feature of this approach is that class-specific PSE of random fragment distributions enables the identification of the molecules sharing with known active compounds a significant number of signature substructures.

Similarity search methods developed for individual compounds are difficult to apply directly for chemical reactions involving many species subdividing by two types: reactants and products. To overcome this problem, Varnek *et al.* [12] suggested to condense all participating in reaction species in one only molecular graph (*Condensed Graphs of Reactions* (*CGR*) [12]) followed by its fragmentation and application of developed fingerprints in "classical" similarity search. Besides conventional chemical bonds (simple, double, aromatic, etc), a CGR contains dynamical bonds corresponding

to created, broken or transformed bonds. This approach could be efficiently used for screening of large reaction databases.

It should be noted that the similarity concepts are widely used in selecting of diverse sets of compounds (see reviews [111-115] and references therein).

3.3. SAR/QSAR/QSPR Models

Simplistic and heuristic similarity-based approaches can hardly produce as good predictive models as modern statistical and machine learning methods able to assess quantitatively biological or physicochemical properties. QSAR-based virtual screening consists in direct assessment of activity values (numerical or binary) of all compounds in the database followed by selection of hits possessing desirable activity. Mathematical methods used for models preparation could be subdivided into *probabilistic* and *regression* approaches. The former assesses a probability that a given compound is active or not active whereas the latter numerically evaluate the activity values. A limited size of this paper doesn't allow us to cite all successful stories related to application of probabilistic and regression models in virtual screening; only some examples will be presented.

Harper *et al.* [116] have demonstrated a much better performance of probabilistic *binary kernel discrimination* method to screen large databases compared to backpropagation neural networks or conventional similarity search. The Carhart's atom-pairs [58] and Nilakantan's topological torsions [66] were used as descriptors in that study.

Aiming to discover new cognition enhancers, Geronikaki *et al.* [117] applied the PASS program [46], which implements a probabilistic Bayesian-based approach, and the DEREK rule-based system [81] to screen a database of highly diverse chemical compounds. Eight compounds with the highest probability of cognition-enhancing effect were selected. Experimental tests have shown that all of them possessed a pronounced antiamnesic effect.

Bender *et al.* [49-53] have applied several probabilistic machine learning methods (naïve Bayesian classifier, inductive logic programming, and support vector inductive learning programming) in combination with circular fingerprints to perform the classification of bioactive chemical compounds and to carry out virtual screening on several biological targets. It has been shown that he performance of support vector inductive learning programming was significantly better than the other two methods [53].

Regression QSAR/QSPR models are used to assess ADME/Tox properties or to detect "hit" molecules capable to bind a certain biological target. A general scheme of building QSAR/QSPR models using fragment descriptors is given in Fig. **2**. Available in the literature fragments based QSAR models for blood-brain barrier [118], skin permeation rate [119], blood-air [120] and tissue-air partition coefficients [120] could be mentioned as examples. Many theoretical approaches of calculation of octanol/water partition coefficient log *P* involve fragment descriptors. The methods by Rekker [121,122], Leo and Hansch (CLOGP) [90,123], Ghose-Crippen (ALOGP) [124-126], Wildman and Crippen [127], Suzuki and Kudo (CHEMICALC-2) [128], Convard (SMILOGP) [129], Wang (XLOGP) [130,131] represent just a few modern examples. Fragment-based predictive models for estimation solubility in water [132] and DMSO [132] are available.

Benchmarking studies performed in references [118-120,133] show that QSAR/QSPR models for various biological and physicochemical properties involving fragment descriptors are, at least, as robust as those involving topological, quantum, electrostatic and other types of descriptors.

3.4. *In Silico* Design

In this section we consider examples of virtual screening performed on a database containing only virtual (still non-synthesized or unavailable) compounds. Generation of virtual libraries is usually performed using combinatorial chemistry approaches [134-136]. One of simplest ways is to attach systematically user-defined substituents R_1, R_2, ..., R_N to a given scaffold. If the list for the substituent R_i contains n_i candidates, the total number of generated structures is $N = \prod n_i$, although taking symmetry into account could reduce the library's size. The number of substituents R_i (n_i) should be carefully selected in order to avoid a generation of too large set of structures (combinatorial explosion). The "optimal" substituents could be prepared using fragments selected at the QSAR stage, since their contributions into activity (for linear models) allow one to estimate an impact of combining the fragment into larger species (R_i). In such a way, a focused combinatorial library could be generated.

The technology based on combining QSAR, generation of virtual libraries and screening stages has been implemented into ISIDA program and applied to computer-aided design of new uranyl binders belonging to two different families of organic molecules: phosphoryl containing podands [137] and monoamides [138]. QSAR models have been developed using different machine-learning methods (multi-linear regression analysis, associative neural networks [139] and support vector machines [140]) and fragment descriptors (atom/bond sequences and augmented atoms). Then, these models were used to screen virtual combinatorial libraries containing up to 11000 compounds. Selected hits were synthesized and tested experimentally. Experimental data well correspond to predicted the uranyl binding affinity. Thus, initial data sets were significantly enriched with new efficient uranyl binders, and one of new molecules was found more efficient than previously studied compounds.

A similar study was conducted for development of new 1-[2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives potentially possessing high anti-HIV activity [141]. This demonstrates universality of fragment descriptors and broad perspectives of their use in virtual screening and *in silico* design.

CONCLUSION

The power of fragment descriptors originates from their universality, very high computational efficacy, simplicity of interpretation, as well as their high diversity and versatility. The latest challenges in chemogenomics and high throughput virtual screening have raised their role in effective processing of huge amounts of relevant data and computer-aided design of new compounds.

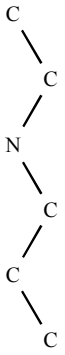
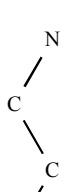
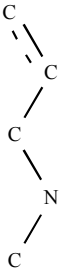
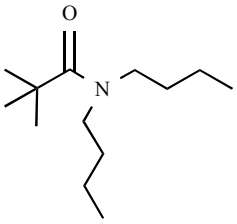
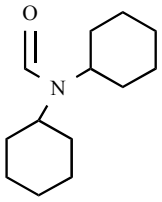
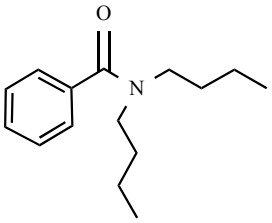
DATASET							
	0	10	1	5	0		-0.222
	0	8	1	4	0		0.973
	0	4	1	2	4		-0.066
	DESCRIPTOR MATRIX (occurrences of fragment)						**PROPERTY**
QSAR/QSPR MODEL	$PROPERTY_{CALC} = -0.36 * N_{C\text{-}C\text{-}C=N\text{-}C\text{-}C} + 0.27 * N_{C=O} + 0.12 * N_{C\text{-}N\text{-}C*C} + ...$						

Fig. (2). A general scheme of building QSAR/QSPR models based on fragment descriptors.

ACKNOWLEDGEMENT

The authors thank Dr I. Tetko and Dr G. Marcou for fruitful discussion.

REFERENCES

[1] Kubinyi, H.; Muler, G. *Chemogenomics in Drug Discovery*; Wiley-VCH Publishers: Weinheim, 2004.
[2] Gorse, A.D. *Curr. Top. Med. Chem*., **2006**, *6*, 3-18.
[3] Walters, W.P.; Stahl, M.T.; Murcko, M.A. *Drug Discov. Today,* **1998**, *3*, 160-178.
[4] Seifert, M.H.; Kraus, J.; Kramer, B. *Curr. Opin. Drug. Discov. Devel*., **2007**, *10*, 298-307.
[5] Cavasotto, C.N.; Orry, A.J. *Curr. Top. Med. Chem*., **2007**, *7*, 1006-14.
[6] Ghosh, S.; Nie, A.; An, J.; Huang, Z. *Curr. Opin. Chem. Biol*., **2006**, *10*, 194-202.
[7] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors.*; Wiley-VCH Publishers: Weinheim, 2000.
[8] Zefirov, N.S.; Palyulin, V.A. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 1112-1122.
[9] Japertas, P.; Didziapetris, R.; Petrauskas, A. *Quant. Struct.-Act. Relat*., **2002**, *21*, 23-37.
[10] Artemenko, N.V.; Baskin, I.I.; Palyulin, V.A.; Zefirov, N.S. *Russ. Chem. Bull*., **2003**, *52*, 20-29.
[11] Merlot, C.; Domine, D.; Church, D.J. *Curr. Opin. Drug Discov. Devel*., **2002**, *5*, 391-399.
[12] Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V.P. *J. Comput. Aided Mol. Des*., **2005**, *19*, 693-703.
[13] Jelfs, S.; Ertl, P.; Selzer, P. *J. Chem. Inf. Model*., **2007**, *47*, 450-459.
[14] Tatevskii, V.M. *Doklady Akademii Nauk SSSR,* **1950**, *75*, 819-822.
[15] Tatevskii, V.M.; Mendzheritskii, E.A.; Korobov, V. *Vestnik Moskovskogo Universiteta,* **1951**, *6*, 83-86.
[16] Bernstein, H. J. *J. Chem. Phys*., **1952**, *20*, 263-269.
[17] Laidler, K.J. *Canadian J. Chem*., **1956**, *34*, 626-648.
[18] Benson, S.W.; Buss, J.H. *J. Chem. Phys*., **1958**, *29*, 546-572.
[19] Free, S.M., Jr.; Wilson, J.W. *J. Med. Chem*., **1964**, *7*, 395-9.
[20] Hiller, S.A.; Golender, V.E.; Rosenblit, A.B.; Rastrigin, L.A.; Glaz, A.B. *Comput. Biomed. Res*., **1973**, *6*, 411-21.
[21] Hiller, S.A.; Glaz, A.B.; Rastrigin, L.A.; Rosenblit, A.B. *Doklady Akademii Nauk SSSR,* **1971**, *199*, 851-853.

[22] Golender, V.E.; Rozenblit, A.B. *Avtomatika i Telemekhanika,* **1974**, 99-105.
[23] Golender, V.E.; Rozenblit, A.B. *Med. Chem. (Academic Press),* **1980**, *11*, 299-337.
[24] Piruzyan, L.A.; Avidon, V.V.; Rozenblit, A.B.; Arolovich, V.S.; Golender, V.E.; Kozlova, S.P.; Mikhailovskii, E.M.; Gavrishchuk, E.G. *Khimiko-Farmatsevticheskii Zhurnal,* **1977**, *11*, 35-40.
[25] Avidon, V.V.; Pomerantsev, I.A.; Golender, V.E.; Rozenblit, A.B. *J. Chem. Inf. Comput. Sci*., **1982**, *22*, 207-214.
[26] Cramer, R.D., 3rd; Redl, G.; Berkoff, C.E. *J. Med. Chem*., **1974**, *17*, 533-5.
[27] Stuper, A.J.; Jurs, P.C. *J. Chem. Inf. Model*., **1976**, *16*, 99-105.
[28] Brugger, W.E.; Stuper, A.J.; Jurs, P.C. *J. Chem. Inf. Model*., **1976**, *16*, 105-110.
[29] Hodes, L.; Hazard, G.F.; Geran, R.I.; Richman, S. *J. Med. Chem*., **1977**, *20*, 469-75.
[30] Milne, M.; Lefkovitz, D.; Hill, H.; Powers, R. *J. Chem. Doc*., **1972**, *12*, 183-189.
[31] Adamson, G.W.; Cowell, J.; Lynch, M.F.; McLure, A.H. W.; Town, W.G.; Yapp, A.M. *J. Chem. Doc*., **1973**, *13*, 153-157.
[32] Feldman, A.; Hodes, L. *J. Chem. Inf. Model*., **1975**, *15*, 147-152.
[33] Willett, P. *J. Chem. Inf. Model*., **1979**, *19*, 159-162.
[34] Willett, P. *J. Chem. Inf. Model*., **1979**, *19*, 253-255.
[35] Willett, P.; Winterman, V.; Bawden, D. *J. Chem. Inf. Model*., **1986**, *26*, 36-41.
[36] Fisanick, W.; Lipkus, A.H.; Rusinko, A. *J. Chem. Inf. Model*., **1994**, *34*, 130-140.
[37] Hodes, L. *J. Chem. Inf. Model*., **1989**, *29*, 66-71.
[38] McGregor, M.J.; Pallai, P.V. *J. Chem. Inf. Model*., **1997**, *37*, 443-448.
[39] Turner, D.B.; Tyrrell, S.M.; Willett, P. *J. Chem. Inf. Model*., **1997**, *37*, 18-22.
[40] Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 1273-1280.
[41] Tong, W.; Lowis, D.R.; Perkins, R.; Chen, Y.; Welsh, W.J.; Goddette, D.W.; Heritage, T.W.; Sheehan, D.M. *J. Chem. Inf. Model*., **1998**, *38*, 669-677.
[42] Petrauskas, A.A.; Kolovanov, E.A. *Perspect. Drug Discov. Design*., **2000**, *19*, 99-116.
[43] Artemenko, N.V.; Baskin, I.I.; Palyulin, V.A.; Zefirov, N.S. *Doklady Chemistry*., **2001**, *381*, 317-320.
[44] Adamson, G.W.; Lynch, M.F.; Town, W.G. *J. Chem. Soc*., *C,* **1971**, 3702-3706.
[45] Hodes, L. *J. Chem. Inf. Comput. Sci*., **1981**, *21*, 132-136.
[46] Poroikov, V.V.; Filimonov, D.A.; Borodina, Y.V.; Lagunin, A.A.; Kos, A.*J. Chem. Inf. Comput. Sci*., **2000**, *40*, 1349-1355.
[47] Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 666-670.
[48] Xing, L.; Glen, R.C. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 796-805.
[49] Bender, A.; Mussa, H.Y.; Glen, R.C.; Reiling, S. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 170-178.
[50] Bender, A.; Mussa, H.Y.; Glen, R.C.; Reiling, S. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 1708-1718.
[51] Glen, R.C.; Bender, A.; Arnby, C.H.; Carlsson, L.; Boyer, S.; Smith, J. *IDrugs,* **2006**, *9*, 199-204.
[52] Rodgers, S.; Glen, R.C.; Bender, A. *J. Chem. Inf. Model*., **2006**, *46*, 569-576.
[53] Cannon, E.O.; Amini, A.; Bender, A.; Sternberg, M.J.E.; Muggleton, S.H.; Glen, R.C.; Mitchell, J.B.O. *J. Comput. Aid. Mol. Design*., **2007**, *21*, 269-280.
[54] Faulon, J.-L.; Visco, D.P., Jr.; Pophale, R.S. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 707-720.
[55] Faulon, J.-L.; Churchwell, C.J.; Visco, D.P., Jr. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 721-734.
[56] Churchwell, C.J.; Rintoul, M.D.; Martin, S.; Visco, D.P., Jr.; Kotu, A.; Larson, R. S.; Sillerud, L.O.; Brown, D.C.; Faulon, J.L. *J. Mol. Graph. Model*., **2004**, *22*, 263-73.
[57] Avidon, V.V.; Leksina, L.A. *Nauchno.-Tekhn. Inf. Ser*., **1974**, *2*, 22-25.
[58] Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci*., **1985**, *25*, 64-73.
[59] Horvath, D. In *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications*; Ghose, A., Viswanadhan, V., Eds.; Marcel Dekker: New York, 2001, p 429-472.
[60] Horvath, D.; Jeandenans, C. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 680-690.
[61] Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. *J. Chem. Inf. Model*., **2006**, *46*, 2457-2477.
[62] Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 391-405.
[63] *MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada*, *www.chemcomp.com*.
[64] Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. *J. Med. Chem*., **2005**, *48*, 6997-7004.
[65] Byvatov, E.; Sasse, B.C.; Stark, H.; Schneider, G. *ChemBioChem*., **2005**, *6*, 997-9.
[66] Nilakantan, R.; Bauman, N.; Dixon, J.S.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci*., **1987**, *27*, 82-85.
[67] Kearsley, S.K.; Sallamack, S.; Fluder, E.M.; Andose, J.D.; Mosley, R.T.; Sheridan, R.P. *J. Chem. Inf. Comput. Sci*., **1996**, *36*, 118-27.
[68] Bemis, G.W.; Murcko, M.A. *J. Med. Chem*., **1996**, *39*, 2887-93.
[69] Bemis, G.W.; Murcko, M.A. *J. Med. Chem*., **1999**, *42*, 5095-9.
[70] Guener, O.F. *Pharmacophore Perception, Development, and Use in Drug Design*.; Wiley-VCH Publishers: Weinheim, 2000.
[71] Langer, T.; Hoffman, R.D. *Pharmacophores and Pharmacophore Searches*.; Wiley-VCH Publishers: Weinheim, 2000.
[72] Wang, J.; Lai, L.; Tang, Y. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 1173-1189.
[73] Kazius, J.; McGuire, R.; Bursi, R. *J. Med. Chem*., **2005**, *48*, 312-20.
[74] Cunningham, A.R.; Rosenkranz, H.S.; Zhang, Y.P.; Klopman, G. *Mutat. Res*., **1998**, *398*, 1-17.
[75] Klopman, G. *J. Am. Chem. Soc*., **1984**, *106*, 7315-21.
[76] Klopman, G.; Rosenkranz, H.S. *Mutat. Res*., **1984**, *126*, 227-38.
[77] Rosenkranz, H.S.; Mitchell, C.S.; Klopman, G. *Mutat. Res*., **1985**, *150*, 1-11.
[78] Klopman, G. *Quant. Struct.-Act. Relat*., **1992**, *11*, 176-84.
[79] Klopman, G.; Rosenkranz, H.S. *Mutat. Res*., **1994**, *305*, 33-46.
[80] Klopman, G.; Chakravarti, S.K.; Harris, N.; Ivanov, J.; Saiakhov, R.D. *SAR QSAR Environ. Res*., **2003**, *14*, 165-180.
[81] Sanderson, D.M.; Earnshaw, C.G. *Hum. Exp. Toxicol*., **1991**, *10*, 261-73.
[82] Gombar, V.K.; Enslein, K.; Hart, J.B.; Blake, B.W.; Borgstedt, H.H. *Risk Anal*., **1991**, *11*, 509-17.
[83] Judson, P.N. *Pestic. Sci*., **1992**, *36*, 155-160.
[84] Judson, P.N. *J. Chem. Inf. Comput. Sci*., **1994**, *34*, 148-153.
[85] Barratt, M.D.; Rodford, R.A. *Curr. Opin. Chem. Biol*., **2001**, *5*, 383-8.
[86] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. *Adv. Drug Deliv. Rev*., **2001**, *46*, 3-26.
[87] Oprea, T.I. *J. Comput. Aided Mol. Des*., **2000**, *14*, 251-64.
[88] Veber, D.F.; Johnson, S.R.; Cheng, H.Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. *J. Med. Chem*., **2002**, *45*, 2615-23.
[89] Hann, M.M.; Oprea, T.I. *Curr. Opin. Chem. Biol*., **2004**, *8*, 255-63.
[90] Leo, A.J. *Chem. Rev*., **1993**, *93*, 1281-1306.
[91] Tetko, I.V.; Livingstone, D.J. In *Comprehensive Medicinal Chemistry II: In silico tools in ADMET*; Testa, B., van de Waterbeemd, H., Eds.; Elsevier: 2006, *5*, 649-668.
[92] Kubinyi, H. *Persp. Drug Discov. Design*., **1998**, *9-11*, 225–252.
[93] Martin, Y.C.; Kofron, J.L.; Traphagen, L.M. *J. Med. Chem*., **2002**, *45*, 4350-8.
[94] *Daylight Chemical Information Systems Inc*., *http://www.daylight.com*.
[95] *Barnard Chemical Information Ltd.*, *http://www.bci.gb.com/*.
[96] *Tripos Inc*., *http://www.tripos.com*.
[97] Jaccard, P. *Bull. Soc. Vaud. Sci. Nat*., **1901**, *37*, 241-272.
[98] Taylor, R. *J. Chem. Inf. Comput. Sci*., **1995**, *35*, 59-67.
[99] Delaney, J.S. *Mol. Divers*., **1996**, *1*, 217-22.
[100] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. *Angew. Chem. Int. Ed*., **1999**, *38*, 2894-2896.
[101] Hull, R.D.; Singh, S.B.; Nachbar, R.B.; Sheridan, R.P.; Kearsley, S.K.; Fluder, E. M. *J. Med. Chem*., **2001**, *44*, 1177-84.
[102] Hull, R.D.; Fluder, E.M.; Singh, S.B.; Nachbar, R.B.; Kearsley, S.K.; Sheridan, R. P. *J. Med. Chem*., **2001**, *44*, 1185-91.
[103] Willett, P.; Barnard, J.M.; Downs, G.M. *J. Chem. Inf. Comput. Sci*., **1998**, *38*, 983-996.
[104] Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 1177-1185.
[105] Ormerod, A.; Willett, P.; Bawden, D. *Quant. Struct.-Act. Relat*., **1989**, *8*, 115-29.

[106] Godden, J.W.; Furr, J.R.; Xue, L.; Stahura, F.L.; Bajorath, J. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 21-29.
[107] Godden, J.W.; Stahura, F.L.; Bajorath, J. *J. Med. Chem*., **2004**, *47*, 5608-11.
[108] Batista, J.; Godden, J.W.; Bajorath, J. *J. Chem. Inf. Model*., **2006**, *46*, 1937-44.
[109] Godden, J.W.; Bajorath, J. *J. Chem. Inf. Comput. Sci*., **2001**, *41*, 1060-1066.
[110] Batista, J.; Bajorath, J. *J. Chem. Inf. Model*., **2007**, *47*, 59-68.
[111] Maldonado, A.G.; Doucet, J.P.; Petitjean, M.; Fan, B.T. *Mol. Divers*., **2006**, *10*, 39-79.
[112] Bajorath, J. *Mol. Divers*., **2002**, *5*, 305-13.
[113] Waller, C.L. *Mol. Divers*., **2002**, *5*, 173-4.
[114] Agrafiotis, D.K.; Myslik, J.C.; Salemme, F.R. *Mol. Divers*., **1998**, *4*, 1-22.
[115] Trepalin, S.V.; Gerasimenko, V.A.; Kozyukov, A.V.; Savchuk, N. P.; Ivaschenko, A.A. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 249-58.
[116] Harper, G.; Bradshaw, J.; Gittins, J.C.; Green, D.V.S.; Leach, A.R. *J. Chem. Inf. Comput. Sci*., **2001**, *41*, 1295-1300.
[117] Geronikaki, A.A.; Dearden, J.C.; Filimonov, D.; Galaeva, I.; Garibova, T.L.; Gloriozova, T.; Krajneva, V.; Lagunin, A.; Macaev, F.Z.; Molodavkin, G.; Poroikov, V.V.; Pogrebnoi, S.I.; Shepeli, F.; Voronina, T.A.; Tsitlakidou, M.; Vlad, L. *J. Med. Chem*., **2004**, *47*, 2870-6.
[118] Katritzky, A.R.; Kuanar, M.; Slavov, S.; Dobchev, D.A.; Fara, D.C.; Karelson, M.; Acree, W.E., Jr.; Solov'ev, V.P.; Varnek, A. *Bioorg. Med. Chem*., **2006**, *14*, 4888-917.
[119] Katritzky, A.R.; Dobchev, D.A.; Fara, D.C.; Hur, E.; Tamm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V.P. *J. Med. Chem*., **2006**, *49*, 3305-14.
[120] Katritzky, A.R.; Kuanar, M.; Fara, D.C.; Karelson, M.; Acree, W.E., Jr.; Solov'ev, V.P.; Varnek, A. *Bioorg. Med. Chem*., **2005**, *13*, 6450-63.
[121] Mannhold, R.; Rekker, R.F.; Sonntag, C.; ter Laak, A.M.; Dross, K.; Polymeropoulos, E.E. *J. Pharm. Sci*., **1995**, *84*, 1410-9.
[122] Nys, G.G.; Rekker, R.F. *Eur. J. Med. Chem*., **1973**, *8*, 521-535.
[123] Leo, A.; Jow, P.Y.C.; Silipo, C.; Hansch, C. *J. Med. Chem*., **1975**, *18*, 865-868.
[124] Ghose, A.K.; Crippen, G.M. *J. Chem. Inf. Comput. Sci*., **1987**, *27*, 21-35.
[125] Ghose, A.K.; Crippen, G.M. *J. Comput. Chem*., **1986**, *7*, 565-577.
[126] Ghose, A.K.; Pritchett, A.; Crippen, G.M. *J. Comput. Chem*., **1988**, *9*, 80-90.
[127] Wildman, S.A.; Crippen, G.M. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 868-873.
[128] Suzuki, T.; Kudo, Y. *J. Comput. Aided. Mol. Des*., **1990**, *4*, 155-98.
[129] Convard, T.; Dubost, J.-P.; Le Solleu, H.; Kummer, E. *Quant. Struct.-Act. Relat*., **1994**, *13*, 34-37.
[130] Wang, R.; Gao, Y.; Lai, L. *Persp. Drug Discov. Design*., **2000**, *19*, 47-66.
[131] Wang, R.; Fu, Y.; Lai, L. *J. Chem. Inf. Comput. Sci*., **1997**, *37*, 615-621.
[132] Balakin, K.V.; Savchuk, N.P.; Tetko, I.V. *Curr. Med. Chem*., **2006**, *13*, 223-41.
[133] Varnek, A.; Kireeva, N.; Tetko, I.V.; Baskin, II; Solov'ev, V.P. *J. Chem. Inf. Model*., **2007**, *47*, 1111-22.
[134] Feuston, B.P.; Chakravorty, S.J.; Conway, J.F.; Culberson, J.C.; Forbes, J.; Kraker, B.; Lennon, P.A.; Lindsley, C.; McGaughey, G.B.; Mosley, R.; Sheridan, R.P.; Valenciano, M.; Kearsley, S.K. *Curr. Top. Med. Chem*., **2005**, *5*, 773-83.
[135] Green, D.V.; Pickett, S.D. *Mini Rev. Med. Chem*., **2004**, *4*, 1067-76.
[136] Green, D.V. *Prog. Med. Chem*., **2003**, *41*, 61-97.
[137] Varnek, A.; Fourches, D.; Solov'ev, V.P.; Baulin, V.E.; Turanov, A.N.; Karandashev, V.K.; Fara, D.; Katritzky, A.R. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 1365-1382.
[138] Varnek, A.; Fourches, D.; Solov'ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. *Solvent Extraction and Ion Exchange*., **2007**, *25*, 433-462.
[139] Tetko, I.V. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 717-728.
[140] Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer, 1995.
[141] Solov'ev, V.P.; Varnek, A. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 1703-1719.

Received: August 16, 2007 **Revised: November 29, 2007** **Accepted: November 29, 2007**