Advanced Exact Structure Searching in Large Databases of Chemical Compounds

Sergey V. Trepalin, Andrey V. Skorenko, Konstantin V. Balakin,* Anatoly F. Nasonov,
Stanley A. Lang, Andrey A. Ivashchenko, and Nikolay P. Savchuk

Chemical Diversity Labs, Inc., 11575 Sorrento Valley Road, San Diego, California 92121

Received August 16, 2002

Efficient recognition of tautomeric compound forms in large corporate or commercially available compound databases is a difficult and labor intensive task. Our data indicate that up to 0.5% of commercially available compound collections for bioscreening contain tautomers. Though in the large registry databases, such as Beilstein and CAS, the tautomers are found in an automated fashion using high-performance computational technologies, their real-time recognition in the nonregistry corporate databases, as a rule, remains problematic. We have developed an effective algorithm for tautomer searching based on the proprietary chemoinformatics platform. This algorithm reduces the compound to a canonical structure. This feature enables rapid, automated computer searching of most of the known tautomeric transformations that occur in databases of organic compounds. Another useful extension of this methodology is related to the ability to effectively search for different forms of compounds that contain ionic and semipolar bonds. The computations are performed in the Windows environment on a standard personal computer, a very useful feature. The practical application of the proposed methodology is illustrated by several examples of successful recovery of tautomers and different forms of ionic compounds from real commercially available nonregistry databases.

INTRODUCTION

As much as the modern pharmaceutical industry moves toward the high-throughput technologies in biology and medicine, there are a number of problems, such as the real-time recognition of tautomeric bonds and different forms of ionic bonds in the large in-house structural databases, that are still problematic today. This problem can be illustrated using the following example. Table 1 clearly indicates that even commercial versions of compound databases from a number of large vendors of exploratory small molecule chemistry for bioscreening often contain significant amount of tautomeric pairs. In several instances, the tautomeric forms of one compound have different prices! Because of license requirements, we keep the vendors anonymous. Thus when compound libraries are purchased for bioscreening, a fair number of identical compounds may be ordered, as many biotechnology companies do not have the special database management tools to identify and eliminate the duplicate but tautomeric structures. This problem is even more acute for the compound library vendors which have to effectively filter out the identical compounds among tens of thousands of samples before synthesis or acquisition. Another major problem related to the existence of tautomeric compound representations is the possible incorrect calculation of molecular descriptors resulting in different values of calculated molecular parameters for identical molecules.

Though less significant, the problem also exists of the potential occurrence of redundant multiple forms of compounds that contain ionic or semipolar bonds in their structures. It is known that the selection of a counterion for ionic compounds plays a very important role in the optimization of the pharmacokinetic properties of a potential drug. But at the early stages of drug discovery (primary bioscreening of large compound sets) the presence of multiple forms of identical compounds usually is redundant and consumes resources. Basically, there is a need for a special software tool that will permit an effective identification of tautomeric compounds and different salts of the same compound in large nonregistration compound databases.

Table 1. Occurrence of Tautomeric Pairs in Commercial Databases of Small Molecule Compounds (Released in 2001−2002)

vendor	ca. size of database	no. of tautomeric pairs[a]	no. of tautomers	no. of basic generic tautomeric structures	% of tautomers
V1	30 000	1	2	1	0.007
V2	50 000	28	56	28	0.112
V4	70 000	60	112	55	0.160
V5	180 000	66	129	64	0.072
V6	120 000	264	513	254	0.428
V7	100 000	1	2	1	0.002

[a] The number of tautomeric pairs for a basic generic structure is equal to the number of all possible nonidentical (e.g. A↔B is identical to B↔A, where A and B are tautomers, and these two conversions constitute only one tautomeric pair) conversions between the different tautomeric forms of this structure.

The problem of tautomerism is very complex, and a detailed analysis goes beyond this publication. First of all, tautomerism in its more general aspect is related to a number of more local phenomena such as the type of migrating group, cationotropic and anionotropic properties, valence tautomerism, and tautomerism relating to migration of the neutral groups in molecules. In some instances, photoisomerism (*cis*−*tr*ans isomerization) is referred to as tautomeric conversion. In the current work, only the most frequently encountered tautomeric possibilities which are important in

* Corresponding author phone: (858)794-4860; fax: (858)794-4931; e-mail: kvb@chemdiv.com.

10.1021/ci025582d CCC: $25.00 © 2003 American Chemical Society
Published on Web 02/22/2003

the chemical science (including the chemistry of physiologically active substances), a subtype of cationotropic tautomerism, namely, the proton migration (prototropic tautomerism) is considered. It should also be emphasized that the developed algorithm does not predict the most favorable tautomeric form, as the tautomeric equilibration is a function of a complex number of variables including concentration, temperature, pressure, solvent type, pH, etc.

It should be noted that several reviews address the theoretical aspects of phenomenon called tautomerism.[1−3] This manuscript describes a useful applied tool for the management of compound databases rather than for the theoretical studies.

METHODS

The exact structure search algorithm is an integrated utility of the ChemoSoft software environment for chemical database management.[4,5] All the tests were performed using a standard personal workstation with the Pentium 1.8 GHz processor on a Windows 2000 platform.

RESULTS

In this work, we report an algorithm and software program for conducting tautomer searches in aromatic and nonaromatic systems with common 1,3-migration as well as with distant migration of hydrogen atom. The tautomeric equilibration with the hydrogen atom migration between the distant atoms (1,5 and more) in acyclic compounds and the tautomeric equilibration related to transformations of molecular scaffold (such as cyclic and acyclic forms in carbohydrates) are not considered here. In a special part of this paper, we elucidate an algorithm for searching chemical structures having ionic and semipolar bonds.

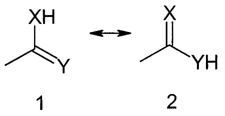
Several reports about tautomer search in chemical databases can be found in open datasources,[6−11] though in most cases, the details of the applied algorithms are not disclosed. Thus tautomer recognition and searching were issues that had to be addressed several decades ago by producers of the global chemical compound registry systems such as Chemical Abstracts Service (CAS) and Beilstein. Such systems require a structure representation by a unique structure diagram or connection table. This task poses serious problems to chemical information systems when the substance has several possible representations, chemically equivalent but structurally distinct. The CAS registry system[10] handles the problem by normalizing (i.e., recognizing the equivalence of) tautomeric and alternating bond structures, replacing the explicit single and double bonds with special tautomer and alternating bonds, and associating the migrating hydrogen in a tautomer with a group of atoms rather than just a single atom. In this case, single/double bond patterns and specific migrating group locations have been replaced by normalized data, and all forms of the tautomeric structure lead to the same registry structure record.

Another approach to the canonicalization and enumeration of tautomers has been described.[7] All conjugated double bonds and single bonds between them are assigned as an undefined bond type both in the query structure and in the structure to be searched. Then the conventional substructural search with alternation of the single and double bonds is carried out. While this procedure identifies all the tautomers in a compound base including the tautomers with distant hydrogen migration, it requires significant computation time and would not be applicable in the cases where real-time scanning of large compound databases when the use of standard personal computers is required.

Canonical Structure. We have found that more than 99% of all compounds for which different tautomeric forms related to hydrogen migration which have been reported in the scientific literature can be unambiguously assigned to a unique canonical structure. In this case, the task of the tautomer search can be reduced to generation of a database of canonical structures. Subsequent conventional structure search using an indexing algorithm enables to significantly decrease the computational time required for searching complex databases.

The term "canonical structure" implies the chemical structure to which the tautomeric forms of a compound can be reduced using strongly defined rules. These rules are similar but not identical to that used in the CAS registry system.[10] In addition, we use an alternative representation of canonical structure that is different to the normalized structure of the CAS registry system, as we do not replace single/double bond patterns by normalized data (such as alternating or tautomeric bonds) but use their standard representations. The canonical structure can be entirely different from the actual chemical structure. In this paper, the canonical structures of the following types of tautomeric equilibration are described: (1) 1,3-migration of hydrogen atom in acyclic compounds, (2) hydrogen migration in five-member aromatic systems, and (3) hydrogen migration in heteroaromatic systems with the loss of aromaticity.

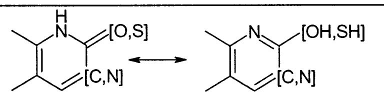
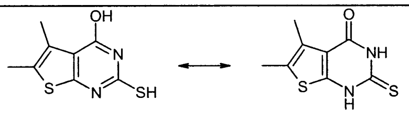
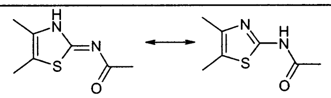
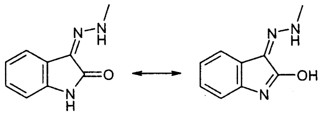
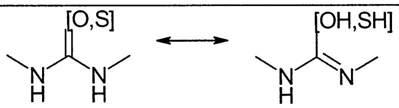
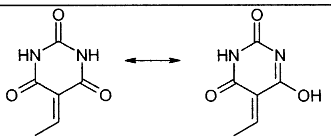
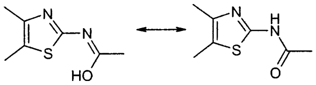
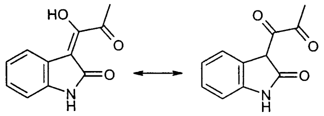
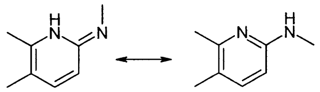
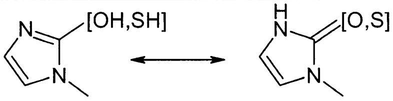
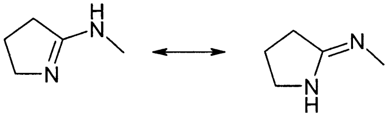
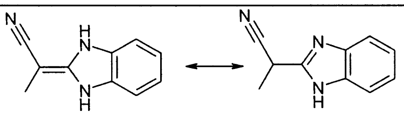
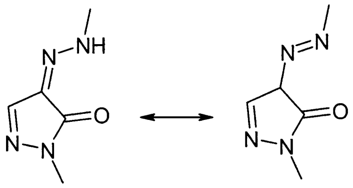
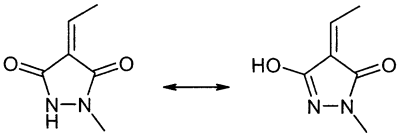
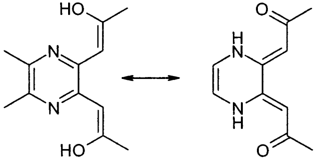
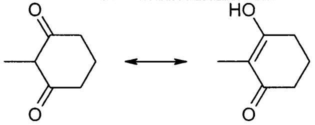
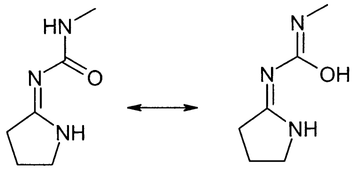
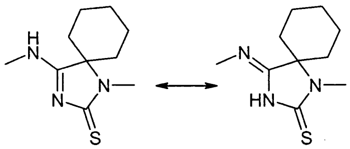
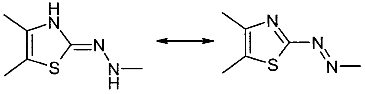
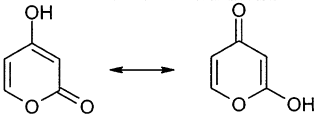
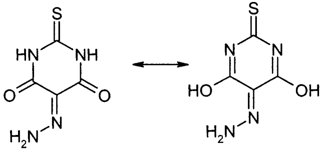
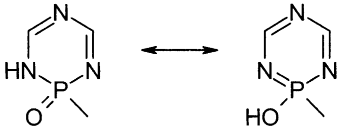
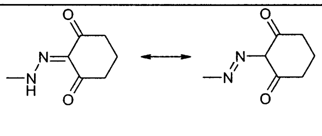
1,3-Migration of Hydrogen Atom in Nonaromatic Compounds. An example of tautomeric equilibration (fragments **1** and **2**) is shown below:



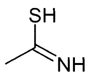
The double bond in structures of this type is assigned to the most electronegative atom: O > N > S > Se > Te > C. If the atoms X and Y are equal, then the double bond is assigned to the atom with the higher substituent according to Cahn−Ingold−Prelog rules (CIP rules). If all the substituents are equal, then the migration of double bond does not lead to tautomer formation and their position is assigned in a random manner. In a special case, when X═Y═C (carbon), the position of double bond is not tautomeric, as in the absence of heteroatom the proton has very low mobility, and 1,3-migration, in general case, leads to distinctly different compounds that can be isolated as individual distinct substances.

It should be noted that in the CAS and Beilstein registry systems,[8,10] X and Y are limited to nitrogen and chalcogen atoms. Keto−enol tautomers are not recognized as such and are registered and named as distinct structures, as are all other variations of HY−C═C ↔ Y═C−CH type exemplified below. Such modification is caused by a wide spread occurrence of redundant tautomeric compound forms related to this kind of tautomerism in the corporate compound libraries (Table 2, the Discussion section).

Table 2. Tautomeric Pairs Most Frequently Encountered in Commercial Databases of Small Molecule Compounds

Entry	Tautomeric Pairs	Number
1		**85**
2		**60**
3		**54**
4		**21**
5		**20**
6		**15**
7		**14**
8		**12**
9		**11**
10		**10**
11		**10**
12		**9**
13		**8**
14		**8**
15		**5**
16		**5**
17		**5**
18		**5**
19		**4**
20		**4**
21		**3**
22		**3**
23		**2**

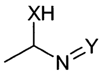
It should also be noted that the priority selection by an electronegativity principle can lead to a hypothetical tautomeric form that might not exist or exists only in a small percentage. For example, fragment **3** used as a canonical form exists only as a minor component of the tautomeric equilibrium.



3

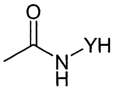
However, this fact does not affect the tautomer search, as both the chemical structure in a database and the query structure are reduced to the same representation.

The tautomeric equilibration with participation of nitrogen atom should be considered separately. In this case, we tried to generate rules of reduction to a canonical structure that lead to chemically legitimate structures



4

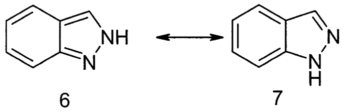
where X,Y are heteroatoms. The structures **4** are usually nitroso-compounds (tautomeric forms of oximes, Y=O) or azo-compounds (tautomers of hydrazones, Y=N). These structures are first analyzed and if the tautomer is found, it is presented in a standard form—the oxime or hydrazone. The found structure can further be modified using the aforementioned electronegativity rules. Thus the generation of a canonical structure for 1,3-migration is an iterative process, which is discontinued after the final structure meets the electronegativity and CIP rules. In a particular case, when X=O, structure **5** is obtained.



5

In addition to the points noted above, it should be noted that these rules are also applicable for cyclic nonaromatic compounds.

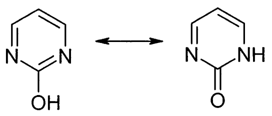
Hydrogen Migration in Five-Member Aromatic Systems. Compounds **6** and **7** represent an example of tautomeric equilibration of this type.



6 7

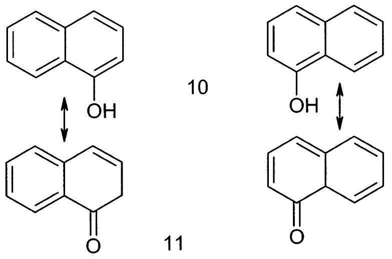
Reduction to a canonical structure is maximally simplified in this case: all heteroatom-bound hydrogen atoms in aromatic ring are removed, and all bonds are considered identical and aromatic. The canonical structure is thus always uniquely defined.

Lactim−Lactam Tautomerism in Aromatic Compounds. Compounds **8** and **9** represent an example of tautomeric equilibration of this type.



8 9

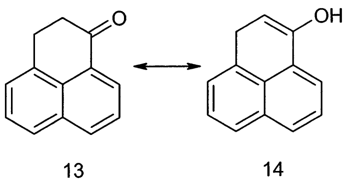
In this case, the generation of a canonical structure using the electronegativity and CIP rules is inapplicable for the reasons of ambiguous alternation of single and double bonds in the cycle (structures **10**−**12**):



10 11 12

1,3-Migration of hydrogen atom in α-naphthol **10** leads to different chemical structures **11** and **12**, depending on the initial positioning of the double bonds. Therefore for tautomeric equilibration associated with the loss of aromaticity, a tautomer with aromatic conjugation is selected. At that, the problem of alternation of single and double bonds may appear, which can be solved using the following algorithm. Alternation of single and double bonds is initiated in a cycle to which a heteroatom is connected through the double bond. If the aromatic system can be generated for this cycle, the resulting structure is considered canonical. If the aromatic cycle cannot be obtained, and the desired cycle is related to a polycyclic fragment, the alternation of single and double bonds is performed for the whole polyaromatic system. In this case, the following iterative procedure is used. At first, a single bond is selected that is bound to a heteroatom in a five-member ring. If such a fragment is absent, the single bond is selected randomly. The double bond connected to the initial single bond is further positioned. During this procedure, an uncertainty can appear concerning the positioning of each next bond. The procedure is repeated until the whole aromatic polycyclic system is generated or until all possible alternation routes and all possible positionings of the first single bond relative to a heteroatom in five-member ring are tested.

Sometimes, it appears impossible to obtain an aromatic cycle using any kind of alternation (structures **13** and **14**).

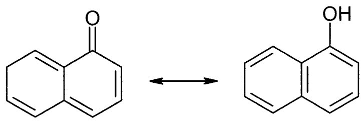


13 14

For such compounds, structure **13** is used as a canonical in accordance with the electronegativity rule.

Sometimes, obviously incorrect representation (intentional or unpremeditated) of a structure is possible with the loss of aromaticity of several cycles (structure **15**). Usually, such incorrect structures are ignored by the chemical registry systems, but their occurrence poses a real problem for the owners of the in-house databases, particularly those collected from different sources.

For the solution of this problem, each acyclic double bond (C=O) is considered ordinary and the number of double bonds in a polycyclic system is increased by 1 for each bond. Subsequently, alternation of ordinary and double bonds is



15 16

carried out for the whole polyaromatic system. If the arrangement of double bonds is found for which the whole system remains aromatic, the novel structure is considered canonical and saved in a database for searching. If the alternation of ordinary and double bonds is impossible, the initial structure is used as a canonical.

Exact Structure Search in Ionic Compounds. A large number of compounds, such as amines and carboxylic acids, can be isolated in the form of their salts. The selection of a counterion plays significant role at the stages of clinical trials and optimization of the pharmacokinetic properties of a drug candidate. But at the early stages of drug discovery usually associated with in vitro screening of large compound sets, the presence of multiple ionic forms of identical compounds is redundant. Therefore, we implemented a special algorithm of reduction to a canonical structure of different forms of ionic compounds. The situation is opposite in the case of the chemical registry systems, where such “normalization” of different salts can lead to erroneous results.

Thus in general case, the type of a cation (e.g. sodium, calcium) or anion (e.g. chloride, bromide) is not essential for the biotesting. However, the convenient exact structure search leads to finding the different structures. Moreover, the salts can often be variously represented: for example, sodium acetate can be drawn either with the covalent Na−O bond or in the form of cation and anion. Both these structures will be considered distinct by the conventional search algorithm.

In this case, the following algorithm is used for reduction to a canonical structure:

(1) if a metal atom is present in a molecule, all covalent bonds with the metal atom are removed (the metals are Li, Be, Na, Mg, Al, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Rb, Sr, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, Cs, Ba, La and lanthanides, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Fr, Ra, Ac and actinoids);

(2) if an ionized metal or ammonium cations are present, they are considered noncharged (an exception is the quarternary ammonium salts; they are reduced to canonical structures without changing the charges on cation and anion); if all hydrogen atoms in the ammonium cation are represented in an explicit form, one hydrogen atom is removed;

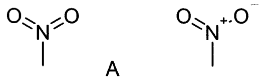
(3) all ionized carboxyl, sulfonyl, hydroxyl, and sulfide groups are considered noncharged;

(4) all unbound atoms are removed; this operation excludes metal cations and simple anions from the search;

(5) if several unbound atoms are present in a molecule, they are searched in a database that contains standard cations and anions; if successfully found, the fragment is removed from the molecule’s representation;

(6) this step is optional: if the molecule still contains several fragments, a maximal weight fragment is used; this option avoids errors related to the incompleteness of the database of standard cations and anions; on the other hand, there is a risk of the loss of the centrally significant fragment.

Semipolar Bonds. A typical example of uncertainty that appear as a result of the occurrence of semipolar bonds is shown:

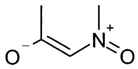


17

Though the correct representation of nitro-group **17** is shown in picture B, the representation A is also frequently used. However, these forms are recognized as distinct structures by many search systems (e.g. ISIS/Base 2.1.1).[12]

We suggest the following solution to this problem. All bound atoms in the structure are analyzed, and if they are oppositely charged, the corresponding charges are increased (for negatively charged atoms) or decreased (for positively charged) by 1, and the bond order between these atoms is also increased by 1 (the ordinary bond is transformed into the double bond, the double − into the triple). Thus the representation B of the nitro-group will be reduced to representation A, and the latter is considered canonical.

This algorithm does not work for compounds which bear the charges localized on distant unbound atoms (structure **18**).



18

The Search Procedure. After all the canonical structures in a database are generated, the indexing algorithm is further applied for the exact structure search. One of these indices is the molecular weight which is stored as a variable of type “single” (4 bytes). Using exhaustive analysis of all molecular possibilities, we have observed that the molecular weight unambiguously defines the molecular formula for a molecule with molecular weight less than 1000. In addition, two indices are calculated (type “integer”, size of each index is equal to 4 bytes) similar to these reported earlier[13] that are sensitive to molecular topology. The obtained 12-byte number plays role of a filter: two structures are identical only if their 12-byte numbers are equal. It should be noted that the equality of these 12-byte numbers is required but an insufficient condition of the identity of two molecules. A sufficient condition is achieved by verification of equality of molecular graphs. However, after analysis of about 53 millions of virtual compounds, we were able to find only two different structures with equal 12-byte numbers. This is related to the fact that the dynamic range of this 12-byte variable is equal to $2^{96} = 7.9 \times 10^{28}$, and the probability of equal indices for differing structures is very low. Therefore, the atom-by-atom verification of molecular graphs identity is possible in our algorithm only as an option.

To accelerate the search, the obtained 12-byte numbers are further sorted in a descending order and a bisection algorithm is used for the exact structure search. The bisection algorithm proceeds in the following way. Suppose N_{min} and N_{max} are the minimal and maximal numbers, and N_t is the target value. Consider then $N_{mid,1}$ is the midpoint of interval $[N_{min}, N_{max}]$. If $N_{mid,1} = N_t$, we found the solution. If $N_{mid,1} <$

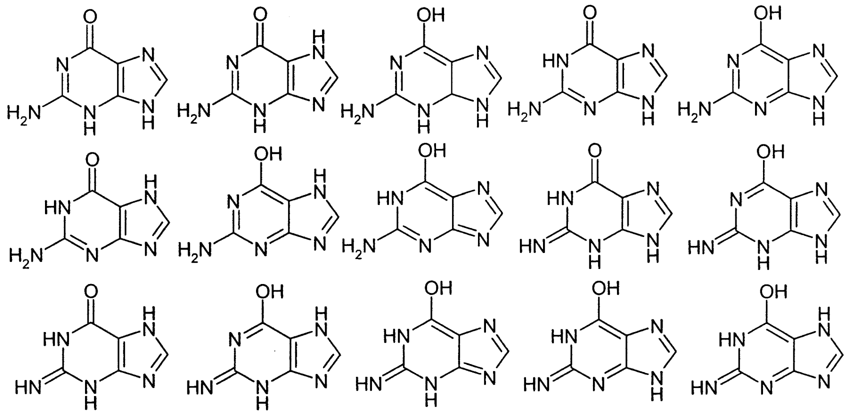
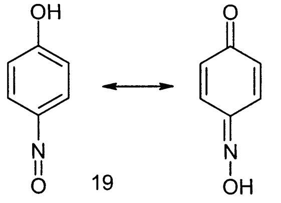


Figure 1. Fifteen tautomeric forms of guanine base used for testing the tautomer search engine.

N_t, the bisection method continues then focusing on interval $[N_{mid,1}, N_{max}]$; and vice versa, if $N_{mid,1} > N_t$, the bisection method continues with interval $[N_{min}, N_{mid,1}]$. In either case the algorithm continues with an interval lowered by a factor of 2. The process is continued until for the *i*th iteration $N_{mid,i} = N_t$. The bisection algorithm convergence is $\log_2 N$, where N is the number of compounds.

Results of Testing. To test the described tautomer search engine, we used 15 tautomeric forms of guanine reported earlier[7] (Figure 1). If any of the structures was queried, all the others could be effectively retrieved. However, it should be noted that for the structures **19** and **20** reported in the same publication, the described algorithm does not work: both structures are considered different. The algorithm cannot generate the canonical structures with a simultaneous rebuilding of the aromatic cycle and aliphatic chains for these particular tautomers.



19 20

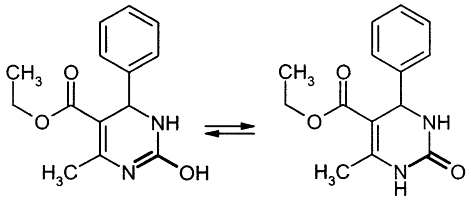
For a database consisting of 100 000 structural entries, approximately 25 min is required for a standard PC (Pentium 1.8 GHz processor) on a Windows 2000 platform to create a database of canonical structures. The time increases linearly with the size of the database. After all the canonical structures are generated, the final search takes a very short time. For example, for a comparison of two 100 000-compound databases that contain both tautomers and ionized compounds less than 3 s are required using the same computer.

DISCUSSION

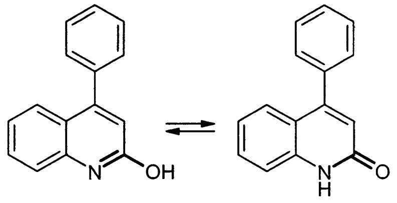
Tautomer Occurrence in Commercial Databases. Table 1 contains data on the frequency of the occurrence of tautomeric pairs in a number of commercial databases from several leading providers of exploratory small molecule chemistry for bioscreening. All these tautomers were revealed using the described algorithm. There are a considerable number of real tautomers in these databases; the most representative examples of tautomeric pairs found in these databases are shown in Table 2.

Examples of Tautomers. In this section, several examples of tautomeric compound forms found with the use of the described methodology are shown. All these examples are taken from the real commercial databases of small molecule compounds available for purchase for bioscreening. Highlighted are the fragments involved in the tautomeric transformation.

Lactim–Lactam Tautomerism. This is a prevailing type of tautomeric conversion found in the commercial databases studied. Two typical examples are presented here as an illustration (compounds **21–24**).



21 22

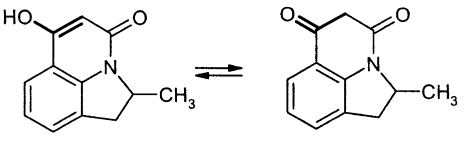


23 24

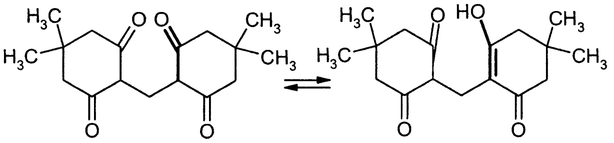
Keto–Enol Tautomerism. This type of tautomerism plays a very important role in the synthetic organic chemistry, and examples of this type are frequently encountered in the databases. As it was mentioned above, this type of tautomerism as well as other HY–C=C↔Y=C–CH conversions are not recognized by the CAS and Beilstein registry systems. Two examples of the keto–enol tautomer conversion in nonaromatic cycles are shown (compounds **25–28**).

Imine–Enamine Tautomerism. Structures **29** and **30** represent an interesting example of 1,4→3,4-dihydropyrimidine conversion within a polycyclic system. Again, these structures are registered by the CAS and Beilstein registry systems as distinct molecules.

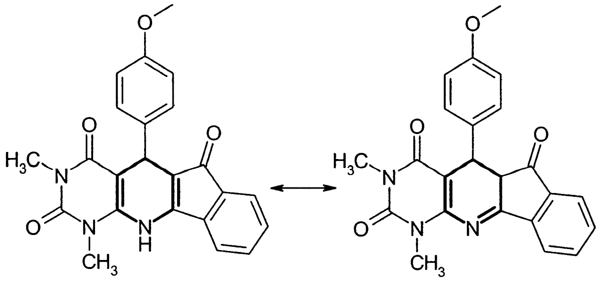
Multiple Tautomeric Forms. Some compounds can be represented in more than two tautomeric forms. Though



25 26

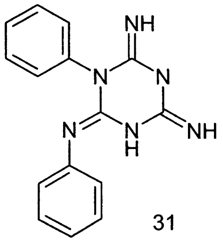


27 28

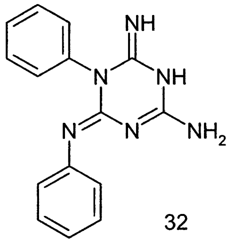


29 30

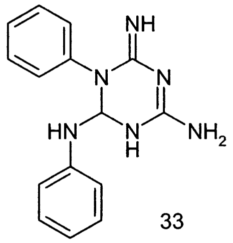
infrequent, examples of such structures are to be found in the compound databases (compounds **31**−**33**).



31

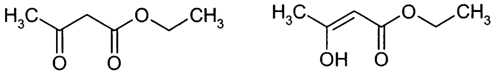


32



33

Preferable Tautomeric Forms. Care should be taken when using the described algorithm for exclusion of hypothetically redundant multiple tautomeric forms from a compound database. In the scientific literature many examples are described of stable tautomeric compound forms that can be isolated in an individual state. A classical example is acetoacetic ester.[1] Two tautomeric forms of this compound can be isolated in different aggregate states:



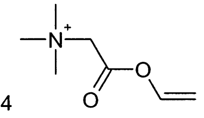
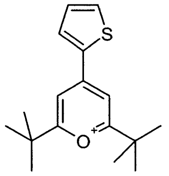
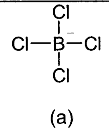
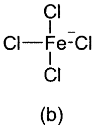
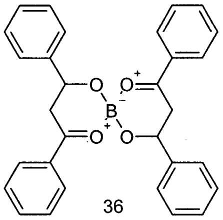
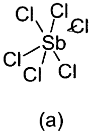
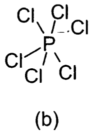
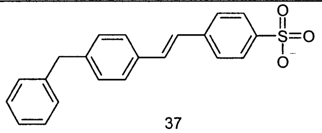
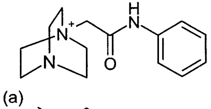
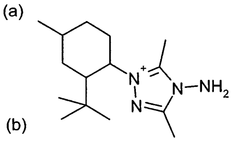
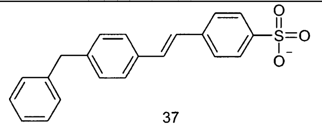
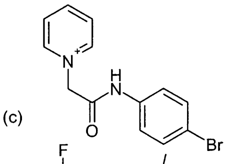
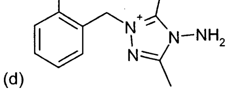
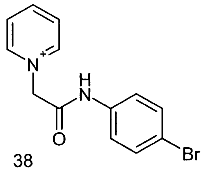
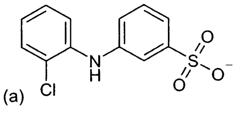
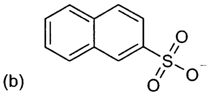
Solid, m.p. = -39 °C Liquid even at -78 °C

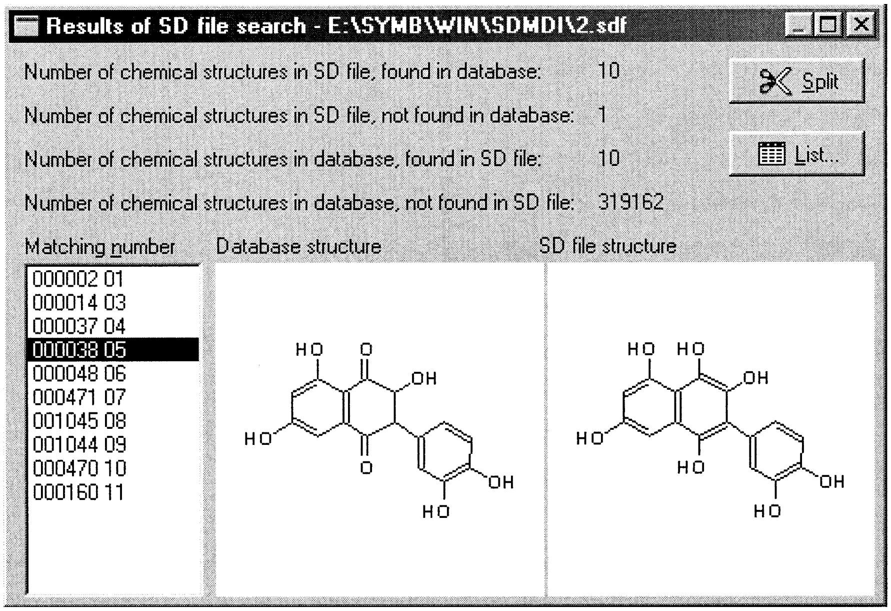
For the computational medicinal chemistry studies, it is desirable to find the most preferable tautomer, as the calculation of molecular properties required for different types of molecular modeling procedures, such as QSAR or docking, can lead to contradictory results for identical compounds. As a result, the predictive power of the models could be decreased. For example, it was shown[14] that the tautomers of 30 known nucleobase analogues docked into the active site of Herpes simplex virus type 1 thymidine kinase achieved different scores as compared to earlier

Table 3. Temperature Influence on the Tautomeric Equilibration of Pentane-2,5-dione[11]

temperature, °C	22	180	275
% of enol form	95	68	44

Table 4. Different Forms of Ionic Compounds Found in Sigma-Aldrich Catalog[17]

№	Form 1	Form 2
1.	 34	Br^- (a) Cl^- (b)
2.	 35	 (a)  (b)
3.	 36	 (a)  (b)
4.	 37	 (a)  (b)
5.	 37	 (c)  (d)
6.	 38	 (a)  (b)



Results of SD file search - E:\SYMB\WIN\SDMDI\2.sdf

Number of chemical structures in SD file, found in database: 10

Number of chemical structures in SD file, not found in database: 1

Number of chemical structures in database, found in SD file: 10

Number of chemical structures in database, not found in SD file: 319162

Matching number
000002 01
000014 03
000037 04
000038 05
000048 06
000471 07
001045 08
001044 09
000470 10
000160 11

Figure 2. Results of a structure data file comparison with an existing database in the ChemoSoft software environment.

screenings in which tautomerism was disregarded. However, multiple factors, such as pH of the solution,[1] concentration,[1] temperature,[15,16] light,[17] nature of solvent,[1,18] intermolecular interactions,[19] etc., can affect the tautomeric equilibration or the presence of one tautomer over the other.

For example, the temperature dependence of tautomeric equilibration for pentane-2,5-dione can be illustrated by Table 3 (the data are given for an aqueous solution).[14]

In another example, the solvent effect on the ground-state energy difference of meridine (enol form) and meridin-12-(13H)-one (keto form) was examined.[15] The results show that in apolar solvents, the enol form is the predominant species, whereas in polar solvents, the enol and keto forms are present in an almost equal amount. Rare tautomeric forms of compounds can also be stabilized by intermolecular forces. A ruthenium(III) complex was recently described[16] with an adenine derivative coordinated as a monodentate ligand through the exocyclic N6 nitrogen; two intramolecular hydrogen bonds stabilize the coordination of the adenine, which is present as a neutral ligand in the rare tautomeric imine form with a unique short C6−N6 distance of 1.293-(3) Å.

The complexity of the recognition of a more preferable tautomeric form of a compound, based on the high degree of influence of a number of variables, makes this a very difficult task, and this problem is not addressed in the current version of this algorithm.

Different Salts of Ionic Compounds. Table 4 demonstrates several examples of pairs of identical ionic compounds with different counterions. All these pairs were found in Sigma-Aldrich catalog[20] with the use of the described algorithm. Of course, in this case, the possibility of occurrence of different salts of compounds is specified by the catalog publishers.

Examples 1−3 (compounds **34a,b**−**36a,b**) represent three organic cations having different anionic parts that could be found in an internal database of standard cations and anions. Examples 4 and 5 (compounds **37a**−**d**) demonstrate the possibility of finding compounds having several counterions with complex structures. Entries 5 and 6 represent an interesting example where the same ionic fragment can play either a role of the central fragment (structure **38**) or of the counterionic part (cationic part of the compound **37c**). In this case, attribution of the fragment **38** to the main or supplementary category was made using the principle of larger molecular weight. Here we have an example of structures for which such a division is very relative.

Computational Platform. All the calculations were carried out on a Windows 2000 platform in the ChemoSoft environment. The application of the ChemoSoft software for the solution of different tasks in the field of computational medicinal chemistry was recently reported.[4,5,21] As an illustration, Figure 2 shows a ChemoSoft window with the typical results of a structure data file comparison with an existing database. The identification numbers of 10 found tautomers are indicated in the left panel. Two tautomeric forms of a compound are shown in the structure boxes.

CONCLUSIONS

In this paper, we report an algorithm and software program for exact structure search in large compound databases. The algorithm enables effective high-throughput search for tautomeric compound forms as well as different salts of ionic compounds in an automated fashion. The approach is based on the generation of a canonical structure, i.e., the chemical structure to which the tautomeric or ionic forms of a compound can be reduced using strongly defined rules. The handling of tautomeric and alternating bond structures is similar to methods of structure normalization used in CAS chemical registry system. The distinctive features of our work are the alternative concept of a "standard", canonical structure; ability to work with tautomeric conversions of HY−C=C ↔Y=C−CH type; ability of simultaneous handling of tautomeric forms; different forms of compounds with ionic and semipolar bonds; and even incorrectly represented structures with the loss of aromaticity. It is important that in contrast to the global chemical registry systems, such as CAS or Beilstein, the management of such a wide number of tasks is available for large corporate databases (up to millions of compounds) in a real time using moderate computational requirements. The limitations of this methodology are related to the complexity of the phenomenon of tautomerism. Thus only one prototropic subtype of cationotropic tautomerism is considered. Within this subtype, the tautomeric equilibration with the hydrogen atom migration between the distant atoms (1,5 and more) in acyclic compounds, and the tautomeric equilibration related to transformations of a molecular scaffold (such as cyclic and acyclic forms in carbohydrates) were not considered. It should also be noted that the proposed algorithm does not provide information concerning the preferable tautomeric form, as the tautomeric equilibration is a function of a complex number of micro- and macroenvironmental factors. At the same time, we tried to implement, where possible, the chemically legitimate representations of canonical structures (as in the case of oximes and hydrazones).

The practical application of the developed methodology is illustrated using the successful recovery of different tautomeric compound forms (including multiple forms) as well as different forms of ionic compounds in commercial databases of organic compounds. We anticipate that the developed algorithm will become very useful for routine analysis of compound databases. For the vendors of exploratory organic chemistry as well as for biotechnology companies involved in high-throughput screening of large compound sets, this tool provides a fast and convenient method for elimination of multiple structural representations of same compounds. For computational medicinal chemistry studies, this tool will provide an early alert on possible redundant information in reference databases that serve as a source of structural data.

ACKNOWLEDGMENT

We thank Dr. Lisa Staggenborg (Brains For Business, Worthington, OH) for invaluable help and discussions.

REFERENCES AND NOTES

(1) *March's Advanced Organic Chemistry: Reaction, Mechanisms, and Structure*, 5th ed.; Smith, M. B., March, J., Eds.; Wiley-Interscience: New York, Chichester, Weinheim, Brisbane, Singapore, Toronto, 2001; Vol. 1, Chapter 2, p 73.

(2) Elguero, J.; Marzin, C.; Katritzky, A. R.; Linda, P. The Tautomerism of Heterocycles. In *Advances in Heterocyclic Chemistry, Suppl. 1*; Academic Press: New York, 1976.

(3) Sugawara, T.; Takasu, I. Tautomerism in the Solid State. *Adv. Phys. Org. Chem.* **1999**, *32*, 219−265.

(4) Trepalin, S. V.; Yarkov, A. V. CheD: Chemical Database Compilation Tool, Internet Server, and Client for SQL Servers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 100−107.

(5) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. Ph.; Ivaschenko A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249−258.

(6) Taylor, K. T.; Hounshell, W. D.; Nourse, J. G.; Christie, B.; Leland, B. A. Structure searching: What You Get Is What You Wanted. MDL Information Systems. Internet presentation; http://www.lib.uchicago.edu/cinf/222nm/presentations/222nm007.pdf.

(7) Sayle, R.; Delany, J. Canonicalization and Enumeration of Tautomers. Materials of EuroMug-99, 28−29 Oct 1999, Cambridge, UK.

(8) CrossFire Structure and Reaction Searching, Manual: Version 2, September 1996, for Beilstein Commander Version 2.1 and CrossFire Server Version 3.x; http://www.mimas.ac.uk/crossfire/docs/pdf/xfss2e.pdf.

(9) Structure Searching, Internet lecture; November 1999; Indiana University; http://www.indiana.edu/∼cheminfo/400css8.html.

(10) Mockus, J.; Stobaugh, R. E. The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18−22.

(11) Staggenborg, L.; Cada, K.; Cross, K.; Deacon, D.; Dixon, L.; King, L.; Smith, H. Chemical Structure Conventions and Searching in the CAS Registry File via SciFinder. Fourth International Conference on Chemical Structures, 2−6 June 1996, Noordwijkerhout, Netherlands.

(12) MDL Information Systems, http://www.mdl.com/.

(13) Hu, C.-Y.; Xu, L. On Highly Discriminating Molecular Topological Index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 82−90.

(14) Pospisil, P.; Ballmer, P.; Folkers, G.; Scapozza, L. Tautomerism of Nucleobase Derivatives and Their Score in Virtual Screening to Thymidine Kinase. 224nd American Chemical Society National Meeting, Boston, MA, August 2002.

(15) Hush, N. S.; Livett, M. K.; Peel, J. B.; Willett, G. D. Variable-Temperature Ultraviolet Photoelectron Spectroscopy of the Keto−Enol Tautomers of Pentane-2,4-dione. *Aust. J. Chem.* **1987**, *40*, 599−609.

(16) Benedict, C.; Langer, U.; Limbach, H.-H.; Ogata, H.; Takeda, S. Observation of Thermal Tautomerism of Thermochromic Salicylideneaniline Derivatives in the Solid State by 15N CPMAS NMR down to Cryogenic Temperatures. *Ber. Bunsen-Ges. Phys. Chem.* **1998**, *102*, 335−339.

(17) Schwoerer, M.; Wirz, J. Photochemical Reaction Mechanisms of 2-Nitrobenzyl Compounds in Solution. I. 2-Nitrotoluene: Thermodynamic and Kinetic Parameters of the aci-Nitro Tautomer. *Helv. Chim. Acta* **2001**, *84*, 1441−1458.

(18) Sato, T.; Kataoka, M. Solvent Effects on Relative Stability of Meridine and its Tautomer: MO Calculations. *Heterocycles* **2001**, *54*, 55−60.

(19) Velders, A. H.; van der Geest, B.; Kooijman, H.; Spek, A. L.; Haasnoot, J. G.; Reedijk, J. Ruthenium(III) Coordination to the Exocyclic Nitrogen of 9-Methyladenine and Stabilisation of the Rare Imine Tautomer by Intramolecular Hydrogen Bonding. *Eur. J. Inorg. Chem.* **2001**, *2*, 369−372.

(20) Sigma-Aldrich Catalog of Rare Chemicals; July 2001.

(21) Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. I.; Savchuk, N. P. Property-Based Design of GPCR−Targeted Library. *J. Chem. Inf. Comput. Sci.*, published on Web, Oct. 31, 2002.

CI025582D