The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening[†]

Frank Oellien,[‡] Jörg Cramer,[‡] Carsten Beyer,[‡,||] Wolf-Dietrich Ihlenfeldt,[§] and Paul M. Selzer*,[‡]

Intervet Innovation GmbH, BioChemInformatics, Zur Propstei, D-55270 Schwabenheim, Germany, and
Xemistry GmbH, Auf den Stieden 8, D-35094 Lahntal, Germany

Received March 27, 2006

In the field of in silico screening, many applications do not automatically consider possible tautomeric states of molecules. However, the detection of new compound candidates might rely on correct structural description, which is important for the perfect fit toward the biologically relevant interactions. In this paper, we present a new exhaustive tautomer enumeration approach implemented by means of the CACTVS software package. The approach contains a set of 21 predefined SMIRKS-based transforms and a powerful transformation engine that is capable of generating most tautomers described comprehensively in the literature or found in databases in the field of medicinal chemistry. User-defined tautomer rules applied to specific structural databases or scientific issues can be implemented easily and used instead of the predefined rules. In addition, we describe the impact of tautomer-enriched databases on pharmacophore screening approaches for human matrix metalloproteinase 8 as an example of a protein-based pharmacophore screening scenario and for human cyclin-dependent kinases as an example of a ligand-based pharmacophore screening approach. In both test cases, as a preprocessing step, we have used our new tautomer enumerator tool for the tautomer enrichment of the screening data sets and have used it as a postprocessing step to remove tautomeric duplicates from the results. We could demonstrate that the tautomer-enriched screening data sets show significant advantages compared to their non-enhanced counterparts. The discrimination between hits and nonhits was significantly better in the case of tautomer-enriched databases. Moreover, it has been proved that tautomer-enhanced databases will lead to a higher number of potential hits.

INTRODUCTION

Tautomerism is an important property of chemical compounds which plays a critical role in several steps of the drug discovery process where computer-aided methods are involved. That especially applies to the unique identification and registration of compounds, the calculation and analysis of their physicochemical properties, and their usage in virtual screening approaches. Because of their different structure representations, tautomers are often recognized and handled as different structures by computer-based applications, leading to different results. Figure 1 shows two tautomeric forms of one compound found in MDL's Available Chemicals Directory database.[1] The two tautomers have been offered by a supplier as different compounds with different prices and availabilities. In addition to such duplicate check issues caused during the storage of compounds, different tautomer forms also may lead to conflicts, when processed by chemoinformatics algorithms. For example, physicochemical properties such as the lipophilicity will return different values, if the structure representation differs. For the two tautomer forms in Figure 1, two different clogP values can be found that differ by one log unit. In addition to the calculation of properties, the unequal structure representation also may have an impact on similarity searches and clustering algorithms.

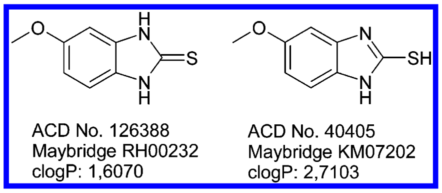


Figure 1. Two tautomer forms of the same compound within a supplier library.[1] clogP values were calculated with Tripos Sybyl 6.92.[3]

The issue of different structure representations of one molecule in compound databases, such as in the case of tautomers, has already been addressed by several companies and academic groups. Some decades ago, Chemical Abstract Service and Beilstein had to implement canonicalization techniques, which were capable of recognizing and searching different tautomer forms, to build their global chemical compound registration systems.[4,5] However, these two approaches only solved the tautomer problem on proprietary databases, whereas other software vendors such as MDL, Daylight, and Chemical Diversity Labs presented algorithms that also allow for the canonicalization of tautomers for in-house databases of pharmaceutical companies.[2,6,7]

While the canonicalization of tautomers is well-known in the drug discovery process, techniques for the enumeration of tautomers have been neglected for a long time. However, the advent of virtual screening methods in the drug discovery process has strongly increased the need for powerful and fast tautomer enumeration techniques. Usually, in silico

[†] Dedicated to Professor Johann Gasteiger.

* Corresponding author phone: +49 6130 948 206; fax: +49 6130 948 517; e-mail: Paul.Selzer@intervet.com.

[‡] Intervet Innovation GmbH, BioChemInformatics.

[§] Xemistry GmbH.

[||] Current address: BASF Aktiengesellschaft, GVC/C - A030, D-67056 Ludwigshafen, Germany.

10.1021/ci060109b CCC: $33.50 © 2006 American Chemical Society
Published on Web 09/08/2006

screening methods consider the presence or absence of directed interactions such as H-bond donors or H-bond acceptors to estimate the binding affinity of a molecule to a protein target. Because these interactions strongly depend on the defined bioactive conformation of a molecule, different tautomeric forms will have a great impact on the affinity scoring functions. The effect of tautomers on the ligand protein binding mechanism has also been described in the literature. Already in 1976, Weinstein et al. published a mechanistic model which suggests the existence of a specific histamine tautomer during the interaction with a protein receptor.[16] Furthermore, recent publications such as Brandstetter et al.,[8] Pospisil et al.,[9] and others[10−15] showed that specifically defined tautomeric states of molecules are present and interact with the active-sites residues of target proteins. These results clearly show that in silico screening approaches can only lead to reliable results if all reasonable tautomers including the correct bioactive tautomer form have been processed and have been taken into account during the screening process.

Although this aspect obviously has been known, tautomerism was neglected while performing virtual screening approaches for a long time. As a result, today, many commonly used applications for in silico screening, including 3D databases and also pharmacophore- and docking-based methods, are not capable of handling tautomeric structure representations. To overcome this problem, a few commercial vendors and academic groups recently have started efforts to develop tools for the enumeration of tautomers. These techniques can be used to enrich molecular data sets with different tautomers before these data sets will be used by virtual screening applications. When we started our work on a tautomer enumeration tool for virtual screening, only two enumeration applications, Agent[17] and Daylight,[2] existed. Interestingly, Daylight's approach had not been noticed by chemoinformaticians as a suitable tool for the virtual screening process at that time. During the following year, three additional applications for the enumeration of tautomers were released by commercial vendors—QUACPAC,[18] Pipeline Pilot,[19] and LigPrep.[20] All initial releases of these five applications showed significant limitations which were sometimes critical in a virtual screening approach. Some limitations still exists even in the newest releases of some of these tools.

Here, we present a new approach based on the CACTVS chemical data management system,[21,22] for the systematic enumeration of tautomers and for the identification of tautomeric duplicates. Our tool has specifically been designed to fulfill the requirements which occur in the virtual screening workflow. In contrast to other tools, our method is not limited to a specific kind of chemistry or a small set of tautomer cases. In fact, our application is capable of generating almost all reasonable tautomer forms found in the literature and chemical databases by using a well-defined set of molecular transformations combined with a powerful transformation algorithm. Furthermore, our approach offers an easy way to embed user-defined tautomer rules instead of the implemented predefined tautomer set to allow a customization of the enumeration process for a specific virtual screening scenario or a specific data set. Finally, the presented tool is also capable of identifying tautomeric duplicates. This option is extremely useful to identify different tautomer forms with different scoring values of one molecule in the resulting ranking lists generated by virtual screening applications. We demonstrated and validated the usability of our new tautomer-enhanced screening approach on two pharmacophore-based in silico screening examples, the human matrix metalloproteinase 8 (MMP8)—a protein-based pharmacophore example—and human cyclin-dependent kinases (CDK)—a ligand-based pharmacophore screening example.

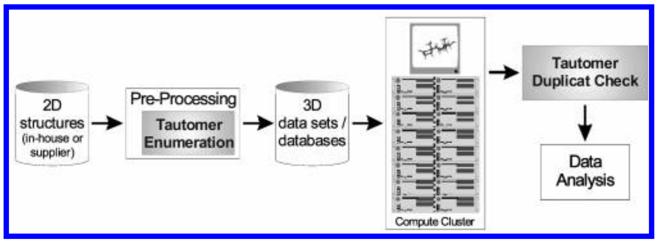


Figure 2. Enhanced virtual screening workflow for pharmacophore or target-based in silico screenings. Tautomer enumeration and tautomer duplicate check (gray squares) have been added to the general virtual screening workflow.

METHODS

Tautomer-Enhanced Virtual Screening. Figure 2 shows a schematic overview of our tautomer-enhanced virtual screening workflow. In general, every virtual screening process starts with 2D structures from an *in-house* database or from supplier catalogues. Depending on the virtual screening method (pharmacophore- or target-based screening), several preprocessing steps have to be performed before the structures can be used by a virtual screening application. Usually, all molecules are converted into 3D structures, and salts and killing fragments are removed. In the case of target-based in silico screenings, the structures are also prepared by using specific ionization rules. The resulting data sets can be used directly as an input file for target-based virtual screening tools or will be stored in specific 3D databases such as those generated by Catalyst,[23] if pharmacophore-based screenings have to be performed. We extended the preprocessing phase with an additional step to generate different tautomeric forms of the structures. In addition to the tautomer enumeration process, we added a postprocessing step to the workflow, which identifies tautomer forms with different scoring values of one molecule in the resulting ranking lists.

To evaluate the impact of tautomeric libraries in virtual screening and to evaluate the usability of the workflow, we applied the tautomer-enhanced workflow to protein-based and ligand-based pharmacophore screenings. In both cases, a diverse compound set which was extracted from the National Cancer Institute database of the year 2000 (NCI 2000 Catalyst database provided by Accelrys Inc., San Diego, CA) was used in addition to known ligands of interest to us. First, the NCI 2000 database was completely converted into 3D by CORINA 3.0,[24] and the Lipinski properties were calculated for all compounds using QSAR+ in Cerius^2.[25] Afterward, a diverse set of 993 molecules was selected from the NCI database, performing the coverage-based selection of Cerius^2 under default conditions (exception: neighborhood radius was set to 2.1). This compound set containing 993 molecules built the decoy database and was used in both pharmacophore screening cases (Figure 3).

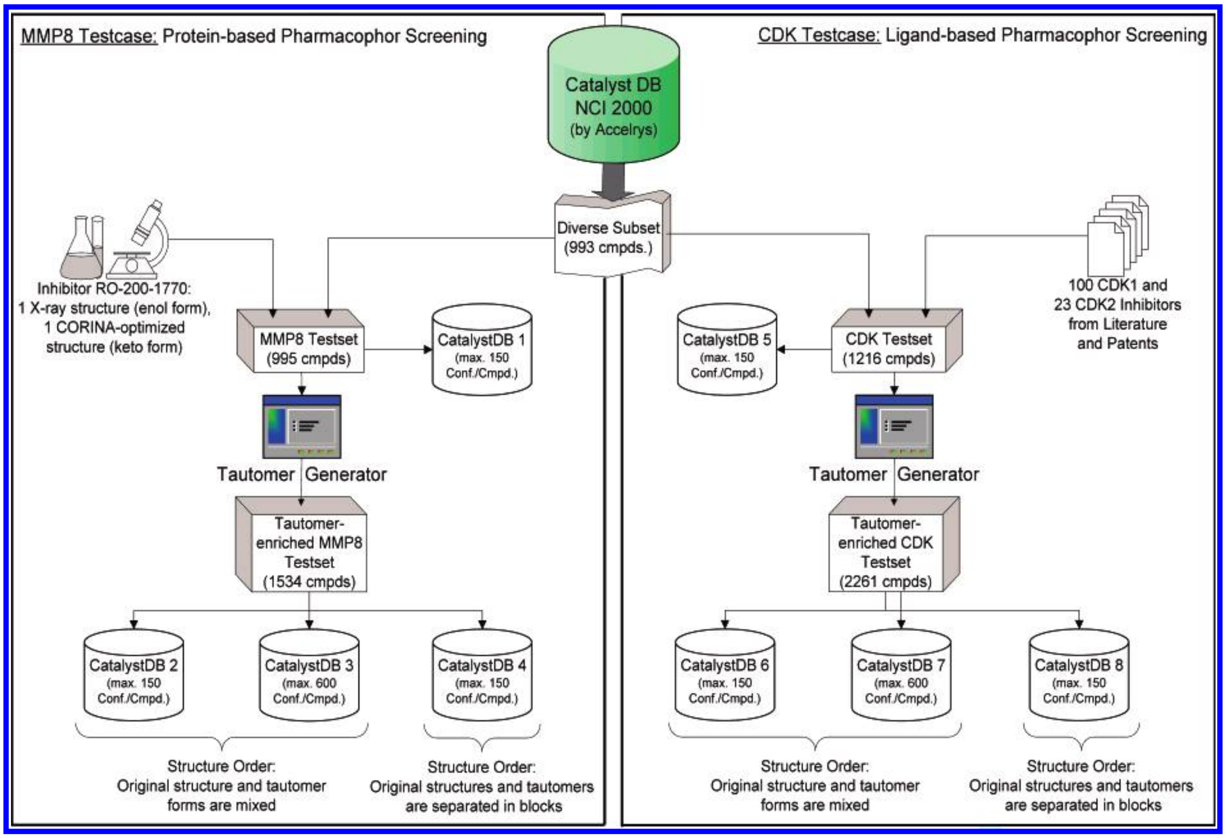


Figure 3. Flowchart showing the generation of the Catalyst databases used in the MMP8 and CDK test cases to examine the impact of tautomer enriched data sets on pharmacophore-based screenings. In the human MMP8 test case, the original conformation of the inhibitor RO-200-1770 (enol form) extracted from an X-ray structure and the usual keto form of the inhibitor (CORINA-optimized) have been added to the test set. A total of 100 human CDK1-specific and 23 human CDK2-specific inhibitors have been added to the diverse subset of the NCI 2000 database in the second test case. In both test cases, the results from the tautomer-free Catalyst databases (CatalystDB 1 and 5) have been compared to the results of their tautomer-enriched counterparts (CatalystDB 2−4, 6−8). In each case, three different tautomer databases have been generated, to analyze the impact of the number of maximal conformers per compound and the order of the original structures and their tautomer forms in the database.

The term “protein-based pharmacophore approach” implies that all essential pharmacophore features can be defined manually by means of a protein−ligand complex X-ray structure of a protein and a ligand in its bioactive conformation. In the protein-based pharmacophore approach, described in this paper, the pharmacophore model was extracted from the X-ray structure of the active enol tautomer form of the barbiturate derivative RO-200-1770 (PDB entry code: 1JJ9) bound to the human MMP8 as reported by Brandstetter et al.[8] The active enol tautomer of the barbiturate inhibitor, found in the X-ray structure, and the CORINA-optimized barbiturate structure were added to the diverse NCI 2000 subset, forming the 995 compounds for the MMP8 test set (Figure 3, MMP8 test case). In contrast to the X-ray structure, the CORINA-optimized counterpart of the inhibitor was drawn in its usual keto form, because this is the conventional tautomer form found in common structure collections used for virtual screening.

In the ligand-based pharmacophore study, 23 human CDK2-specific and 100 human CDK1-specific inhibitors, identified by a literature and patent search,[26] were added to the decoy database, building the 1216-compound-strong CDK test set (Figure 3, CDK test case). In contrast to the protein-based MMP8 approach, a qualitative CDK2-specific pharmacophore model was obtained by calculating the common features for all 23 CDK2 inhibitors using the HipHop module of Catalyst.[23] The conformational space has been defined by the best eight CDK2 inhibitors, and the feature space was limited to the best active and selective ligand. All selected features were taken under default conditions.

After preparing both initial test sets, we used our tautomer enumeration tool to build the corresponding tautomer-enriched versions of the two structure collections. In the MMP8 test case, the tautomer enrichment led to an increased test set with 1534 compounds including six tautomer forms for the two added inhibitor forms. In contrast, the CDK test set has increased to 2261 structures containing 733 tautomer forms of the 123 CDK inhibitors. In both cases, the collection without tautomers and the tautomer-enriched version were transformed into Catalyst databases that provided a basis for our screening approaches (Figure 3). The Catalyst database builds were performed under default conditions with the exception that the first conformation of each structure used the 3D coordinates generated by CORINA instead of those generated by Catalyst itself. Thus, the conformational sampling of Catalyst used the incoming CORINA 3D coordinates as a starting point for all other conformers. Moreover, the maximum number of possible conformers was increased from 100 to 150. For both test cases, the search was performed using the *best algorithm*, as our number of maximum possible conformations was expected to limit the quality of the X-ray and ring-structure resemblances. For additional test purposes, we generated a different version of the tautomer-enriched Catalyst databases by varying the database building conditions. On one hand, we increased the maximum number of possible conformers to 600 for a separate second search (Figure 3, databases 3 and 7). On the other hand, we changed the order of the original structures and their tautomer forms in the databases. In the one scenario, each original structure was followed by its tautomeric forms (Figure 3, MMP8: databases 2 and 3, CDK: databases 6 and 7). Alternatively, we generated databases where all original structures and their tautomer forms were split and separated into two structure blocks, whereas the complete set of all original structures was followed by the collection of all generated tautomers (Figure 3, databases 4 and 8).

By comparing the tautomer-free Catalyst databases with their tautomer-enhanced counterparts, we analyzed the impact of tautomer enrichment on both pharmacophore screening approaches. The MMP8 test case was meant to analyze, in detail, whether our approach is capable of reconstructing the bioactive form of the inhibitor RO-200-1770 found by X-ray analysis starting from the corresponding CORINA-optimized keto tautomer form. Thereby, the reconstruction implied the generation of the bioactive enol form by our tautomer generator and the generation of the bioactive conformation by Catalyst. The CDK test scenario has been used to examine the effect of the tautomer approach on a common, daily work pharmacophore screening setup. Moreover, the aim of both test cases was to observe whether (a) Catalyst is able to reidentify the highly active compounds in a tautomer-sensitive fashion, (b) the fit values significantly discriminate between the different hits and nonhits, and (c) the hit rate of known actives is increased (independent of the discrimination via good geometric scores) and whether the found set of hit compounds provides a broader structural diversity, if tautomer-enriched databases are used.

Tautomer Enumeration and Unique Tautomer Identification. The presented tautomer enumeration process and the tautomer-sensitive duplicate check are based on the chemical data management system CACTVS from Xemistry.[21,22] CACTVS is based on a core library written in ANSI C[27] that functions as an object-orientated data manager. The library defines a rich set of predefined properties and chemical objects such as molecules, molecule ensembles, atoms, and bonds as well as methods managing relations between the above objects. The implemented property set can be extended with user-defined object properties at runtime. To enable convenient access to the core functionalities, a Tcl[28] scripting layer has been added on top of the basic library, which contains a large set of chemistry-specific commands in addition to the standard language capabilities.

Tautomer Transformation Rules. In this study, we used a set of 21 tautomer rules that encode a wide range of typical 1,2-, 1,3-, 1,5-, 1,7-, 1,9-, and 1,11 H shifts (Table 1). The tautomer transforms are based on Daylights SMIRKS line notation,[29] which originally had been designed to describe reaction transforms. CACTVS supports the advanced version of this encoding that allows the creation and deletion of atoms within a transform. Furthermore, the SMIRKS parser was enhanced with a set of useful extensions. All transformations are performed by the CACTVS Tcl command “ens transform”, which will be described in the next section. By systematically combining the 21 tautomer transformations using the CACTVS command, we were capable of enumerating all of the tautomer forms we had found in the literature[30] and during our work in the drug discovery process.

The first two rules (Table 1) address keto−enol tautomerism of ketones and hydroxides and their sulfur, selenium, and tellurium analogues. While rule 1 considers simple enol−keto cases with hydrogen 1,3 migrations, rule 2 has been implemented to handle 1,5 H shifts (long-distance keto−enol tautomerism). The attributes x0 and x1, which have no counterpart in the original SMIRKS syntax, indicate the number of heteroatoms substituted to the carbons.

The next two transformations in the transformation table (rules 3 and 4) implement 1,3 H shifts of imines. Rule 3 describes the hydrogen migration of aliphatic imines, whereas rule 4 also matches aromatic substructures containing an aliphatic hydrogen acceptor or donor carbon. Furthermore, this carbon atom is not allowed to be in a ring system or to be bound to a heteroatom. The terminal nitrogen in rule 4 is not allowed to have other heteroatom neighbors.

To handle hydrogen shifts between aromatic or aliphatic heteroatoms in aromatic heterosystems, we have implemented seven rules (rules 5−11). Rules 5 and 6 control 1,3-hydrogen migrations, whereas rules 7−11 handle long-range H shifts. Rule 5 describes a transformation, where the central carbon atom has to be in a ring system with 6π electrons. This condition is defined by the e6 extension, which also has no counterpart in the SMIRKS line notation. The additional constraints imposed by these transforms, besides simple element type and bond order patterns, serve the purpose of suppressing the generation of classes of unlikely high-energy tautomer forms, reducing the result set in size and speeding up the computation. One terminal atom in the matching pattern has to be nitrogen, whereas the other one can be nitrogen or oxygen. Figure 4 shows an exemplary tautomer case, which is handled by this rule.

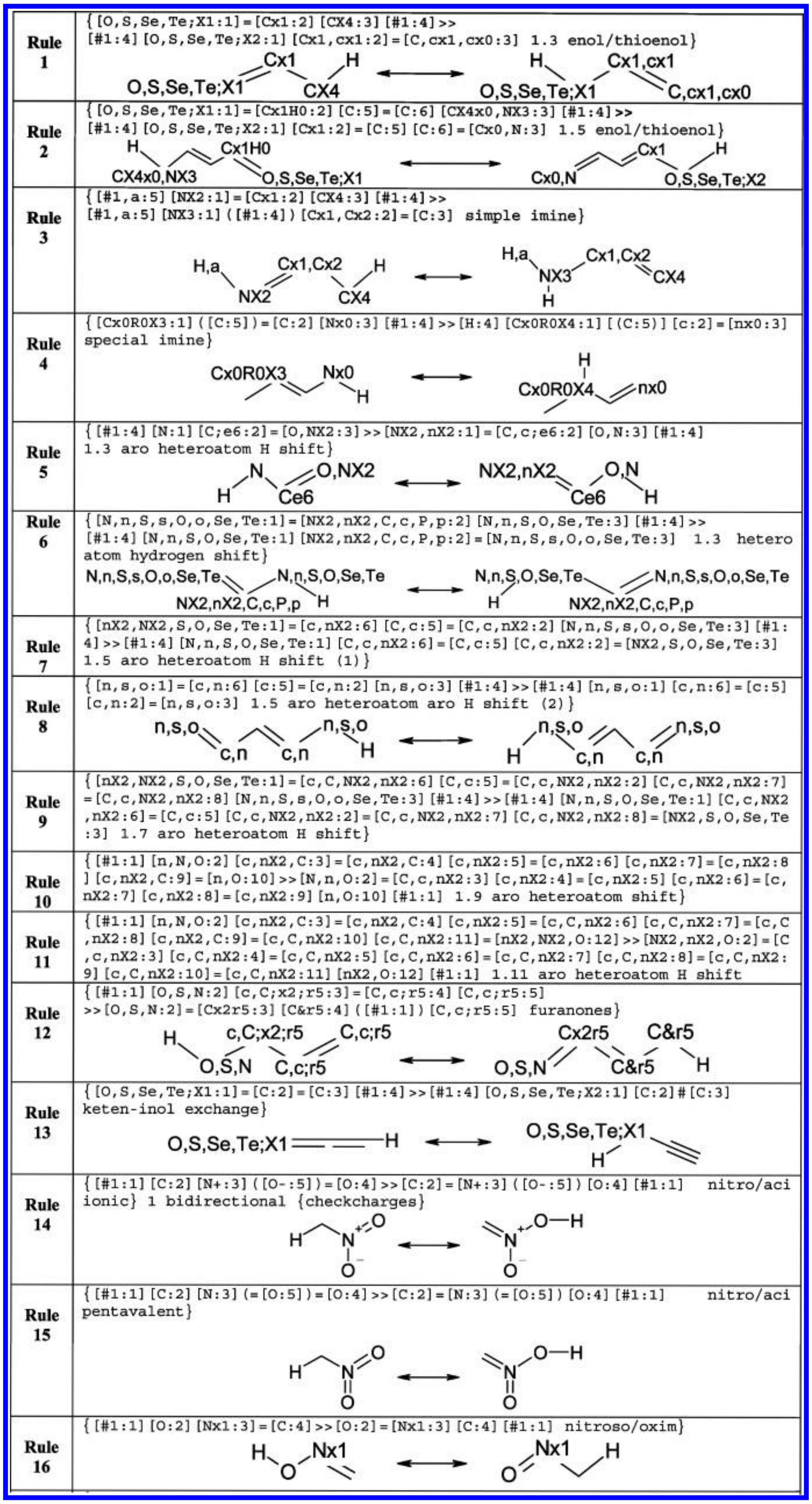
In contrast to rule 5, all atoms in the matching substructure of rule 6 can be aliphatic or aromatic. Moreover, the central atom can also be nitrogen or phosphorus, whereas the terminal atoms of the matching pattern also allow sulfur, oxygen, selenium, and tellurium during the substructure matching. The three tautomer forms in Figure 5 are addressed by rule 6.

The next two tautomer rules (rules 7 and 8) describe 1,5 H shifts. While rule 8 only addresses aromatic systems, rule 7 also can deal with specific aliphatic structures. In addition, rule 7 also accepts selenium and tellurium as terminal atoms. Using two specific rules such as 7 and 8 instead of one global rule limits the number of aromatic systems with aliphatic side chains addressed by these rules and therefore limits the number of unreasonable tautomer forms. Figures 6 and 7 show examples for the two tautomer cases.

The last three aromatic transformation rules (rules 9−11) manage very long hydrogen migrations within aromatic heteroatom systems. While rule 9 addresses 1,7 H shifts in aromatic and aliphatic systems containing carbon, nitrogen, oxygen, sulfur, selenium, and tellurium as terminal donor and acceptor atoms (Figure 8, structures 1 and 3), rules 10 (1,9 H shifts, Figure 9) and 11 (1,11 H shifts) are limited to terminal nitrogen and oxygen atoms.

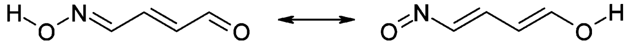
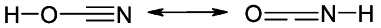
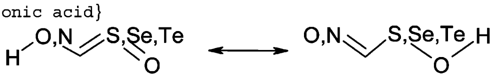
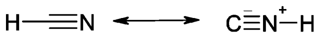
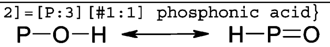
Tautomerisms of furanones and furanone-like molecules are handled by rule 12. Other than oxygen, the terminal atom can also be nitrogen or carbon. Furthermore, the substructure

Table 1. Overview of the 21 Extended SMIRKS Transforms and Additional Individual Flags of CACTVS's 'ens transform' Command Used in the Presented Study[a]



Rule 1	{[O,S,Se,Te;X1:1]=[Cx1:2][CX4:3][#1:4]>> [#1:4][O,S,Se,Te;X2:1][Cx1,cx1:2]=[C,cx1,cx0:3] 1.3 enol/thioenol}
Rule 2	{[O,S,Se,Te;X1:1]=[Cx1H0:2][C:5]=[C:6][CX4x0,NX3:3][#1:4]>> [#1:4][O,S,Se,Te;X2:1][Cx1:2]=[C:5][C:6]=[Cx0,N:3] 1.5 enol/thioenol}
Rule 3	{[#1,a:5][NX2:1]=[Cx1:2][CX4:3][#1:4]>> [#1,a:5][NX3:1]([#1:4])[Cx1,Cx2:2]=[C:3] simple imine}
Rule 4	{[Cx0R0X3:1]([C:5])=[C:2][Nx0:3][#1:4]>>[H:4][Cx0R0X4:1][(C:5)][c:2]=[nx0:3] special imine}
Rule 5	{[#1:4][N:1][C;e6:2]=[O,NX2:3]>>[NX2,nX2:1]=[C,c;e6:2][O,N:3][#1:4] 1.3 aro heteroatom H shift}
Rule 6	{[N,n,S,s,O,o,Se,Te:1]=[NX2,nX2,C,c,P,p:2][N,n,S,O,Se,Te:3][#1:4]>> [#1:4][N,n,S,O,Se,Te:1][NX2,nX2,C,c,P,p:2]=[N,n,S,s,O,o,Se,Te:3] 1.3 hetero atom hydrogen shift}
Rule 7	{[nX2,NX2,S,O,Se,Te:1]=[c,nX2:6][C,c:5]=[C,c,nX2:2][N,n,S,s,O,o,Se,Te:3][#1: 4]>>[#1:4][N,n,S,O,Se,Te:1][C,c,nX2:6]=[C,c:5][C,c,nX2:2]=[NX2,S,O,Se,Te:3] 1.5 aro heteroatom H shift (1)}
Rule 8	{[n,s,o:1]=[c,n:6][c:5]=[c,n:2][n,s,o:3][#1:4]>>[#1:4][n,s,o:1][c,n:6]=[c:5] [c,n:2]=[n,s,o:3] 1.5 aro heteroatom aro H shift (2)}
Rule 9	{[nX2,NX2,S,O,Se,Te:1]=[c,C,NX2,nX2:6][C,c:5]=[C,c,NX2,nX2:2][C,c,NX2,nX2:7] =[C,c,NX2,nX2:8][N,n,S,s,O,o,Se,Te:3][#1:4]>>[#1:4][N,n,S,O,Se,Te:1][C,c,NX2 ,nX2:6]=[C,c:5][C,c,NX2,nX2:2]=[C,c,NX2,nX2:7][C,c,NX2,nX2:8]=[NX2,S,O,Se,Te :3] 1.7 aro heteroatom H shift}
Rule 10	{[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2:5]=[c,nX2:6][c,nX2:7]=[c,nX2:8][c,nX2,C:9]=[n,O:10]>>[N,n,O:2]=[C,c,nX2:3][c,nX2:4]=[c,nX2:5][c,nX2:6]=[c, nX2:7][c,nX2:8]=[c,nX2:9][n,O:10][#1:1] 1.9 aro heteroatom shift}
Rule 11	{[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2:5]=[c,C,nX2:6][c,C,nX2:7]=[c,C ,nX2:8][c,nX2,C:9]=[c,C,nX2:10][c,C,nX2:11]=[nX2,NX2,O:12]>>[NX2,nX2,O:2]=[C ,c,nX2:3][c,C,nX2:4]=[c,C,nX2:5][c,C,nX2:6]=[c,C,nX2:7][c,C,nX2:8]=[c,C,nX2: 9][c,C,nX2:10]=[c,C,nX2:11][nX2,O:12][#1:1] 1.11 aro heteroatom H shift
Rule 12	{[#1:1][O,S,N:2][c,C;x2;r5:3]=[C,c;r5:4][C,c;r5:5] >>[O,S,N:2]=[Cx2r5:3][C&r5:4]([#1:1])[C,c;r5:5] furanones}
Rule 13	{[O,S,Se,Te;X1:1]=[C:2]=[C:3][#1:4]>>[#1:4][O,S,Se,Te;X2:1][C:2]#[C:3] keten-inol exchange}
Rule 14	{[#1:1][C:2][N+:3]([O-:5])=[O:4]>>[C:2]=[N+:3]([O-:5])[O:4][#1:1] nitro/aci ionic} 1 bidirectional {checkcharges}
Rule 15	{[#1:1][C:2][N:3](=[O:5])=[O:4]>>[C:2]=[N:3](=[O:5])[O:4][#1:1] nitro/aci pentavalent}
Rule 16	{[#1:1][O:2][Nx1:3]=[C:4]>>[O:2]=[Nx1:3][C:4][#1:1] nitroso/oxim}

Table 1. (Continued)

Rule 17	{[#1:1][O:2][N:3]=[C:4][C:5]=[C:6][C:7]=[O:8]>> [O:2]=[N:3][c:4]=[c:5][c:6]=[c:7][O:8][#1:1] nitroso/oxim via phenol} 
Rule 18	{[#1:1][O:2][C:3]#[N:4]>>[O:2]=[C:3]=[N:4][#1:1] cynanuric acid} 
Rule 19	{[#1:1][O,N:2][C:3]=[S,Se,Te:4]=[O:5]>>[O,N:2]=[C:3][S,Se,Te:4][O:5][#1:1] formamidinsulfonic acid} 
Rule 20	{[#1:1][C0:2]#[N0:3]>>[C-:2]#[N+:3][#1:1] isocynanide} 1 bidirectional {checkcharges checkaro} 
Rule 21	{[#1:1][O:2][P:3]>>[O:2]=[P:3][#1:1] phosphonic acid} 

[a] Current versions of the CACTVS toolkit are using slightly modified standard rule sets.

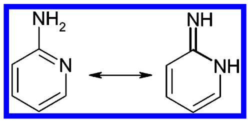


Figure 4. 1,3-Migration of hydrogen in 2-aminopyridine.

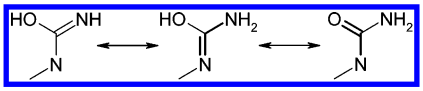


Figure 5. Three tautomer forms being addressed by tautomer rule 6.

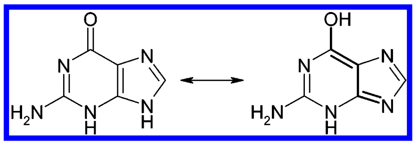


Figure 6. Two guanine tautomer forms transformed by rule 7.

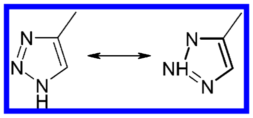


Figure 7. Two triazole tautomers addressed by tautomer rule 8.

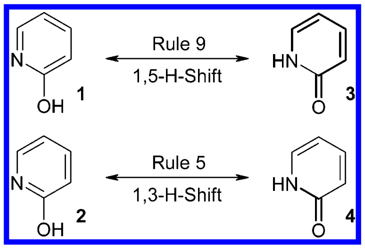


Figure 8. Two tautomer transformations of 1,2-pyridone. The two amide forms 1 and 2 differ only in ambiguous alternating single and double bonds transforming to the same iminole form using different rules.

defines that all remaining carbons have to be in a five-membered ring system.

To handle keten−inol tautomerism, we have implemented rule 15. As in the keto−enol cases described above, the terminal atom of the substructure can be oxygen, sulfur, selenium, or tellurium.

Rules 14−17 implement transformations to handle hydrogen shifts in C−N=O systems. The tautomerism of nitro groups is handled by rules 14 and 15. Rule 14 transforms ionic nitro groups, and the other maps nitro groups with pentavalent nitrogen. Nitroso−oxim tautomerism is managed by rules 16 and 17. While rule 16 only works on simple nitroso−oxim groups, rule 17 can handle nitroso−oxim tautomerism via phenol systems (Figure 10).

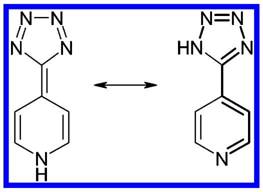


Figure 9. Two tautomer forms addressed by the 1,9 H shift of rule 10.

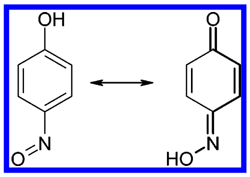


Figure 10. Nitroso−oxim tautomerism via a phenol system.

The remaining four rules (18−21) have been designed to handle some special events such as H shifts of cyanuric acids, formamidinsulfonic acids, phosphonic acids, and isocyanides.

CACTVS Transformation Command. In light of the definition of SMIRKS, it becomes obvious that any single transform can only be used to describe one structure manipulation like the transformation of the enol form of a molecule to the keto form. Because of this simple nature, it is difficult to realize a complete enumeration process, even if a large number of different SMIRKS will be used. We have overcome this limitation by means of automatic processing options of the CACTVS Tcl scripting command “ens transform”, which provides a large number of options to manipulate, loop, and control the transformation process.

The general form of this command is as follows:

ens transform *ehandle SMIRKSlist ?direction? ?reactionmode? ?selectionmode? ?flags? ?overlapmode? ?excludesslist? ?maxstructures? ?timeout?*

We used the following specific form for our tautomer enumeration:

ens transform *ehandle SMIRKSlist* bidirectional multistep all {checkaro preservecharges setname} none “” maxstructures

The command applies a list of one or more SMIRKS transforms (SMIRKSlist) to a molecule ensemble (identified by its handle ehandle) and returns a list of ensemble handles of transformation products. The transformation products are filtered for duplicates. The original structure is never returned, and if a transform set does not match any of the SMIRKS, an empty list is returned. In the case of the tautomer enumeration process, the SMIRKS list contains all transformations described in Table 1, and the command will return a list of all possible tautomers.

In addition to the SMIRKS transform list, the command recognizes the optional parameters direction, reactionmode, selectionmode, flags, overlapmode, excludesslist, maxstructures, and timeout. These parameters can be set in a global manner that will apply to all SMIRKS transformations executed by the command instance, but some of the above can also be defined individually by adding additional local parameters to single SMIRKS transforms within the SMIRKS list (Table 1, rules 16 and 22). If defined, local parameters will be used instead of the global counterparts. A detailed description of all possible parameter options can be obtained elsewhere.[31]

The direction parameter describes how the SMIRKS transform has to be read. When the bidirectional mode is used, both sides of the transform scheme will be matched and, if the match is successful, transformed to the other side. By using this option, both sides of the 21 tautomer rules (Table 1) can be used to identify substructures in the original structure. For example, rule 1 in Table 1 will match and transform both compounds with enol groups and structures with keto groups. The next parameter, the reactionmode parameter, will determine how the possibility of multiple matches of transform substructures in the target molecule is handled. In the multistep mode, all transform products will be generated by systematically applying the transforms to all structures and constantly resubmitting the results, until no new compounds are generated. Like the reactionmode parameter, the command parameter selectionmode has no counterpart in the SMIRKS line parameters. When the all mode is used, all transformations are applied to all result structures. This process is repeated until no additional, structurally distinguishable result ensembles are generated. Other possible values of this parameter have also been used in other contexts, such as combinatorial library enumeration, chemical syntheses planning, or exhaustive rule-based structure generation.

The full set of the resulting tautomers is returned by the command as a list of ensemble object handles. The direction, reactionmode, and selectionmode parameters are the most important parameters during the tautomer enumeration process. Only by using these options is a complete enumeration of all possible tautomer forms guaranteed, which is not possible by using the SMIRKS transforms directly as a one-step process (see Results section).

In addition to these three major parameters, some additional flags are used. In our command configuration, the flags option gives a list of three additional flags, checkaro, preservecharges, and setname, that will influence the matching of the transform substructures. The checkaro flag influences the way uppercase elements will be handled by the aromaticity checking routine. In contrast to Daylights standard implementation, our tool in the configuration used here considers uppercase elements as undefined with respect to aromaticity. However, if the checkaro flag is set, nonaromaticity checking for uppercase atoms is switched on and uppercase atoms in the transform pattern can only match aliphatic atoms in the target ensemble. Atoms specified as aromatic (lowercase atoms) will always be checked for aromaticity in the target ensemble per default. If the preservecharges flag is set, charges are not modified during transformation. By default, the charge of matched atoms is set to the charge of the matching atom in the transform template, as long as the atom has sufficient electrons to allow the charge change. The setname flag appends the names of applied transformations and their direction of application to a compound name property. The three global flags are used for every tautomer rule described in Table 1, except for rules 16 and 22. The latter two rules contain their own flag lists, which are used instead of the global one. Both rules use the checkcharges flag. In contrast to Daylights standard implementation, the default tool configuration employed here will ignore charge specifications for matching. By setting checkcharges, formal charges on the match side of the transform must exactly match the charges on the matched structure. Furthermore, rule 16 has no checkaros flag, and therefore, aromaticity checking for uppercase atoms within the transform string is disabled.

The overlapmode parameter determines if a transform substructure consisting of multiple disconnected fragments can match common target structure atoms or bonds. The default mode, none, prohibits the overlap of substructure fragments, both on atoms and on bonds. Because our transforms do not consist of multiple fragments on either side, the setting of this parameter is of no major relevance. For example, this parameter can also be used to specifically select intra- or intermolecular reaction transforms. The excludesslist parameter was not used in our tautomer enumeration approach. Finally, the maxstructures and timeout options specify the maximum number of tautomers returned by the command and the maximum processing time spent.

To enumerate tautomers and to generate a canonical tautomer structure, we limit this number to 1000 tautomers per target structure. If 1001 structures were reported, the application of the transform rules would not have been performed exhaustively. Therefore, potentially unreported structures are the result of long, complex sequences of transforms and thus are usually of hypothetical interest rather than of practical value and observability. Similarly, tautomers in databases tend to be registered in such a way that they are only one or, at maximum, two transforms apart, so that, for practical purposes, tautomer canonicalization by selecting a canonic tautomer form with a structure rated as reasonable works well, because all registered forms do arrive at the canonic form within a few transforms.

Because of the exhausting combination mechanism, sometimes, subsets of the resulting tautomer list of a specific rule are identical with the resulting subset of another rule. To limit the number of tautomer duplicates produced and then discarded during the enumeration process, we used two or more specific rules instead of one global rule. Nevertheless, some duplicates are still generated but are removed by the duplicate check routine in the “ens transform” algorithm and are not reported in the final result set.

Tautomer-Insensitive Hash Codes. CACTVS can compute structure hash codes that are different for every tautomer structure with different bond connectivity. These tautomer-sensitive hash codes are used, for example, by the internal duplicate check routine within the “ens transform” command to identify identical tautomer representations during the

Table 2. General Rating Rules for the Identification of the Most Reasonable Tautomer Form

structure fragment	rating points
each aromatic ring system	+100
each double bond between carbon and heteroatoms	+1
each double bond between oxygen and nitrogen (in either element combination)	+2
each C═N[OH] fragment	+4
each methyl group (applying a penalty to structures with terminal double bonds)	+1
each double bond between carbon and oxygen	+2
each P−H, S−H, Se−H, and Te−H bond	−1

enumeration process and to remove them from the result list. However, in many cases, it is necessary to identify different tautomer structures of a molecule as duplicates. This feature is important during the registration of structures into databases but is also needed in our virtual screening approach to detect duplicates in our ranking lists after the screening has been done. CACTVS can address this problem by computing a tautomer-insensitive hash code.

As in the enumeration process described, CACTVS first generates all reasonable tautomer representations for a given structure. In the second step, this tautomer list will be used to identify a canonic tautomer. Because the calculation of preferred tautomer forms is time-consuming, difficult, and depends on many aspects such as the solvent, environment, dipole−dipole repulsion, electronic, and thermodynamic effects, we have implemented a fast, empirical, rule-based rating algorithm to detect the most reasonable tautomer. The rating system has been established by analyzing many sets of tautomers and the known preferred tautomer members included in these data sets. Table 2 shows the general rating rules obtained by this analysis.

The resulting tautomer form with the best rating is defined as the unique tautomer, and the corresponding hash code is used as the tautomer-insensitive hash code. If more than one structure reaches the best score, the tautomer with the largest hash code is chosen as the unique tautomer form. Because of the systematic enumeration of all tautomer forms, this approach is independent from the starting structure. With very high reliability, it results in the same unique preferred tautomer representation, and is guaranteed to do so, if the transform list was applied exhaustively. Generally, we do not claim that the implicitly selected canonic tautomer is the lowest in energy in absolute terms, because for the computation of a unique structure hash code this is not a requirement. The rule set was only designed to yield a reasonable structure which will not trigger immediate comments from concerned chemists who happen to come across depictions of the structure of the canonic tautomer.

RESULTS

Tautomer Enumeration. One major point that needs to be kept in mind when using SMIRKS transforms is the issue of alternating single and double bonds in aromatic systems. In CATVS, these bonds are internally represented as single and double bonds for purposes of bond electron counting. Because of this fact, it is, for example, impossible to implement a single general tautomer rule that is capable of transforming the two amide forms shown in Figure 8 (structures 1 and 2) to the corresponding iminole forms (structures 3 and 4). Instead, one rule has to be defined for the 1,7 H shift of the upper transformation (Figure 4, structures 1 and 3) and one for the hydrogen 1,3 migration (Figure 8, structures 2 and 4).

As already mentioned above, our enumeration approach does not require the separate implementation of tautomer transformations for every tautomer form. The capability of the “ens transform” command to systematically combine all SMIRKS transforms and to apply them to the target structures as well as to the resulting tautomer forms recursively allows for the generation of tautomer forms, which cannot be directly obtained from the starting structure with a single transform. To prove this approach, we did set up a test script to enumerate all possible 15 tautomer forms of guanine (Figure 11). In this case study, only two rules (rules 6 and 7, Table 1) are necessary to fulfill the task. However, if we look separately at these two rules, it becomes obvious that rule 6 (1,3 heteroatom hydrogen shift) can only generate the tautomer forms 2, 4, 5, 6, 7, 9, 10, 11, and 12 starting from tautomer form 1, whereas rule 7 is only capable of generating forms 3 and 5. Though, the tautomer form 5 has been generated by both rules but will occur only once in the result set, because the generated duplicate is automatically removed by the duplicate check routine of the “ens transform” command. However, the tautomer forms 8, 13, 14, and 15 cannot be generated by only one of the two rules. These four guanine tautomers have to be generated through the combination of both rules independent of the order of

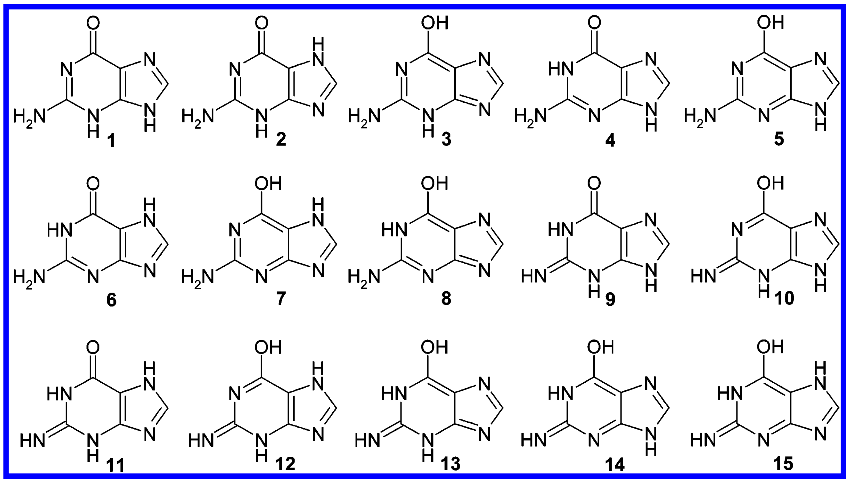


Figure 11. All possible tautomer forms of guanine.

Table 3. Comparison of Data Set Sizes of the MMP8 and CDK Inhibitor Test Sets and the Diverse Subset of the NCI 2000 Database before and after the Tautomer Enrichment (see also Figure 3)

data set	number of structures before tautomer generation	number of structures after tautomer generation
MMP8 inhibitors	2	6
CDK inhibitors	123	733
NCI 2000 subset	993	1528

the used rules by using the respective individual tautomer result of the first rule as a new starting structure of the second rule and vice versa. This gives an impressive insight of the capabilities and the power of our tautomer enumeration approach.

The number of ligands and conformers is a critical aspect within virtual screening approaches, because it significantly increases the time and CPU resources required for any study. Therefore, we also analyzed the data set sizes of our two screening approaches before and after tautomer enrichment (Table 3).

The majority of all structures contained in the MMP8 and CDK test sets was imported from the diverse subset of the NCI 2000 database. Representative for common screening libraries, we analyzed the enlargement of this specific structure subset obtained by our tautomer generator, and we observed a moderate enrichment factor of 1.5. A database enlargement by this factor is not a critical issue for virtual screening applications and can be realized easily on common Linux cluster systems. In contrast to the diverse NCI subset, both inhibitor subsets showed a different behavior during the tautomer enumeration. In the MMP8 test case, six tautomer forms could be generated starting from the two initial inhibitor structures, resulting in an above-average enrichment factor of 3. This observation was no surprise, because we did know that a specific tautomer form was responsible for the activity and that RO-200-1770 is capable of forming different tautomer states. While the above-average number of tautomers for the MMP8 inhibitor was known from the beginning, the tautomer enrichment factor for the CDK inhibitors has been expected to be similar to the NCI 2000 subset. Surprisingly, the CDK inhibitor subset shows a more dramatic increase during the tautomer generation than that observed for the MMP8 inhibitors: 733 tautomer forms could be generated starting from the initial 123 CDK inhibitors, leading to an enrichment factor of 6 (Table 3).

The multiplier between the original and the tautomer enriched data sets only gives a very general overview about the situation during the tautomer process. We could observe that the performance of the tautomer generator can strongly differ (up to 10 times) between different screening databases, even if they had the same enrichment factor. Therefore, the multiplier is an inapplicable value to estimate the CPU time. In fact, the CPU time mainly depends on the structures included in the data sets and the number of possible tautomer combinations for each structure. To impart an approximate estimation of the tautomer generator performance, we have measured the CPU time needed to convert a subset of 1000 compounds from the Maybridge Screening Collection (Autumn 2005): The tautomer process has taken 4 min and 42 sec on an AMD Opteron Processor 250 (2.4 GHz) and generated a corresponding tautomer data set with 2905 tautomers.

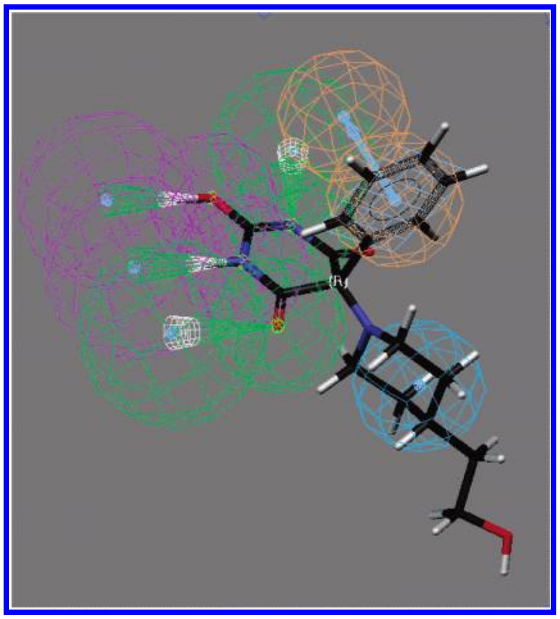


Figure 12. Protein-based pharmacophore model manually derived from the barbiturate derivative RO-200-1770 and MMP8 complex.[8] The features are colored for the two H-bond donors in green, for the two H-bond acceptors in magenta, for the hydrophobic feature in blue, and for the ring aromatic π interaction in brown.

Protein-Based Pharmacophore Screening. By using different conditions during the database building process, three Catalyst databases (Figure 3, databases 2−4) have been generated for the tautomer-enriched MMP8 data set and one database has been generated for the tautomer-free test set (Figure 3, database 1). Furthermore, a pharmacophore model was derived manually from the X-ray structure of RO-200-1770 and has been used as a query on all four Catalyst databases (Figure 12). In a first step, we compared the tautomer-free database 1 (Figure 3) and its tautomer-enriched counterpart (Figure 3, database 2), which had been build under the same conditions. When database 1 was used, the pharmacophore query shown in Figure 12 retrieved 29 hits, whereby all hits had to map all six pharmacophore features. The X-ray structure of the barbiturate RO-200-1770 mapping all six feature centers scored best with a geometric fit value of 7.2 in a range of 0−7.2, whereby each single feature had a maximum contribution of 1.2 to the scoring value. The dimensionless geometric fit value is provided by Catalyst and decreases hyperbolically distance-dependent on the feature center. Values below a geometric hit score of 4 were estimated to be not interesting hits as those missed on average almost two features. The highest-ranked conformer found for the CORINA-optimized keto form of the inhibitor had a best fit value of 3.2, whereas the best nonbarbiturate molecule from the test collection had a high scoring value of 6.2. Only three hits showed best fit values higher than 4. On average, the molecules scored 3.0 for the fit value, and the standard deviation was 1.7.

When performing the search on the tautomer-enriched database 2 (Figure 3), we retrieved 30 hits. Using CACTVS' tautomer-invariant duplicate check routine, we identified and removed eight tautomeric duplicates from the result set, and only 22 structurally unique hits remained. Again, the X-ray conformation of the barbiturate scored highest with 7.2 for the geometric fit value, while the keto form retrieved by the CORINA-optimized structure had a best fit value of 3.2. However, in contrast to database 1, the best conformer of the enol tautomer which was generated by our application using the CORINA-optimized keto form has a higher value (best fit 6.3) than the best nonbarbituric structure showing only a best fit of 5.5. The comparison of the two databases revealed that the inhibitor RO-200-1770 would not have been found ranked high without taking into account the tautomeric states of database 2. Therefore, one may take a closer look at the "*wrong*" nonbarbituric structures, before examining the low-ranked RO-200-1770, instead of using databases without tautomers. These results show that the recovery of the active compound is possible, because our enumeration application was able to generate the correct tautomer form, and Catalyst was capable of generating a conformer of this tautomer form, which almost corresponds to the X-ray structure. Surprisingly, all top hits (five molecules) but only 60% (22 hits versus 29 hits in database 1) of the hits retrieved by the nontautomeric approach could be found within database 2. Usually, it should be assumed that a database containing another database as a subset should find all of the hits that can be identified by performing a search on the subset only. Ideally, the bigger tautomer-enhanced data set should result in an increased number of hits. While the tautomer enumeration process obviously had a positive impact on the identification of the bioactive inhibitor form, the critical decrease of the hit retrieval rate between the nontautomeric database 1 and the tautomer-enriched database 2 cannot be explained at this time.

To analyze this issue, we tested in an additional step whether the limitation of the number of conformers during the Catalyst database building process had an impact on the unsatisfying hit retrieval rate. Therefore, we used database 3 which allowed a maximum of 600 conformations per compound instead of database 2 (Figure 3) which was limited to a maximum of 150 conformers per compound. However, a comparison of databases 1 and 3 led to the same unsatisfying results, indicating that the number of conformers has no impact on the hit retrieval rate and also will not improve the overlap of the search results (hit retrieval rate) for the search for the tautomeric and the nontautomeric data set.

Finally, we changed the order of the original structures and their tautomeric counterparts in the database. While in databases 1−3 each original structure was followed by its tautomeric forms, database 4 (Figure 3) contained all original structures separated as a block from their tautomer counterparts. Thereby, the tautomer block was added behind the block of original structures. The Catalyst database building conditions were the same as those for database 1. This approach should prove an oral communication with other Catalyst users suggesting Catalyst 4.9 contains a critical error: Catalyst's hit retrieval depends on the order of the structures in the input data set. In fact, our setup was capable of proving this lack. Using the reordered data set as Catalyst input, we were now able to identify RO-200-1770 as the most active inhibitor, but moreover, the hit retrieval rate was significantly better than that in the case of database 2 or 3. As expected, we could retrieve all of the hits found by using the nontautomeric database, and in addition, even 10% more hits were retrieved by using the tautomer-enriched counterpart. A deeper analysis of the additional compounds found has not been performed.

Ligand-Based Pharmacophore Screening. The CDK test scenario was performed parallel to the protein-based screening approach in a similar way. One Catalyst database (Figure 3, database 5) was generated starting from the tautomer-free test set, whereas three Catalyst databases (Figure 3, databases 6−8) were built for the tautomer-enriched data set by varying the building conditions. Furthermore, 10 qualitative pharmacophore models were derived from the human CDK2 inhibitors. The best pharmacophore hypothesis (Figure 13) retrieved 55% of the selective CDK2 inhibitors in the nontautomeric database 5 (Figure 3), whereas all molecules with selectivity values higher than 4 against human CDK2 were mapped. Taking into account the average overall pharmacophore hypotheses, 40% of the known selective inhibitors were retrieved. In total, 81 hits were extracted from the nontautomeric database.

To compare the tautomer-free and the tautomer-enhanced databases, we performed the same searches on the tautomer-enriched database 2 (Figure 3). When the best qualitative pharmacophore hypothesis was used, only 61 structurally unique hits could be isolated after removing tautomeric duplicates from the result set. As in the MMP8 test case before, this dramatic decrease of the hit retrieval rate (54%) could not be explained at this time. The average hit rate overall hypotheses also showed a decrease of 66% for the known selective inhibitors with a standard deviation of 10%. However, the comparison of databases 5 and 6 reveals also a major benefit of the tautomer-enhanced approach. Using the tautomer-enriched test case (database 6), we observe a much better discrimination between hits and nonhits and CDK2 and CDK1 inhibitors. While the tautomer-free database 5 only retrieved 55% of the CDK2 inhibitors, the tautomer-enhanced database 6 was capable of identifying 17% more CDK2 inhibitors (72% CDK2 inhibitors found).

As in the protein-based pharmacophore approach, changing the maximum number of conformers up to 600 conformers per compound (database 7) did not solve the hit retrieval issue. However, rearranging the order of the original structures and their tautomer forms within the Catalyst input file (Figure 3, database 8) led to an increase of the hit retrieval rate. As observed in the MMP8 approach, all hits found by the tautomer-free version could be identified, and moreover, 100% more hits were retrieved by using the tautomer database 7. Moreover, the already observed better discrimination between hits and nonhits of the tautomer-enriched databases 6 and 7 could still be improved significantly in the reordered tautomer-enriched database (database 7). A deeper analysis of the additional compounds found has not been performed.

DISCUSSION AND CONCLUSION

Tautomer-Enhanced Virtual Screening. To demonstrate the advantages of tautomer-enriched databases compared to their nonenhanced counterparts, we used a protein-based and a ligand-based pharmacophore screening approach. The X-ray-based pharmacophore example has been used to validate the usability of our tautomer enumeration approach in the virtual screening process. Usually, pharmacophore features for a specific target are not known in the early stages

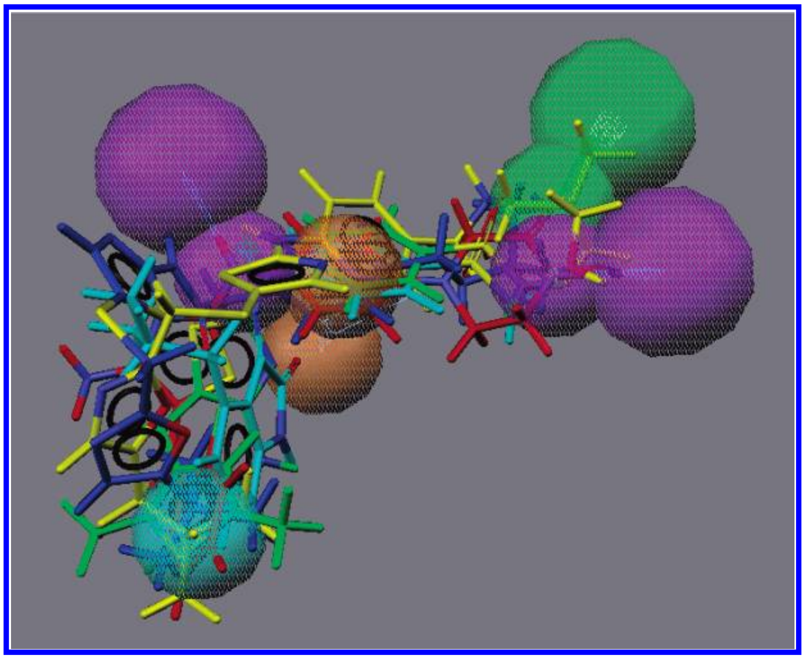


Figure 13. Mapped molecules of the training set toward the qualitative pharmacophore model derived from human CDK2 inhibitors. The pharmacophore model addresses all common features of the training set compounds. The features are colored for the H-bond donor in green, for the H-bond acceptors in magenta, for hydrophobic in blue, and for ring aromatic in brown. All nitrogen atoms are shown in blue and oxygen atoms in red, whereas the molecules are colored in dark blue, light blue, green, yellow, and red.

of a project and have to be identified by building a pharmacophore model derived from a data set with active and less active ligands. In contrast to this usual workflow, we needed a protein-inhibitor scenario with well-known pharmacophore features to validate our results. Another important requirement for the validation scenario was the existence of a specific tautomer form, which has been identified as the active form of the inhibitor and which usually will not be found in common data sets for virtual screening. The barbituric inhibitor RO-200-1770 of the human matrix metalloproteinase 8, described by Brandstetter et al.,[8] fulfilled these requirements. In contrast, the ligand-based pharmacophore approach (CDK2 pharmacophore model) was performed to demonstrate the usability of our approach in a common, nonartificial pharmacophore screening workflow.

Both screening processes reveal significant advantages of the tautomer-enriched data sets compared to their non-enhanced counterparts. On one hand, we could prove that data sets enriched with tautomers are necessary to identify inhibitors, whose activity depends on a specific tautomeric form, which is not included in common structure data sets used for virtual screening. In addition, our comparisons revealed that the discrimination between hits and nonhits was significantly better in the case of tautomer-enriched databases. Finally, we could demonstrate that tautomer-enhanced databases will lead to a higher number of hits, which is the main reason for using tautomer-enriched databases in virtual screening workflows. From oral communication with experienced Catalyst users, we became aware that the hit retrieval of Catalyst (version 4.9) seems to depend on the order of the molecules in the input data set, which is obviously a critical issue of the application. We could prove this theory with the reordered databases 4 (MMP 8) and 8 (CDK) and, therefore, could exclude that the tautomer-enrichment process has a critical impact on the hit retrieval rate in general. It should be mentioned that Accelrys has already fixed that problem in Catalyst 4.10; however, we have not checked this improvement.

The Canonicalization and Enumeration Tool. *Tautomer Enumeration.* With the success of virtual screening techniques, effective methods for the preparation of structure data sets are in demand. While many virtual screening applications address conformations and stereoisomerism, methods to handle ionization states and especially tautomerism are usually not included. Therefore, virtual screening data sets usually need to be enriched with different protonation states and tautomer forms before being used by a standard target-based or pharmacophore-based in silico screening tool. As we started our work on a tautomer enumeration tool meant to be used for virtual screening, only two enumeration applications were available. While Agent[9] has found big interest in the molecular modeling community as a promising tool within the virtual screening workflow, Daylight's[2] much older algorithms for the enumeration and canonicalization of structures have not been noticed as reasonable tools for this specific research field at this time. After we had finished the implementation of our tautomer application, other enumeration tools had been released by several commercial vendors.[18−20] One major restriction of all initial releases of these tools was that these approaches were limited to a specific kind of chemistry, handling only a small number of tautomer transformations. In many cases, olefinic, keto−enol tautomers, as all other variations of $HY{-}C{=}C$, and tautomer systems of $Z{=}C{=}QH \leftrightarrow HZ{-}C{\equiv}Q$ like keten−inol tautomers are not addressed by these approaches. Moreover, hydrogen shifts between distant atoms (1,5 and more) like the nitroso−oxim tautomerism via a phenole system cannot be handled by almost all algorithms. In addition, the tautomer rules often were implemented directly into the source code of the application and therefore could not be extended or modified easily by the users themselves. In addition to these restrictions, to our knowledge, none of the applications mentioned above is capable of performing a tautomer-sensitive duplicate check. However, this technique is useful to identify duplicate tautomer forms generated during the enumeration process and in particular is a suitable tool during the analyses of the virtual screening ranking lists.

By the combination of the "ens transform" algorithm with a set of extended SMIRKS, which considers a large variety of tautomer rules up to long-range 1,11 H migrations, our approach produces exhaustive tautomer collections considering more tautomer types than any other application we are aware of does. Thereby, the tautomer rule set has been tuned to handle all important tautomer forms found in the literature and encountered during our day-to-day work. As a result, our application is capable of handling Z═C═QH ↔ HZ−C≡Q conversions or very long-range hydrogen shifts of aliphatic and aromatic structures up to 1,11 migrations, which are not addressed by other applications.

Furthermore, this architecture clearly separates the tautomer rules from the implemented transformation engine and therefore has several advantages. The predefined transformation set can be easily modified or extended by the user without the necessity of changes on the transformation algorithm. In principle, additional classes of tautomerism, such as ring-side chain tautomerism, which significantly differs from proton migrations (prototropic tautomerism), can easily be handled by adding one or more additional SMIRKS transforms to the set. In addition to these modifications, the transformation process itself can directly be influenced by a large variety of options that come with the "ens transform" command. Second, the user can define their own set of tautomer rules instead of using the predefined set. This is useful, if the user does not need all tautomer forms during the virtual screening or a complete set requires too much CPU time. Finally, this approach gives the user the opportunity to define focused transformation rules for a specific virtual screening project.

Tautomer Canonicalization. While it is easy to identify differently ranked 3D conformers of one molecule in current pharmacophore-based or docking-based screening applications, the identification of differently ranked tautomer forms of the same molecule is not automatically possible. Therefore, tautomeric duplicates have to be identified manually by the cheminformatician, if tautomer-enriched data sets have been used during the virtual screening. This process can become very awkward, if the resulting ranking lists contain several hundred structures. Our tautomer tool used in the presented virtual screening workflow is not only capable of enriching common virtual screening data sets with their tautomer forms but also allows an automatic identification of the differently ranked tautomer forms after the virtual screening has been completed.

Tools that allow a unique identification of chemically equivalent but structurally distinct molecule representations such as tautomers also have been developed by other groups and vendors.[2,4−7] Unfortunately, details about the implemented algorithms are often not disclosed. However, to our knowledge, all existing approaches address this problem by generating a canonical structure. Depending on the implemented canonicalization algorithm of these applications, the tautomer forms handled can differ and are limited to a specific kind of tautomer transformations. For example, many canonicalization applications show the same limitations as already described for the tautomer enumeration tools. Usually, tautomer systems such as Y═C−CH ↔ HY−C═C or Z═C═QH ↔ HZ−C≡Q are not addressed.

In contrast to the described approaches, we do not generate a canonical structure representation. Instead of reducing the given tautomer form, we first generate all other reasonable tautomer structures for the given substance. This process will also be performed by the "ens transform" command in the same way as in the enumeration scenario. In the next step, a standard tautomer form is identified by a rule-based approach. The resulting unique tautomer form is found independently of the starting structure representation.

Our transformation algorithm is completely independent from the chemistry. Therefore, the tautomer duplicate check routine is also not limited to a specific kind of chemistry and includes tautomer classes such as Z═C═QH ↔ HZ−C≡Q conversions or very long-range hydrogen shifts of aliphatic and aromatic structures up to 1,11 migrations, which cannot be handled by other approaches.

Preferred Tautomeric Forms. Ideally, tautomer-enriched data sets should only contain reasonable tautomer forms. On one hand, the elimination of unreasonable structures leads to smaller database sizes after the enumeration process and therefore saves computing time during the virtual screening. On the other hand, high-scored but unreasonable structures will not occur in the ranking lists of virtual screening applications. However, the prediction of reasonable tautomer forms is difficult in several ways. First, the tautomeric equilibration depends on multiple factors, such as thermodynamic factors, solvent dependencies, the pH of solution, dipole−dipole repulsions, and intramolecular H bonds. Second, the types and positions of substituents with different electronic properties as well as the conjunction of the molecule change the transformation rate significantly toward a preferred molecular constitution. Finally, steric constraints such as ring closures of sugars and ionization at a defined pH range destabilize the balance of tautomeric forms. This issue becomes more complex especially if we take into account the conditions and interactions in the active site of the protein, which may prefer another tautomer form than that found in an aqueous solution of the compound. New releases of some enumeration approaches tried to address this problem recently.[9,20]

In this study, we did not limit the enumeration process to reasonable tautomer forms. All virtual screening examples presented in this work used the standard set of tautomer transformation rules described above, which leads to an exhaustive enumeration process. Although we generate a high number of rather unlikely and probably not stable tautomer forms with this approach, the results of our virtual screening examples are already very promising. The definition of a problem-focused tautomer rule set instead of the predefined transformation set will probably lead to an increase of the result quality and can be done easily by using our tool. In other cases, it might be a better solution to generate all tautomer forms, rather than losing some important tautomers during the screening process because of a too-restrictive transformation rule set. Therefore, we believe that this issue should be addressed on a case-by-case basis by the user.

Finally, we want to mention that the most reasonable tautomer form can be generated by our approach using our rule-based rating mechanism. Usually, the resulting preferred representation corresponds closely to the preferred tautomer forms found for many tautomer sets in the literature. Unlike energy-based approaches, this rule-based mechanism is very fast.

ACKNOWLEDGMENT

The authors wish to thank Oliver Sacher (Molecular Networks GmbH) and Tadashi Hidaka (Takeda Pharmaceutical Ltd.) for their numerous hints and suggestions during the refinement of our tautomer rules. Mr. Hidaka was not personally involved in this project but supplied Xemistry, in the context of a separate tautomer-identification project at Takeda, with a rich set of complex structure test cases which were used to optimize the standard toolkit tautomer transform rule set.

REFERENCES AND NOTES

(1) MDL Available Chemicals Directory. http://www.mdl.com/products/experiment/available_chem_dir/index.jsp (accessed Dec 2004).

(2) Sayle, R.; Delany, J. Canonicalization and Enumeration of Tautomers. Materials of EuroMug-99, October 28−29, 1999, Cambridge, U. K.

(3) *SYBYL 6.92*; Tripos Associates Inc.: St. Louis, MO.

(4) 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 (accessed Dec 2004).

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

(6) MDL Information Systems. http://www.mdl.com (accessed Dec 2004).

(7) Trepalin, S. V.; Skorenko, A. V.; Balakin, K. V.; Nasonov, A. F.; Lang, S. A.; Ivashchenko, A. A.; Savchuk, N. P. Advanced Exact Structure in Large Databases of Chemical Compounds, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 852−860.

(8) Brandstetter, H.; Grams, F.; Glitz, D.; Lang, A.; Huber, R.; Bode; W.; Krell, H. W.; Engh, R. A. The 1.8-Å Crystal Structure of a Matrix Metalloproteinase 8-Barbituric Inhibitor Complex Reveals a Previously Unobserved Mechanism for Collagenase Substrate Recognition *J. Biol. Chem.* **2001**, *276*, 17405−12.

(9) Pospisil, P.; Kuoni, T.; Scapozza, L.; Folkers, G. Advanced Virtual Docking: The Issue of Tautomers. *Helv. Chim. Acta* **2002**, *85*, 3237−50.

(10) Sadowski, J. *Abstracts ACS National Meeting*, 224th National Meeting of the American Chemical Society, Boston, MA, August 18−22, 2002; American Chemical Society, Washington, DC, 2002.

(11) Mazurek, A.; Osman, R.; Weinstein, H. Molecular Determination for Recognition of Triazole and Tetrazole Analogs of Histamine at H_2-Receptors. *Mol. Pharmacol*. **1986**, *31*, 345−350.

(12) Duke, N. E. C.; Codding, P. W. Structural and Molecular Modeling Studies of Quinazolinone Anticonvulsants. *Acta Crystallogr., Sect B* **1993**, *49*, 719−726.

(13) Hernánandez, B.; Orozco, M.; Luque, F. J. Role of the Tautomerism of 2-Azaadenine and 2-Azahypoxanthine in Substrate Recognition by Anthine Oxidase. *J. Comput.-Aided Mol. Des*. **1997**, 153−162.

(14) Wu, C. S.; Huang, J. L.; Sun, Y. S.; Yang, D. Y. Mode of Action of 4-Hydroxyphenylpyruvate Dioxigenase Inhibition by Triketone-Type Inhibitors. *J. Med. Chem.* **2002**, *45*, 2222−2228.

(15) Yu, Q.; Zhu, X.; Holloway, H. W.; Whittaker, N. F.; Brossi, A.; Greig, N. H. Anticholinesterase Activity of Compounds Related to Geneserine Tautomers. N-Oxides and 1,2-Oxazines. *J. Med. Chem.* **2002**, *45*, 3684−3691.

(16) Weinstein, H.; Chou, D.; Johnson, C. L.; Kang, S.; Greem, J. P. Tautomerism and the Receptor of Histamine: A Mechanistic Model. *Mol. Pharmacol*. **1976**, *12*, 738−745.

(17) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer-Aided Drug Design. *J. Recept. Signal Transduction* **2003**, *23* (4), 361−371.

(18) *QuAC PAC*; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com/products/applications/quacpac.html (accessed Dec 2004).

(19) *Pipeline Pilot*, v. 3.0; Scitegic: San Diego, CA. http://www.scitegic.com (accessed Dec 2004).

(20) *LigPrep*; Schrödinger: Portland, OR. http://www.schrodinger.com/docs/2004−1/final_pdf/ligprep/lp15_user_manual.pdf (accessed Dec 2004).

(21) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. J. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.

(22) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. J. CACTVS: A Chemistry Algorithm Development Environment. In *Proceedings of the 15th Symposium on Chemical Information and Computer Sciences/20th Symposium on Structure−Activity Relationships*; Machida, K., Nishioka, T., Eds.; Kyoto University: Kyoto, Japan, 1992; pp 102−105.

(23) *Catalyst 4.9*, release 4.9.1.; Accelrys Inc.: San Diego, CA, 2004.

(24) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(25) *Cerius2 Modelling Environment*, release 4.9; Accelrys Inc.: San Diego, CA, 2004.

(26) Beyer, C.; Cramer, J.; Selzer, P. M. Oral & Poster Presentation at the 18th Darmstädter Molecular Modelling Workshop 2004, May 18−19, 2004, Erlangen, Germany.

(27) Kernighan, B. W.; Dennis, M.; Ritchie, D. M. *Programmieren in C, Zweite Ausgabe, ANSI C*; Carl Hanser Verlag: München, Germany, 1990

(28) Ousterhout, J. K. *Tcl and the Tk Toolkit*; Addison-Wesley: Reading, MA, 1994.

(29) Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M.; Delany, J. J. Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1161−1172.

(30) Elguero, J.; Marzin, C.; Katritzky, A. R.; Linda, P. *The Tautomerism of Heterocycles*; Academic Press: New York, 1976.

(31) *CACTVS Tcl Command Reference*; Xemistry GmbH: Lahntal, Germany, 2002. http://www.xemistry.com/academic/tcl_reference.pdf (accessed Dec 2004).

CI060109B