JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

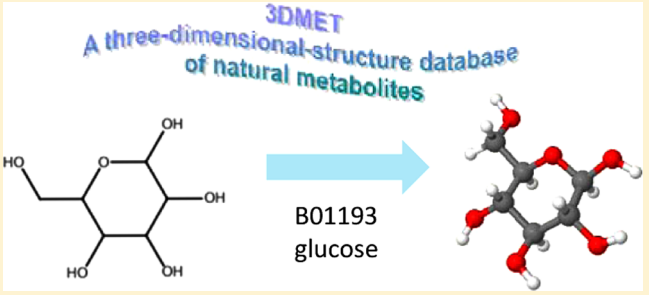
# Three-Dimensional Structure Database of Natural Metabolites (3DMET): A Novel Database of Curated 3D Structures

Miki H. Maeda*,[†] and Kazumi Kondo[‡]

[†]Biomolecular Research Unit, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

[‡]Qs' Research Institute, Otsuka Pharmaceuticals, Co., Ltd., 463-10 Kagasuno Kawauchi-cho, Tokushima, 771-0192, Japan

**ABSTRACT:** A database of 3D structures of natural metabolites has been developed called 3DMET. During the process of structure conversion from 2D to 3D, we found many structures were misconverted at chiral atoms and bonds. Several popular converters were tested in regard to their conversion accuracy. For verification, three canonical strings were also tested. No procedure could satisfactorily cover all the structures of the natural products. The misconverted structures had to be corrected manually. However, a nonnegligible number of mistakes were also observed even after manual curation, so a self-checking system was developed and introduced to our work flow. Thus, the 3D structures in our 3DMET database were evaluated in two steps: automatically and manually. The current version includes most of the natural products of the KEGG COMPOUND collection [http://www.genome.jp/kegg/compound/] and is searchable by string, value range, and substructure. 3DMET can be accessed via http://www.3dmet.dna.affrc.go.jp/, which also has detailed manuals.




## INTRODUCTION

Ongoing genome research has resulted in a huge number of DNA/protein sequences being determined, but the function of about half of them have not yet been identified. Annotation of genome sequences is generally based on the similarity of known protein sequences. In the case of the rice genome project, the annotatable threshold of similarity is 60% identity.[1] As the result, about half the rice genes estimated to be open reading frames were not annotated. A new method for protein annotation with weak similarity or no similarity to known proteins is needed when only sequences are known.

To estimate protein function, we can apply other information: experimental results from physiology, biochemistry, and/or structural biology. One candidate information to bridge between sequence and function is 3D structure. About 80 000 entries are now available from the Protein Data Bank.[2] It contains both sequence and atomic coordinate information. Determination of 3D structures has become a popular technique for analysis of proteins. When using experimental methods such as X-ray crystallography or NMR, the 3D structures of transcriptional products without homology to known proteins can be identified. If the sequence homology of two proteins is more than 40%, 3D structures can be estimated by similarity modeling.[3] Thus, if 3D structures can be applied to the functional annotation of proteins, more proteins of unknown function will be estimated.

How can we relate 3D structures to biochemical or physiological function? One answer is by predicting their natural ligands. Protein−ligand interaction is an essential component of protein function. In other words, to discover natural ligands becomes a key piece of information for estimating protein function. This interaction can be approachable by two computational methods: docking simulation and protein−protein docking, distinguished by the ligand size. The two methods are based on the same concepts but are technically distinguished: Protein−protein docking requires a large computer resource, whereas docking simulations entail relatively small costs. There are more computer programs for docking simulation than for protein−protein docking. Therefore, we targeted estimation of small ligands by using docking simulation. When this project started, we could not find any database where 3D structures of natural compounds was publicly available. As the result, we developed a new database, called the three-dimensional structure database of natural metabolites (3DMET).

At the beginning of database construction, we thought that we could simply convert 2D structures of natural products databases to 3D structures. However, the results showed that the original 2D structures were not the same as the converted 3D structures, especially at chiral atoms and bonds. Therefore, we had to check the correspondence of structures before and after conversion. We can apply several methods to judge whether two structures are the same or not. One method is superposition of two compounds. For example, "Fit atoms" or "Match atoms" of SYBYL[4] can superpose two structures if two compounds are completely the same. Otherwise, a program using graph theory may be applicable if some chiral information

**Received:** July 5, 2012
**Published:** January 7, 2013

   dx.doi.org/10.1021/ci300309k | *J. Chem. Inf. Model.* 2013, 53, 527−533

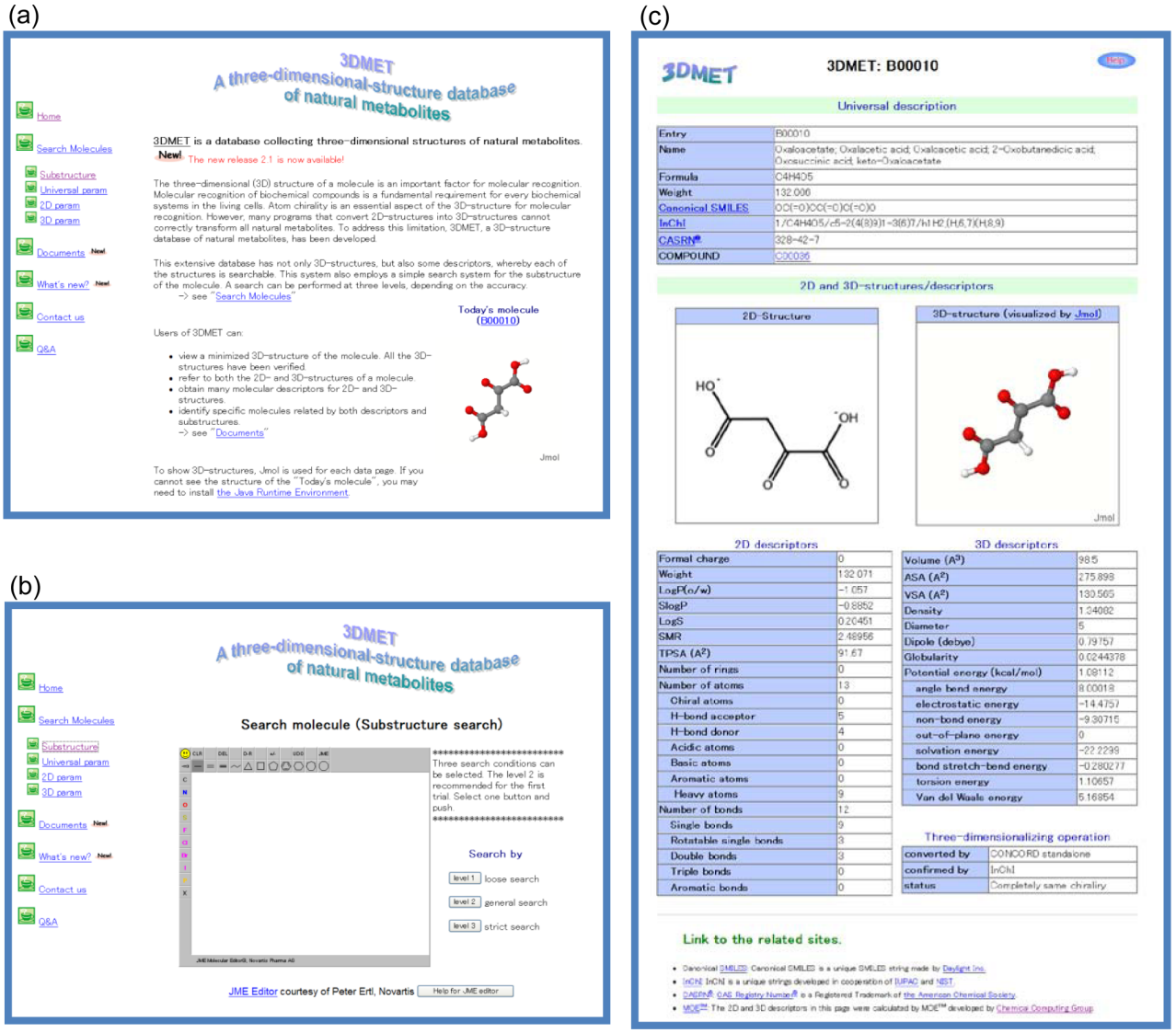



**Figure 1.** Web pages of 3DMET: (a) top page of the Web site, (b) substructure search page, and (c) an example of the data page.

is included in the program. In both cases, much time will be spent when many compounds need to be compared.

Line notation offers another method. "SMILES (simplified molecular input line entry specification syntax)"[5,6] is a widely used line notation of compounds. One compound can be described by various kinds of SMILES strings. Therefore, "canonical SMILES"[7] is also provided by the developer to make only one string for molecules with chiral information. WLN (Wiswesser line notation)[8] and SLN (SYBYL line notation)[9] are algorithms developed from almost the same idea. "InChI (IUPAC International Chemical Identifier)"[10] is a recently developed method. This is a little different from the earlier line notations. SMILES, WLN, and SLN just produce a string of characters from an input atomic coordinate file. However, InChI makes a structure related string and an InChI-key after estimation of chemical characteristics, e.g., tautomerization. Thus, two tautomers derived from one compound can be judged to be the same. If two line notation strings before and after conversion correspond, the two structures are judged to be the same compound. This approach is faster than superposition, and we chose this comparison to check structures.

## MATERIALS AND METHODS

**Evaluation of 2D−3D Converters.** Two-dimensional mol files of the Kyoto Encyclopedia of Genes and Genomes (KEGG) COMPOUND collection,[11] made as a part of the KEGG LIGAND database,[12] was used for 2D−3D conversion. Release 29 data of C00001−C14124 (12 161 entries) was downloaded on Nov. 29, 2005. Structures containing R (residue), X (halogen), and n (repeat) were deleted for the sample data set because their structures were not fully identified. The 2D structure of each COMPOUND entry file was converted into a 3D structure by CONCORD sketch/standalone,[13] CORINA (ver. 3.4),[14] or MOE (2004 and 2005).[15] The converted structures were energy minimized with the SYBYL molecular modeling system[4] for post-CONCORD structures and with MOE for the others. For the purpose of evaluation including atom and bond chirality, 2D and 3D structures were compared with InChI,[10] canonical SMILES,[6] and aRSChirality, a module of MOE described the next section. For all of minimizations reported here, the TRIPOS and MMFF94x force fields were employed in SYBYL and MOE, respectively. Minimization calculations were run with default program settings.


dx.doi.org/10.1021/ci300309k | J. Chem. Inf. Model. 2013, 53, 527−533

**Confirmation by Canonical Strings.** For confirmation of two structures, we used two kinds of canonical strings to describe the compound structures, i.e., InChI[10] and canonical SMILES.[6] Concerning phosphate, P═O and P−OH are clearly distinguished in the canonical SMILES strings. However, the two moieties are equal because the hydrogen of P−OH is released in aqueous solution. Therefore, this chiral tag was removed from the canonical SMILES strings for our verification operation. A string converted by using these programs we tested defines one structure. So if at least one of the two strings is conserved between before and after the conversion, the converted 3D structures were deemed to be "correct".

**Structure Files of the Database.** The 2D structure files of metabolites were obtained from the COMPOUND database.[11] These files are described in MDL mol format.[16,17] In the COMPOUND database, there are many entries of non-natural metabolites and structures not defined as unique molecules. Many of the non-natural entries are metabolites of artificial drugs or herbicides. Because no such compounds exist under normal conditions in a natural living body, they should be deleted from a natural metabolite data set.

The initial metabolite data set was made of entries corresponding to one of the following KEGG pathway descriptions: (1) included in the KEGG pathway map no. 1.1 to 1.10, (2) listed in the "compounds with biological roles" table, (3) listed in the "Lipids" table, or (4) listed in the "Phytochemical compounds" table. Next, a secondary data set was made from the initial data set by removing the entries not defined as unique structures, i.e., the entries with nonatomic symbols such as "R (residue)", "n (repeat number of substructure)", and "X (halogen)". Some entries with halogens (F, Cl, and Br) were also deleted from the data set because these compounds with halogens are not encountered in ordinary conditions.

The 2D structure of each file in the metabolite data set was converted into a 3D structure by CONCORD standalone[13] and energy minimized by SYBYL.[4] Our preliminary analysis about the structure conversion indicates that a significant number of structures had been misinterpreted at chiral atoms when compared to the original 2D mol files. Therefore, the chiral conservation of two structures before and after conversion should be verified by the procedure described above in the Confirmation by Canonical Strings section. The compounds that could not be correctly converted by the above procedure were manually corrected and energy minimized by MOE.[15] All of the "correct" structures for the last data set were collected in the database. Thirty structure-based descriptors calculated by MOE were also added to the database.

**Database System.** The 3DMET system is constructed on the PostgreSQL[18] as a relational database management system. Its web interface is implemented using PHP[19] and HTML (Figure 1a−c). Three searching methods are provided on the web site: string, value range, and substructure. The string search and value range search are executed by the function of PostgreSQL. For the substructure search, the JME molecular editor applet[20] is used for a graphical input (Figure 1b). Our simple substructure search is based on a search for a SMILES string made by the applet. We provide three levels of the search differing in regard to database description and search keys made by the editor. The three levels are as follows: level 3, ignore atom chiralities; level 2, ignore atom chiralities (level 3 set), charge, and subset definition; level 1, ignore atom chiralities, charge, subset definition (level 2 set), and bond types.

The data page for each compound is separated into four sections: universal description, 2D structure and descriptors, 3D structures and descriptors, and conversion status. Thirty descriptors are listed with 2D and 3D structures (Figure 1c). The 3D structure is visualized by Jmol.[21,22] To fully utilize this Web site using the JME editor and the Jmol viewer, the Java Runtime Environment is required. On the data page, the 2D−3D converter, the verification strings, and verification status for the visualized 3D structure are also shown in the conversion status column labeled "Three-dimensionalizing operation".

**Structure-Checker.** Structure-Checker, an in-house structure confirmation system, was developed using HTML and PHP. The outline of the data handling in the server is as follows: Two input files described in the MDL SD/mol format are uploaded from PC to the server and transferred to InChI and/or canonical SMILES strings (Figure 2). The InChI strings are separated by the sections and resulting canonical SMILES and InChI strings are displayed.

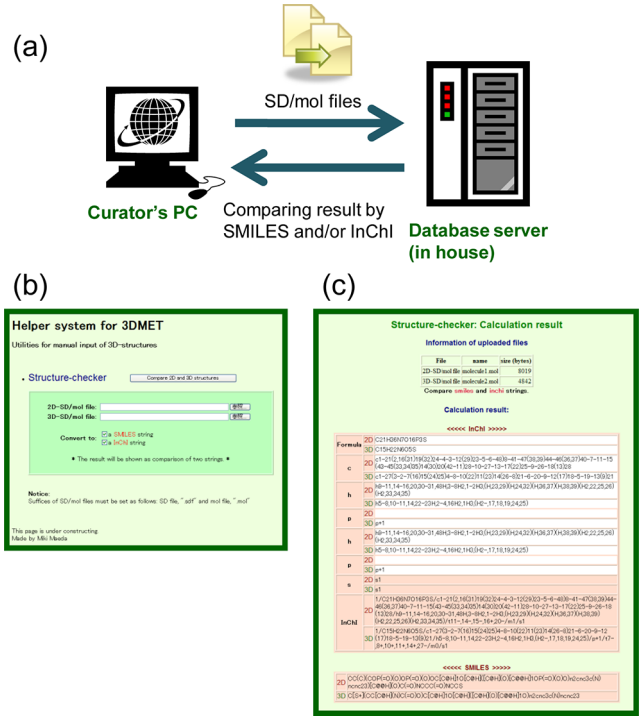



**Figure 2.** Structure-Checker: (a) outline, (b) the top page, and (c) an example of the results page of the system.

## RESULTS AND DISCUSSION

**Accuracy of 2D−3D Converters.** Initially we planned automatic construction by using CONCORD sketch. When the conversion for all entries of COMPOUND had finished, we checked more than 100 of the resulting structures and found that many structures were improperly converted. Results of our tests are shown in Table 1. Table 1 shows the results of the five converters evaluated by three verification methods. InChI and SMILES are better than aRSChirality, which is a module of MOE. The module program makes a chiral tag string from a molecule by the atom order in the file. In some conversion cases, atom orders were different in pre- and postconverted files. The reason for weak detection by aRSChirality could be the difficulty of the correspondence of the atoms. Canonical SMILES is better in the case of CONCORD standalone, and


dx.doi.org/10.1021/ci300309k | J. Chem. Inf. Model. 2013, 53, 527−533

**Table 1. Accuracy of 2D−3D Converters and Verification Methods**

| converter | version | verification methods: InChI | canonical SMILES | aRSChirality |
|---|---|---|---|---|
| CONCORD | standalone | 6533 (76%) | 6605 (77%) | 2984 (35%) |
| CONCORD | sketch | 3323 (39%) | 3648 (42%) | 2505 (29%) |
| CORINA | 3.4 | 5388 (63%) | 6390 (74%) | 2454 (29%) |
| MOE | 2004 | 3447 (40%) | 3464 (40%) | 2429 (29%) |
| MOE | 2005 | 3319 (39%) | 3321 (39%) | 2414 (28%) |

[a]Each program was evaluated using 8601 entries of COMPOUND.[11] The percents shown are the percentages of correct conversions.

InChI is better in the case of CORINA. To learn the reason, the errors were checked in detail. In most cases of unmatched SMILES, no strings were produced for the structures. Because these molecules were symmetric structures or multiring complexes, the canonicalization step had failed. The other SMILES strings, i.e., unique SMILES by SYBYL 6.9 and by MOE 2005 were also tested, but canonical SMILES of Daylight achieved the highest detection (data not shown). Consequently, we chose canonical SMILES for structure verification.

All the converters could make correct 3D structures up to 80% yields. Two programs, CONCORD standalone and CORINA, yielded relatively better results than the other three, CONCORD sketch, MOE 2004, and MOE 2005. Another program, Ligprep, gave also the same yield as CONCORD standalone and CORINA (data not shown). Because we recognized CONCORD standalone and CONCORD sketch were similar, we planned to use CONCORD sketch at first. However, the two programs were very different in accuracy. According to the vendor, this difference is caused by the purpose of each module. CONCORD sketch needs manual correction in the graphical user interface. Comparison of the results from CONCORD and CORINA showed unconvertible structures were slightly different. Compared to the other conversion programs, MOE 2004 and 2005 did not perform well. However, a more recent version of MOE seemed to be an improvement. For example, version 2009, chiral errors in sugar were less observed. Some corrections in the chiral detection module seemed have been added. On the basis of our results, we now use MOE (2009/2010/2011) for manual development and correction of the 3D structures.

At first, the conversion error was unpredictable because 2D−3D converters were already utilized for drug design. As shown by the examples in Figure 3, the errors frequently occurred in the structures with large (more than 10 atoms) complex rings (Figure 3a), sugars (Figure 3b), and steroids (Figure 3c). Sometimes errors were found in amino acids. For the standalone programs (CONCORD standalone and CORINA), all errors were in regard to stereoisomerism at atoms and bonds. Thus, this conversion problem may not be so serious because many synthetic compounds have only a few chiral atoms. However, since our interest is in 3D structures of metabolites, the accuracy of the programs was insufficient.

**Overview of the 3DMET History.** The current release of 3DMET is 2.1. From the initial release of 3DMET (2005) to the current, structures are developed based on the KEGG COMPOUND database[11] entries. In the earliest release, 2D mol files were converted by CONCORD sketch, energy minimized by SYBYL 6.9, and confirmed by InChI and/or SMILES. The number of entries was 1123. Next, in the release 2.0 (2008), the converter program was changed to CONCORD standalone and minimized by SYBYL7.2, and 5920 entries were available. The remaining COMPOUND entries, mistaken by automatic conversion, were manually corrected and confirmed after release 2.0. The current release is 2.1. It has 8581 entries derived from the COMPOUND database from Feb. 6, 2011. 3DMET is accessible at http://www.3dmet.dna.affrc.go.jp/. Use of the database on the site is freely available. For restrictions and availability to download bulk data, users should contact to us via the above Web site. General information is shown on the Web site.

**Structure-Checker.** During the process of the database development, automatic conversion could not treat all COMPOUND entries as we have described above. Hence, we started manual curation and correction of the converted 3D structure files after publishing release 2.0. We are developing new entries by the following steps: (1) Two-dimensional mol files are three-dimensionalized by MOE, (2) converted 3D files are checked and corrected by chemically educated curators, (3) corrected 3D files are rechecked visually by using Structure-Checker (illustrated in Figure 2a).

The input files to the system are two SD/mol files (Figure 2b). After uploading the files, two structures are compared by their canonical strings selected on the uploading page, InChI and/or SMILES. If no files and/or empty files are found, an error message page will be returned. The resulting page is shown in Figure 2c. In the part labeled "Output of SMILES", two canonical SMILES strings are shown in parallel because differences are easily distinguishable by the string length and by checking the symbol of "@", "\", and "/" when using canonical SMILES. In the part labeled "Output of InChI", InChI strings are separated and compared as columns. If strings in the two columns differ, the background color is white, whereas the

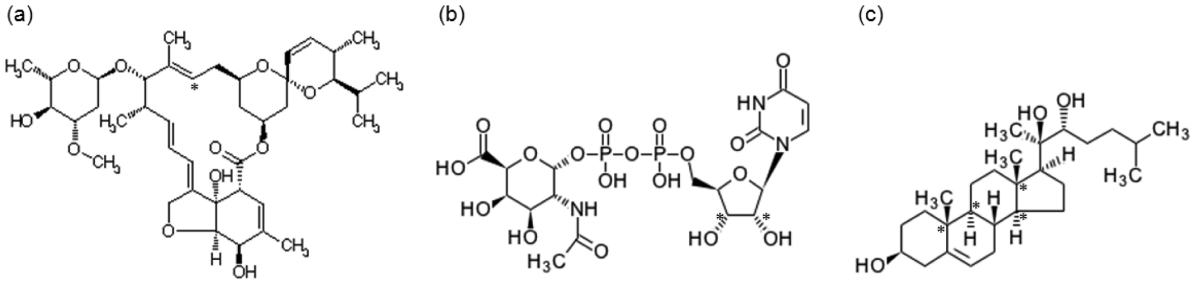



**Figure 3.** Typical structures incorrectly handled by commercial 2D−3D converters: (a) C11965, avermectin B1b monosaccharide; (b) C13952, UDP-*N*-acetyl-D-galactosaminuronic acid; and (c) C05501, 20$\alpha$,22$\beta$-dihydroxycholesterol of KEGG COMPOUND entries. Asterisks on the molecular graphics show improperly assigned stereocenters.


dx.doi.org/10.1021/ci300309k | J. Chem. Inf. Model. 2013, 53, 527−533

background color is orange when two strings correspond (Figure 2c). By using this system, it became easier to compare two strings. It is difficult to estimate the atom number of InChI from just the InChI string. Thus, we also use w-InChI, the InChI program working on the Windows PC with a graphical interface, to identify the atom number.

**How to Make Correct 3D Structures.** Our present collection policy of the structural data of 3DMET is as follows: (1) In the case of conversion from 2D structures provided by another database source, resulting 3D structures should be same as in the original database. If no chiral definition was originally given for any atoms/bonds, any chiral description after automatic conversion is permitted. (2) In the case of development from a paper resource, all the isomers should be produced for a compound with less than three unknown chiral centers. Because our interest is in natural ligands, it is not desirable to miss any natural isomers. Thus, at least one of the isomers should exist naturally for our purpose. Detailed information and update logs about the database are available on the 3DMET Web site. The Web site will be updated as the occasion demands.

In release 2.0, all 3D structures in 3DMET were automatically verified by InChI and canonical SMILES because both strings are needed (Table 1). A weak point in the process is dealing with two canonical strings especially for complex molecules consisting of more than two rings. In many cases of such compounds, one string could not be produced from 3D structures. Improperly converted structures were manually corrected as described above and recorded in release 2.1. During manual correction, many mistakes were found even in the structures after manual curation. These errors were mainly caused by having two or more chiral errors in a file. Curators missed correcting more than one mistake per molecule. That is the reason why we developed an in-house system named "Structure-Checker" for comparing 2D and 3D structures. By using this self-checking system on the resulting structures, mistakes were significantly reduced. Subsequently, the conversion for all natural products defined in the Materials and methods section in COMPOUND (Jan 2011) was completed. In July 2011, the ftp site for the KEGG database including COMPOUND database started requiring payment. Recent extension of the database relates mainly to medicinal chemicals. Therefore, the future resource of metabolites for 3DMET will be other databases and/or books.

**Current Database System.** Three searching methods are provided on the Web site. String search and value range search are performed as a function of postgreSQL. String search can be applied to general descriptions such as name, formula, InChI, etc. The calculated descriptors are searchable by a value range. Substructure searching was introduced starting with Release 2.0. Substructure searching is based on the correspondence of canonical SMILES strings.

There are three kinds of structure search algorithms: graph based, fingerprint based, and SMILES based. Each of these algorithms has its own merit. The graph-based method detects all structures with strictly the same substructures. The fingerprint method can detect similar but not necessarily the same structures. The SMILES-based search method is the most rough but is very fast and simple.

Considering the power of our server, we chose the SMILES method. But some improvements were added. By ignoring characters in the SMILES string, the search becomes stricter on the order from level 3 to 1. Some results comparing the three levels are shown in Table 2. By this small improvement, some similar structures can be detected. It may be enough to find some similar structures of the target in the natural products. We know the limitation of the SMILES method. Thus, the graph-based method and/or the fingerprint-based method will be introduced on a future version of the Web site.

**Table 2. Examples for the Three Levels of the Substructure Search**

| | query = benzene | | query = phenol | |
|---|---|---|---|---|
| level | hit number | example | hit number | example |
| level 3 | 565 | | 30 | |
| level 2 | 730 | 4-aminobenzoate (B00134) | 317 | fomonectin (B00192) |
| level 1 | 815 | cyclohexylamine (B00135) | 351 | cyclohesanone (B00106) |
| | query = 1,3-butadiene | | query = *n*-octane | |
| level | hit number | example | hit number | example |
| level 3 | 142 | | 193 | |
| level 2 | 221 | retinal (B00095) | 222 | undecaprenyl phosphate (B00093) |
| level 1 | 3322 | L-lysine (B00013) | 433 | arachidonic acid (B00061) |

The search results can be displayed as a data page: 30 descriptors (Table 3) are listed with 2D and 3D structures (Figure 1c). In the former release, MDL Chime was used as the 3D visualizer. However, the program worked only on the Windows PC. The Jmol applet introduced to the web page starting with release 2.0 permits 3D viewing of the compound on Windows, Macintosh, and Linux computers. The 2D−3D converter, the verification strings, and verification status (including manual curation) for the 3D structure are shown in the independent column of the data page.

**Table 3. Searchable Descriptors on the 3DMET Server**

| category | descriptor information |
|---|---|
| general descriptors | entry, name, formula, molecular weight, InChI, canonical SMILES, CASRN |
| 2D descriptors | formal charge, weight, logP(o/w), SlogP, logS, SMR, TPSA, number of rings, all atoms, chiral atoms, H-bond acceptors, H-bond donors, acidic atoms, basic atoms, aromatic atoms, heavy atoms, all bonds, single bonds, rotatable single bonds, double bonds, triple bonds, aromatic bonds |
| 3D descriptors | volume, ASA, VSA, density, diameter, dipole, globularity, potential energy |

[a]Names of 2D and 3D descriptors shown above are those used in MOE. Detailed information is on the 3DMET website: http://www.3dmet.dna.affrc.go.jp/.

**Significance of a Metabolite Database of 3D Structures.** Database construction of metabolite structures serves two purposes: a searchable collection of metabolites and a basic library for drug design. The 3DMET database is a resource for both purposes.

For metabolomics research, the physicochemical properties of the metabolites are very important. On the basis of the current results of the metabolomics project, the importance of small metabolites is gradually becoming clear. For example, it


dx.doi.org/10.1021/ci300309k | J. Chem. Inf. Model. 2013, 53, 527−533

has been found that the amounts of many kinds of metabolites fluctuate depending on the environment.[23,24] However, natural compounds have been handled in bioinformatics research as only graphical nodes of the biological pathway. In such a case, all nodes are equal on the pathway maps, no matter how the compound corresponding to each node is distinguishable physically and chemically. At the same time, it is well-known that some properties of compounds are related to their biological roles in a cell. One famous example is nonspecific membrane transport. This phenomenon is concerned with the physicochemical properties such as the polar surface area (PSA) and the octanol−water partition coefficient (LogP) of compounds,[25,26] but not local or global structures. In addition to experimentally determined values of descriptors, descriptors predicted by calculation are widely used. For example, computed CLogP[27,28] and SLogP[29] values correspond well with experimental LogP values. Likewise, TPSA[30] values are calculated for polar surface area (PSA). In general, these predicted descriptors are calculated based on the 2D structures of molecules. Thus, a collection of 2D structures is useful for such purposes.

On the other hand, 3D structures are important in drug design research because they are necessary for virtual screening and for structure-based drug design (SBDD).[31] Databases of 2D/3D structures such as the KEGG COMPOUND collection,[11] ChemBank,[32] ChEBI,[33] ZINC,[34] and KNApSAcK[35] are available. Many pharmaceutical companies have their own compound libraries (database and material set) containing commercial and in-house synthetic compounds for finding leads to new medicines. When using a 3D structure database as a screening library, the structures are often directly applied. However, the accuracy of the 3D structure is important. If a database has 2D only structures, some three-dimensionalization steps are necessary. Previously constructed 3D structure databases have been generated by such converters.[34,36] During the construction processes, we found that widely used commercial 2D−3D converters are not so reliable on stereochemical structures with many chiral atoms and bonds, especially for natural products. It is impossible to estimate a priori the rate of accurate structures in them. Hence uncertainty clouds attempts at virtual screening. Some further operations including manual inspection and correction are necessary if such databases are to be used reliably. In addition, when metabolite structures are needed, hypothetical compounds should be distinguished from natural products. 3DMET meets the need for a novel resource available now with reliable 3D structures of natural products.

## CONCLUSION

A 3D structure database of natural metabolites (3DMET) has been developed. Release 2.1 became available in February 2012. This release covers 3D structures of most of the natural products in the KEGG COMPOUND collection.[11] For verification of the 3D structures, automatic and manual curation performed on every 3D structures. The database contains physicochemical properties derived from the 2D and 3D structures frequently used in chemoinformatics research. By using these descriptors, physicochemically similar compounds can be identified. On the data page, these descriptors are shown and 2D and 3D structures are displayed.

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: mmaeda@nias.affrc.go.jp. Telephone/fax: +81-29-838-7010.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The JME molecular editor applet was kindly supplied by Dr. Peter Ertl of the Novartis Institutes for BioMedical Research. We would like to thank Mr. Hisataka Numa of National Institute of Agrobiological Sciences, and Ms. Mie Saiki and Ms. Yumiko Hodotsuka for their technical support. The descriptor calculation by MOE was performed by using the computer resources of the Computer Center of the Ministry of Agriculture, Forestry and Fisheries. This work was supported by a grant-in-aid for Publication of Scientific Research Results (2118065 and 238062).

## REFERENCES

(1) The rice annotation project. Curated genome annotation of *Oryza sativa ssp. japonica* and comparative genome analysis with Arabidopsis thaliana. *Genome Res.* **2007**, *17*, 175−183.

(2) Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlić, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukich, B.; Zardecki, C.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.* **2011**, *39*, D392−D401.

(3) Cavasotto, C. N.; Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today* **2009**, *14*, 676−683.

(4) *SYBYL*, Tripos Inc., St. Louis, MO, USA, 2006; http://www.tripos.com/.

(5) Weininger, D. SMILES, A chemical language and information-system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(6) *SMILES*, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA, USA, 2006; http://www.daylight.com/.

(7) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(8) Wiswesser, W. J. The Wiswesser Line Formula Notation. *Chem. Eng. News* **1952**, *30*, 3523−3526.

(9) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(10) *The IUPAC International Chemical Identifier*, version 1.02. http://old.iupac.org/inchi/ (accessed July 2012).

(11) KEGG COMPOUND. http://www.genome.jp/kegg/compound/ (accessed Nov 2005).

(12) Goto, S.; Nishioka, T.; Kanehisa, M. LIGAND: Chemical database for enzyme reactions. *Binformatics* **1998**, *14*, 591−599.

(13) Pearlman, R. S. Rapid generation of high quality approximate 3-D molecular structures. *Chem. Des. Auto. News* **1987**, *2* (1), 1,5-7.

(14) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(15) *Molecular Operating Environment (MOE)*, version 2004.03 and 2005.06, Chemical Computing Group, Montreal, Canada, 2004 and 2005; http://www.chemcomp.com/.

(16) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. A. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(17) *CTfile Formats*; Accelrys, Inc., San Diego, CA, USA, 2003; http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php.
(18) PostgreSQL, http://www.postgresql.org/ (accessed July 2012).
(19) PHP, http://www.php.net/ (accessed July 2012).
(20) Ertl, P. *JME molecular editor*; Novartis A. G.: Basel, Switzerland, 2002; http://www.molinspiration.com/jme/index.html.
(21) Murray-Rust, P.; Rzepa, H. S.; Williamson, M. J.; Willighagen, E. L. Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 462−469.
(22) Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ (accessed July 2012).
(23) Hirai, M. Y.; Klein, M.; Fujikawa, Y.; Yano, M.; Goodenowe, D. B.; Yamazaki, Y.; Kanaya, S.; Nakamura, Y.; Kitayama, M.; Suzuki, H.; Sakurai, N.; Shibata, D.; Tokuhisa, J.; Reichelt, M.; Gershenzon, J.; Papenbrock, J.; Saito, K. Elucidation of gene-to-gene and metabolites-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. *J. Biol. Chem.* **2005**, *280*, 25590−25595.
(24) Soga, T.; Baran, R.; Suematsu, M.; Ueno, Y.; Ikeda, S.; Sakurakawa, T.; Kakazu, Y.; Ishikawa, T.; Robert, M.; Nishioka, T.; Tomita, M. Differential metabolomics reveals ophthalmic acid as an oxidative stress biomarker indicating hepatic glutathione consumption. *J. Biol. Chem.* **2006**, *281*, 16768−16776.
(25) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.
(26) Subramanian, G.; Kitchen, D. B. Computational models to predict blood-brain barrier permeation and CNS activity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 643−664.
(27) Leahy, D. E.; Taylor, P. J.; Wait, A. Model Solvent Systems for QSAR Part I. Propylene Glycol Dipelargonate (PGDP). A new Standard Solvent for use in Partition Coefficient Determination. *Quant. Struct.-Act. Relat.* **1989**, *8*, 17−31.
(28) Leo, A. J. Calculating log Poct from structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(29) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(30) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(31) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078−1082.
(32) Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N. J.; Schreiber, S. L.; Clemons, P. A. ChemBank: A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* **2008**, *36*, D351−D359.
(33) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2008**, *36*, D344−D350.
(34) Irwin, J. J.; Shoichet, B. K. ZINC − A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(35) Nakamura, Y.; Asahi, H.; Altaf-Ul-Amin, Md.; Kurokawa, K.; Kanaya, S. *KNApSAcK*. http://kanaya.naist.jp/KNApSAcK/ (accessed July 2012).
(36) Beck, B.; Horn, A.; Carpenter, J. E.; Clark, T. W. Enhanced 3D-databases: A fully electrostatic database of AM1-optimized structures. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1214−1217.