

Encoding and Decoding Graphical Chemical Structures as Two-Dimensional (PDF417) Barcodes

M. Karthikeyan*

Information Division, National Chemical Laboratory, Pune - 411 008, India

Andreas Bender

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Received August 4, 2004

A wide range of molecular representations exist today, ranging from human-readable structural diagrams over line notations such as Wiswesser Line Notation (WLN) and SMILES to several dozen computer-readable file formats. Still, to encode molecular structures in a computer-readable way for inputting structures in computer systems those formats are not the method of choice since they are not easily and faultlessly readable via optical recognition. In the present study a two-dimensional (PDF417) barcode representation of molecular structures in SMILES format is explored that enables the user to read and input molecular structures into computer systems in a fully automated fashion. A Lempel-Ziv-Welch (LZW) based compressed version of SMILES is suggested for cases where the size of the structure exceeds the storage capacity of PDF417 barcodes. Alternatively, the compact ACS format may be employed as a structural representation. The input via barcodes is fast, practically error free due to the 2D barcodes used which employ error correction and fully automatic. A Web application interface is developed which is able to interpret these barcodes and export them as optimized 3D chemical structures. Applications of this representation range from keeping automated storage systems to Web-based tracking systems of molecular samples. The National Chemical Laboratory, Pune, employs 2D barcode encoded structures for in-house repository management, where barcodes can also be used for querying the database for similar or substructures of the query structure.

INTRODUCTION

The input of chemical structures into computer systems is a ubiquitous task in situations where compounds and compound mixtures are handled, in the chemical and pharmaceutical industries as well as in related ones. A wide range of molecular representations exist today, ranging from human-readable structural diagrams over line notations such as Wiswesser Line Notation (WLN) and SMILES to several dozen computer-readable file formats. All of these representations have their merit in particular areas, such as human readability, line-based storage systems, and storage of associated data in computer file formats. Still, to encode a molecular structure in a computer-readable way (which also includes some degree of error correction) that enables the user to enter a structural formula in an automated fashion, those formats are not the method of choice. In the present study a two-dimensional (PDF417) barcode representation of molecular structures in the SMILES format is explored that enables the user to read and input molecular structures in a fully automated fashion. Barcodes are chosen due to their ability to deal with transmission errors and their track record in nearly every area of information encoding. The rationale behind the use of barcodes is outlined in the following paragraphs.

In the chemical as well as the pharmaceutical industry, tracking of samples or materials which move across the organization play an important role, for example between the organic chemist who synthesizes the compound, the analytical chemist who determines identity and purity of it,

and the medicinal chemist who evaluates activity of the substance via biological assays. Finally, the sample has to be stored and its identity has to be noted on the surface of the container. Since the chemical structure is the 'international language of chemistry'^{1,2} the naming of a sample by its chemical structure seems to be the most intuitive choice. Still, there is a need for the development of automation tools for creating this information in a computer-readable format and thus in an automated fashion, which will accelerate structure based chemical inventory management. Error-free input of data, here molecular structures, is of utmost importance for this task.

A large amount of inventory information today is communicated and stored in digital format, which is generally prone to technical errors and thus prone to the loss of valuable information. In conventional (noncomputerized) inventory systems, handwritten data are delivered physically or electronically with human interference, thereby inadvertently introducing transmission errors in the original data as well. Both of these error-causing situations apply not only to inventory systems in general but also to molecular inventory systems in particular. To counter information loss, the utilization of error control codes^{3,4} is a well-known digital signal processing technique which protects digital data against errors that may occur during transmission and/or storage (up to a certain extent that is, depending on the particular method used). In inventory systems, one of the frequently employed data entry and tracking tags used is the barcode technology that is able to accommodate error detection and error correction to varying extents. Examples

* Corresponding author e-mail: karthi@ems.ncl.res.in.

are supermarkets billing systems, the tracking of valuable mail items, or library systems. Most consumables sold today contain some type of barcode for automatic product identification and pricing to enhance efficiency in both billing and inventory tracking. This methodology can also be directly applied to molecular structures, as presented in this article.

The first barcode alphabets were developed in the 1970s, starting with the Universal Product Code (UPC).⁶ Other commonly employed alphabets are Code 39, Interleaved 2 of 5, and Code 93 (for details see ref 5). Those early barcode standards have in common that they are arranged in a linear (one-dimensional) notation, so that the symbols are encoded and decoded along only one axis or direction. Data are stored in the form of "symbol characters" (the entirety of which forms the "symbolology" or alphabet) that are generally parallel arrangements of alternating strips of lower reflectivity ("bars") and strips of higher reflectivity ("spaces"). A unique pattern of bars and spaces corresponds to a particular symbol character and thus to a particular data value.

More recently, barcodes were introduced that represent data in a two-dimensional arrangement, most prominently the Portable Data File (PDF) 417 barcode.^{7,8} Thus, where one-dimensional codes are only able to store data in the range of several dozen bytes, the PDF417 barcode definition is able to store up to 1.1 kB of data in the same area of space. This allowed barcode applications in novel areas such as the storage of personal access data, travel related documentation, and other security applications.⁹ Today wireless tags using radio frequency (RF) are partly replacing existing paper based barcodes in inventory systems, however retaining to some extent former information encoding methods.

As digital data storage is becoming increasingly affordable it is now possible to encode more information in ever more compact data formats. Thus one has to make a balanced decision on selecting technologies considering cost of storage media such as plain paper, magnetic stripes, RF tags, etc. and corresponding decoding hardware required for their application as well as, on the other hand, employees' time savings by using those media on a day-to-day basis. In academic as well as commercial libraries barcodes are extensively used for inventory management. Barcoding is nowadays also finding its way into the chemical laboratory environment¹⁰ where it is used for sample and inventory tracking of frequently used chemicals based on internal catalog numbers or identifiers.

Thus, in this publication we propose an encoding methodology for chemical structures as commercial barcodes. This paper describes an integrated approach which takes advantage of both automation technologies such as commercial barcoding tools as well as existing molecular structure representation formats for the description of the structure in the first place. We demonstrate that two-dimensional PDF417 barcodes are one of the most practical methods for inputting molecular structures into computer systems in a fully automated and less error-prone fashion.

MATERIAL AND METHODS

1. General Barcodes. Barcode systems resemble human languages to a surprising extent, by employing start codons, stop codons, and a set of defined symbols to convey meaningful expressions or data. This is true even up to the level of error correction: in some—but not all—communi-

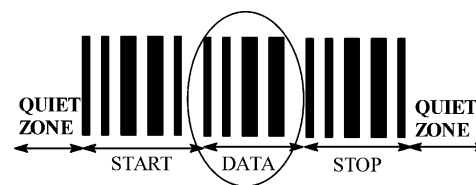


Figure 1. Typical linear barcode with data contents, consisting of a data part, which is enclosed by start and end zones and adjacent quiet zones.

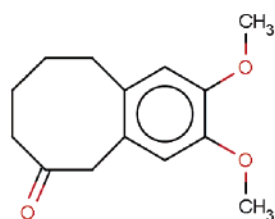


Figure 2. Sample 2D barcode in PDF417 format. Again, a data block is enclosed by start and end zones. Still, due to its two-dimensional nature this barcode format is able to store about 1 kB of data, as opposed to a few dozen bytes in the case of one-dimensional barcodes.

cations the original information content can be restored, for example due to the order of words the original meaning can be inferred, or at least the person who listens is able to realize that there is "something wrong" with the information he or she has just heard, which is the equivalent of error detection in computerized data transmission.

Linear barcodes usually contain a set of black bars in varying width separated by white spaces encoding alphanumeric characters (or purely numerical information in the case of more restricted symbolologies). Each barcoding specification contains both header data and footer data aligned with either side of the encoded data. These barcode elements are illustrated in Figure 1.

The earliest barcode symbology, called Universal Product Code (UPC),⁶ in one of its subspecifications (UPC-A) was able to encode 12 numerical-only values, thus accommodating 11 data positions plus one checksum number, equivalent to roughly 37 bits of data. Another symbology that is only able to encode numerical values is Interleaved 2 of 5, which is able to use both bars and spaces to encode information in a slightly denser format. More flexible by being able to encode alphanumeric values are, among others, the symbolologies Code 39 and Code 128, which are also able to encode information of variable length. A larger amount of encoded information is bought at the expense of longer barcodes, which still leads to practical limits such as the width of the barcode reader which determines a practical maximum barcode width. These linear barcodes are usually used on commercial product packagings. Some of the advanced linear barcodes also encode data verification code for safe and accurate decoding, such as the Code 39 symbology. The type and amount of data to be encoded varies depending on the selected barcode specification; usually about 10–15 characters including white space can be encoded on a linear barcode. Thus, one-dimensional barcodes are not able to cope with the storage demand of molecular structures.



SMILES: COc2cc1CCCCC(=O)Cc1cc2OC

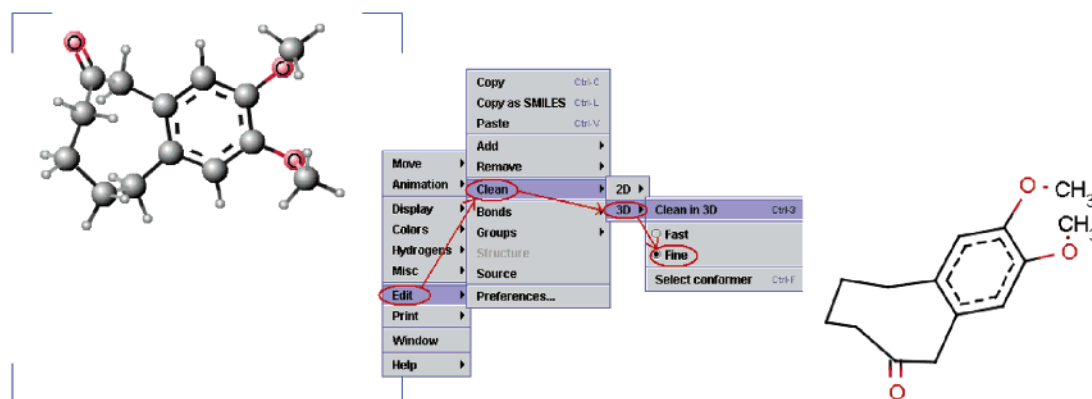


Figure 3. Upon decoding of the SMILES structure from barcode format, the three-dimensional structure of the molecule can automatically be retrieved via a molecular builder (here MarvinViewer).

To increase the density of stored data, “multirow” or “stacked” symbolologies have been developed, such as PDF417.^{7,8} Stacked symbolologies generally employ several adjacent rows of bars and spaces of variable width. Figure 2 shows a typical PDF417 symbol and its structural elements. Every PDF417 symbol contains a minimum of three to a maximum of 90 rows. Each symbol character, as the smallest unit containing information, is formed from four bars and four spaces. Each bar or space is of the width of one of up to six “modules” subject to the constraint that each symbol character has a total width of 17 modules. Each symbol character is able to represent a numerical value in the range of 0 to 928, and the number of encoded data symbol characters in the data region cannot exceed 928.

PDF417 symbology employs error correction values, which are generated using the Reed-Solomon error control code algorithm.³ (For details of the data format see again ref 5.) Here suffice it to say that various error correction levels exist, which, at the one extreme, do not allow for actual error correction but at least its detection and allow the storage of up to 1850 characters. At the other extreme, only 830 characters can be stored in one barcode, but up to 50% of the code may be destroyed and still all of the original information content can be recovered from the error correction algorithm. In either case, about 1 kB of storage space renders two-dimensional barcodes much more suitable for the storage of structural chemical information than their one-dimensional counterparts.

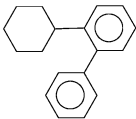


In the case presented here, the molecular structures are stored in the SMILES format. To accommodate more data,

especially large molecular structures where the input data size exceeds the data storage limit of the barcode, the Lempel-Ziv-Welch (LZW)^{11,12} compression was implemented. Alternatively a more compact molecular data representation than the SMILES format can be employed, the ACS format.^{13,14} Coding of the original SMILES format, LZW-compressed SMILES, and the ACS format are discussed in the following paragraphs.

2. Barcoding Chemical Structures as SMILES and Back-Conversion. In chemical informatics and computational chemistry various methods and standards have been developed for the representation of chemical structures in a computer readable format.^{1,2} Plain text file formats containing both atom coordinates and atom-to-atom connectivity are frequently used for data transfer between computer programs, such as the “mol” file format¹⁵ for single molecules and the “SDF” file format¹⁵ for multiple molecules. In recent times, the coordinate-free, single-line “Simplified Molecular Input Line Entry Systems” (SMILES) format developed by Daylight¹⁶ became increasingly popular. Additionally, since no atomic coordinates are stored, it is a much smaller molecular format than other formats such as SD format or XYZ format. In the most straightforward encoding scheme used, the SMILES string is encoded directly in PDF417 format. For the generation of SMILES strings in the first place several options are feasible.

Manual input of SMILES for even moderately complicated structures with more than about 20 atoms is a tedious task not only for novice users. Usually structure-drawing programs are used in this situation, such as programs such as

Table 1. Illustration of the Depict¹⁷ Program To Convert Barcoded Structural Information to Chemical Structures

Compound in Structural and SMILES representation	Barcode representations for SMILES and ACS format
 2-Cyclohexyl-biphenyl [A] <chem>C1=CC=CC=C1C2=CC=CC=C2C3CCCCC3</chem>	 2D barcode for [A]4c  2D barcode for [A]4d

Depict,¹⁷ JchemPaint,¹⁸ and CDK,¹⁹ which output SMILES strings from graphical chemical structures. In Table 1 an example of a chemical structure and the corresponding SMILES is given; in addition also the barcode representation is shown which was generated directly from the SMILES string.

In the opposite direction the barcode is first decoded to standard SMILES format. Going from the SMILES representation to a graphical molecular structure, standalone programs such as Jmol,¹⁹ MarvinView,²⁰ or ChemDraw²¹ accept SMILES strings for input and display clean 2D chemical structures from connection table data. (2D-) Molecular coordinates are generated dynamically through built in heuristic algorithms. If three-dimensional structures are required, programs such as Corina²² use both rules and/or molecular mechanics force fields to generate 3D coordinates from SMILES format. In our case we utilize the MarvinViewer applet²⁰ to generate 3D chemical structures directly from PDF417 barcodes. Hence encoding linear representations such as SMILES or an equivalent format such as barcodes comes in handy for automatic and efficient molecular input. In the example given in Figure 3 the 3D structure generated directly from SMILES using Marvin is shown. The MarvinViewer also adds implicit hydrogens for the given SMILES string.

The Marvin toolkit²⁰ was chosen for structural representation since it neatly integrates a structural database, JChemBase, and the actual viewer, Marvin. Querying JChemBase using information provided by barcodes has proven to be quick and reliable in practice. JChemBase at the same time implements structural searching capabilities by employing an underlying Oracle database. This includes (sub-)structural searching as well as similarity searching, while the definition of descriptors for similarity searching can be chosen from a variety of graph-based and pharmacophore-based approaches.

In addition to recovering the original structures, SMILES based barcodes can also be used for inputting small chemical structures or substructures (fragments) as query structures into large chemical structure databases. Adopting conventional identification number-based barcoding systems would in this case require a database lookup for the fragment that corresponds to the ID, a step that is not necessary if the structure is encoded explicitly.

While the encoding of SMILES strings as PDF417 barcodes is feasible for most "small molecules", barcoding of chemical structures using plain text file formats such as the MOL format, which contains molecular coordinates and connection table data, are not feasible due to data storage constraints. But also a succinct linear representation in

SMILES format will eventually exceed even the capacity of the two-dimensional barcode format, in particular if higher error correction levels are used, since the length of a SMILES representation of a structure is roughly proportional to the molecular size (number of atoms plus branching, bond types and stereochemistry). One way to solve this problem is to compress the SMILES using Lempel-Ziv-Welch (LZW) compression.^{11,12}

3. LZW—Compressed SMILES. Lempel-Ziv-Welch (LZW) compression^{11,12} is a well-known lossless data reduction technique which is today widely used for example in text compression. LZW replaces parts of strings with tokens which refer to dictionary entries. The dictionary is initialized with 256 entries, one for each possible ASCII character. Successively, both the dictionary entries themselves as well as, if new strings are encountered, the number of dictionary entries is increased until all of the original text is replaced by integer numbers. These integers can, with the help of the dictionary, be back-translated into the original text without data loss. Data reduction varies widely, depending on the repetition of dictionary entries (for example words or phrases). Typically about 50% compression can be achieved on long English texts but up to about 80% on highly repetitive sequences.

To decode barcodes from LZW compressed SMILES the conversion program first decompresses the data from an ASCII array into original SMILES, before interpreting them in MarvinViewer.

As an alternative to LZW compression, a more compact format for encoding chemical structures can be used in the first place, which is the ACS format.^{13,14} This format can be seen as another alternative, space-saving alternative to SMILES in cases where the barcode-encoded SMILES exceeds storage capacity of the PDF417 barcode format.

4. Compact ACS Format for Barcoding Applications. To reduce the amount of data that has to be encoded on the barcode, a template based chemical structure encoding method was developed, the ACS file format.^{13,14} This method is based on the Computer Generated Automatic Chemical Structure Database (CG-ACS-DB)¹⁴ originally developed to create a virtual library of molecules through enumeration from a selected set of scaffolds and functional groups. Scaffolds and groups are stored in Automatic Chemical Structure (ACS) format as a plain text file. In this ACS format most commonly used chemical substructures are represented as templates (scaffolds or functional groups) through reduced graph algorithm along with their interconnectivity rather than atom-by-atom connectivity information.



Figure 4. Sample vials with attached barcodes. For size comparison, an AA type battery is shown on the right-hand side of the picture.

Table 2. Customized 2D Barcodes (PDF417) for Chemical Structures Generated by ICBC^a

Number of Columns	5	3	2	1
2D BARCODE				
PDF417 Format				

^a The number of columns of the barcodes can be customized to suit the technical requirements of the user.

This method of representing substructures or molecular fragments irrespective of their molecular complexity by means of a unique identifier for each substructure or fragment causes compactness of the molecular structure, thus enabling us to encode structures of considerable size in barcode format. This approach to compression is analogous to other text based compression methods such as the LZW algorithm, but it is tailored to specific chemical requirements (definition of fragments similar to functional groups and the importance of intrafragment connectivity). On the other hand, not the whole molecular information (or rather its SMILES representation) is now encoded in the barcode, so a back-translation from ACS format into SMILES format is necessary.

An illustration of the ACS data format, ACS barcode format, and SMILES string is shown in Table 1, together with the corresponding PDF417 barcode representations. The full definition of the ACS format can be found in refs 13 and 14.

5. Hardware Used. Linear barcodes generated in Code39 and Code128 format from ACS format were generated by the Internet Compatible Barcoding (ICBC, see Supporting Information) Programs and tested and optimized using the Welch Allyn SCANTEAM 3400 CCD Long-Range barcode scanner. 2D barcodes in PDF417 format were tested and optimized using the Welch Allyn 4410 image scanner.

RESULTS AND DISCUSSION

1. Conversion of SMILES to Barcode Format. Conversion of SMILES format to barcode format is straightforward as it employs only the standard PDF417 encoding algorithm.

Barcodes from ACS format can be directly generated as linear (Code 39 or Code 128 format) or two-dimensional barcodes (PDF417 format) using the Internet Compatible Bar Coding (ICBC) program developed at the National Chemical Laboratory, Pune. The barcode output format of the program can be customized, as shown in Table 2. (This program is accessible via a Web interface at <http://www.moltable.com/barcode/barcode2str.jsp>).

The encoded SMILES based barcodes will be extracted to give the original SMILES again (or the ACS representation of the structure, if compression was used in the first place). This structure can in turn be used as an input structure to generate molecular descriptors for computational studies or can be used to retrieve relevant information from public Internet resources or from other available commercial databases. If SMILES/ACS data is retrieved, then it is interpreted as a single molecule with a valid connection table. 2D atomic coordinates are automatically generated for visualization.

At the National Chemical Laboratory, Pune, a database of molecular structures in combination with barcodes for inventory tracking of molecular samples is employed. Incoming samples are, after removal of duplicates, submitted to an Oracle database. After scanning the barcode, the database entry is accessed, which enables the user to perform exact searching, substructural searching, or similarity searching of the database. Sample vials with attached barcodes are shown in Figure 4. For comparison, an AA type battery is shown on the right-hand side of the picture.

DB_ID: 9999
 Molecular Weight: 517.57798
 Molecular Formula: C₃₁H₂₇N₅O₃

Submitter: Dr M Karthikeyan
 e-mail: m.karthikeyan@ncl.res.in
 Institute: NCL-Pune
 Institute Code: ncl-pune
 Weight of Sample: 10 mg
 Date of Submission: 01-01-2005

Figure 5. Querying the database via barcode input. The scanned PDF417 barcode is read, which queries the database for associated information about the sample in storage as well as displays the molecular structure. Arbitrary metadata can be stored in the database such as purity, supplier, or bioactivity data.

232142 MF: C₃₈H₄₂N₄O₂S₄ MWT: 1067.01664

229670 MF: C₅₄H₉₁N₇O₁₄ MWT: 1062.33912

231743 MF: C₅₆H₈₇N₇O₁₄ MWT: 1082.32876

229251 MF: C₄₉H₆₁F₃N₁₆O₁₀ MWT: 1091.1056896

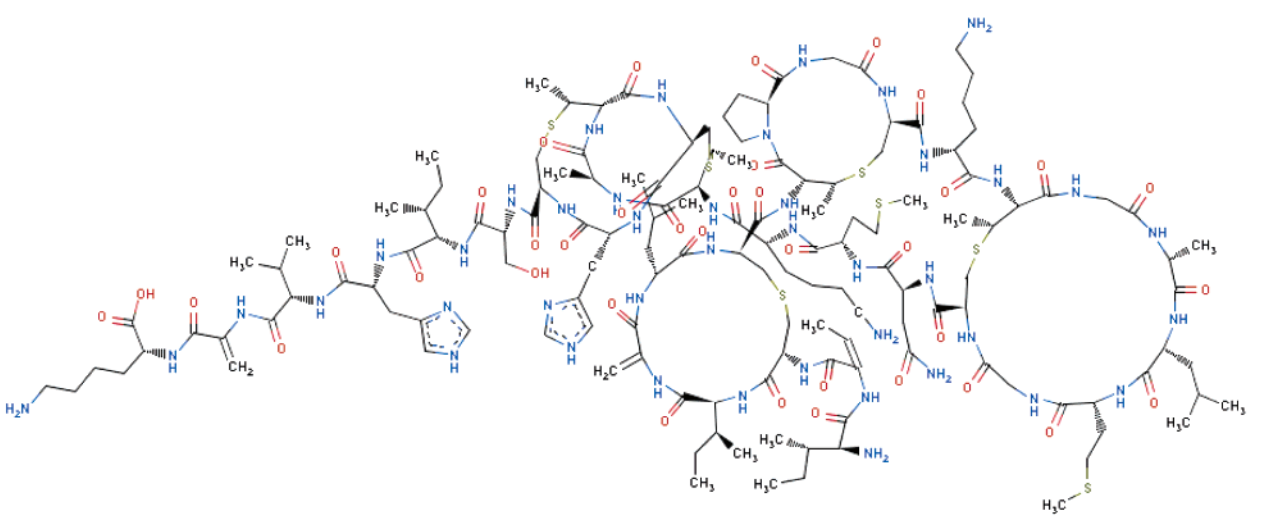
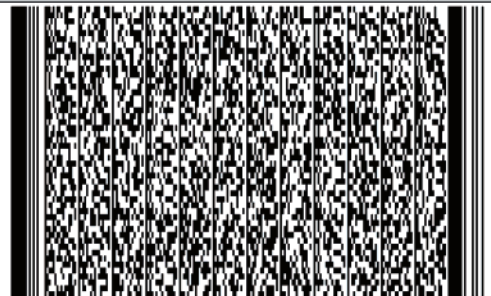
Figure 6. Browsing the database in query mode, here searching for all structures of molecular weight between 1050 and 1100 atomic units. The output format can be user-defined and also include customized additional information.

Screenshots of the Web-interface are shown in Figures 5 and 6. In Figure 5 the database is queried using PDF417 barcode input. While the structure is directly encoded in the barcode, the associated identifier is used to query the database for additional information which is customizable and might also include fields such as purity, supplier, or bioactivity information. In Figure 6 the database is used in manual query

mode, retrieving in this case all structures from the database whose molecular weight is between 1050 and 1100 atomic units.

2. LZW—Compressed SMILES. A case study showing the effect of LZW packing on size of the resulting SMILES is shown in Table 3. The oligopeptide which requires 582 alphanumeric characters in SMILES format is reduced to

Table 3. LZW-compressed SMILES Format Case Study^a

Chemical structure

SMILES (582 characters)
<chem>CC[C@H](C)[C@H](N)C(=O)NC(=C/C)C(=O)N[C@H]1CSC[C@@H](NC(=O)[C@@H](CC(C)C)NC(=O)C(=C)NC(=O)[C@@H](NC1=O)[C@@H](C)CC)C(=O)N[C@H]2[C@@H](C)SC[C@@H](NC(=O)CNC(=O)[C@@H]3CCCN3C2=O)C(=O)N[C@H](CCCCN)C(=O)N[C@H]4[C@@H](C)SC[C@@H](NC(=O)CNC(=O)[C@@H](CCSC)NC(=O)[C@@H](CC(C)C)NC(=O)[C@@H](C)NC(=O)CNC4=O)C(=O)N[C@H](CC(N)=O)C(=O)N[C@H](CCSC)C(=O)N[C@H](CCCCN)C(=O)N[C@H]5[C@H](C)SC[C@@H]6NC(=O)[C@H](NC(=O)[C@H](C)NC5=O)[C@@H](C)SC[C@@H](NC(=O)[C@@H](Cc7c[nH]cn7)NC6=O)C(=O)N[C@H](CO)C(=O)N[C@H]([C@H](C)CC)C(=O)N[C@H](Cc8c[nH]cn8)C(=O)N[C@H](C(C)C)C(=O)NC(=C)C(=O)N[C@H](CCCCN)C(=O)O</chem>
SMILES Barcode (PDF417)

LZW Compressed SMILES (263 ASCII characters equivalent)
4 48 67 5 176 67 4 0 72 5 208 49 16 1 2 16 64 93 3 32 79 16 144 93 3 48 40 16 129 3 16 81 17 2 145 9 16 161 17 16 145 20 4 48 41 4 241 29 3 49 22 16 49 24 17 193 26 5 208 40 4 240 67 2 128 61 17 224 67 3 209 1 18 33 27 18 17 24 17 225 37 17 81 41 18 177 50 19 17 53 19 65 19 18 97 28 19 113 60 18 49 57 18 97 65 19 17 0 18 225 45 4 48 49 19 33 10 3 32 67 2 225 18 16 160 52 21 17 5 3 81 13 16 48 93 3 97 17 18 241 66 17 113 27 19 81 28 17 224 79 3 97 76 19 161 61 18 113 64 2 145 73 21 225 62 20 49 39 22 225 63 18 161 69 22 209 59 23 1 106 23 33 50 22 209 111 23 81 96 22 161 73 20 144 52 20 48 53 20 241 84 5 208 55 24 80 56 21 112 64 5 208 57 18 113 96 23 209 65 20 33 0 3 145 56 23 209 59 22 145 107 19 1 116 22 113 115 19 97 120 25 97 156 23 97 105 16 225 21 23 177 161 23 209 71 26 144 55 20 48 56 4 48
Compression Study (Statistics)
Original SMILES: 582 characters; Compressed SMILES: 263 characters Compressed Data : 45.34 % of actual SMILES, Compression Ratio1:2.21




^a LZW compression is able to reduce the storage space to less than 50% of the original data which renders PDF417 2D barcodes able to encode structures of considerable size (about four times the size of the structure given in this example).

263 characters using the template-based packing algorithm. This illustrates the size of molecules which can be encoded on a single 2D PDF417 barcode. Note that the maximum molecular size which can be encoded is about 4-fold the size of the oligopeptide shown in Table 3, since about 1kB of information can be stored in PDF417 format. This is equivalent to a maximum storage of roughly 2000 ASCII characters or several hundred atoms in SMILES format

(where multiple bond orders and branched structures require additional storage space).

3. Illustration of the ACS Format. An example of the original CG-ACS-DB and the barcode-compatible ACS format are given in Figure 7 for 2-cyclohexylbiphenyl. Only connectivity between fragments is stored, thus giving a very compact molecular representation, encoding the whole structure in the string `_4#11.1.1.1.11.1.2.6` if the barcode-

Table 4. Sample ACS Formats and Corresponding 2D Barcodes (PDF417)^a

Data Type	2D Barcode (PDF417)
ACS Barcode Format e.g., <code>_4#11.1.1.1.11.1.2.6</code>	
ACS Data Format (CG-ACS-DB) e.g., <code>1 4 0 0 0 0 2 11 1 1 1 3 3 11 1 2 6 4</code>	
ACS to SMILES e.g., <code>C1(C2(C3CCCCC3)=CC=CC=C2)=CC=CC=C1</code>	

^a The ACS format requires less space than the SMILES format if it is converted to barcode format, which is reflected in the different size of the barcodes.

Table 5. Illustration of the ACS Format and the Associated Space Savings^a

SMILES representation	ACS representation
C1CC1	_1#
C1CCC1	_2#
C1CCCC1	_3#
C1CCCCC1	_4#
N	_7#
N(CCC)	_9#
C1CCOCC1	_10#
...	
C1(C2CCOCC2)CC1	_1#10.3.1.1
C1(C2CCCC2)CCCC1	_3#4.1.1.1
C1(C2CCOCC2)CCCCC1	_4#10.3.1.1
C1(C2(C3CCOCC3)CCOCC2)CC1	_1#10.3.1.1.10.1.2.1

^a In particular in case of low connectivity higher space savings are achieved. At the top of the table single-letter equivalents of frequently encountered molecular fragments are shown. At the bottom of the table examples of structures assembled from multiple fragments are given.

compatible ACS format is employed. In this case study, 15 templates (as shown in Figure 8) are numerically encoded in the program for virtual library enumeration along with replaceable atom information. If required the ACS format encoded with individual templates can be interpreted like SMILES as standard atom-bond data if interaction with other programs is required. SMILES generation from ACS format for a template T1(T2(T3-T3)T2)T1 would for example give rise to the SMILES string C1(C2(C3CCCCC3)=CC=CC=C2)=CC=CC=C1. As illustrated in Figure 7 and Table 4 the bar code compatible ACS format is used for deriving barcodes, eliminating redundant information from the original CG-ACS-DB database format, thus reducing its size considerably.

Generally the compression rate of the ACS format, compared to the original SMILES strings, depends on the connectivity of the fragments to be encoded. For example, the SMILES structure C1(C2CCCCC2)CCOC1 (the ether

CG-ACS-DB: Enumerated data format for 2-Cyclohexyl-biphenyl

`L1,T1,0,0,0,0 L2,T2,T2_ca1,L?,T?_ca?,d1, L3,T3,T3_ca1,L?,T?_ca?,d?...
dn,Ln,Tn,Tn_ca1,L?,T?_ca?`

e.g., `1 4 0 0 0 0 2 11 1 1 1 3 3 11 1 2 6 4`

ACS: Barcode Compatible Data Format

`_T1#T1.Tca1.L1.Lca2.T3.Tca3.L2.Lca4..... T[n].Tca[n].Ln.Lca[n]`

e.g., `_4#11.1.1.1.1.11.1.2.6`

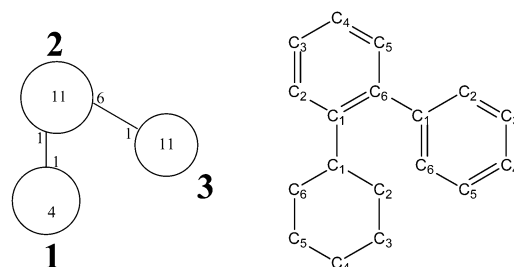


Figure 7. Illustration of the ACS barcode encoding. Instead of atom information, template based packing is performed, which stores the type of fragments and the connectivity information between them.

of bicyclohexane) would in ACS format be represented as `_4#4.1.1.1` where 1 = C1CCCCC1. Thus, in the case of low connectivity very high compression rates are achieved. In case of higher connectivity, the compression rate decreases slightly. In case of higher connectivity, such as tetracyclohexane—SMILES representation C1CCCCC1C2CC(C3CCCCC4)CCC3)CCC2—the ACS representation would be `_4#4.1.1.1.4.1.2.3.4.1.3.3`, representing the type and connectivity of the molecular fragments. Still, the ACS format, depending on the particular template alphabet used, is generally smaller than the corresponding SMILES code. This brevity is bought at the expense that back-translation

Table 6. Illustration of Linear Formats Used To Represent Chemical Structures

linear notation	sample data
WLN	L66J BMR& DSWQ 1N1&1
ROSDAL	1=-5=-10=5,10-1,1-11N-12=17=12,3-18S-19,18=20O,18=21O,8-22N-23,22-24
ICIS	T1 × 1 y1 × 2 y2 s d T2 × 1 y1 × 2 y2 s d T3 × 1 y1 × 2 y2 s d
IPX	1 0 0 0 2 1 1 1 0 3 2 1 2 0
CCML	A[0,0]A[1,a1-a1]B[1,a2-a1]C[2,a4-a1]D[4,a3-a1] or AABCD (default)

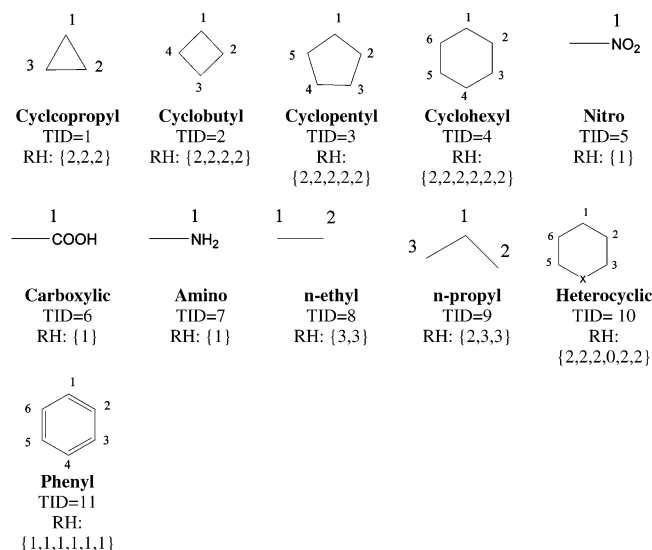


Figure 8. List of numerically encoding templates used for virtual library generation.

of the fragment code is necessary. Here also some examples of the ACS format are given, shown in Table 5. For the smallest set of fragments, single alphabet symbols are used. More complex structures whose parts are identical to previously derived template structures are represented by a set of fragment template structures.

Several other commonly employed line notations are for comparison shown in Table 6.

CONCLUSIONS

Inputting molecular structures into computer systems for applications such as inventory tracking of molecular samples is of major importance for the chemical and pharmaceutical industries today. Here we describe the methodology to represent chemical structures in a 2D (PDF417) barcode format.

This format facilitates the automated, error-free, and fast input of molecular structures into computer systems. The barcode format is easily back-translatable into SMILES format. In case of large structures which require more space than is available on PDF417 barcodes the template-based ACS format can be used, thus being able to encode molecular structures of up to the size of small macromolecules in a single barcode. Alternatively, the SMILES string can be LZW-compressed which reduces the data to be encoded by a factor of roughly two. This facilitates the storage of small macromolecules up to the size of several hundred atoms (roughly 2000 characters of SMILES ASCII code) in barcode format.

Given the ever-increasing capacity of storage media, it will soon be possible to add additional computed and experimentally measured molecular properties to the information stored on the barcode. Today a database back-end is used which is able to retrieve additional sample data such as purity, supplier, and bioactivity. The database fields as well as the output format are fully user-customizable.

ACKNOWLEDGMENT

M.K. thanks NISSAT (DSIR), New Delhi, India for financial support (No. NI/SI/035/2001), the Department of Science and Technology for a BOYSCAST fellowship, and Prof. Alex Tropsha (University of North Carolina at Chapel Hill, U.S.A.) for his able guidance. A.B. thanks the Gates Cambridge Trust and Unilever for funding and Robert C. Glen for his friendly and helpful support.

Supporting Information Available: Java source code for LZW compression of SMILES or equivalent formats. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) *Chemical Structure Systems*; Ash, J. E., Warr, W. A., Willett, P., Eds.; Ellis Horwood: Chichester, 1991.
- (2) Jurs, P. C. *Computer Software Applications in Chemistry*; Wiley: New York, 1996.
- (3) Lin, S.; Costello, D. J. Error control coding. Fundamentals and applications. In *Prentice Hall Computer Applications in Electrical Engineering Series*; Prentice Hall: NJ, 1983.
- (4) MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, U.K., 2003.
- (5) Palmer, R. C. *The Bar Code Book: Comprehensive Guide to Reading, Printing, Specifying, and Applying Bar Code and Other Machine-Readable Symbols*; Helmers Publishing: New Hampshire, 2001.
- (6) Savir, D. and Laurer, G. J. The characteristics and decodability of the Universal Product Code symbol. *IBM Syst. J.* **1975**, *1*, 16–34.
- (7) AIM 1994. Uniform Symbolology Specification: PDF417; 1994.
- (8) Symbol Technologies, Inc., Holtville, New York, U.S.A.
- (9) Noore, A.; Tungala, N.; Houck, M. M. Embedding biometric identifiers in 2D barcodes for improved security. *Computers Security* **2004**, *23*, 679–686.
- (10) Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Denisenko, S. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1281–1282.
- (11) Ziv, J.; Lempel, A. A Universal Algorithm for Sequential Data Compression. *IEEE Trans. Inform. Theory* **1977**, *23*, 337–343.
- (12) Welch, T. A. A Technique for High Performance Data Compression. *IEEE Comput. Mag.* **1984**, *17*, 8–19.
- (13) Karthikeyan, M.; Krishnan, S.; Steinbeck, C. Text based chemical information locator from Internet (CIL) using commercial barcodes. 223rd American Chemical Society Meeting – Orlando Florida, U.S.A., March 2002.
- (14) Karthikeyan, M.; Uzagare, D.; Krishnan, S. Compressed Chemical Markup Language for compact storage and inventory applications, 225th ACS Meeting - New Orleans, March 23–27, 2003.
- (15) Molecular Design Ltd., San Leandro, U.S.A., <http://www.mdli.com>.
- (16) DAYLIGHT, DAYLIGHT Inc., Mission Viejo, California, U.S.A. <http://www.daylight.com>.
- (17) DAYLIGHT, DAYLIGHT Inc., Mission Viejo, California, USA. <http://www.daylight.com/daycgi/depict>.
- (18) Steinbeck, C.; Krause, S.; Willighagen, E. JChemPaint – Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. *Molecules* **2000**, *5*, 93–98.
- (19) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493–500.
- (20) Csizmadia, F. JChem: Java Applets and Modules Supporting Chemical Database Handling from Web Browsers. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 323–324. <http://www.chemaxon.com>.
- (21) ChemDraw, CambridgeSoft, Cambridge, MA.
- (22) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic 3-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

CI049758I