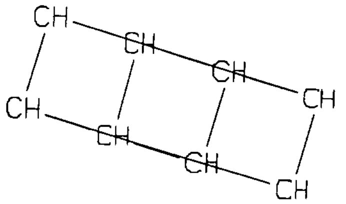


Figure 8. Depiction of cubane, illustrating poor coordinate selection for a condensed multicyclic system.

Yet, there are a few types of structures for which DEPICT does not provide adequate graphics. This is the case for condensed multicyclic systems which cannot be dealt with easily without drawing curved lines. An example of this pitfall is cubane (Figure 8). The cubane structure is drawn correctly but the attempt to make "as many perfect polygons as possible" forces two bonds to fall directly on top of others, leading to a confusing picture. A general solution to such problems is currently under investigation.

SUMMARY

DEPICT is a computer algorithm that produces a graphical display of any chemical structure for which the linear notation language SMILES can be generated. Atomic coordinates are computed by evaluating angles between atoms while treating the structure as a tree with fixed length edges. The graphic representation is then derived from the coordinates.

The most important aspects of DEPICT are illustrated in the four examples in Figure 7. These include chain positioning, ring system representation, aromaticity indication, and display of atomic properties such as charge.

No explicit graphical information is required as input to DEPICT. Since most connection table formats and other linear notations can be converted to SMILES, DEPICT can be used to display structures stored in such formats. Furthermore, DEPICT is ideal for display of structures which are not stored in a database, e.g., novel structures generated by a computer.

ACKNOWLEDGMENT

I thank Steven Burns and Arthur Weininger for assistance in programming and Joseph L. Weininger for editorial assistance.

REFERENCES AND NOTES

(1) Goodson, A. L. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 212.
(2) Kalbfleisch, W.; Ohnacker, G. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 176.
(3) Shelley, C. A. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 61.
(4) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(5) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.
(6) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1972; Chapter 6 (Partitions).
(7) Balducci, R.; Pearlman, R. F. Novel Algorithms for the Rapid Perception of the Unique, Optimal Set of Rings. Unpublished results.

Computerized Retrieval of Information on Biosynthesis and Metabolic Pathways[1]

SANDOR BARCZA,* LAWRENCE A. KELLY, and CHRISTOPHER D. LENZ
Sandoz Research Institute, East Hanover, New Jersey 07936

Received February 22, 1990

Biosynthetic metabolic pathways were analyzed, and a hierarchy of attributes was constructed. Representation of the knowledge base on metabolic conversions was effected in terms of this hierarchy of attributes and the chemical structures of molecules participating in metabolic conversion steps. A prototype database was constructed with the MACCS and DATACCS programs, already in use for storage, searching, and reporting of chemical structures, chemical-biological data, and chemical reactions at Sandoz.[2,3] Key data in the new metabolic conversions database are the enzyme name and classification, effectors, inhibitors, literature reference, etc. Participating molecules, if known and under 255 heavy atoms, are stored and diagrammed as stereostructures. The crucial data on metabolic conversions are represented by "From" and "To" datatypes. All the data are exact match and range searchable for text and numbers. Thus, precursors and progenitors of compounds can be found. Structures are match and substructure searchable. This tool is a useful and very flexible complement to metabolic charts. It in itself can be used to report and graph conversion steps and sequences.

INTRODUCTION

Living organisms are probably the most complex entities of the universe. Man attempts to describe, document, and understand organisms for several reasons: academic knowledge for its own sake; understanding, so that the organism can be influenced, controlled, repaired—e.g., hybridization of corn and curing of diseases; and to transfer the ingenuity found in nature to products of man, e.g., preceptrons, sensory systems, robotics, *biomimetic* organic *syntheses*, etc.

The *description* of organisms occurs at several levels and with different armamentaria: ecosystems, anatomy, physiology, biochemistry, biophysics, quantum biology, etc. An outstandingly important level of description of constituents and processes of living organisms is that of the chemical *transformation of molecules in the body*. These interconversions make up the *metabolic pathways*, *biosynthetic pathways*, replication of genetic material, etc. The main pathways have been classified into broad (and in some cases fuzzy) sets of catabolism—involving degradation and energy release—and anabolism—involving energy enrichment and construction.

The pathways of biosynthesis and metabolism form a highly complex information system. In order to make this information computer storable and retrievable, the issues of *representation* of knowledge must be addressed. The description should be readily understandable or at least serve some utilitarian purposes. It should be preservable on paper and readable, transformable, writable, and usable by computer. As a practical pedagogic matter, it should be easily taught, which requires graphical representations as well.

The System. A knowledge base of a biological system is typically too complex for easy comprehension and memory. Further, application of the knowledge often becomes quite

0095-2338/90/1630-0243$02.50/0 © 1990 American Chemical Society

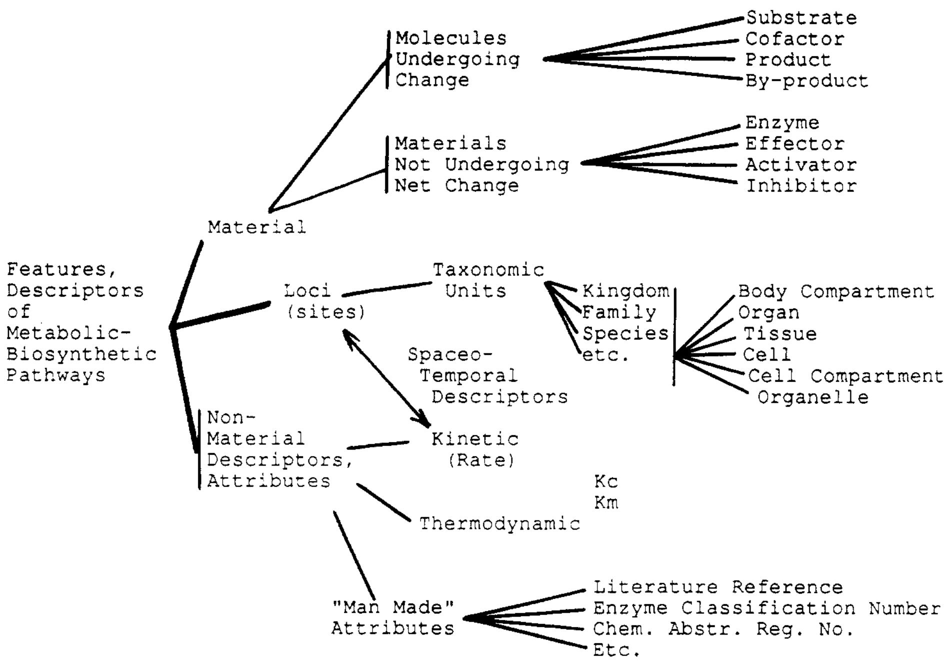


Figure 1. Organization of the attributes of biosynthetic–metabolic pathways.

specific, involving a small well-chosen subset of the total description. For these reasons, the description should be specifically *searchable*, by certain criteria along certain parameters, etc. Viewing a knowledge base as a multidimensional array of data, one should be able to isolate slices, planes, space angles, lines, etc. from this space. All these theoretical and practical considerations argue forcefully for a computerized *database* (DB) management system, or what constitutes the computerized knowledge base of an expert system for the representation of metabolic pathways.

Features and Requirements. At this point we should increase our resolution of analysis of *metabolic pathways* and emphasize the *essential* specifics of the problem. The system consists of entities that are *molecules*, which in bulk make compounds (hopefully clearly defined single compounds rather than unknowns, or with a distributed composition, such as man-made polymers). The compounds get chemically *interconverted*. Each conversion can be thought of as a *linkage* and cross reference between two compounds. Sometimes, and with some arbitrariness or convention, these linkages are assigned a direction, i.e., become vectorial. There can be more than two molecules recognized in a metabolic transformation, and in fact, the most typical case is $A + B \rightarrow C + D$. These can be called substrate, cofactor, product, byproduct, etc. and are molecules with *do undergo chemical change*. Another class of particpants consists of molecules that *do not undergo any net chemical change* in the metabolic conversion. These may be broadly classified as catalysts or enzymes and effectors, inhibitors, etc.

Furthermore, a complete description of the metabolic conversion of molecules demands additional items that are not material entities at all, rather, they are *attributes*, such as rate or equilibrium constant. Further concepts or auxiliary pieces of information are man-made, primarily the literature reference.

Thus, the main *descriptors* of metabolic conversion are the chemicals that do change, the chemicals that do not change, spaceo-temporal descriptors, nonmaterial attributes, property descriptors, etc. For an organized overview these are grouped in Figure 1. The exact names, definitions, and distinctions are subject to discussion.

The ultimate aim is to *describe* all *metabolic conversions* that occur in an organism. To the extent that any random or highly transient aspects exist, this seems an impossible task at present. Another factor is genetic diversity. Given enough scrutiny (high enough resolution), all individuals differ, including their metabolic pathways. This is further complicated by acquired traits, differing exposure to xenobiotic chemicals, enzyme induction, etc. Nevertheless, knowledge advances and is gainfully applied starting from the simpler (oversimplified), progressing to the more complex, more complete.

The most important requirement is to create a system which can represent metabolic pathways in a useful way and is *extendable*, so that it will accept future data *when* they become available.

Nature of the Information. Metabolic pathways of an organism, even at low resolution, form a colossal *network*. In mathematical terms, it can be regarded as a graph[20] where the nodes (points, crossings) are compounds and the edges (connecting lines) are conversions. Even if one ignores cofactors, enzymes, byproducts, inhibitors, etc., this network is highly complex and multidimensional, and almost certainly nonplanar in the topological sense.

Desired Representation. Although great strides have been made in three-dimensional computer stereographics, for the foreseeable future, paper (i.e., two dimensions) will remain a practical vehicle for communicating knowledge. We hope to use 3-D tools in the future for displaying networks. It is a great challenge to *represent complex networks* effectively on paper. By effectively it is meant the following: if objects (molecules) are connected by lines (pathways), there should be a minimum of line crossings, confusion, and crowding, and the length of lines should be minimized, or express some toplogical distance, conceptual distance, or dissimilarity or other useful feature (e.g., reciprocal rate). Multidimensional scaling is one technique of charting points in a lower, e.g., 2-D, space with minimum injustice to their (multidimensional distance) relationships. Conceivably, some choice of attributes and multidimensional scaling could "squeeze" the network of molecules onto paper with minimum injustice. However, line crossings cannot be avoided in a nonplanar graph, no matter how the nodes are rearranged. (A graph may rapidly become nonplanar as nodes acquire high connectivity.)

In the face of all these difficulties, the *metabolic charts* in use are truly outstanding efforts and accomplishments. Ingenious use of color, symbol conventions, breaking up of the overall network at strategic connections to avoid line crossings, and repeated display of those substrates that are frequently involved make thse charts (especially the Boehringer-Mannheim[4]) very useful tools. No matter how successful are the DBs created for metabolic pathways, there will be, as long as paper is used, an important complementary role for metabolic biosynthetic charts. They may be increasingly prepared with computer assistance, and ultimately they may be prepared from and *used together with* the very *DBs* we are discussing here.

Analogous Problems. There exist *analogues to metabolic networks* (and it is worth pondering their properties and how they are treated):

- The network of organic (mostly synthetic) chemicals, petroleum and coal derived, etc.
- Chemicals (chemical interconversions) in a chemical plant with many products.
- Movement, flow of chemicals (or other goods, for that matter) (rate of interconversion ≅ rate of movement or volume).
- Flow of goods in the economy of a country.
- Flow of money in the economy or the world.
- Network of activities to accomplish something (e.g., land a man on the moon or develop a drug to market).
- Network of pipelines to pump fluids.
- Traffic of a metropolis.
- Relationships among subsets of sciences or concepts (e.g., as determined by co-citation frequencies).
- Relationship of concepts in a human mind.
- Neuronal networks.
- Electronic circuit networks.

Metabolic networks are defined as the appropriate joining of metabolic *steps*. The definition of a step has some arbitrariness. We *define a step* as the conversion of one isolable chemical, whose existence is well established, to another, ignoring transition states and fleeting intermediates. No doubt, additional stable intermediates will be discovered in the future.

Alternatives and Reasons for Nonadoption (at This Time). Some alternate possibilities for representing metabolic pathways are as follow:

(1) A graph in which the nodes are the *reactions* and the compounds they interconvert are the edges. In the DB version, the substrate, etc. would be listed as attributes under each reaction as main entry.

This is a less favored possibility, because of lack of analogy with metabolic charts, other concepts of reference (e.g., chemical factory), and lower manageability, in general. It may be useful for some purposes in the future.

(2) A system mainly built around the *enzymes.* Precedents exist,[17] but without chemical structure searchability. To the extent that one enzyme only performs one reaction, this is equivalent to 1 above.

There are variants which place more emphasis on the material aspects of enzymes:

(2a) Register *enzymes as compounds*, and all else as data. This is declined because (1) the small molecules that the enzymes operate on are much better known as compounds and are much more easily handled by well-established structure retrieval tools; (2) the structures of many enzymes are nebulous; and (3) enzymes are often simply named after the substrate or product molecule.

(2b) *Hierarchical classification* according to the chemistry of the catalyzed reaction, e.g., ligases, esterases, kinases, redox enzymes etc.

This may be a useful adjunct to the DB described here and may be added as an additional tier later. It is in fact implicit in the enzyme class numbers, which we include.

Table I

	fixed DTs	flexible DTs
general features	max 1 line max 20 char	max 300 lines per cpd max 120 char per line
numeric	max 2 real nos.	max 2 nos. + comment to 120 char
formatted	max 10 fields	max 10 fields + comment to 120 char
text	max 20 char (any printing char)	text to max 120 char per line

(2c) Registration of enzymes as special compounds *in addition* to the low molecular weight compounds. With current capability this may be done as an additional tier, by name, and in the future by structure of sequence, as the structure of more enzymes will be known and as macromolecular retrieval capability grows.

A good, flexible information retrieval system allows retrieval of the information in many ways, to the point where alternate representations become much less of an issue. The important step is to get the data into computer readable form in a *versatile DB.* Then *transformations* of the data are possible into new arrangements. A case in point is storing the enzyme classification numbers as data in our DB allows retrieval by group or individual EC numbers.

MATERIALS AND METHODS

Software. The DB described herein was created using the MACCS (Molecular Access System)[5,6] DB management system and the ancillary structure-data transfer and report-generating program DATACCS (Data Access System, recently merged into MACCS)[7] for the following reasons:

(1) This software is chemical compound oriented, and compounds are the best defined entities of known metabolic pathways. It has superb molecular structure storage and search capabilities, including stereochemical distinctions.

(2) The same software is already familiar and in use at Sandoz Research Institute and at about 100 other organizations for chemical structure and data storage,[2,3] and uniquely at Sandoz Research Institute for the storage and searching of chemical reactions.[2]

(3) The programs, as well as the hardware, can therefore be shared.

Properties of the Software. A DB can be created by the user of the MACCS program. Such a DB is a conceptual matrix of compounds (cpds) × datatypes (DTs). Unique DTs are the registry numbers (RNs) and the coded (stereo)structures. The latter are searchable for match and by substructures and various formula [range] searches.[8]

The rest of the DTs involve numeric and alphanumeric data, and the types and their features are tabulated in Table I.

Features and Capacities of DTs. Data from the compact, fixed DT's are automatically displayed with the structure upon retrieving a compound and are very rapidly searchable.

New compounds and data can be added at any time and are immediately viewable and searchable by those authorized. Data are searchable for exact match, or match of a substring of characters or for ranges of numbers or characters. Searches can also be done on designated columns (zones) in the lines of data. "AND" and "OR" combinations of searches are both feasible. Posing several search queries for several zones on the same line automatically performs an AND search. Posing them consecutively and merging hitlists constitutes an OR search.

Input to the DB occurs by drawing structures with a light pen or mouse.[8] Data are input from the keyboard. Structures and data can also be imported from files and other DBs, e.g., an in-house database of commercial chemicals or proprietary compounds, using the DATACCS program[7] (now M-II module of MACCS). The latter is a structure and data-transfer and -reporting program which can move data between keyboard input, files, tables, and DBs, and to the screen or to hardcopy.

No additional hardware or software acquisition was necessary to create and use the metabolic pathways DB. We wanted and are using "off the shelf" software. At the time this work was undertaken, only MACCS was available to the authors. Even today, there is still no tool available that is ideally suited to metabolic biosynthetic pathways. A program specifically designed to store reactions, such as REACCS[9] would be more suited. Although we plan to create a REACCS and/or ChemBase version of our database, even REACCS is unable at this time to (1) sufficiently and easily distinguish the *roles* that substances play, e.g., substrate, cofactor, effector, inhibitor, product, byproduct, etc. and (2) suitably store and allow searching over multiple steps. Technology to handle and search multistep reactions, reaction sequences, and networks only now seems to be emerging.[19]

Hardware. Prime 550, now Vax 8530 computer, Versatec V80, Calcomp 1012 and Hewlett-Packard color pen plotter, Apple Laser Writer, Imlac II high-resolution graphic terminal (most beneficial for structure drawing, report from design, and viewing), Retrographics VT-640 and 650 terminals with light pen, and IBM-PC and compatible Compaq 286, 386, and NEC-APC microcomputers with mouse. The time-sharing and device connections are mostly via local area networks.

Sources of Data. Metabolic conversions and structures. The biochemical pathways DB was mainly configured according to the chart assembled by Gerhard Michal for Boehringer Mannheim GmbH.[4] Where there were questions concerning the complete stereochemistry, the *Atlas of Stereochemistry* by Klyne and Buckingham was used as a reference.[10] Recently a book on enzyme inhibitors appeared.[18]

Fixed DTs and Flexible Text DTs. The CAS NUMBER DT is the Chemical Abstracts Registry Number, which is available from the registry directly or in journals such as *Biochemistry* published by the American Chemical Society.

The Aldrich number was obtained from the current version of the Aldrich Chemical Co.'s *Fine Chemicals Catalog*.[11]

The ECLASNO DT which contains the enzyme classification number generated by the International Union of Biochemistry was obtained from the 1984 version.[12] The literature references provided by this source were also entered directly into the DT REF as one of the references to a step.

Information was in general checked with sources such as *Biochemistry* by Lehninger[13] or *Review of Physiological Chemistry* by Harper et al.[14] The text DT allows sufficient room for substantial additional documentation.

RESULTS AND DISCUSSION

Configuring the DB. The prototype DB was configured by applying the experience gained earlier with the creation and use of the main chemical–biological in-house DB utilizing the MACCS program. For this purpose, it was valuable to have configured and filled many biological DTs[3] and DTs to accomplish storage of chemical reactions.[2]

The DTs in the bioconversion DB were set up on the basis of the analysis of attributes of biosynthetic-metabolic steps. An index is shown in Table II.

Fixed DTs. Configured at the time of DB creation, two of these 2 × 20 characters were devoted to express the From information, i.e., the RNs or abbreviated names of those compounds *from* which the compound displayed/retrieved is formed in any metabolic step.

Provision was also made via two additional DTs to express the To information, i.e., the RNs (names) of compounds *to* which the current compound is converted in any metabolic step.

Table II. Index of Data Types (DT)[a]

	FIXED DATATYPES
1	EXTREG [@@@-000000IIIII]
2	CAS.NUMBER [000000-00-0]
3	ALDRICH.NUMBER [%%%%%%%]
4	FROM.A
5	FROM.B
6	TO.A
7	TO.B
	FLEXIBLE DATATYPES
8	SYNPATH -SUBSTR-COFAC,ECLASNO,ENZY,EFF+BYPRO+
9	REF LITERATURE REF
10	FULL.MOLNAME SYSTEMATIC NAME, ALSO SYNONYMS
11	SUBSTRATE SUBSTRATE, STARTING COMPOUND, REGNO, NAME
12	COFACTOR COFACTOR, OTHER PARTNER, REGNO, NAME
13	SPECIES ORGANISM (TYPE) (FAMILY), SPECIES, ETC, OR GENERAL
14	SITE BODY COMPARTMENT, PHASE, ORGAN, TISSUE, CELL, ORGANELLE
15	ENZYME ENZYME ABBREV, NAME
16	ECLASNO ENZYME CLASSIFICATION NO.INT.UNION BIOCHEM,eg,EC 4.1.3.16
17	EFFECTOR EFFECTOR, METAL ION, ETC.
18	BYPROD BY-PRODUCT, SECOND PRODUCT, PARTNER OF PRODUCT, REGNO, NAME
19	RATE RATE CONSTANT OR OTHER RATE DESCRIPTOR
20	INHIB INHIBITOR, [FEEDBACK, ALLOSTERIC], REPRESSOR
21	ACTIVATOR ACTIVATOR (DRUG OR OTHER), BY REGNO OR NAME
22	PATH MAJOR PATH MEMBERSHIP, E.G. KREBS CYCLE
23	COMMENTS NOTES, ETC. NOT ACCOMMODATED ELSEWHERE, SPECIAL CIRCUMST.

[a] Further explanation of contents and definitions occurs in the text.

Although the information on metabolic conversions is specified and recoverable from the From data above, it is convenient and efficient to store and display a compound's offsprings as well as precursors.

Other fixed DTs that were configured are more of the "man-made" type: Chemical Abstracts RN, properly formatted, a chemical supplier (typically Aldrich) catalog number, and an "external" RN with letter(s) and digits. The latter is potentially useful to number the compounds in lettered groups, independently of the automatically generated internal RN of the substances.

Some fixed field data are exemplified below:

```
MACCS: Find Name = CHOLESTEROL
** SEARCHING FILE
** REFERENCE LIST: 1
40: CHOLESTEROL
Formula: C27 H46 O
Cas.number: 000057-88-5
Aldrich.number: C07520-9
From.a: 039 OR 038
From.b: 039
```

Flexible DTs. Provision is made via text DTs to store literature reference(s) for each compound, and the full molecule name or synonyms, should the 80-character "molname" that is associated and retrieved with the structure not suffice.

The crucial biosynthetic–metabolic information is stored in the following DTs:

Substrate: This is the main precursor of the current compound, mentioned by internal RN and/or name.

Cofactor: An auxiliary participant in the metabolic conversion step, often supplying the energy needed, mentioned by RN and/or name, e.g., ATP.

Byproduct: Internal RN and/or name of the molecule, if any, that accompanies the formation of the current compound in a metabolic conversion step. Sometimes formed from the cofactor supplying energy for the step, e.g., AMP.

Enzyme: Enzyme needed for forming the current product in the metabolic conversion step, by full or abbreviated name, preferably IUBC name.[12]

Enzyme Classification Number: The hierarchical number of the enzyme in the classification and nomenclature of enzymes as recommended (1984 version) by the International Union of Biochemistry.[12]

Site: The body compartment, phase, organ, tissue, cell, organelle, etc. in which the transformation and enzyme operate, e.g., mitochondria.

Species: The species or family, etc. in which this product by this step is forming, e.g., mammals.

Effector: Allosteric effectors.

Activator: Agents which activate the enzyme by covalent bonding or otherwise (but not allosteric effectors).

Inhibitor: RN and/or name of agent which inhibits the step to form the product, repressor, allosteric or feedback inhibitor.

Rate: Quantitative information about the step forming the product; Michaelis–Menten constant, K_m; turnover number; or other quantitative rate, equilibrium, etc. information.

Path: Membership of the step, to form the product in question, in a group of steps forming together a certain metabolic pathway, e.g., cholesterol biosynthesis, Krebs cycle, glycolysis, etc., identified by naming the pathway.

Literature Reference: Most relevant reference(s) where information was taken from about the formation of the product shown.

Additionally, a DT was set up to accommodate a very compact organization of the most essential information about step(s) leading to a product: One line of "zoned"[3] text per path leading to the product. In specified columns, it contains the registry identifying numbers of the substrate, cofactor, the enzyme class number, enzyme name, effector, and byproduct.

This compact arrangement is useful for summary reporting, reporting in a small form, and searching since fields (zones) within the same line of data within a DT can be conveniently AND searched even in case of multiline entries.[2,3]

Filling the DB. The DB is open-ended with respect to:

(1) adding compounds

(2) adding data

(3) creating more datatypes

It is planned to be continually enlarged.

The entries as of this writing number approximately 400 structures and their (inter)conversions and other data filled into about 4160 (DT) × (compound) combinations. Together with the structures, the DB occupies about 2250 Vax blocks.

Typical additions occurred by:

(1) Choosing the next most important multistep pathway to add (e.g., cholesterol biosynthesis).

(2) Registering the structures (products) along this pathway, preferably in the order of biosynthetic sequence.

(3) Adding the interconversion data.

(4) Adding other data. Some vacancies or "dummies" were deliberately left between major pathways or branches.

Drawing and registration of structures requires about 1–3 min, depending on complexity. A (bio)chemist's skill is desirable to avoid mistakes in (stereo)structures. Input/registration of data takes only negligibly more time than typing it. The time to build a new path is proportional to the number of steps.

Inserting an intermediate discovered later into a path is like the normal addition of more compounds and data to the DB, except that the To and From information for the immediate precursor(s) and progenitor(s) has to be updated. A minor inconvenience is that the newcomer is at the end of the DB.

Linking already known (registered) cpds by newly discovered enzymes/conversions is simply done by adding data (enzyme, etc.), including the new To and From pointers.

Pathways and groups of compounds which have been accommodated thus far in the DB:

Structures of functional groups in the pyruvate dehydrogenase complex
Cholesterol biosynthesis
Embden–Meyerhoff pathway (glucose → lactate)
Citric acid cycle
Fatty acid synthesis
Phosphatides
Sidepaths of the Embden–Meyerhoff path
Mesaconate pathway and glutamate group of amino acids
Histidine metabolism
Purine biosynthesis and purines
Urea cycle
Glutamate decarboxylation pathway
Pyrimidines

Multiple paths leading up to and away from a "central" compound are registered in the order (e.g., 7 = central compound):

$$\begin{array}{l} 1 \rightarrow 2 \rightarrow 3 \rightarrow 7 \rightarrow 8 \\ \quad 4 \rightarrow 5 \rightarrow 7 \rightarrow 9 \rightarrow 10 \\ \qquad 6 \rightarrow 7 \rightarrow 11 \rightarrow 12 \rightarrow 13 \end{array}$$

Self-Documenting DB. One compound was dedicated for which the data entered under each DT is the definition of that DT. Another compound serves to illustrate typical data, since no "real" compound has all data available. This built-in documentation is accessible by MACCS in the same way as real data, thus, where and when the user can best use them.[2]

Searching the DB. The main types of searches are chemical structure-related and data searches. It is a particular strength of this DB/program that powerful structure/substructure searches can be carried out in combination with data searches.

Structure Match. After drawing a structure with light pen or mouse, or otherwise calling a structure, e.g., from a file, "Find Current" will find the structural matches from the DB. The matching structure will appear with name and other essential (Fixed field) information. The latter may be optionally turned off.

Isomers to a structure may be located, and the program tells which hit is a diastereomer, an enantiomer, a racemate, or an identical match.

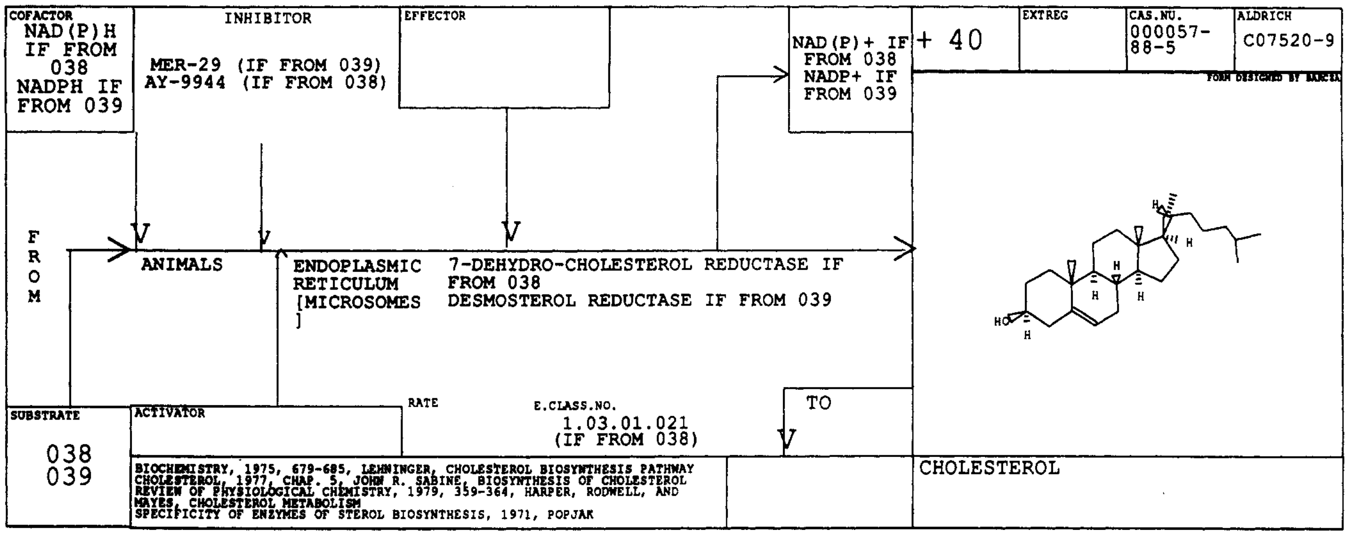
Formulas can be searched exactly or by ranges of numbers of the elements, for all or some of the elements.

Substructure searches are the most powerful structure-related searches. The query can be a substructure drawn or a called-up drawing template or file or a previously hit structure or its fragment. One may draw multivalued bonds or (hetero) elements (atom lists), e.g., "Any", into the query. Then the substructure search finds all products whose structure embeds the query.

Names searches find the product of matching name or all names which embed a truncated name (truncation on right, left, or both).

Display and manipulation of search results. In "Search mode" the names and RNs of the products hit are shown. This may be optionally turned off, in which case only the total count is shown. The RNs of the hits are placed in a "Current" list, from which they may be written into a (permanent) listfile for reuse. The structures in the Current list may be viewed, further searched, or plotted with captions in "Plot mode" or with selected data in flexible ways by the DATACCS (M-II) module.

Examples of structure, formula, and name-related searches are included in Table III. Substructure search with the template THF (tetrahydrofuran) hits all the furanoses. Similarly, "Phos", the phosphate template hits all phosphates, including di- and triphosphates. Name fragment search for @meva@ (@=truncation "wildcard" for names) hits the mevalonate-related products.



COFACTOR
NAD(P)H IF FROM 038
NADPH IF FROM 039

INHIBITOR
MER-29 (IF FROM 039)
AY-9944 (IF FROM 038)

EFFECTOR

NAD(P)+ IF FROM 038
NADP+ IF FROM 039

+ 40

EXTREG

CAS.NU.
000057-88-5

ALDRICH
C07520-9

FORM DESIGNED BY BARCZA

FROM

ANIMALS

ENDOPLASMIC RETICULUM [MICROSOMES]

7-DEHYDRO-CHOLESTEROL REDUCTASE IF FROM 038
DESMOSTEROL REDUCTASE IF FROM 039

SUBSTRATE
038
039

ACTIVATOR

RATE

E.CLASS.NO.
1.03.01.021
(IF FROM 038)

TO

BIOCHEMISTRY, 1975, 679-685, LEHNINGER, CHOLESTEROL BIOSYNTHESIS PATHWAY
CHOLESTEROL, 1977, CHAP. 5, JOHN R. SABINE, BIOSYNTHESIS OF CHOLESTEROL
REVIEW OF PHYSIOLOGICAL CHEMISTRY, 1979, 359-364, HARPER, RODWELL, AND MAYES, CHOLESTEROL METABOLISM
SPECIFICITY OF ENZYMES OF STEROL BIOSYNTHESIS, 1971, POPJAK

CHOLESTEROL

Figure 2. Compact report form showing structure of the product with arrows from substrate(s), inhibitor(s), etc. and branching to byproduct. The species, site, and enzyme(s) are listed under the main arrow. Literature reference (bottom) and other "man-made" data are also shown. This form embodies the main organizational unit of the DB, namely a *step to form a product*. Forms can be plotted stacked to show sequence of steps.

Data searches can also be performed in "Data mode" where the corresponding data are displayed as soon as hit during the search.

Data can be range searched by supplying a range of strings or numbers. Automatic truncation is implied with search strings. This may be overruled.

Many *examples of text string searches* are shown in Table III, in order to illustrate the contents of the DB at this stage and to show the manner of searching.

For example, searching DT "Enzyme" for string "KIN" hits 68 products that are formed by enzymes whose names include "kinase". Searching for author "Bloch" in DT Reference gave the desired five references. String "choles" searched in DT Path gave the 21 compounds along the cholesterol biosynthesis pathway. The *range search* of enzyme classification numbers (Eclasno) 1.01.01.001–1.99.99.999 yielded the 87 products formed by oxidoreductases.

Since a search for a string occurs over an entire line of data or only in a certain range of columns, if so specified (column search), the last example could also be performed by searching for string "1" in column 1.

Searches on this prototype DB of ~400 structures and ~4K filled datafields (many multiline) take negligible time (≲5 s, Vax), whether (sub)structure or data searches.

Boolean logic combinations of searches can be accomplished in several ways:

(1) As is shown in Table III, searching for effector "Mg" produced 112 hits. Doing the search for cofactor "ATP" only on this "Active list" retained 49. Repeating the process for byproduct "AMP" retained 3 hits on the intersection of the three sets.

(2) Searching for several strings with column range designations (column search) will imply AND logic (intersection) within the same line; e.g., searching for a combination of a substrate, cofactor, enzyme classification number, and effector within the compact DT Synpath only hits products that are formed in a step involving all of those partners.

The two methods can also be used to produce OR logic, if the searches are done consecutively on the same original reference list and the resultant listfiles are combined (merged).

(3) Completely general Boolean logic combinations are available by combining list files with the respective operators, nested to any degree. For example, searching DT enzyme for the string "reduc" hit 35 products, presumably formed by reductases. The hitlist for enzyme classification numbers beginning with 1 (oxidoreductases) had 87 products (Eclas1). Eclas1 but not reduc = 56. Reduc but not Eclas1 = 4. Both (intersection) = 31. Exclusive OR = 60. Either (union) = 91. Spot checks revealed that Eclas1 hits also included (correctly) dehydrogenases.

Table III. Summary of Searches[a]

DATATYPE	QUERY OR STRING	HITS
STRUCTURE	TEMPL>PHOS	132
"	TEMPL>THF	60
FORMULA	S(1-5)	1
PATH	CHOLES	21
COFACTOR	ATP	66
BYPROD	AMP	8
EFFECTOR	Mg	112
>COFACTOR	ATP	>49
>>BYPROD	AMP	>>3
COFACTOR	ATP	66
>EFFECTOR	Mg	>49
>>BYPROD	AMP	>>3
COFACTOR	CoA	5
ENZYME	MEV	2
MOLNAME	@MEV@	3
ENZYME	REDUC	35
ECLASNO	1.01.01.001 TO 1.99.99.999	87
ENZYME	OXID	18
"	PHOS	106
"	TRANSF	86
"	KIN	68
"	CHOLIN	>5
"	ESTER	4
ACTIVATOR	AMI	5
"	GLUTAM	2
INHIB	CHLOR	1
REF	POPJAK	5
"	COREY	2
"	BLOCH	5

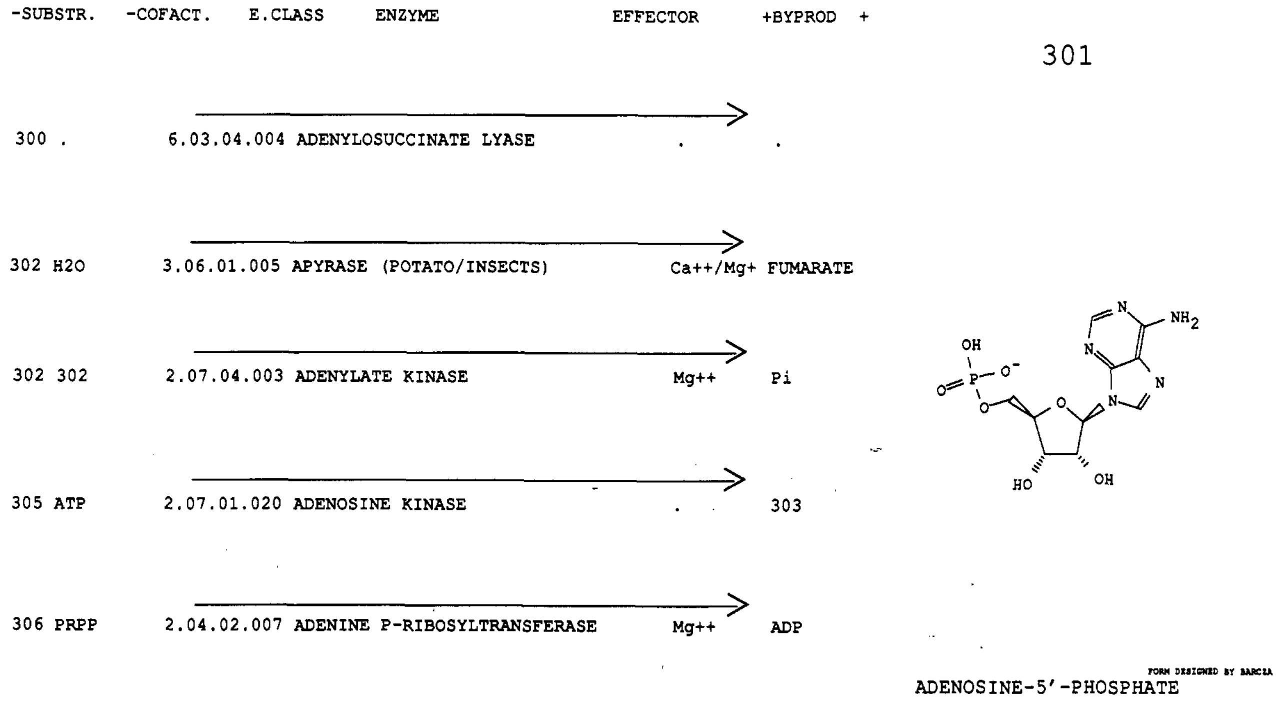
(> indicates nested search (intersection)).

[a] For each search, the type of information searched, the query or search string used, and the number of hits obtained are shown. Some of the searches were performed on a reference list less than the total DB. Those include the symbol >.

The most *typical search strategies* for combined searches include narrowing down the hitlist in Search mode by successive searches, and then either inspecting the data in Data mode or producing DATACCS (M-II) reports containing both structures and (selected) data.

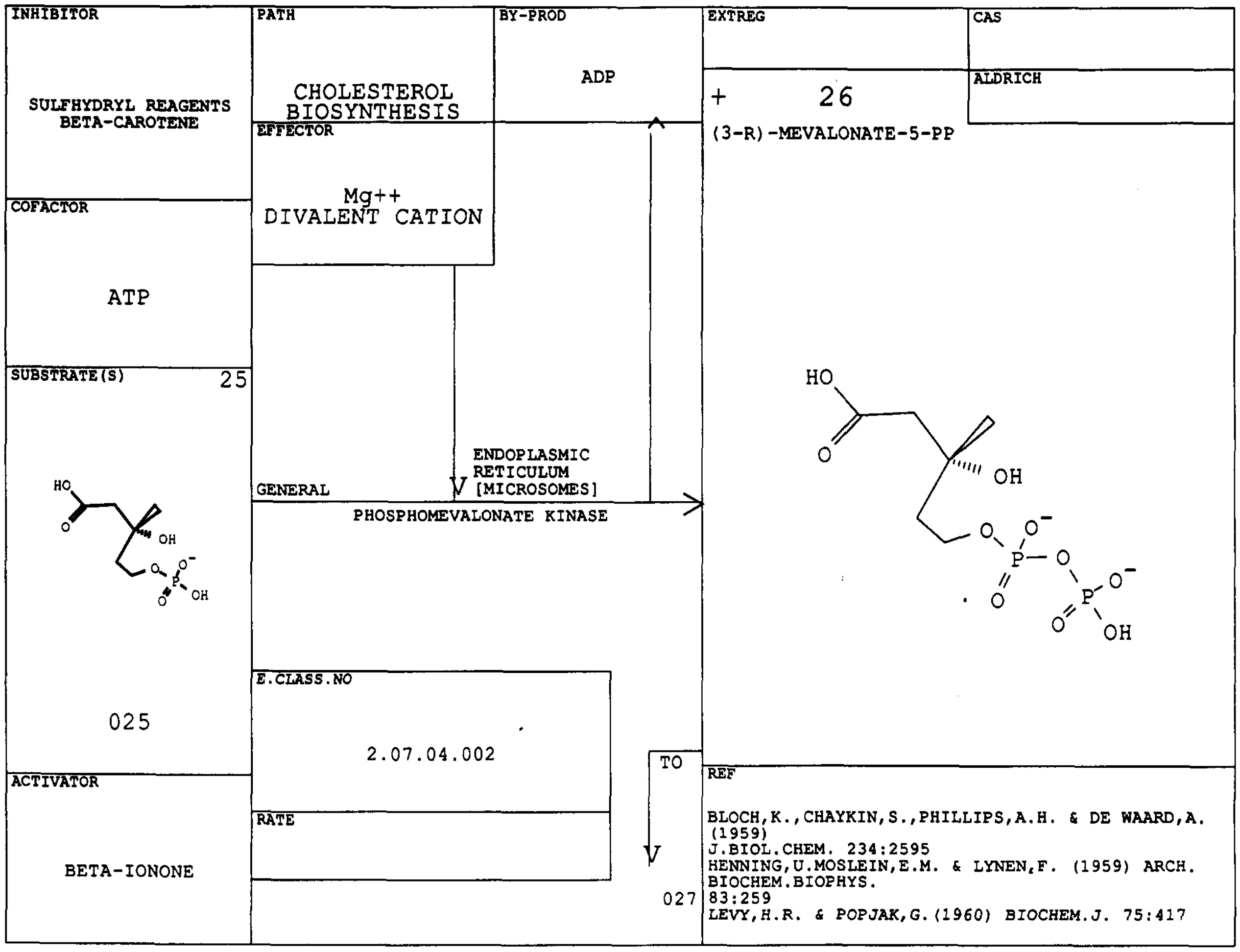
Parent and offspring searches are a special capability of the biosynthesis–metabolic pathways DB. To find out what substrates lead to the formation of a particular product, inspect DT Substrate or the compact DT Synpath or the even more compact DTs From.A and From.B together. The structures and data may be inspected after finding the RNs of the parents listed in these DTs.

There are also several ways of determining what products are formed from a particular substrate: (1) Search DT



```
-SUBSTR.   -COFACT.   E.CLASS      ENZYME                    EFFECTOR      +BYPROD  +

                                                                                          301

300 .        6.03.04.004 ADENYLOSUCCINATE LYASE                  .       .

302 H2O      3.06.01.005 APYRASE (POTATO/INSECTS)           Ca++/Mg+ FUMARATE

302 302      2.07.04.003 ADENYLATE KINASE                     Mg++     Pi

305 ATP      2.07.01.020 ADENOSINE KINASE                       .      303

306 PRPP     2.04.02.007 ADENINE P-RIBOSYLTRANSFERASE         Mg++     ADP

                                                                        FORM DESIGNED BY BARCZA
                                                                  ADENOSINE-5'-PHOSPHATE
```

Figure 3. Display of multiple pathways leading to a product. Each line of data and arrow corresponds to one path. The lines are taken by DATACCS from the compact DT Synpath, in which designated zones of columns show the RN of the substrate, number or name of cofactor, enzyme class number, enzyme, effector, and byproduct. The data in any or all of these zones are coherently AND searchable. Coherence means that matching strings *within* but not over different lines will produce a hit. (The original reports on the Calcomp or HP plotter are in color. The multiple pathways form has alternating red and blue lines.)



INHIBITOR
SULFHYDRYL REAGENTS
BETA-CAROTENE

PATH
CHOLESTEROL
BIOSYNTHESIS

BY-PROD
ADP

EXTREG

CAS

ALDRICH

+ 26
(3-R)-MEVALONATE-5-PP

EFFECTOR
Mg++
DIVALENT CATION

COFACTOR
ATP

SUBSTRATE(S) 25

025

GENERAL
ENDOPLASMIC
RETICULUM
[MICROSOMES]

PHOSPHOMEVALONATE KINASE

E.CLASS.NO
2.07.04.002

RATE

ACTIVATOR
BETA-IONONE

TO
027

REF
BLOCH,K.,CHAYKIN,S.,PHILLIPS,A.H. & DE WAARD,A. (1959)
J.BIOL.CHEM. 234:2595
HENNING,U.MOSLEIN,E.M. & LYNEN,F. (1959) ARCH. BIOCHEM.BIOPHYS.
83:259
LEVY,H.R. & POPJAK,G.(1960) BIOCHEM.J. 75:417

Figure 4. Report on biosynthesis of a *product* showing also the *structure* of the most important (first-listed) *substrate*. The main arrow points from substrate to product, branching to byproduct, and is annotated with species, site, and enzyme names(s). Other data for agents acting on this step are shown elsewhere. The bottom also has an arrow pointing to the *next* product(s), near the literature reference.

Substrate for a particular RN, (2) Search the columns of the compact DT Synpath where the substrate numbers are given, (3) Search both DTs From.A and From.B for the given substrate number. In each of the above, the searches hit those products that are formed from the given substrate. (4) Alternatively, for the molecule given as substrate, inspect the DTs To.A and To.B. This shows directly the RNs of the products formed from the given compound as substrate. For example, for RN 30 = squalene, DT To.A contains 31, which is the number for oxidosqualene.

There is presently no simple way to search over multiple steps or to "jump" over (unknown) intermediates. However, all the members along a path can be called up by searching DT Path with the appropriate string, e.g., purine. Since multiple Path membership has been provided for, overlaps and "crisscrosses" are properly handled.

As the examples show, many searches can be done in several ways. This is partly a result of our experimentation in this prototype DB, and partly it is intentionally built-in flexibility.

REPORTING

There are many, to some extent complementary, ways of reporting structures and data on graphic and nongraphic terminals, plotters, and printers. "Find molname:cholesterol," for example, displays the structure with fixed data, e.g., RNs of precursor(s) and progenitor(s), if any (From and To data), as shown earlier.

Molecule names and numbers are shown as they are hit in a search in Search mode. Searching or "Find"-ing data in Data mode displays the data. Data can be transferred for a list of compounds hit in a search from the DB into a standalone file, which can be printed. Captioned structures can be plotted in Plot mode.

Customized, *combined structure + data reporting* forms were designed for this DB with DATACCS. Figure 2 shows a compact form reporting most pertinent data with a structure. It is convenient to stack these forms corresponding to the biosynthetic sequence (path).

In order to show multiple pathways to a compound, the form in Figure 3 was designed with multiple arrows, under each arrow a different line of compact data being displayed, showing substrate number, cofactor, enzyme and class number, effector, and byproduct—one line per path to the compound.

A command procedure was written on the Prime computer, which, after a simple command by the user, plotted forms for the list of compounds given on which not only the structure of the compound in question (product) but the structure of the most important, first-listed precursor (substrate) is also shown, along with enzyme, cofactor, and other data (Figure 4). This is the closest to which MACCS can be brought to act as an automatic reaction (transformation) display system, provided the From (substrate) data are configured as done here.

Other report forms were also configured, and easy modifications can be made during sessions at graphic terminals to suit many purposes.

DT documentation/definitions are available by the same means, since they are stored as data under a special, reserved "compound".

FUTURE PLANS AND RAMIFICATIONS

A reaction retrieval system[7] offers additional possibilities and conveniences for storage and retrieval of metabolic conversions; our plans include transforming the DB to REACCS and/or ChemBase format.

The need to input more compounds, paths, inhibitors, rates, etc. is obvious and is only limited by manpower, and to some extent by availability and acceptance of (quantitative) data. We seek collaboration with designers of metabolism charts, DB vendors, and other researchers in the field.[18]

Important but challenging extensions would be to combine the DB with an expert system for metabolite prediction[15,16] and to calculate fluxes[20] and concentrations of compounds, with and without perturbations, e.g., by drugs. The latter is especially difficult because of complex mathematics for open systems and scarcity of data.

Irrespective of ambitious long-term goals, we have shown the potential of the DB, as configured, as a convenient storage and retrieval tool for a variety of applications, e.g., what agents catalyze or inhibit a given (type of) metabolic conversion, representation of metabolic disorders as inhibition of steps, following up on the possible consequences of drugs affecting metabolism, etc. A key and superior feature of this DB is the combined sub- or stereostructure-data searchability.

ACKNOWLEDGMENT

The support of the software and hardware by Y.-m. Tseng, M. H. Myers, and H. L. Lukas and the constructive reviewing of the manuscript by L. D. Boyajy and R. V. Coombs are greatly appreciated.

REFERENCES AND NOTES

(1) Presented in part at the 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 1986, paper CINF 24; and at the 76th Annual Meeting of the American Society of Biological Chemists, Washington, DC, June 1986, Abstract 1603; Fed. Proc., *Fed. Am. Soc. Exp. Biol.* **1986**, *45*, 1756.

(2) Barcza, S.; Mah, H. W.; Myers, M. H.; Wahrman, S. S. Integrated Chemical-Biological-Spectroscopy-Inventory-Reactions Preclinical Database. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 198.

(3) Barcza, S.; Kelly, L. A.; Wahrman, S. S.; Kirschenbaum, R. E. Structured Biological Data in the Molecular Access System. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 55.

(4) Michal, G. *Biochemical Pathways*; Universitaets Druckerei: Wurtzburg, West Germany, 1982.

(5) MACCS, DATACCS, REACCS, and ChemBase are programs and trademarks of Molecular Design Ltd., Inc., 2132 Farallon Dr., San Leandro, CA 94577.

(6) Wipke, W. T.; Dill, J. D.; Peacock, S.; Hounshell, W. D.; Marson, S. Search and Retrieval Using an Automated Molecular Access System. Presented at the 182nd National Meeting of the American Chemical Society, New York, Aug 1981.

(7) Dill, J. D. DATACCS—An Interface from MACCS to Other Software Systems. *Abstracts of Papers*, 187th National Meeting of the American Chemical Society, St. Louis, MO, April 1984; American Chemical Society: Washington, DC, 1984; CINF 46.

(8) Anderson, S. Graphical Representation of Molecules and Substructure—Search Queries in MACCS. *J. Mol. Graphics* **1984**, *2*, 83.

(9) French, S. E. REACCS Applied to a Corporate Database. *Abstracts of Papers*, 187th National Meeting of the American Chemical Society, St. Louis, MO, April 1984; American Chemical Society: Washington, DC, 1984; CINF 45.

(10) Klyne, W.; Buckingham, J. *Atlas of Stereochemistry*; Oxford University Press: New York, 1974.

(11) *Aldrich Fine Chemicals*; Aldrich Chemical Co., Inc.: Milwaukee, WI.

(12) Webb, Edwin C., Ed. *Enzyme Nomenclature*; Academic Press, Inc.: Orlando, FL, 1984.

(13) Lehninger, A. L. *Biochemistry*; Worth Publishers Inc.: New York, 1975.

(14) Harper, H. A.; Rodwell, V. W.; Mayes, P. A. *Review of Physiological Chemistry*; Lange Medical Publications: Los Altos, CA, 1979; *Steroids*; Steroloids Inc.: Wilton, NH, 1980.

(15) Sieber, W. CASP—Synthesis Planning in Industrial Research. (Discussed CAMP-computer-assisted metabolite prediction.) *Abstracts of Papers*, 191st National Meeting of the American Chemical Society, New York, April 1986; American Chemical Society: Washington, DC, 1986; ORGN 88.

(16) (a) Spann, M. L.; Chu, K. C.; Wipke, W. T.; Ouchi, G. Use of Computerized Methods to Predict Metabolic Pathways and Metabolites. *J. Environ. Pathol. Toxicol.* **1978**, *2*, 123. (b) Wipke, W. T.; Ouchi, G. I.; Chou, J. T. Computer-Assisted Prediction of Metabolism. In *Structure Activity Correlation as a Predictive Tool in Toxicology: Fundamentals, Methods, and Applications*; Goldberg, L., Ed.; Hemisphere: New York, 1983; pp 151–169. (c) Darvas, F. Predicting Metabolic Pathways by Logic Programming. *J. Mol. Graphics* **1988**, *6*, 80.

(17) (a) Seressiotis, A.; Bailey, J. E. MPS: an Artificially Intelligent Software System for the Analysis and Synthesis of Metabolic Pathways. *Biotechnol. Bioeng.* **1988**, *31*, 587. (b) Sel'kov, E. E.; Goryanin, I. I.;

Kaimachnikov, N. P.; Shevelev, E. L.; Yanus, I. A. Factographic Data Bank on Enzymes and Metabolic Pathways. *Stud. Biophys.* **1989,** *129,* 155. (c) Goryanin, I. I.; Shevelev, E. L.; Yanus, I. A. *Ibid.* **1989,** *129,* 165.

(18) Zollner, H. *Handbook of Enzyme Inhibitors*; VCH Publishers: Weinheim, Germany, 1989.

(19) Moock, T., et al. Multistep Reaction Schemes in the Reaction Access System. To be presented at the Second International Meeting on Chemical Structures, Noordwijkerhout, The Netherlands, June 3-7, 1990.

(20) (a) Schauer, M.; Heinrich, R. Quasi-Steady-State Approximation in the Mathematical Modeling of Biochemical Networks. *Math. Biosci.* **1983,** *65,* 155. (b) King, R. Bruce The Flow Topology of Chemical Reaction Networks. *J. Theor. Biol.* **1982,** *98,* 347. (c) Beretta, E.; Vetrano, F.; Solimano, F.; Lazzari, C. Graph Theory for Chemical and Biochemical Networks I. Tree Graphs and their Qualitative Stability Properties under Law Mass Hypothesis. *Stud. Urbinati, Fac. Farm.* **1976,** *49,* 7. (d) Graph Theory for Chemical and Biochemical Networks II. Loops and Cycles Graphs and a Perturbative Approach to the Cycles Arising from Enzyme Reactions. *Ibid.* **1976,** *49,* 46.

Benzenoid Series Having a Constant Number of Isomers. 2. Topological Characteristics of Strictly Peri-Condensed Constant-Isomer Benzenoid Series

JERRY RAY DIAS

Department of Chemistry, University of Missouri, Kansas City, Missouri 64110-2499

Received March 5, 1990

Our previous constant-isomer series are supplemented and extended. Two distinct classes of constant-isomer benzenoid groups have been identified. One class is topologically unique and the other forms a pairwise topologically equivalent class. An extention to our previous algorithm which led to these additional results is presented. Circulenes, benzenoid-related molecular systems having holes, have no constant-isomer series.

INTRODUCTION

It is well known that as the number of carbons in alkanes increases so does the number of isomers.[1] However, we previously demonstrated that there exist special benzenoid series in which the number of isomers remain constant as the number of carbons increase.[2] The recent availability of the benzenoid isomer table of Stojmenović and co-workers has allowed us to extend these results.[3]

RESULTS AND DISCUSSION

Conceptual Tools. Using the formula periodic table for benzenoids (Table PAH6 = Table 1 in ref 4) and the excised internal structure concept, several new strictly peri-condensed benzenoid series possessing an identical number of isomers have been identified.[5] Strictly peri-condensed benzenoids have all their internal carbon vertices mutually connected and collectively make up their excised internal structures. Pyrene ($C_{16}H_{10}$) has ethene, coronene ($C_{24}H_{12}$) has benzene, and ovalene ($C_{32}H_{14}$) has naphthalene ($C_{10}H_8$) as excised internal structures, and they are strictly peri-condensed benzenoids that are members of the one-isomer series found in Tables I and II. All constant-isomer strictly peri-condensed benzenoids are found on the extreme left-hand edge of Table PAH6 and are devoid of adjacent or proximate bay regions (fjords and coves). This translates into benzenoids with perimeters having everywhere two-carbon-atom gaps or greater, which means that all these even carbon strictly peri-condensed benzenoids can serve as excised internal structures for other successor (larger) strictly peri-condensed benzenoids. Other strictly peri-condensed benzenoids on this edge possess some isomers with doublet bay regions (coves) and are antecedents, but not members, of constant-isomer series. Herein, we present a refinement of our previously published algorithm[2,4-6] that allows us to identify this subset of benzenoids that cannot serve as excised internal structures. The use of the excised internal structure in generating the $C_{22}H_{12}$ constant-isomer series is illustrated in Figure 1 of refs 4 and 5.

Table I. Constant-Isomer Series of Even Strictly Peri-Condensed Benzenoids

series	no. of isomers	series	no. of isomers
(C_6H_6)		$C_{76}H_{22}$	12(4)
$C_{24}H_{12}$	1	$C_{126}H_{28}$	
$C_{54}H_{18}$		$C_{188}H_{34}$	
$C_{96}H_{24}$		...	
...		$C_{90}H_{24}$	27(12)
$C_{10}H_8$	1	$C_{144}H_{30}$	
$C_{32}H_{14}$		$C_{210}H_{36}$	
$C_{66}H_{20}$		...	
$C_{112}H_{26}$		$C_{106}H_{26}$	38(19)
$C_{170}H_{32}$		$C_{164}H_{32}$	
...		$C_{234}H_{38}$	
$C_{16}H_{10}$	1	...	
$C_{42}H_{16}$		$C_{124}H_{28}$	38(19)
$C_{80}H_{22}$		$C_{186}H_{34}$	
$C_{130}H_{28}$		$C_{260}H_{40}$	
...		...	
$C_{22}H_{12}$	2(1)[a]	$C_{142}H_{30}$	$\underline{133}$[b]
$C_{52}H_{18}$		$C_{208}H_{36}$	
$C_{94}H_{24}$		$C_{286}H_{42}$	
$C_{148}H_{30}$		...	
...		$C_{162}H_{32}$	$\underline{199}$
$C_{30}H_{14}$	3(1)	$C_{232}H_{38}$	
$C_{64}H_{20}$		$C_{314}H_{44}$	
$C_{110}H_{26}$		...	
$C_{168}H_{32}$		$C_{184}H_{34}$	$\underline{199}$
...		$C_{258}H_{40}$	
$C_{40}H_{16}$	3(1)	$C_{344}H_{46}$	
$C_{78}H_{22}$		...	
$C_{128}H_{28}$		$C_{206}H_{36}$	$\underline{428}$
...		$C_{284}H_{42}$	
$C_{50}H_{18}$	7(2)	...	
$C_{92}H_{24}$		$C_{230}H_{38}$	$\underline{616}$
$C_{146}H_{30}$		$C_{312}H_{44}$	
...		...	
$C_{62}H_{20}$	12(4)	$C_{256}H_{40}$	$\underline{616}$
$C_{108}H_{26}$		$C_{342}H_{46}$	
$C_{166}H_{32}$		...	
$C_{236}H_{38}$		$C_{282}H_{42}$	$\underline{1265}$
...		$C_{372}H_{48}$	
		...	

[a] Number of less stable diradical isomers are given in parentheses. [b] Sum of diradical and nonradical isomers are underlined.

Constant-Isomer Series of Strictly Peri-Condensed Benzenoids. Tables I and II present all known benzenoid constant-isomer series in which the latter are now reported for the first time. Each table reveals the same distinct pattern in which the number of isomers alternates between singlet and doublet occurrence. Also, the pattern for the progressive increase in the first member formula of each series should be

0095-2338/90/1630-0251$02.50/0 © 1990 American Chemical Society