ESSESA: An Expert System for Structure Elucidation from Spectra. 3. LNSCS for Chemical Knowledge Representation

HONG HUIXIAO* and XIN XINQUAN
Department of Chemistry, Nanjing University, Nanjing 210008, People's Republic of China

Received March 27, 1991

LNSCS (Linear Notation System of Chemical Structures) is a support for chemical knowledge representation in the form of a linear unambiguous code for processing and manipulating chemical structure information. Based on principles of molecular graph theory, LNSCS can notate chemical structures using a very small number of natural rules and can be very easily used by chemists. LNSCS code is an unambiguous but not unique linear character string. By a transformation program it can be easily changed into a computer's internal code. LNSCS may be used as an interface between computer and user in chemical knowledge-based systems and chemical substructure search systems. It has now been adopted by ESSESA as a tool to exchange chemical structure information between system and user.

INTRODUCTION

To build chemical information processing and manipulating systems, we must deal with chemical structures. Because a computer cannot accept and understand structural formulas and nomenclature generally used in chemistry, many linear notation systems for computer application have been designed.[1-9] Wiswesser[10] described the historical development of chemical nomenclature from the beginning of chemistry as a rudimentary science to the start of the computer era.

Many linear notation systems have been developed for their application systems. Some of them are ambiguous,[11] and most of them have many rules to encode complex molecular structures. Consequently they cannot be used easily by chemists to build up their own application system. The LNSCS system, based on the principles of molecular graph theory, is developed for processing chemical information in application of artificial intelligence. The first aim of the LNSCS system is to build an interface between the chemist and computer in ESSESA, an expert system for structure elucidation from spectra analysis,[12,13] by which the chemist describes molecular structures and substructures to assist the computer to revise rules about spectra analysis in the knowledge base of ESSESA.

LNSCS is similar to SMILES,[1] but there are many differences in encoding and implementation. For example, SMILES can encode aromatic compounds with Kekulé structural formulas or other structural formulas, but LNSCS only permits the use of the Kekulé structural formulas because we think the use of 'c' in SMILES may create ambiguity in some cases such as 'Sc' which may be considered as 'S' and 'c' or 'Sc'. Differences also include some encoding rules in LNSCS which do not exist in SMILES, such as use of the underscore character to impress code, encoding geometrical and conformational isomers, compounds containing isotopes and coordination compounds, and so on. Generally LNSCS is easier to learn than SMILES, and it can encode almost all chemical structures. The differences in implementation between LNSCS and SMILES will be given in the next paper of this series.

When we designed ESSESA we intended it to be used to treat the spectral analysis and structure elucidation of coordination compounds and to be applied to conformational analysis; so LNSCS was developed to be able to represent all information about these chemical structures. SMILES has not been reported to be able to support chemical knowledge representation about these types of chemical structures.

The LNSCS code is an unambiguous but not unique character string. Rules for generating LNSCS code are very few and natural and a chemist can easily write out the LNSCS code of a molecule from the structural formula. A machine-friendly and machine-independent ancillary program which understands LNSCS code and transforms it to a computer's unique internal code has been developed. In this paper we will describe the methodology and encoding rules used in the LNSCS system. The generation of the computer's unique internal LNSCS code and its algorithm are the subjects of following papers.

THEORETICAL BACKGROUND OF LNSCS

LNSCS is a linear code, based on the principles of molecular graph theory, which represents a chemical structure by a linear string of characters, similar to natural language. A chemical structure can be considered as a colored graph Gx, in which the nodes are related to atoms and the edges are related to chemical bonds:

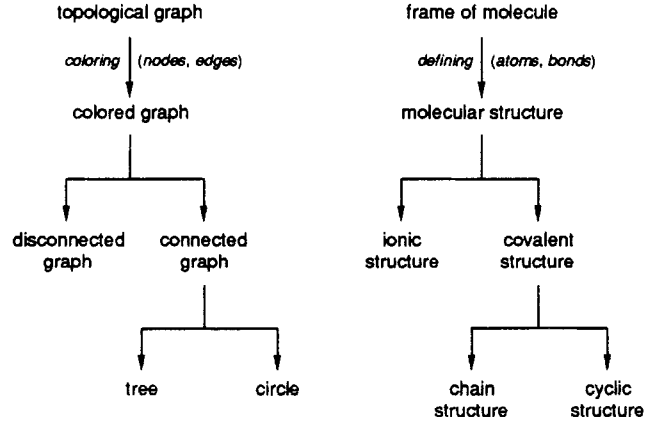
$$Gx(X, U, Xx, Xu) \quad (1)$$

where X is a set of nodes in the graph (the kinds of atoms in the chemical structure), U is the set of edges in the graph (the kinds of chemical bonds in the chemical structure), Xx and Xu are the coloring functions on X and U, respectively.

Graph G(X, U) is the topological graph obtained from the colored graph Gx(X, U, Xx, Xu). In chemistry, the colored graph Gx may reflect the structure of a compound, and the topological graph G can reflect the structure frame of a compound.

In accordance with the above idea, the LNSCS system notates the chemical structure by describing the related colored graph. By definition of the nodes and edges in the related topological graph, the colored graph can be generated. In the LNSCS system, stereochemistry may be described by suffix terms of the nodes and edges.

The relationship between a colored graph and a chemical structure can be seen in the following:



0095-2338/92/1632-0116$03.00/0

Table I. Symbols Used in LNSCS

symbols	meaning
all element symbols	atoms in structure
= # :	double, triple, and coordinate bonds, respectively
()	branches
0 1 2 3 4 5 6 7 8 9	ring-closure atom or number of same branches, valences, and charges
+ -	charge kind
.	free valence
*	chiral atom
| ~ / \	axial, equatorial bond, and *Z*, *E* configuration, respectively
[]	denote ion or macro-atom that is not a single atom
,	symbol between numbers of ring-closure on same atom
&	isotope atom
r s	*R* and *S* configuration
-	the same nodes

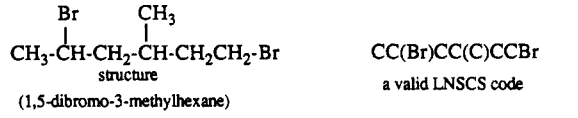
ENCODING RULES

The LNSCS system was designed to represent on a computer the two-dimensional valence-oriented picture of a chemical structure which chemists draw by a series of characters that ends with a ⟨RETURN⟩ key on the computer's keyboard. No attempt is made to represent any particular three-dimensional arrangement of atoms in a chemical structure: the conformation of the molecule. In a LNSCS code, hydrogen atoms may be omitted (hydrogen-suppressed graph) or encoded (hydrogen-complete graph), and all single bonds are omitted. All aromatic structures are specified directly in a Kekulé formula. The encoding rules, which are very few and natural, are given in this section.

Symbols Used in LNSCS. The symbols used in the LNSCS system and the corresponding meaning are shown in Table I. Some examples of the encoding rules will now be given.

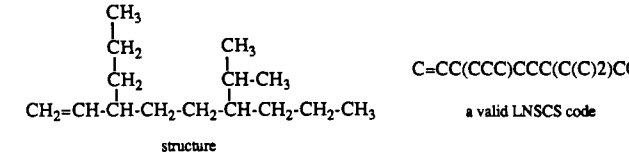
Chain Structures. When encoding a chain structure, we can start from any terminal of the structure. The atoms that have the same connectivities and valences may be encoded with an underscore followed by a number to denote how many occurrences there are of this type of atom. For example, 1-bromo-4-chlorobutane may be represented by two equally valid LNSCS codes:

If there are branches in the chain structure, the branches are denoted by enclosures in parentheses following the atom to which the branch is attached. An example is

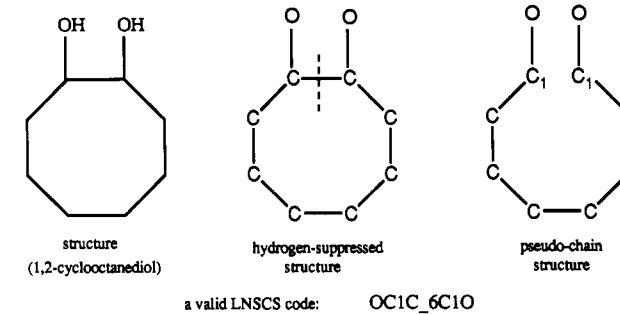


If there are several identical branches attached to the same atom, the branches can be represented as above. An alternative method is that the number of the same branches is denoted by a digit following the parentheses, as in the following example.

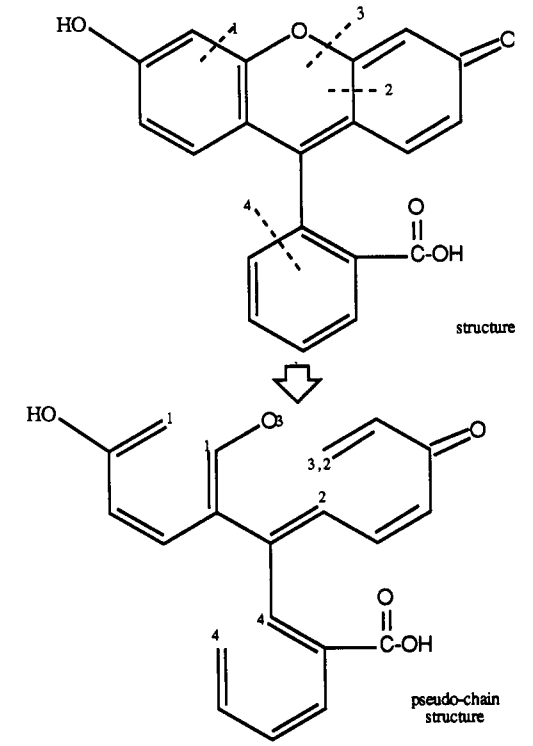
If a branch has smaller branches, the smaller branches should be nested or stacked in the branch, and if the smaller branches have yet smaller branches, the latter should be nested or stacked in the smaller branches, and so on. An example is



Cyclic Structures. In order to generate the LNSCS code of a cyclic structure, we first break chemical bonds, usually single bonds of rings in the cyclic structure, resulting in a connected pseudo-chain structure in which the atoms linked together by the broken bonds are denoted by numbers. Then we encode this pseudo-chain structure using the rules for encoding chain structures described above, with a number to denote the atoms that should be closed in restoring the cyclic structure in the application system based on the LNSCS code. The procedure to encode a cyclic structure is illustrated in the following example.



One atom may have more than one ring closure (or ring opening) number. The numbers are separated by a comma, as in the following example.

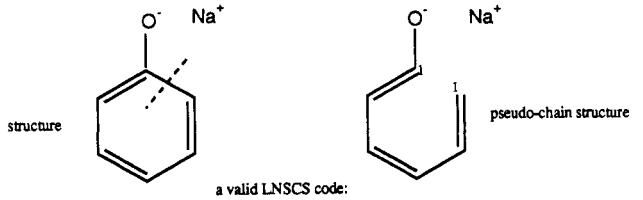


a valid LNSCS code

C1=C(O)C=CC(=C1O3)C(=C2C=CC(=O)C=C2,3)C4=C(C(O)=O)C=CC=C4

Ionic Compounds. Ionic compounds can be considered as disconnected structures that may be encoded individually. Then these codes of individual structures are added together arbitrarily, inserting a space between them, to generate a complete code. When we encode a disconnected structure, the ion may be closed in brackets, denoting the charge kind by the symbol "+" or "−" attached to the element symbol. The

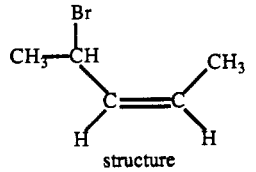
charge on an ion is specified by a number following the charge symbol "+" or "−", for example, sodium phenoxide:



a valid LNSCS code:
[Na+] [O-]C1=CC=CC=C1

Here the [O−] is a macroatom. For convenience, we consider a structure unit which is not a single atom as a macro atom. For example, the benzene ring also may be encoded as a macro atom [C6H5], so the LNSCS code of sodium phenoxide has other valid forms [Na+]O−[C6H5], Na+-O−[C6H5], and so on. The LNSCS system should recognize codes of the type of [Na+][O−]C1=CC=CC=C1 as being synonymous with codes such as [Na+]O−[C6H5].

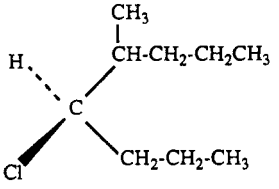
Isomers. In the LNSCS system, configuration isomers are distinguished by use of the suffixes "/" (cis or Z isomer) and "\" (trans or E isomer) attached to the last encoded atom of a double bond. We give one example to illustrate the rules for generating the LNSCS code of configuration isomers.



structure
(4-bromo-cis-2-pentene)

a valid LNSCS code:
CC(Br)C=C/C

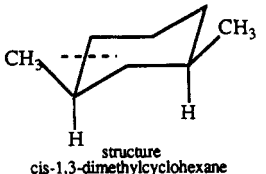
Chemical structures that have chiral atoms may be encoded by use of the suffixes "r" and "s", attached to the symbol of the chiral atom, "*", in the LNSCS system:



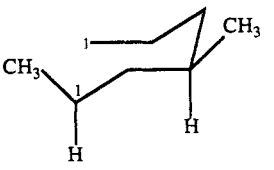
structure
(R)-4-chloro-2-methyloctane

a valid LNSCS code:
CCCC*r(Cl)C(C)CCC

Conformation isomers about cyclohexane rings are encoded by use of the suffixes "|" (axial bond) and "~" (equatorial bond) attached to the atom connected to the ring. Two examples follow to illustrate the encoding rules.

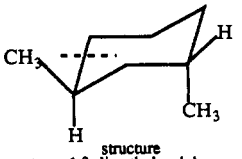


structure
cis-1,3-dimethylcyclohexane
(both methyls on equatorial)

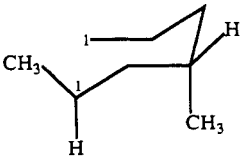


pseudo-chain structure

a valid LNSCS code: C1CCC(C~)CC1C~



structure
trans-1,3-dimethylcyclohexane
(one methyl on equatorial)



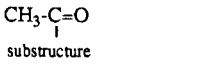
pseudo-chain structure

a valid LNSCS code: C1CCC(C|)CC1C~

All these LNSCS codes of different stereoisomers may be understood and transformed into internal code by a computer program, the LNSCS compiler.

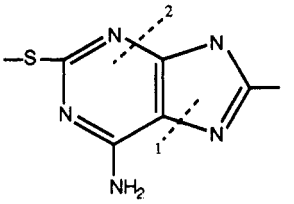
Substructures. In artificial intelligence systems for processing chemical structural information, substructures (also called partial structures) must be manipulated. Examples are substructure searching systems, systems for organic synthesis design, systems for molecular structure elucidation, and so on. A substructure is a structural unit that contains free valences. In graph theory, a substructure can be considered as an incomplete graph or subgraph that can grow to generate a complete graph by some method such as in our expert system for structure elucidation, ESSESA. The LNSCS notation system provides the function to encode and manipulate substructures.

The procedure for encoding substructures is similar to that for complete structures, but the free valences (or partial bonds) must be specified by the symbol "." suffixed to the atom that contains the free valences. The number of free valences on an atom can be specified by a digit attaching to the symbol ".". This procedure is illustrated in the following examples.

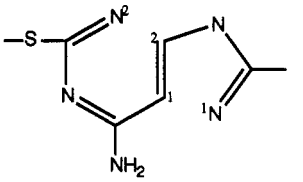


substructure

a valid LNSCS code: CC.=O



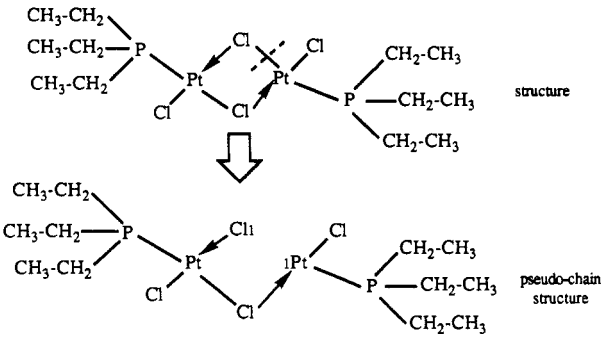
structure



pseudo-chain structure

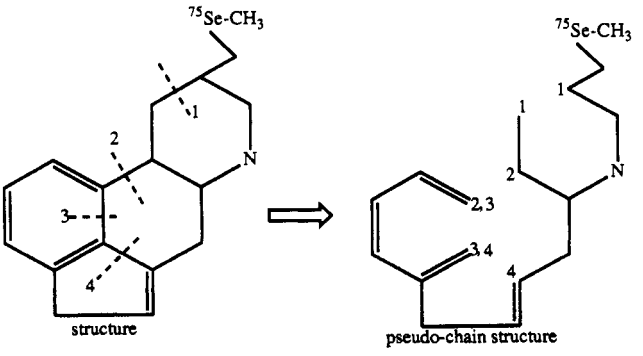
a valid LNSCS code: N2=C(S.)N=C(N)C1=C2NC.=N1

Coordination Compounds. The LNSCS system has the ability to encode coordination compounds. The coordination bonds are distinguished from general chemical bonds by use of a symbol ":" between atoms linked with a coordination bond, e.g.



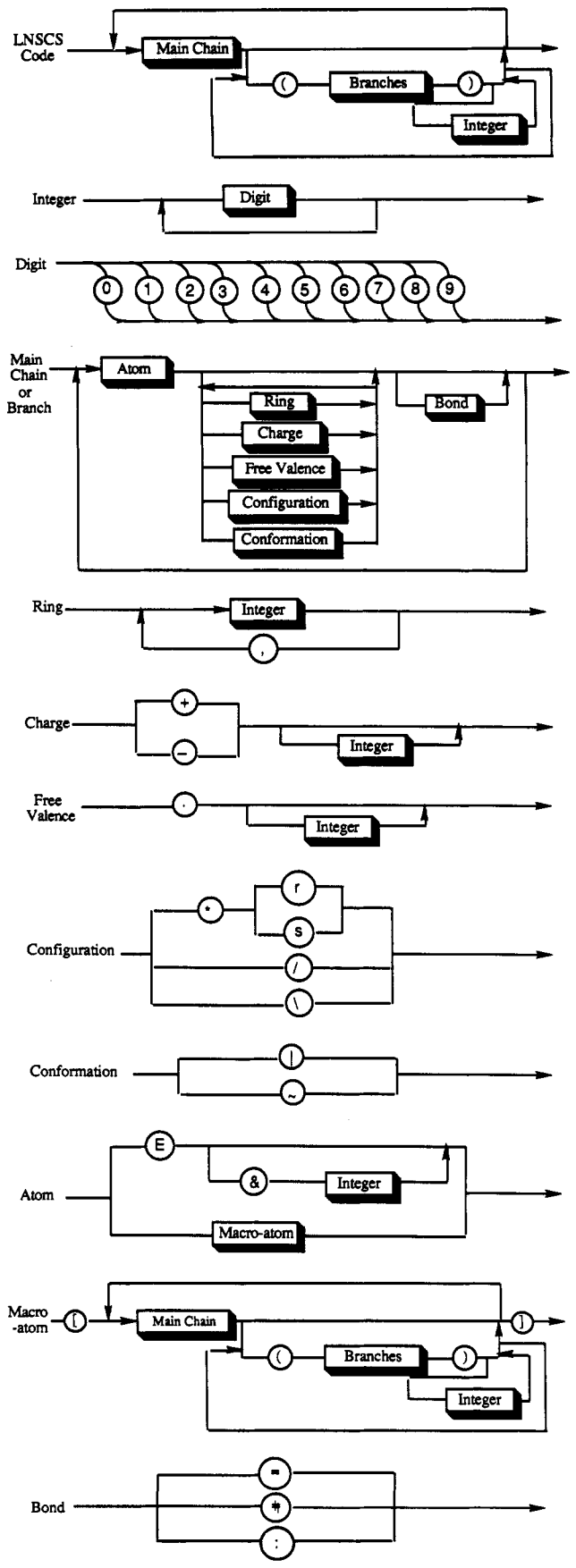
a valid LNSCS code:
Cl1:Pt(Cl)(P(CC)3)Cl:Pt1(Cl)P(CC)3

Compounds Containing Isotopes. Encoding structures that contain isotopes is handled in the LNSCS system by use of the symbol "&" suffixed to the isotopic atom and followed by a number to specify the isotope. The following example illustrates this procedure.



a valid LNSCS code:
CSE&75CC1CNC(C2C1)CC4=CCC(=C3,4)C=CC=C2,3

Scheme I



Grammar of LNSCS. We will show the LNSCS code generation grammar with the following grammatical sketch (Scheme I), which illustrates the construction of LNSCS code and the relationship between grammatical components of LNSCS code, and formulates the standard. In the scheme, the symbols in circles are valid elements in LNSCS code, and E is an element symbol. The grammatical components in frames should be further specified by other components.

DISCUSSION

The representation of chemical structures has been proved to be of immense importance to the organization and manipulation of chemical knowledge on a computer, and it is on these representations that both manually and automatically operated chemical information systems have, without exception, become established. Several notations for structure representation have been developed. Each representation derives from the two-dimensional structural formula and is sufficient to provide a simple characterization of that structure. Some notations produce ambiguous codes, and most of them have many encoding rules and are not easy to learn for chemists. It is very hard to generate unique codes from some of these notation systems. Generally, ambiguous representation provides only a partial description of a chemical structure, so generation of the correct complete structure is not usually possible.

In chemistry, a molecular structure is considered as a colored graph, so coloring is related to topology. To color a node is to decide atomic features of a node which can specify the number of atoms that it should be bonded to, that is to say, the color of a node decides the connectivity in topology. As we know, all atoms in a compound have a regular connectivity, the valence. Because a hydrogen atom has one connectivity and most compounds contain hydrogen atoms, it is very convenient to omit the hydrogen atoms in encoding. The number of hydrogen atoms in a node can be obtained from the calculation of atomic feature and edge color in the LNSCS code. Single bonds are also omitted. The single bonds can be obtained from the LNSCS code because it is assumed that an atom is linked with its neighbor atom by a single bond if there is no bond color between them.

The LNSCS system, based on the principles of molecular graph theory, is believed to be a facility that can generate unambiguous codes because every atom and bond is specified in LNSCS. It has a very small set of rules to obey for encoding a complex structure, so it is very easy for chemists to use it to develop their own system. It may be understood in a very fast, compact manner by a LNSCS compiler, thereby satisfying the computer objectives of time and space savings.

Originally, LNSCS was developed to provide a user–computer interface in our expert system for structure elucidation, ESSESA. Beyond this objective, it is valuable for implementation of a wide variety of computer-oriented chemical information processing systems.

ACKNOWLEDGMENT

We owe a deep debt of gratitude to Dr. Chris Marshall for redrafting the manuscript. We also thank Dr. Wendy A. Warr for encouraging the writing of this paper.

REFERENCES AND NOTES

(1) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(2) Wilcox, C. S.; Levinson, R. A. A Self-Organized Knowledge Base for Recall, Design and Discovery in Organic Chemistry. In *Artificial Intelligence in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1986; pp 209–230.
(3) Dyson, G. M.; Lynch, M. F.; Morgan, H. L. A Modified IUPAC-Dyson Notation System for Chemical Structures. *Inf. Storage Retr.* **1968**, *4*, 27–83.
(4) Dubois, J. E.; Sobel, Y. DARC System for Documentation and Artificial Intelligence in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 326–333.
(5) Read, R. C. A New System for Designation of Chemical Compounds. 1. Theoretical Preliminaries and the Coding of Acyclic Compounds.

J. Chem. Inf. Comput. Sci. 1983, 23, 135-149.
(6) Warr, W. A. Diverse Uses and Future Prospects for Wiswesser Line-Formula Notation. J. Chem. Inf. Comput. Sci. 1982, 22, 98-101.
(7) IUPAC Nomenclature of Organic Chemistry, Sections A-F and H; Pergamon: Oxford, 1979.
(8) Rush, J. E. Status of Notation and Topological Systems and Potential Future Trends. J. Chem. Inf. Comput. Sci. 1976, 16, 202-210.
(9) Wiswesser, W. J. A Linear-Formula Chemical Notation; Crowell: New York, 1954.
(10) Wiswesser, W. J. Historic Development of Chemical Notations. J. Chem. Inf. Comput. Sci. 1985, 25, 258-263.
(11) Communication, Storage and Retrieval of Chemical Information; Ash, J. E., Chubb, P. A., Ward, S. E., Welford, S. M., Willett, P., Eds.; Ellis Horwood: Chichester, 1985.
(12) Huixiao, H.; Xinquan, X. ESSESA: An Expert System for Elucidation of Structures from Spectra. 1. Knowledge Base of Infrared Spectra and Analysis and Interpretation Programs. J. Chem. Inf. Comput. Sci. 1990, 30, 203-210.
(13) Huixiao, H.; Xinquan, X. ESSESA: An Expert System for Structure Elucidation from Spectra. II. A Novel Algorithm of Perception of the Linear Independent Smallest Set of Smallest Rings. Anal. Chim. Acta, in press.

Hausdorff Dimension as a Quantification of Local Roughness of Protein Surfaces

CARL-DIETER ZACHMANN and JÜRGEN BRICKMANN*

Institut für Physikalische Chemie, Technische Hochschule Darmstadt, Petersenstrasse 20, D-6100 Darmstadt, Germany

Received August 27, 1991

Based on structural information from the Brookhaven Protein Data Bank, the contact surfaces of the proteins lysozyme, trypsin, and BPTI (bovine pancreatic trypsin inhibitor) with a spherical test particle of the size of a water molecule have been calculated and systematically analyzed. It is our purpose to establish (i) self-similarity as a statistical concept for the characterization of surface roughness and (ii) the Hausdorff dimension as a measure of the local surface complexity. It is found that the proteins statistically show self-similarity within a yardstick range 1.2 Å < R < 20 Å, and that this concept also holds reasonably for parts of the surface which are not too small.

INTRODUCTION

Up to now, there was no way to predict the three-dimensional structure of a protein with a given sequence of amino acids based only on the knowledge of a molecular force field, which—at least in principle—can be derived from first principles. Additional information from structural investigations (X-ray, NMR) for fragments are necessary for any successful attempt. Consequently, the building principles of protein structures presently can not be completely understood on the basis of first principles (in contrast to small- and medium-sized molecules). A systematic analysis of 3D protein data may help to understand the building principles which have been working during the evolution of these biopolymers. There are several arguments for the statement saying that one of these principles is related to the surface complexity of these molecules.[1-5]

The "surface roughness" (or geometrical complexity) of biological macromolecules influences the transport of substrate molecules from the aqueous bulk phase to the surface as well as the migration on the surface. While by an increase of surface roughness the first process is speeded up, the second is slowed down.[1,2] The roughness of molecular surfaces also plays a role in intermolecular recognition. Recognition is related to geometrical selectivity and to the specificity of point interactions. One molecule can only be positively identified by a second one when two conditions are fulfilled, namely

(i) that the shapes of the molecular repulsion surfaces fit in the contact region

(ii) that specific point links via directed hydrogen bonds can be formed as a consequence of reasonable proton donor-acceptor arrangements

This paper deals with the quantification of "surface roughness". It has been argued[1-5] that the concept of fractals may be helpful for the characterization of the surface complexity of proteins. These surfaces seem to be self-similar within a certain yardstick range. Moreover, it has been suspected[1,2] that the selectivity of a receptor site may be related to local fractality of the protein surface in the receptor-site region.

The purpose of this paper is to re-examine whether protein surfaces really exhibit self-similarity, i.e., whether the concept of fractals holds at least in a statistical sense. If the answer is yes, we shall attempt to use it to quantify the roughness (i.e., geometrical complexity) of such a surface as a local property (a quantity which can be assigned to a local surface area of the protein). The fractal concept can only be used for correlation studies describing the selectivity of an active-site region if it can be shown to be meaningful to describe a local property of the surfaces.

Some work has been done to establish the fractal-surface dimension of proteins. Lewis and Rees[3] calculated the contact surfaces $A(R)$ of some proteins with spherical test molecules (hard spheres) following the algorithm of Connolly[8] and using sphere radii between R = 1.0 Å and R = 3.5 Å. They determined the fractal dimension D from the average gradient in a log (A)/log (R) plot and found D = 2.44, 2.44, and 2.43 for lysozyme, ribonuclease A, and superoxide dismutase, respectively. The curves in their diagrams are, however, not straight lines, as expected for self-similar objects.[5-7] Consequently, from the work of these authors, it seems questionable whether the concept of fractals is applicable in this case, and, in particular, whether the "local" fractal dimensions calculated by the authors are really meaningful. Their conclusion that molecular regions involved in the formation of tight complexes (such as antibody-combining regions) appear to be more irregular (with high-D values) than regions involved in the formation of transient complexes (such as active sites) is open to discussion. Åqvist and Tapia[4] re-investigated the surface fractality as a guide for studying protein-protein interactions. Starting from the same type of surface generation and surface analysis as done by Lewis and Rees,[3] they indeed found linear plots in the log (A)/log (R) diagrams (with standard deviations of the regression coefficient < 0.02) for a range 1.5 Å < R < 3.0 Å. The authors obtained a fractal dimension of D = 2.19 for lysozyme, which is at variance with the results of Lewis and Rees (D = 2.44), but in remarkable agreement with

* Author to whom correspondence should be addressed.

0095-2338/92/1632-0120$03.00/0