(4) Golender, V. E.; Drboglav, V. V.; Rosenblit, A. B. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 196.
(5) Razinger, M. *Theor. Chim. Acta* **1982**, *61*, 581.
(6) Herndon, W. C. *Inorg. Chem.* **1983**, *22*, 554.
(7) Herndon, W. C. *Tetrahedron Lett.* **1974**, 671.
(8) Herndon, W. C. *J. Chem. Doc.* **1974**, *14*, 150.
(9) Davis, M. I.; Ellzey, M. L., Jr. *J. Comput. Chem.* **1983**, *4*, 267.
(10) (a) Randić, M.; Davis, M. I. *Int. J. Quantum Chem.* **1984**, *24*, 69. (b) Randić, M.; Brissey, G. M.; Wilkins, C. W. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52. (c) Randić, M. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171.
(11) Shelley, C. A.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 171.
(12) (a) Shelley, C. A.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247. (b) Munk, M. E.; Christie, B. D. *Anal. Chem. Acta*, **1989**, *212*, 57.
(13) Gibbons, A. *Algorithmic Graph Theory*; Cambridge University: London, 1985.
(14) Wilkinson, J. H. *The Algebraic Eigenvalue Problem*; Clarendon, Oxford, 1965.
(15) Jennings, A. *Matrix Computation for Engineers and Scientists*; John Wiley & Sons: New York, 1977.
(16) Balasubramanian, K.; Liu, Xiaoyu *J. Comput. Chem.* **1988**, *9*, 406.
(17) Householder, A. S. *The Theory of Matrices in Numerical Analysis*; Blaisdell: New York, 1964.
(18) Householder, A. S.; Bauer, F. L. *Numer. Math.* **1959**, *1*, 29.
(19) Wilkinson, J. H. *Comput. J.* **1960**, *3*, 23.
(20) Ortega, J. In *Mathematical Methods for Digital Computers*; Ralson, Antony, Wilkf, Herbert S., Eds.; Wiley: New York, 1967; Vol. II, pp 65-93.
(21) Corneil, D. G.; Gotlieb, C. C. *J. Assoc. Comput. Mach.* **1970**, *17*, 51.
(22) Read, R. C.; Corneil, D. G. *J. Graph Theory* **1977**, *1*, 339.
(23) Herndon, W. C.; Ellzey, M. L., Jr. *Tetrahedron* **1975**, *31*, 99.
(24) Liu, X. Y.; Balasubramanian, K.; Munk, M. E. *J. Magn. Reson.* **1990**, *87*, 547.
(25) Herndon, W. C. In *Chemical Applications of Topology & Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983.
(26) Wipke, W. T.; Dyott, T. M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140.
(27) Heath, J. R.; O'Brien, S. C.; Zhang, Q.; Liu, Y.; Curl, R. F.; Kroto, H. W.; Tittle, F. K.; Smalley, R. E. *J. Am. Chem. Soc.* **1985**, *107*, 7779.
(28) Balasubramanian, K.; Liu, X. Y. *Int. J. Quantum Chem. Symp.* **1988**, *22*, 319.

# Simple and Fast Search System for Closely Related Proteins

SHIN-ICHI NAKAYAMA,* KATSUKO SUGAI, YURIKO HOTATE, and MASAYUKI YOSHIDA

University of Library and Information Science, Tsukuba-city, Ibaraki, 305 Japan

Received March 14, 1990

A method for measuring a probability distance from the Euclidean distance between proteins, which were expressed by use of all possible paris of amino acids as descriptors, is described. A system to search for closely related proteins by the probability distance between proteins was constructed, and its advantage over other systems is discussed.

## INTRODUCTION

Some properties of a little-studied protein would be predicted from the knowledge of well-characterized proteins of similar amino acid sequences. Likewise, evolutionary relationships among biological species would be revealed by sequential similarities between proteins. Thus, if an amino acid sequence of a new protein were revealed, it is necessary that one should search for proteins with similar sequences.

Meanwhile, sequence data of many proteins have been accumulated, and various methods to search for similar proteins from sequence data have been developed. In most of them sequential similarities were measured on the basis of the maximum match between amino acid sequences.[1] This process is, however, very time-consuming. It takes 8 h to search for a protein of 200 residues with 2600 proteins (about 500 000 residues) in the NBRF protein library by a computer program SEQHP on the VAX11/750 computer. A recently developed program FASTP greatly cut short the execution time to 2-5 min under the same condition.[2] However, the number of proteins included to the NBRF database has increased steadily to 5251 (1 384 621 residues) in March 1988, and development of a faster search method is now required.

Previously, we have presented a method for clustering proteins using all possible pairs of amino acids as descriptors.[3] This paper describes the simple and fast search system for proteins based on an improved similarity measurement.

## A METHOD FOR MEASURING PROBABILITY DISTANCE

All possible pairs of amino acids, which numbered 400, were assigned as the descriptors that reflect one-dimensional structures of proteins. Protein $i$ ($P_i$) was then expressed by a set of descriptor values as $P_i = (x_{i1}, x_{i2}, ..., x_{ik}, ..., x_{i400})$, where $x_{ik}$ was the frequency of occurrence of the $k$th descriptor in protein $i$ and was readily derived from its amino acid sequence.

The Euclidean distance $d_{ij}$ between proteins $i$ and $j$ was then calculated by eq 1. The $d_{ij}$ value is just a number, and how

$$d_{ij} = \sqrt{\sum_{k=0}^{400} (x_{ik} - x_{jk})^2} \quad (1)$$

close the distance is between proteins $i$ and $j$ is not known from its magnitude. Ideally, one would prefer to know the exact probability of obtaining a given $d_{ij}$ value, but that is prohibitively difficult.

As an alternative to determining the exact probability, it is possible to estimate it experimentally. Thus, a pair of random amino acid sequences $i$ and $j$ ($A_i$ and $A_j$) having the same lengths as proteins $i$ and $j$ was produced, and the Euclidean distance $ad_{ij}$ for that pair was calculated. In a way similar to the above the distance $ad_{ij}$ was calculated for 200 pairs of $A_i$ and $A_j$ produced independently. A plot of the $ad_{ij}$ values thus obtained against the frequency of the pairs showed a normal distribution, from which a mean distance $m_{ij}$ and a standard deviation $\sigma_{ij}$ were calculated. The division of the difference between $d_{ij}$ and $m_{ij}$ by $\sigma_{ij}$ gives the probability distance $Z_{ij}$ between proteins $i$ and $j$ (eq 2).

$$Z_{ij} = (d_{ij} - m_{ij})/\sigma_{ij} \quad (2)$$

## EMPIRICAL EQUATIONS FOR THE MEAN DISTANCE AND THE STANDARD DEVIATION

In the calculations of $Z_{ij}$, estimation of $m_{ij}$ and $\sigma_{ij}$ takes a long time. Thus, empirical equations for them were constructed.

The data sets 1, 2, and 3 of $m_{ij}$ and $\sigma_{ij}$ were produced experimentally for type 1 pairs of $A_i$ and $A_j$ with shorter

0095-2338/90/1630-0269$02.50/0 

**Table I.** Coefficients of $\mu_{ij}$ and $s_{ij}$ for the Data Sets 1, 2, and 3

| | set 1 | set 2 | set 3 |
|---|---|---|---|
| $\mu_{00}$ | 3.3892487 | 6.6351717 | $3.38333116 \times 10^{1}$ |
| $\mu_{01}$ | $7.9541142 \times 10^{-2}$ | $5.3654354 \times 10^{-2}$ | $1.15272066 \times 10^{-1}$ |
| $\mu_{02}$ | $-6.7982143 \times 10^{-5}$ | $-1.4312088 \times 10^{-6}$ | $-2.52407914 \times 10^{-5}$ |
| $\mu_{03}$ | $6.2460412 \times 10^{-8}$ | $1.7846269 \times 10^{-10}$ | $2.8730991 \times 10^{-9}$ |
| $\mu_{10}$ | $7.9535125 \times 10^{-2}$ | $6.7785081 \times 10^{-3}$ | $1.15287054 \times 10^{-1}$ |
| $\mu_{11}$ | $-6.2910563 \times 10^{-4}$ | $-6.1431681 \times 10^{-5}$ | $-2.36552442 \times 10^{-4}$ |
| $\mu_{12}$ | $1.4574492 \times 10^{-6}$ | $2.4584777 \times 10^{-8}$ | $9.4461177 \times 10^{-8}$ |
| $\mu_{13}$ | $-1.2986281 \times 10^{-9}$ | $-2.9016720 \times 10^{-12}$ | $-1.12609663 \times 10^{-11}$ |
| $\mu_{20}$ | $-6.7934145 \times 10^{-5}$ | $-1.962442 \times 10^{-5}$ | $-2.52517566 \times 10^{-5}$ |
| $\mu_{21}$ | $1.4572008 \times 10^{-6}$ | $7.1361116 \times 10^{-8}$ | $9.4467905 \times 10^{-8}$ |
| $\mu_{22}$ | $-4.8987950 \times 10^{-9}$ | $3.5621390 \times 10^{-11}$ | $-4.70471639 \times 10^{-11}$ |
| $\mu_{23}$ | $5.1781213 \times 10^{-12}$ | $4.8670078 \times 10^{-15}$ | $6.20740409 \times 10^{-15}$ |
| $\mu_{30}$ | $6.2374816 \times 10^{-8}$ | $1.2214509 \times 10^{-7}$ | $2.87493221 \times 10^{-9}$ |
| $\mu_{31}$ | $-1.2978019 \times 10^{-9}$ | $-1.9938645 \times 10^{-10}$ | $-1.12624348 \times 10^{-11}$ |
| $\mu_{32}$ | $5.1764861 \times 10^{-12}$ | $8.4080221 \times 10^{-14}$ | $6.20765561 \times 10^{-15}$ |
| $\mu_{33}$ | $-6.1537341 \times 10^{-15}$ | $-1.0634755 \times 10^{-17}$ | $-8.7127683 \times 10^{-19}$ |
| $s_{00}$ | $5.3289168 \times 10^{-1}$ | $1.1223335 \times 10^{-1}$ | 3.62991649 |
| $s_{01}$ | $-5.4842297 \times 10^{-3}$ | $6.8629681 \times 10^{-4}$ | $5.7088236 \times 10^{-3}$ |
| $s_{02}$ | $3.7530962 \times 10^{-5}$ | $-2.6535666 \times 10^{-6}$ | $2.6535666 \times 10^{-6}$ |
| $s_{03}$ | $-6.2604170 \times 10^{-8}$ | $4.0951818 \times 10^{-11}$ | $-3.53228296 \times 10^{-10}$ |
| $s_{10}$ | $-5.4835288 \times 10^{-3}$ | $4.1608679 \times 10^{-3}$ | $-5.73940808 \times 10^{-3}$ |
| $s_{11}$ | $1.6005519 \times 10^{-4}$ | $5.2293393 \times 10^{-7}$ | $1.37550368 \times 10^{-5}$ |
| $s_{12}$ | $-8.7199606 \times 10^{-7}$ | $-1.3361045 \times 10^{-9}$ | $-6.6590816 \times 10^{-9}$ |
| $s_{13}$ | $1.3415682 \times 10^{-9}$ | $1.8723881 \times 10^{-13}$ | $8.99922093 \times 10^{-13}$ |
| $s_{20}$ | $3.7526791 \times 10^{-5}$ | $4.59054475 \times 10^{-6}$ | $2.66627368 \times 10^{-6}$ |
| $s_{21}$ | $-8.7198218 \times 10^{-7}$ | $-3.88127682 \times 10^{-8}$ | $-6.65931855 \times 10^{-9}$ |
| $s_{22}$ | $4.8014431 \times 10^{-9}$ | $2.3635944 \times 10^{-11}$ | $3.4405541 \times 10^{-12}$ |
| $s_{23}$ | $-7.3700086 \times 10^{-12}$ | $-3.3595643 \times 10^{-15}$ | $-4.83355468 \times 10^{-16}$ |
| $s_{30}$ | $-6.2598329 \times 10^{-8}$ | $-2.5092981 \times 10^{-8}$ | $-3.54420995 \times 10^{-10}$ |
| $s_{31}$ | $1.3415496 \times 10^{-9}$ | $9.2616171 \times 10^{-11}$ | $8.99388252 \times 10^{-13}$ |
| $s_{32}$ | $-7.3700135 \times 10^{-12}$ | $-5.24312592 \times 10^{-14}$ | $-4.83124037 \times 10^{-16}$ |
| $s_{33}$ | $1.1306997 \times 10^{-14}$ | $7.518518 \times 10^{-18}$ | $6.9953675 \times 10^{-20}$ |

lengths ($i, j$ = 50–400 at 50 intervals), type 2 ones with shorter and longer lengths ($i$ = 50–400 at 50 intervals, $j$ = 500–4000 at 500 intervals), and type 3 ones with longer lengths ($i, j$ = 500–4000 at 500 intervals), respectively. Since the $m_{ij}$ and $\sigma_{ij}$ values depend on the magnitude of $i$ and $j$, empirical equations for them were assumed to be expressed as functions of $i$ and $j$ (eqs 3 and 4). The coefficients were then determined

$$m_{ij} = \mu_{00} + \mu_{01}j + \mu_{02}j^2 + \mu_{03}j^3 + \mu_{10}i + \mu_{11}ij + \mu_{12}ij^2 + \mu_{13}ij^3 + \mu_{20}i^2 + \mu_{21}i^2j + \mu_{22}i^2j^2 + \mu_{23}i^2j^3 + \mu_{30}i^3 + \mu_{31}i^3j + \mu_{32}i^3j^2 + \mu_{33}i^3j^3 \quad (3)$$

$$\sigma_{ij} = s_{00} + s_{01}j + s_{02}j^2 + s_{03}j^3 + s_{10}i + s_{11}ij + s_{12}ij^2 + s_{13}ij^3 + s_{20}i^2 + s_{21}i^2j + s_{22}i^2j^2 + s_{23}i^2j^3 + s_{30}i^3 + s_{31}i^3j + s_{32}i^3j^2 + s_{33}i^3j^3 \quad (4)$$

from the respective data sets by the method of least-squares, and the results are shown in Table I.

The $m_{ij}$ and $\sigma_{ij}$ values calculated by eqs 3 and 4 agree fairly well with those obtained experimentally as shown in Table II. Thus, eqs 3 and 4 were confirmed to be useful.

## A SEARCH SYSTEM OF CLOSELY RELATED PROTEINS

All the data of amino acids sequences of proteins were taken from the Protein Sequence Database Release 16.0 of the National Biomedical Research Foundation (NBRF). The data were expressed as sets of descriptor values and were stored in a file. Similarly, a datum of an amino acid sequence of protein $i$ of interest was also converted to a set of descriptor values. The probability distances $Z_{ij}$ between protein $i$ and each protein $j$ in the database were then measured and their values were ranked in increasing order.

The programs used in the study were written in language C for the SONY NEWS-840 computer, which had 68020 CPU (16.67 MHz) with 68881 (16.67 MHz) coprocessor.

**Table II.** $m_{ij}$ and $\sigma_{ij}$ Values Calculated and Experimentally Obtained

| chain lengths | | $m_{ij}$ | | $\sigma_{ij}$ | |
|---|---|---|---|---|---|
| $i$ | $j$ | calcd | exp | calcd | exp |
| 500 | 500 | 31.4 | 30.0 | 1.225 | 1.220 |
| 500 | 1000 | 45.8 | 46.8 | 1.100 | 1.107 |
| 500 | 1500 | 67.0 | 67.0 | 0.980 | 1.034 |
| 500 | 2000 | 90.1 | 89.8 | 1.204 | 0.987 |
| 500 | 2500 | 113.8 | 114.1 | 0.900 | 0.957 |
| 500 | 3000 | 138.3 | 139.0 | 0.933 | 0.933 |
| 500 | 3500 | 162.7 | 163.5 | 0.872 | 0.897 |
| 500 | 4000 | 187.3 | 186.6 | 0.787 | 0.845 |
| 1000 | 1000 | 44.5 | 46.3 | 1.723 | 1.574 |
| 1000 | 1500 | 55.7 | 55.5 | 1.470 | 1.723 |
| 1000 | 2000 | 73.8 | 72.0 | 1.640 | 1.649 |
| 1000 | 2500 | 95.5 | 93.6 | 1.237 | 1.450 |
| 1000 | 3000 | 118.2 | 117.9 | 1.229 | 1.227 |
| 1000 | 3500 | 141.7 | 142.6 | 1.221 | 1.080 |
| 1000 | 4000 | 165.6 | 165.6 | 1.204 | 1.109 |
| 1500 | 1500 | 54.7 | 55.3 | 2.074 | 2.023 |
| 1500 | 2000 | 64.0 | 64.2 | 2.211 | 2.039 |
| 1500 | 2500 | 80.7 | 79.9 | 1.752 | 1.800 |
| 1500 | 3000 | 100.6 | 100.1 | 1.764 | 1.652 |
| 1500 | 3500 | 122.6 | 122.6 | 1.480 | 1.466 |
| 1500 | 4000 | 145.2 | 145.2 | 1.375 | 1.430 |
| 2000 | 2000 | 63.3 | 64.9 | 2.433 | 2.212 |
| 2000 | 2500 | 71.3 | 72.8 | 2.198 | 2.220 |
| 2000 | 3000 | 86.7 | 86.6 | 1.967 | 2.118 |
| 2000 | 3500 | 105.4 | 104.9 | 1.649 | 1.960 |
| 2000 | 4000 | 126.5 | 126.4 | 1.944 | 1.800 |
| 2500 | 2500 | 70.7 | 72.5 | 2.625 | 2.448 |
| 2500 | 3000 | 78.0 | 78.6 | 2.777 | 2.538 |
| 2500 | 3500 | 92.4 | 91.3 | 2.339 | 2.468 |
| 2500 | 4000 | 109.8 | 110.4 | 2.179 | 2.212 |
| 3000 | 3000 | 77.7 | 77.2 | 2.938 | 2.827 |
| 3000 | 3500 | 84.2 | 83.4 | 2.784 | 2.895 |
| 3000 | 4000 | 97.5 | 98.2 | 2.617 | 2.659 |
| 3500 | 3500 | 83.7 | 82.7 | 3.158 | 3.149 |
| 3500 | 4000 | 90.4 | 90.8 | 3.342 | 3.134 |
| 4000 | 4000 | 89.5 | 89.2 | 3.450 | 3.628 |

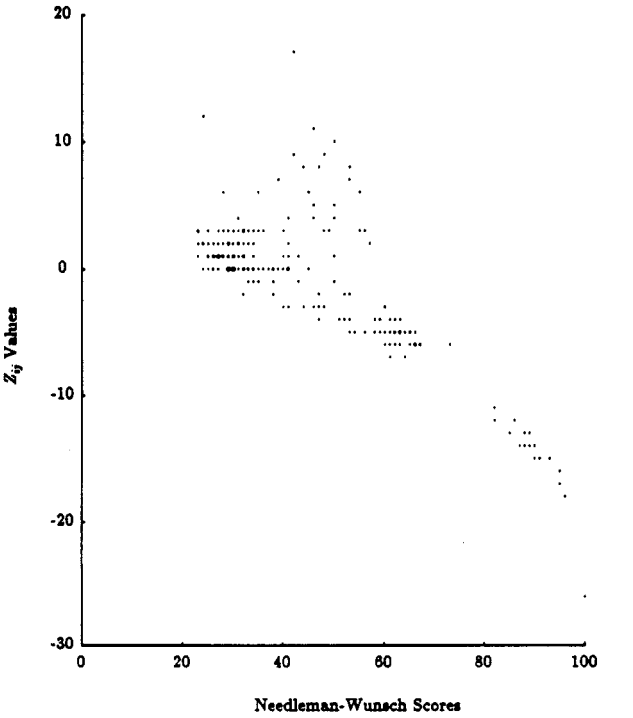


**Figure 1.** Correlation between the $Z_{ij}$ values and the Needleman–Wunsch scores.

## RESULTS AND DISCUSSION

To evaluate the validity of the $Z_{ij}$ values, they were compared with Needleman–Wunsch scores for the pairs produced

**Table III.** $Z_{ij}$ Values between Real Proteins and Pseudo Ones Produced by Adding Random Amino Acid Sequences

| length of seq added[a] | $Z_{ij}$ | | | |
|---|---|---|---|---|
| | CCCS[b] | LWLV6[c] | O4RTPB[d] | TVMVGM[e] |
| 0 | -26 | -25 | -24 | -31 |
| 0.5 | -11 | -13 | -14 | -14 |
| 1 | -9 | -10 | -11 | -8 |
| 1.5 | -7 | -8 | -8 | -6 |
| 2 | -6 | -6 | -6 | -4 |
| 2.5 | -5 | -5 | -5 | -4 |
| 3 | -4 | -5 | -4 | -4 |
| 3.5 | -3 | -4 | -3 | -4 |
| 4 | -3 | -4 | -3 | -4 |

[a] Expressed as ratios of lengths of random amino acid sequences added to those of real proteins. [b] Cytochrome *c*. The sequence length is 111. [c] ATPase, a chain. The sequence length is 248. [d] Cytochrome P450. The sequence length is 491. [e] Kinase-related transforming protein. The sequence length is 746.

**Table IV.** Similar Proteins of $Z_{ij}$ Values Less Than -11 to Protein A27776

| $Z_{ij}$ | protein nos. | names |
|---|---|---|
| -16 | SLONA1 | protamines (salmines) AI and AII |
| -16 | IRTR59 | protamines CII |
| -16 | IRTR1A | protamine (iridine) IA |
| -14 | IRTR2 | protamine (iridine) II |
| -13 | IRTR1B | protamine (iridine) IB |
| -12 | IRTR42 | protamines pRTP242 and pTP8 |
| -11 | IRTRC3 | protamines CIII and pRTP43 |

from the first protein with the following 150 ones in the protein database, and the results are shown in Figure 1.

Apparently the $Z_{ij}$ values decrease linearly with the Needleman-Wunsch scores in a field of $Z_{ij}$ values less than ca. -5 and the scores more than 60, but not in other fields.

The $Z_{ij}$ values between a real protein and pseudo ones produced by adding some lengths of random amino acid sequences to the end of the real one were calculated as shown in Table III. The $Z_{ij}$ values increase with increasing lengths of random amino acid sequences.

These results indicate that the distance $Z_{ij}$ can be used in the search of closely related proteins as well as the Needleman-Wunsch scores. As an example, the results of a search for similar proteins to protein A27776, Pretamin C11-Rainbow trout, are listed in Table IV.

In the next the execution time of the calculation of $Z_{ij}$ was examined.

Random amino acid sequences of lengths 100, 250, 500, 1000, and 3000 were produced, and the time to calculate $Z_{ij}$ between those sequences and proteins in the database was measured. It was found that the time was almost constant at 38 s, irrespective of the lengths. The total execution time needed to calculate $Z_{ij}$ and then rank proteins in increasing order of $Z_{ij}$ was 45-60 s.

In conclusion, the present system gives results similar to the previous ones, but quickly, in the search of closely related proteins, and thus it is found to be useful.

## REFERENCES

(1) Needleman, S. D.; Wunsch, C. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443-453.
(2) Lipman, D. J.; Pearson, W. R. Rapid and Sensitive Protein Similarity Search. *Science* **1985**, *227*, 1435-1441.
(3) Nakayama, S.; Shigezumi, S.; Yoshida, M. Method for Clustering Proteins by Use of All Possible Pairs of Amino Acids as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 72-78.

# Requirements for and Challenges Associated with Submission of Machine-Readable Manuscripts†

MARIANNE C. BROGAN‡ and LORRIN R. GARSON*
Publications and Operational Support Divisions, American Chemical Society, 1155 16th Street NW, Washington, D.C. 20036

Received April 10, 1990

Computer-assisted composition, first used for primary journals in the 1960s, became an important component of publishing in the 1970s and the dominant production method of the 1980s. During the last half of this decade, the availability of word-processing tools and the affordability of computer systems have made electronic submission of manuscripts common in many publishing applications and have intensified discussions of the practicality of electronic submission for other applications. Authors and publishers share a common interest but have different perspectives on requirement and expectations. Utilization of electronically submitted material at any stage of the publication process would significantly impact that stage, with peer review, technical editing, and production being the principal targets for change. Challenges exist in the translation of tables, mathematical expressions, chemical structures, and other graphics from a variety of word processors to formats required for composition. Database applications demand the additional requirement of data-element identification.

## INTRODUCTION AND BACKGROUND

Sixteen years ago at a meeting of the American Chemical Society (ACS), one of us (L.R.G.) was asked by an ACS member (who had just acquired a word processor): "When can I submit my manuscripts to the ACS on diskette?" The response was, "Well, I'm not sure, but it doesn't seem to be terribly difficult to me." Since 1974, that same question has been posed at every ACS meeting, at many journal editorial advisory board meetings, and on many other occasions.

Approximately 12 years ago, most manuscripts submitted to the ACS were prepared by typewriter. Today, a majority of manuscripts are prepared on a wide variety of word-processing systems. Virtually all scientific and technical publishers use high-end, computer-controlled composition systems. Thus, both authors and publishers produce and process manuscripts in machine-readable forms. Unfortunately, those forms are not compatible, which is a major barrier to practical, cost-

† Presented at the Symposium on Electronic Methods of Document Preparation and Information Exchange, Division of Chemical Information, 198th National Meeting of the American Chemical Society, Miami Beach, FL, Sept 14, 1989.
‡ Journals Department, P.O. Box 3330, Columbus, OH 43210.

0095-2338/90/1630-0271$02.50/0 