Polymer Information: Storage for Retrieval, or Hide and Seek? Wrapping Up†

CARLOS M. BOWMAN
The Dow Chemical Company, Midland, Michigan 48667

STUART M. KABACK*
Exxon Research and Engineering Company, P.O. Box 121, Linden, New Jersey 07036

Received July 7, 1991

It has been made clear from the preceding papers that many of the problems peculiar to polymer patent searching will not be solved until the linking between databases is improved. Further, it is clear that polymers are significantly different from "ordinary" chemicals and that searching of polymeric materials cannot be done with the same techniques which are used for normal organic structures. Some suggestions concerning search system requirements are offered.

This contribution encompasses some remarks regarding that part of the Symposium not submitted for publication; the organizer's thoughts regarding an overall theme; and a significant contribution to the discussion period, subsequently reworked for publication. Not to be published at this point is the contribution from Chemical Abstracts Service, whose authors felt it was premature to publish what was essentially a progress report. For the record let it be noted that the subject matter of their presentation focused on the development of polymer class terms and of linkages between monomer-based and SRU-based polymer representations in the CAS Registry.

To the organizer (S.M.K.) an important theme that surfaced again and again was the desirability of more sophisticated linking systems which would enable us to distinguish among similar and closely related systems, to pick up contextual relationships in complex systems. To some extent this is a reflection of my own bias, and expressed both by my own contribution and by the selection of participants for the Symposium. But I believe that this need for linkages transcends any personal bias and reflects one of the most important areas for development in polymer information today.

The discussion period elicited some interesting comments, most significant of which came from Carlos Bowman. His comments were provocative, and I do not necessarily agree with all of his viewpoints, but they certainly provide valuable food for thought. Dr. Bowman has kindly consented to expand on his impromptu remarks, which are reproduced verbatim here. I believe that they provide a fitting conclusion for the Symposium.

POLYMERS ARE NOT ORDINARY CHEMICALS (C.M.B.)

After listening to the papers presented at the Symposium "Polymer Information: Storage for Retrieval, or Hide and Seek?" I am prompted to raise several points that I believe are essential if we are ever to provide the type of system that those seeking information about polymers really need.

Undeniably, polymers are chemicals, but they are different in a number of ways, and until these differences are recognized storage and retrieval systems are doomed to be cumbersome and will not yield the needed information. I am concerned that current systems have simply tried to apply the very same techniques that have been successful with single chemical substances without recognizing the differences. It is not difficult to characterize a single chemical. All that is needed is an accurate characterization of its chemical structure, and then one can easily discriminate between one material and another. The usual one-to-one correspondence between a chemical structure and a systematic name has provided additional avenues for precise and complete retrieval.

Recognition of the need to group substances according to their chemical characteristics prompted the development of schemes for retrieving substances with similar functionality. In the early days this was done by coding each compound by some predetermined set of codes and retrieving by looking for the presence or absence of the appropriate codes. This worked well until the number of substances increased, and it was realized that such coding schemes did not provide much information about the connectivity between the fragments. To remedy this situation, these schemes were modified with varying degrees of success. Others invented more detailed schemes that coded complete structures in a manner similar to chemical nomenclature. The symbols used were much abbreviated, and the resulting notations provided reasonable representation of the structure. These developments were useful but were an artifact of the limitations of the computing technology of the time.

As computers became larger, more powerful, and the cost of storage decreased, schemes were developed for representing the chemical structure in its entirety with a graphical interface that allowed the user to use the universal language of chemists, structural diagrams. Questions are framed using substructure drawings, and results are presented in a form familiar to the chemists. This approach has been very successful, and most systems today provide for this type of retrieval. Unfortunately, there are still some large information systems that have not yet made this paradigm shift and are not as useful as they could be.

We face with polymers the need for a further paradigm shift. Structures alone cannot be used to uniquely and unambiguously identify a polymer or a polymeric mixture. The main reason for this is that the actual structure of a polymer is usually not known. If we could accurately represent the entire atomic arrangement of a polymer, we could be assured of accuracy, but such a task is impossible.

While it may be possible to draw a structure for the starting materials (monomers) or the structure repeating unit (SRU), we are unable to detail the exact number of repeating units or the arrangement of the monomers within the polymer. Thus we need to turn to other information about the polymer to accurately characterize it. This information can be divided into several general categories.

Chemical Structure. We need to detail all we know about the chemical structure of the monomers and/or the SRU.

Molecular Mass. An indication of the size of the molecules in the polymer will further distinguish a polymer. Care must be taken to distinguish the type of molecular mass being reported, such as, number average, weight average, etc.

Molecular Arrangement. Information about the tacticity or other stereo arrangements is essential. In the case of copolymers the type of copolymer should be stated, such as,

† Conclusion of the Symposium of this name, presented at the 201st National Meeting of the American Chemical Society, Atlanta, GA, April 16, 1991.

0095-2338/91/1631-0482$02.50/0

alternating, block, graft, random, etc. End groups should also be reported. If the polymer has been chemically modified, the type of modification and its extent is needed.

Physical Properties. Physical properties such as transition temperatures, swell index, and viscosity give information about the nature of the polymer. The polymer should also be characterized as to whether it is thermoplastic or a network-type polymer.

Preparation. Further information that is relevant to the characterization of a polymer is the method used for its preparation. An indication as to whether it was synthesized by bulk, solution, or emulsion polymerization is important. Also whether it is anionic or cationic, etc. Catalysts and solvents used should also be indicated.

Postprocessing. Processing of a polymer after its preparation can affect its properties. Information as to such processing also needs to be included.

By enumerating all these different types of information I have tried to make the point that we have to view polymers as significantly different from "simple" organic chemical substances. We must provide for a means of representing polymers and their mixtures in such a way that all this information can be associated together and searched in conjunction with one another.

Such a change will require a major shift in the way we index substances and in the software that will be needed to retrieve them. However, unless we do so we will continue to struggle with large numbers of unwanted hits and missing important information.

The papers in the Symposium show that of the major vendors, only one is really starting to think about approaching polymer information in a manner somewhat different from "simple" organic substances. Others have not yet made the shift to chemical structures, much less to integrated information. Coding schemes are not made better by increasing the number of codes. A different approach is needed.

The work reported by Molecular Design Limited shows promise as a start toward integrating information in such a way that accurate searches can be made. Unfortunately, such an approach is not being pursued by any of the major vendors. Although individual companies and scientists can use this approach for their own files, the methodology must be applied to the open literature and eventually to back files. The prospect for this is not very promising.

We must shift out of our present thinking and approach polymers as complex collections of information. If we adjust ourselves to this concept, innovative solutions to the problem will emerge and in years to come we will be able to retrieve information with the precision that we can now obtain for "simple" substances.

Random Walks: Computations and Applications to Chemistry

A. S. SHALABI

Department of Chemistry, Faculty of Sciences, United Arab Emirates University, P.O. Box 17551, Al-Ain, United Arab Emirates

Received December 18, 1989

The concept of random walks (self-returning and self-avoiding walks) in molecular graphs is reconsidered. An algorithm that allows the manual calculation of self-avoiding walks is proposed. Neither self-avoiding walks or self-returning walks provide a totally reliable basis for property prediction, but self-returning walks are generally superior. A program which calculates a similarity matrix and which is dimensioned for up to 50 × 200 entries has been written in Basic for use on IBM PCs, and its use in the determination of similarities of different structures is described. A method is suggested for the calculation of the diagnostic power of a graph theoretical invariant in characterizing structures. Ring closure effects in relation to graph coloring problems are investigated, and applications to total π and ω-π electron energies and physicochemical properties are considered.

INTRODUCTION

Characterization of structures is a central problem in chemistry and mathematics.[1-5] As a basis for characterization, some different graph theoretical invariants have been examined. The concept of random walks (RW) developed by Randić[6] as a graph theoretical invariant deserves special attention because it provides a unique characterization for atom environments, is easily applied, and allows the verification by means of reconstruction algorithms. Definition of the concept of similarity, which may apply to a selected property, a dominant property, or the overall features of the system under investigation[7-9] requires the selection of a graph theoretical invariant upon which comparisons among structures can be based.

In this paper, some conceptual definitions for random walks are proposed. These suggest a pencil-and-paper algorithm for the calculation of self-avoiding walks (SAW) in simple molecular graph structures that have real structural correspondences.[10] The potentials in characterization of molecular graphs of self-avoiding walks and self-returning walks (SRW) are compared. A similarity matrix program has been written to facilitate such comparisons. These comparisons have allowed the measurement of the relative diagnostic power of SAWs and SRWs, and it has been observed that, as predicted by Randić,[6] the SRWs give a superior performance. The calculations of SAWs and SRWs were accompanied by the observation of some subtle relationships between the ring closure process and the coloring system of the graph structures investigated.

These calculations and similarity comparisons have been compared to the similarities derived from the measured physical and chemical properties of two groups of structurally related chemicals. The first group consists of the isomers of hexane, and the second is a group of alkylbenzenes. This study can be considered an application of one of the Randić similarity measures to two sets of chemical compounds.

METHODS

There is no general agreement upon the terminology that is used for many graph theoretical concepts, and subsequently,

0095-2338/91/1631-0483$02.50/0