(22) (a) Gutman, I.; Trinajstić, N. *Chem. Phys. Lett.* **1972**, *17*, 535. (b) Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, C. F., Jr. *J. Chem. Phys.* **1975**, *62*, 3399.
(23) Randić, M. *J. Chromatogr.* **1978**, *161*, 1.
(24) Hosoya, H.; Kawasaki, K.; Mizutani, K. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 3415.
(25) Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419.
(26) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226.
(27) Kier, L. B. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 7.
(28) Kier, L. B. *Acta Pharm. Jugosl.* **1986**, *36*, 171.

End-User Searching of CAS ONLINE. Results of a Cooperative Experiment between Imperial Chemical Industries and Chemical Abstracts Service

WENDY A. WARR*,† and ANGELA R. HAYGARTH JACKSON‡

Information Services Section, Imperial Chemical Industries PLC, Pharmaceuticals Division, Alderley Park, Macclesfield, Cheshire, SK10 4TG, United Kingdom

Received September 24, 1987

Chemical Abstracts Service staff trained 88 Imperial Chemical Industries chemists to use CAS ONLINE over a 2-year period in a cooperative experiment between the two organizations. The effectiveness of the training, the problems encountered by end-users, the usage made of CAS ONLINE, the impact of end-user searching on information scientists, and the attitudes of management were studied.

BACKGROUND

Late in 1982, after 10-years' experience of online interactive searching by ICI information scientists and librarians, one of us gave a paper that stated categorically that the end-user will search online.[1] This opinion was just beginning to obtain credence in the literature.[2-5] However, the ICI view was mainly based on developments within the company in which a policy was being introduced to allow scientists, and others, to access internal company databases themselves, interactively, without the intervention of an intermediary who would interrupt the flow of scientific thought.[6] It was recognized that given suitable terminal and network facilities, most scientists would also wish for direct access to the external databases, representing the published literature.

However, at that time, the online search services were even less "user-friendly" and the telecommunication links less reliable than today. It was recognized that mastery of the intricacies of successful searching took time, training, and experience to acquire. There was top management concern relating to the costs involved in both the end-user scientist's time spent searching and the actual online charges. In addition, management was concerned about the mechanisms for managing this new online information resource, in that a free-for-all approach, in effect an open check, was not an acceptable policy. The information scientists were worried that their knowledge and skill in accessing online services would no longer be required.

Other companies have considered the same issues.[7,8]

INITIATION OF THE ICI CAS ONLINE COOPERATIVE EXPERIMENT

Informal discussions were started with Chemical Abstracts Service (CAS) in 1984. ICI felt that chemists were the most suitable end-users to be trained first because they already had significant experience of using an in-house, interactive system,[6] and their needs for external information could be substantially met by access to the online databases provided by CAS. ICI information scientists were accessing these databases under DARC[9] and CAS ONLINE[10] in the ratio of about 1 to 2 in cost terms, but the company wished to train chemists in one system only. The likely development of both search services was an important issue, but the offer by CAS to train groups of ICI chemists to use CAS ONLINE helped to clinch the matter.

JOINT CAS AND ICI OBJECTIVES FOR THE COOPERATIVE EXPERIMENT

The aim of the cooperative experiment between CAS and ICI was to gather information on end-user chemists' needs and usage of CAS ONLINE. The results would, it was hoped, assist CAS in their developments and marketing of CAS end-user chemists' to industry and would assist ICI in the provision and management within the company of CAS ONLINE as a resource to chemical innovation and chemistry in general. The joint specific objectives of CAS and ICI were as follows:

(1) To determine how effective CAS training was for both end-users and the information scientists who supported them and to assess the appropriateness of related documentation and user manuals.

(2) To find what problems end-users encounter.

(3) To determine the number and type of searches done by an average end-user.

(4) To study the effectiveness of end-user searching.

(5) To determine the types of searching problems that cause an end-user to seek help from an information scientist.

(6) To study the reaction of information scientists to end-user searching.

(7) To study ICI management reaction, including the opinions of both chemistry and information managers on the cost effectiveness of end-user searching and the developments needed from CAS if further progress were to be made.

(8) To learn how CAS might better serve the needs of end-user scientists.

MAIN TERMS OF THE COOPERATIVE EXPERIMENT

The experimental period was 25 months from December 1984 to December 1986. The main terms were as follows:

(1) ICI was to make two advance lump-sum payments to the Royal Society of Chemistry (U.K. marketing agent for CAS ONLINE) against the use of CAS ONLINE by ICI in the U.K.

† Manager.
‡ Former Manager.

0095-2338/88/1628-0068$01.50/0

(2) CAS staff were to train approximately 50 ICI chemists to use CAS ONLINE. In practice, on several courses CAS staff were supported by expert staff from the Royal Society of Chemistry.

(3) ICI was to provide in-house expert information scientist CAS ONLINE searchers to support and help ICI chemists.

(4) ICI was to supply detailed quarterly reports to CAS concerning the use of CAS ONLINE by ICI chemists.

(5) CAS staff were to conduct personal interviews with the chemists they had previously trained to ascertain more information about the practicality and utility of end-user searching of CAS ONLINE and to identify changes that would make CAS ONLINE a better service for end-users.

(6) ICI was to provide an unique login ID for each chemist using CAS ONLINE and would request that the chemists should not share or trade login IDs.

RESULTS OF THE COOPERATIVE EXPERIMENT

CAS Training. CAS staff trained a total of 88 ICI chemists working in the U.K., most of whom were Ph.D chemists working in research, process development, and patent departments. This was an ICI Group project, and participating chemists came from various divisions of ICI with 51 from Pharmaceuticals Division, 24 from Plant Protection Division, 10 from Organics Division, and 3 from other ICI Divisions.

CAS staff ran four sets of courses, with from 7 to 10 participants per course, in November 1984, November 1985, June 1986, and December 1986. Courses were run at both Pharmaceuticals Division and at Plant Protection Division sites on every occasion except December 1986, when an additional course (the "Train-the-Trainer" one described later) was run at the Pharmaceuticals Division site. All the trainees considered the quality of training by CAS staff to be excellent. The main criticisms related to the CAS documentation following the early courses. This was generally considered to be too voluminous and detailed and without appropriate indexing. When major CAS enhancements (e.g., variable groups, name, and formula searching) were introduced in March 1985, an ICI information scientist found it necessary to abstract the CAS documentation and to issue an ICI end-user summary. In November 1985 CAS introduced the MENU facility: a menu of "buttons" to facilitate the graphic input of structures. ICI again wrote its own end-user documentation. CAS had written its documentation with professional searchers in mind, not end-users.

ICI has many more than 88 scientists potentially interested in searching CAS ONLINE, and it was deemed sensible for ICI information scientists eventually to take over the role of training further chemists. CAS designed a Train-the-Trainers course to teach information scientists how to train end-users. We are unable to comment on the effectiveness of this formal CAS course because what was actually run in ICI in December 1986 was an informal event specifically geared to ICI needs. CAS and ICI trainers had worked together for over 2 years, and the ICI staff had therefore already absorbed most of the expertise needed.

End-User Problems. As ever with the benefit of hindsight, it may be that the experiment started some 6 months too early. Fortuitously, the experiment started just as ICI was making a major change from the use of terminals to the installation of IBM PCs. Ready access to a terminal (or terminal emulator) for some chemists was thus delayed. Also, it took longer than we envisaged to install the local area network in research departments. The experiment showed that chemists trained to use CAS ONLINE and the in-house databases must have ready access to appropriate equipment and that most of those who did not search after training did not have easy access to a terminal or were "too busy".

It should also be noted that MENU facilities for structure building did not appear until November 1985. The first set of chemists trained were taught to build structures as on nongraphics terminals using English language commands, the so-called text structure input method.

The main problems encountered by chemists in the early stages of the experiment arose from the use of equipment and, in particular, communication links and the inability at the time to store logging-on procedures. These were the main problems involving help from the information scientists. It was apparent from the early stages of the experiment that end-user chemists are less patient with system disconnects than information scientists. Reliable online search services and telecommunications are essential if end-users are to be encouraged to run their own searches.

These problems, and similar ones encountered in the use of in-house systems, precipitated the development of an "End-User Support" team of information scientists. The concept of end-user support is an extension of the "Help Desk" idea used by many vendors of online systems. The end-user at ICI should not be expected to deduce whether the failure of his system is due to a software bug, a firmware fault, a hardware failure, local area network downtime, or whatever. Each of these problems is ultimately the responsibility of a different team within ICI or may even require intervention by someone outside the company (for example, an engineer from the hardware supplier). The ICI scientist wants one "port of call" with his problem. He telephones someone in the End-User Support team, and the information scientist who answers either solves the problem himself or promptly finds someone else who can produce a solution. In some cases, of course, the problem is actually routed to an information scientist of the online searching unit (for example, if the scientist is confused by CAS ONLINE tautomerism conventions).

This is an oversimplification of the role played by the information scientist in end-user support because it covers only his responsibilities in relation to CAS ONLINE. Nevertheless, the formation of an End-User Support team was a very significant development.

Two information scientists at Pharmaceuticals Division specialize in end-user support for CAS ONLINE as part of their duties. Excluding their heavy involvement in the initial teething-trouble stage, end-user support over the 2 years totaled about 2–3 man-weeks on documentation, 2 man-months on training, 8 man-days on "retraining" (intensive help to those who found searching very difficult at first), and 88 man-days on generalized help desk tasks. (By 1987 the need for support in general has probably halved.)

Chemists' Acceptance and Use. The chemists were extremely keen to participate in the experiment and to attend the CAS training courses. In fact, the enthusiasm and competition to attend training courses were such that the chemistry line management had to make the selection.

The chemists found references online that they had not found in hard-copy *Chemical Abstracts*. At first they found searching more time-consuming than going to an information scientist intermediary to do the search for them, but the advantage of do-it-yourself searching was the ability to make immediate search and topic refinements.

The types of searches made by chemists included (1) chemical structure display and substructure search, for synthetic methods, for lead exploration, and for prior art searching; (2) exact chemical structure matching, for novelty checking and to find preparative methods; (3) keyword search, for biological activity and synthetic methods; (4) author search; (5) chemical name match searches for references.

By December 1986 it was clear that CAS-trained chemists at Pharmaceuticals Division could be divided into three groups

Table I. Usage of CAS ONLINE by Pharmaceuticals Division Chemists in the Last Quarter of 1986

	no. of substructure searches	no. of family/exact searches	no. of CA File searches
14 chemists in first group	34	29	20
15 chemists in second group	124	44	30
10 chemists in third group	9	4	15

of approximately equal size. The heavy users run 5 searches a month or more. (One ran 63 searches between September and December 1986.) The average user runs 5–15 searches in three months. The infrequent user runs one search a month or less.

Of the 88 people trained in all ICI Divisions, only 5 have never used CAS ONLINE after training.

Physical chemists are infrequent users; synthetic organic chemists are heavier users.

The first group of Pharmaceuticals Division chemists to be trained continued to be average or infrequent users over the whole 2 years, whereas the second group contains many heavy users. This is almost certainly because the first group were management-selected, whereas the enthusiasts found their way onto the second course a year later. A comparison of three groups (the fourth was not trained until December 1986) for the last quarter-year of the experiment is shown in Table I.

The first group (trained in November 1984) includes one chemistry manager who has transferred to ICI Americas and one scientist who has been promoted to a managerial job in another ICI Division. Neither have searches recorded in Table I. The group also includes one patent agent and one scientist who has become a trainee patent agent. Both of these are infrequent users.

The second group includes 2 process development chemists who have never used CAS ONLINE and 1 physical chemist who makes low usage of it, but the other 12 members of the group are average or heavy users.

The third group was not trained until June 1986, and Table I does not show an established pattern for this group.

Gradual growth of end-user searching and the effect it had on searching by intermediaries is seen in Plant Protection Division's figures, illustrated in Figure 1. Plant Protection Division chose their first set of users on grounds of computer aptitude or regular usage of other systems.

Organics Division users were probably the most cost conscious. The research chemists at that division use CAS ONLINE for browsing or testing out ideas, but one process development chemist there who turned to CAS ONLINE to solve specific problems found the cost effectiveness of his searching more obvious.

At first CA File was used more than the Registry File. We wondered if this was due to what more than one user called the "complex and time-consuming procedure of building structures for the Registry File". However, by December 1986 time spent searching CA File had fallen to only one-third of the total usage time. Heavy users make more use of substructure search than do average users, who have a higher proportion of CA File accesses. The figures in Table I also suggest that inexperienced users (see third group) make proportionally heavier use of CA File. Of course, the usage of CA File and Registry File should depend on what *type* of information is needed, not on chemists' prejudices about ease of use, but we suspect that our figures do not reflect the true need for use of CA File information.

There have been complaints about graphics structure input. Many chemists report that they use only text structure input because graphics input is "harder work and slower". Others do not use graphics input because "it is of poorer quality than our in-house system". More than one chemist wishes to actually input CAS substructure searches using the in-house system.

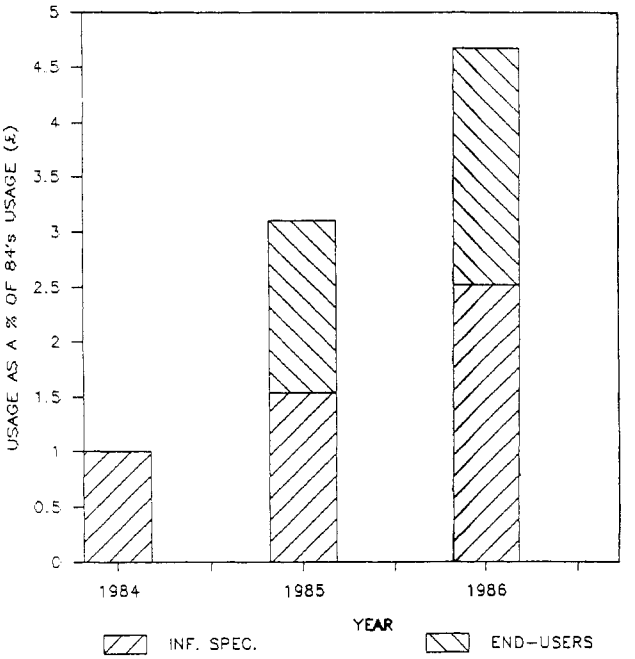

Figure 1. CAS ONLINE usage at ICI Plant Protection Division.

End-users request fewer offline prints than information scientists do. Unlike the intermediary, they can select the references of greatest interest, and they tend to browse through the hits online. Chemists like the immediacy of online hits. Having done a search, enthusiastic chemists want to view their results straightaway.

Role of Information Scientists and Their Reaction to the Experiment. In general throughout the experiment it was stressed that "a little knowledge can be a dangerous thing". The chemists accepted the concept of a two-tier level of searching[1] whereby searches affecting business or project decisions, exhaustive prior art searches, and complex searches were best carried out by an expert information scientist. It was expected that the information scientists would continue to be asked to do complex searches. To some extent this has proved to be true. However, it is difficult to draw conclusions from searches that are never seen in the Information Unit, and some information scientists at Pharmaceuticals Division are concerned that they do no searches at all for many CAS ONLINE end-users. Certainly, patent agents will always demand that information scientists do a full search when a patent application is being prepared, whether or not the chemist claims he has carried out a search. On the whole, chemists are less concerned about comprehensiveness than information scientists are. Chemists *do* appreciate their limitations, but in many cases they are happy with a selection of leading references or a more restrictive search than the information scientist would have done. Buntrock[8] suspects that the chemist often really does need the comprehensive search.

Information scientists are invaluable for end-user support, for supplying information on system updates, and for supplying user-orientated documentation. The chemists realized that the information scientists really had considerable expertise, and as a result so far as chemists were concerned the status of the information scientists increased. Furthermore, the improved knowledge of online searching acquired by chemists helped in query definition for complex online searches, and an appreciation of the capabilities of online searching produced more interesting and challenging questions for information scientists. The number of information scientists has increased due to the increased visibility of the importance of online

searching. Fears among information scientists that end-user searching would make their jobs redundant have been dispelled, and job satisfaction has been enhanced for those who are prepared to accept a different type of role.

Plant Protection Division's figures clearly show that information scientists will not lose their jobs or their online searching skills for lack of practice. The figures show that (in terms of costs) CAS ONLINE usage more than trebled between 1984 and 1985. There was virtually no end-user searching in 1984. In 1985 almost exactly half the total costs were incurred by information scientists and half by end-users. In 1986, usage (in terms of costs) was more than 4 times that of 1984. Information scientists were incurring about two-thirds of the costs. Information scientists spent well over twice as much on CAS ONLINE in 1986 as in 1984. This is illustrated in Figure 1.

Management Reaction. At first some chemistry managers were cautious. There was concern relating to the time chemists would spend accessing online systems and the possible costs.

After the end of the experiment the chemistry manager most involved at Pharmaceuticals Division sent us a report stating that CAS ONLINE is ranked alongside the most important in-house system (i.e., the internal system for searching chemical and biological data) in terms of value to the end-user. He claims that usage will continue to increase.

He emphasizes that two developments should now be sought. First, recording of sessions should be permissible without charge. Second, in his own words "...a user-friendly front-end, preferably widely applicable to other systems, would be a real advantage and would facilitate the general acceptability of online searching by chemists. Such a system should allow the construction of search queries offline which would help contain costs".

In summary, he says that his attitude to the experiment over the years has not changed. He always thought it was a good idea, and this has been confirmed by experience.

(At this point it should be mentioned that the chemistry manager may not have known CAS policy on downloading. CAS does permit certain types of downloading at no charge. For example, a searcher may capture and edit an electronic copy of search results to produce a paper copy for his files. For an annual fee, long-term storage and reuse of CAS data is also permitted through a special downloading license.)

A second chemistry manager is more cautious about costs. He supports the use of CAS ONLINE but is concerned that some chemists spend more on CAS ONLINE per year than they do on consumable stores (e.g., purchasing chemical intermediates). He says the CAS ONLINE facility is not cheap, and the chemistry department needs to exercise management control and develop a departmental policy on end-user searching.

It is significant that in Pharmaceuticals Division, Research Department is not bearing the costs of CAS ONLINE. Information Services Section pays. There are a number of reasons for this, among them administrative considerations and the enthusiasm of the authors of this paper. In the other divisions, chemistry management pays and is much more cost conscious.

Information managers in the company are in favor of end-user searching if the appropriate safeguards are applied.

One of them states that, as an Information Manager, his job is to manage the dissemination of information within the company in an efficient and effective way such that it is used ultimately at the right time in the right place to benefit and to influence business decisions. *How* that is achieved requires getting a balance of all the people, political, financial, and technical issues. We must offer the company a solution that gives the maximum benefit to ICI and is seen to be so in the eyes of everyone (senior managers, users, and online information service staff). His staff know they cannot ever become more expert on the subject matter than lawyers are on law, economists on economics, and so on. However, not all of the subject experts will learn all the relevant online systems, and they will always depend on general systems experts for a service. "We all drive cars but we all take taxis and other forms of transport when it is most relevant or convenient".

An information manager more closely involved with the CAS ONLINE experiment says that he is in favor of end-user searching of CAS ONLINE and the company benefits. However, it is important to establish the boundaries between experts and end-users. (To this end both Pharmaceuticals Division and Plant Protection Division have produced a two-page set of guidelines that were sent to all chemistry managers and end-users.) In his opinion the experiment has enhanced the jobs of information scientists involved with CAS ONLINE.

CONCLUSIONS

CAS ONLINE has become one of the most vital computer systems accessible by ICI's scientists (some would say *the* most vital one). It is no longer used on an experimental basis—new terms were agreed for 1987, and CAS ONLINE is now part of the research culture. The demand for training is high, and many chemists are quite status conscious about being able to use the system. Five 2-day courses have been run at Pharmaceuticals Division in 1987. Refresher courses and training for new features are supplied on an ongoing basis.

ISSUES FOR THE FUTURE

The method of paying for CAS ONLINE usage is somewhat complex and not necessarily equitable to all ICI Divisions. The complexity relates to the combining of various CAS services and products used unequally by the various ICI Divisions. A new way of handling this needs to be negotiated for 1988. We need to devise a sensible pricing structure that encourages volume usage.

Thought will have to be given to letting end-users have access to other databases on STN. (STN International is a network that offers a number of scientific and technical databases including those of CAS.) Cost is not the only issue here. The problems of end-user searching of nonchemical files are different from those of end-user access to CAS ONLINE. ICI biologists, who would need access to many databases, not all of them on STN, might be unwilling to learn multiple search languages and would have problems choosing the best database(s) for each search.

The issue of downloading from CAS ONLINE has yet to be tackled, and this opens up questions about the user interface in general. One of us has already published views on the increasing number of graphic structure entry protocols that people are faced with learning.[11] It would obviously be in ICI's interests if the end-user could have just one method for accessing internal and external chemical structure and reaction databases, for producing reports and documentation, and for communicating electronically. Different companies would have varying views on which interface should be used. CAS ONLINE is of sufficient importance to ICI that CAS compatibility should be a major factor in the choice of systems in future.

CAS has responded to demands for a more user-friendly interface to CAS ONLINE in the form of the new front-end STN Express, announced in December 1987, which incorporates Hampden Data Services's chemical structure drawing interface, already an integral part of a range of PC products. One of the advantages that ICI has gained from establishing a fruitful working relationship with CAS has been the opportunity to participate in the alpha-testing phase of STN

Express. This development opens a new phase in chemical structure searching that should be the subject of another publication.

ACKNOWLEDGMENT

We thank Chemical Abstracts Service for their collaboration and help throughout this experiment. The number of ICI information staff involved with the experiment is large. Without their support the experiment could not have taken place, and their involvement is most gratefully acknowledged. In particular, we would thank Denise Ledgerwood, Colin MacBean, Malcolm Wilkins, Graham Cousins, and Duncan Adshead. Finally, we thank the end-user chemists and the chemistry management for their enthusiasm and interest in the project.

REFERENCES AND NOTES

(1) Haygarth Jackson, A. R. "Online Information Handling—the User Perspective". *Online Rev.* **1983**, *7*(1), 25-32.
(2) Meadow, C. T. "Online Searching and Computer Programming, Some Behavioral Similarities (Or ...Why End-Users Will Eventually Take Over the Terminal)". *Online (Weston, Conn.)* **1979**, *3*(1) 49-52.
(3) Richardson, R. J. "End-User Online Searching in a High Technology Engineering Environment". *Online (Weston, Conn.)* **1981**, *5*(4), 44-57.
(4) Faibisoff, S. G.; Hurych, J. "Is There a Future for the End-User in Online Bibliographic Searching?" *Spec. Libr.* **1981**, *72*, 347-355.
(5) Haines, J. S. "Experience in Training End-User Searchers". *Online (Weston, Conn.)* **1982**, *6*(6), 14-19.
(6) Adamson, G. W.; Bird, J. M.; Palmer, G.; Warr, W. A. "Use of MACCS within ICI". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 90-92.
(7) Buntrock, R. E.; Valicenti, A. K. "End-Users and Chemical Information". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 203-207.
(8) Buntrock, R. E. "Chemical Searching for People Who Hate Chemical Searching". *Database* **1985**, *8*(2), 82-83.
(9) Attias, R. "DARC Substructure Search System: A New Approach to Chemical Information". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102-108.
(10) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. "The CAS ONLINE Search System. 1. General System Design and Selection, Generation and Use of Search Screens". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93-102.
(11) Warr, W. A. "Online Access to Chemical Information: A Review". *Database* **1987**, *10*(3), 122-128.

Method for Clustering Proteins by Use of All Possible Pairs of Amino Acids as Structural Descriptors

SHIN-ICHI NAKAYAMA,* SATOKO SHIGEZUMI, and MASAYUKI YOSHIDA

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

Received September 21, 1987

Proteins were represented as vectors, of which components were all possible pairs of amino acids. From a distance matrix between any pairs of proteins thus represented, several clusters corresponding to connected components were generated. Application of this method to three different sets of proteins showed that it was suitable for clustering closely related proteins with respect to the sequential similarity defined by Dayhoff.

INTRODUCTION

Since sequence data of proteins are believed to have been retained dynamically over evolutionary processes, much attention has been paid to exploring the evolutionary relationships among biological species by sequential similarity between proteins. As a result, various methods of measuring the sequential similarity have been designed.[1] Most of these methods, however, include laborious steps of aligning all or parts of protein sequences to measure the sequential similarity, and thus, a similarity matrix for a large quantity of proteins is not easily obtainable.

Meanwhile, Nishikawa and Ooi expressed proteins as points in a composition space of amino acids and classified them into four groups of intra- and extracellular enzymes and nonenzymes according to the analysis of distribution of points.[2] The method is simple, but composition of amino acids alone is not sufficient to represent structural features of proteins. We expressed proteins using all possible pairs of amino acids as structural descriptors and clustered them on the basis of an easily obtainable distance matrix.

METHOD OF CLUSTERING

If a protein of chain length n were divided to $n - 1$ binary fragments, a set of occurrence counts for each species of binary fragments would form a specific pattern to the protein. The pattern of the protein, however, should differ from that of another one unless the two proteins have the same structure. In some instances the fragments found in one protein may not be included in another one. Thus, all possible pairs of amino acids, which numbered 400, were taken as descriptors, and protein i (P_i) was then represented by a set of descriptor values as $P_i = (x_{i1}, x_{i2}, ..., x_{i400})$, where x_{ik} is the occurrence count for the kth descriptor of the ith protein and is readily derived from the one-dimensional structure of protein i.

Although many different methods are available to cluster a set of proteins represented above, most known methods use distance measurements between each pair of proteins in the set. Thus, for a data set comprising n proteins, a symmetric $n \times n$ distance matrix was generated, the elements of which, d_{ij}, were the distance values between each pair of proteins i and j. In the present work, the Euclidean distance measure was chosen because of its wide use in many areas.[3]

The Euclidean distance measure is considerably affected by scaling factors, and standardization of data is common practice. However, since the descriptors used here were similar in property and standardization was apt to reduce between-group discrimination, the distance measured was used without further standardization.

To produce clusters among proteins the d_{ij} values were ordered by an algorithmic two-dimensional sorting operation to give a rearranged distance matrix.[4]

The clustering process generally consists of fixing a threshold T value in the d_{ij} values and grouping all pairs of objects whose d_{ij}s are less than a chosen threshold. Obviously, the d_{ij} values measured here are just numbers that are the complex function

0095-2338/88/1628-0072$01.50/0