Outline of Activities as Distributed Bioinformation Centre, University of Poona

A.S. Kolaskar,
Distributed Information Centre,
University of Poona,
Pune, INDIA.

In recent years several technological advances in the area of life sciences in general and biotechnology in particular have triggered quantitation in biology. Large numbers of computer readable databanks are becoming available, either in the public domain or through private vendors. Databanks on nucleic acids sequences, protein sequences, and protein crystal structures have proved to be of enormous use in modern biology.

India has realised the importance of making available such databanks to their researchers, entrepreneurs and planners. In 1985-86, the Task Force in Bioinformatics was formed to plan and set up a nationwide Bioinformatics Network.

The Distributed Information Network was established with the financial support from the Department of Biotechnology, Government of India, by identifying nine centres in different parts of the country.
Each centre was assigned a specific task and each is connected to the other through the National Informatics Centre Network (NICNET). NICNET is a government network, having its mother earth station at Delhi. It uses the Indian Communication Satellite INSAT-1 D to communicate with every district head-quarter (about 500 districts).

The Bioinformatics Network is superimposed on this government network and allows biotechnologists to access information. All these centres also have access to INTERNET. The following is a brief report of the activities at our Distributed Information Centre in the University of Poona.

The mandate of this centre includes the following main tasks:

(1) Create automated systems for storing and analysing information about animal viruses and animal cell tissue and organ cultures;

(2) Perform research into advanced methods of computer based information processing for analysing structure and function of biological macromolecules;

(3) Facilitate the use of databases and software required by biotechnology and life sciences researchers and students;

(4) Co-ordinate efforts in gathering biotechnology information worldwide and

(5) Train biotechnologists in particular, and life scientists in general, on the use of computerised databanks and networking.

The Distributed Information Centre at Pune, with the limited staff available, has carried out data and software development in two steps:

(I) Contacts are created with database managers; updates from those who are working in non-profit institutions and whose products are available in public domain through international agencies such as CODATA (Committee on Data for Science and Technology of the International Council of Scientific Unions), GenBank, Protein Identification Resource (PIR), Protein Structure Databank (PDB — three dimensional structure of biological macromolecules) are obtained at regular intervals either on magnetic tape or on CD Rom from USA. Similarly, Nucleic Acids Sequence Data and Swiss-Prot (Protein Sequence Data) are updated regularly with the help of European Molecular Biology Laboratory (EMBL), Heidelberg.

Several other databanks are obtained at regular intervals from international agencies in computer readable form and in printed version form.

(II) Similarly, to analyse this data, software packages which run on IBM Compatible PC-AT systems or on MicroVAX-II are obtained along with their updates.

In-house databank creation activity:

Internationally there are few efforts to create and update computerised information in the area of animal viruses. It was therefore decided to create Computerised Animal Virus Information System.

Presently, we have World's largest computerised animal virus data which is available in two formats: (a) DBASE III+ format; (b) MicroIS format (MicroIS is a database management system developed for microbiologists by the scientists at Microbial Systematics Section, National Institute of Dental Research [NIDR], National Institutes of Health [NIH], USA). There are several unique features which exists in our databank. Information is also stored in the pictorial form by using powerful data compression routines.

The open ended numerical coding system has been developed which can be used for objective taxonomic classification of viruses. In fact the ICTV (International Committee on Taxonomy of Viruses) has appreciated the work done by this centre.

At present information on 537 arboviruses is coded and available to the users. In addition, information on more than 200 viruses from other classes is also available.

The coding system that we have developed is very similar to the one developed by M. Krichevsky and R. Colwell for microbes (RKC code). As far as possible most of the information is coded in the binary form.
Only when information cannot be put in the binary form, are numerical and character codes used. Due to this numerical coding, the space occupied by the data becomes very small and, during retrieval, one can get answers to the queries with almost zero noise.
In this process we have also created a computerised dictionary of standard terminologies which gives a large number of synonyms and acronyms.

In order to create animal cell culture information system we have sent questionnaires to a large number of scientists in this country.

Unfortunately, very few scientists responded to our questionnaire (about 31), and only some three scientists were found to have developed their own cell lines and are using extensively cell lines for their own research work.

We have therefore recently started to collaborate with the American Type Culture Collection (ATCC), USA; European Culture Collection (ECACC), UK and RIKEN, Japan to prepare computerized animal cell culture information system.

It may be mentioned that Microbial Strain Data Network (MSDN) allows online access to ATCC and ECACC cell lines.

Protein secondary structure database

In collaboration with scientists at the Science University, Tokyo, we have developed a database of protein secondary structures.

Objectively, secondary structures in proteins were defined and in problem situations visual inspections were carried out to avoid errors. Using protein crystal structure databank, protein structures were analysed and the information on secondary structures of the protein was organised in free format. Necessary software to analyse such a databank was developed with Prof. A. Tsugita, Science University, Tokyo.

This databank is supplied to all users who receive the protein sequence databank from Protein Identification Resource (PIR).

In-house sofware development activity

In the last few years several computer programs have been developed to analyse nucleic acid sequences, protein sequences and three dimensional structures of the proteins. These programs include prediction of T-helper cell binding peptides and antigenic determinants. Patterns of amino acids having specific properties can also be picked up.

Our major achievements in the software development area are the development of multiple alignment of sequences package called PRAS on the parallel computer of the Centre for Development of Advanced Computing (C-DAC) as well as on PC-AT compatible systems.

We have also developed a single transputer based nucleic acid and protein sequence data analysis system which is as powerful as any other product in the market for this type of work. This software package is also available on a PC-AT platform.
We have applied for copy rights for these products.

Training:

This centre realised right from its inception, that the users will use our facilities only if they are given training in Bioinformatics. During the last few years the following courses have therefore been organised:

(1) National Workshop in Bioinformatics

(2) National CODATA Conference 1989

(3) Short Term Training Course on "Use of Computers in Databanks for Biotechnologists"

(4) International Microbial Strain Data Network Training Course on "The Use of Computers in Culture Collections and Microbiology Laboratories".

Participants with various backgrounds have attended these courses.

The last training course, which was organised with the financial aid from the United Nations Environment Programme (UNEP) and our Department of Biotechnology, received a very wide acceptance.

Users:

A large number of users are using the facilities of this centre which include Xeroxing, Telefaxing, Electronic mail, Analysis of data, Bibliographic search, and Project preparation. The users belongs to various categories — students, established research scientists, teachers and industrialists. International mail is also used widely by visiting scientists to the University of Poona and the staff of the DIC. In short, the centre is helping to improve the quality of scientific research in the area of Biotechnology and in training students and scientists in new computer culture and the facility is created are used extensively. Some of the quality research papers produced using these facilities are given in Table 1.

Any further information can be obtained from:
Prof. A.S. Kolaskar,
Officer-in-charge,
Bioinformatics, Distributed Information Centre,
University of Poona,
PUNE - 411 007, INDIA.
Telephone No. 0091 212 335039, 330195
FAX No. +91 212 330087
BTGOLD 10075: DBT0295
INTERNET: kolaskar@icgeb.trieste.it.

Table 1

List of research papers published by bioinformatics scientists using DIC facilities:

1. P.S. Naik and A.S. Kolaskar, 1990 — Distributed Information Centre and The Animal Virus Information System, *Indian J. Virol.*, **6**, 32-37.

2. M. Kutubuddin, A.S. Kolaskar, M.M. Gore, S.N. Ghosh and K. Banerjee, 1991 — Prediction of helper T Cell epitopes in envelope (E) Glycoprotein of Japanese encephalitis, Westnile and Dengue viruses, *J. Mol. Immunology*, **28** (1/2) 149-154.

3. H. Suzuki, A.S. Kolaskar, S. Samuel, A. Tsugita, 1991 — A protein secondary structure database (PSS), *Protein Sequences and Data Analysis*, 4 97-104.

4. A.S. Kolaskar and S.L. Samuel, 1991 — Analysis of Inverted Repeats in primary structure of proteins, *Protein Sequences and Data Analysis*, 4 105-110.

5. R.V. Kulkarni, S. Date, B. Kulkarni and A.S. Kolaskar, 1991 — PRAS: parallel alignment of sequences package, *Advanced Computing '91*, **1** 392-399.

6. A.S. Kolaskar and U. Kulkarni-Kale, 1992 — A sequence alignment approach to pick up conformationally similar protein fragments, *J.Mol.Biology*, **223** 4, 1053-1061.