Nucleic Acids Research, 1998, Vol. 26, No. 1 145–147

Molecular Probe Data Base (MPDB)

Maria Giuseppina Campi[1,*], Paolo Romano[1], Luciano Milanesi[3], Domenico Marra[1], Maria Assunta Manniello[1], Beatrice Iannotta[1], Gabriella Rondanina[1], Elena Grasso[1], Tiziana Ruzzon[1] and Leonardo Santi[1,2]

[1]National Institute for Cancer Research and [2]Department of Clinical and Experimental Oncology, University of Genoa, Largo R. Benzi 10, I-16132 Genoa, Italy and [3]Advanced Biomedical Technologies Institute, National Research Council, Via Fratelli Cervi 53, Milan, Italy

Received October 3, 1997; Accepted October 6, 1997

ABSTRACT

In this paper, the current status of the Molecular Probe Data Base (http://www.biotech.ist.unige.it/interlab/mpdb.html) is briefly presented together with a short analysis of its activity during 1997. This has been performed by statistically evaluating the 'logs' of the Internet servers that are used for its distribution with reference to the geographical origin of the requests, the words that were utilized to carry out of the searches and the oligonucleotides that were retrieved. Planned enhancements of this database are also described. They include a revision of its data structure and, even more relevant, of its data management procedures.

INTRODUCTION

Molecular Probe Data Base (MPDB) was set up in 1992 and has been made available to research scientists since 1993 by means of Network Information Retrieval (NIR) tools, such as WAIS (Wide Area Information Servers) and Gopher (1,2). It makes available on-line data on synthetic oligonucleotides, the vast majority of which is used for human pathology diagnostics, that have been successfully employed and tested in laboratories and have proven to be reliable and adequate.

Over the years, the database has changed its data structure according to the needs of the end users; the information contained has also been continuously updated (3), so that it has increased both in quality and quantity reaching a total amount of 4142 at the end of September 1997.

MPDB is currently managed by using the Oracle relational database management system. On a periodical basis varying from 2 to 4 months, a flat file archive is automatically created and indexed by means of WAIS. The flat files and indexes can then be reached and searched through the IST Gopher server.

MPDB was also included in the SRSWWW (Sequence Retrieval System on WWW) system, that is devoted to the integration of molecular biology databases by using the existing cross references. In this context MPDB is called MOLPROBE and can be searched at the EBI SRSWWW site (see contacts section).

Over the last year, the only remarkable change to the database structure has been the addition of a new field associated with bibliographic reference and regarding its Medline code number (when available). This field is obviously very useful for the deepening of knowledge on a special oligonucleotide through bibliographic searches. It will also be used in future versions of the database to directly link to Medline public sites available on-line, such as PubMed. The most recent version of MPDB has also been added to SRS5, the new release of the Sequence Retrieval System.

ANALYSIS OF SEARCHES AND ACCESSES

In order to verify how properly MPDB has been used by Internet users, a survey was carried out on the information that is recorded by the server after each access to the database.

Data were taken from the Gopher and WAIS 'log file'. The former includes, among other information, the Internet name or IP address of the calling machine and the retrieved file or directory; the latter includes, a part from the Internet name or address, the 'seed words' (keywords used for the query) and the 'document id' (internal pointer to oligonucleotides' detailed descriptions within MPDB) of retrieved oligonucleotides.

The period from January to September 1997 was taken in examination for the purpose of this analysis. Information of interest was extracted from the log file by means of automatic procedure, mostly Unix shell scripts.

Seed words were analyzed in order to verify the appropriateness of performed searches. Each seed word was then classified with reference to MPDB contents as proper, generic, questionable or improper.

Proper searches were carried out by means of seed words that were both linked to oligos research area and non generic (c-myc, HIV). Instead, generic searches were associated with generic terms pertaining to biomedical sciences, such as 'human' and 'cancer'.

Questionable searches were associated with terms not clearly identifiable, such as acronyms of unknown meaning (d11, ef-tu). Finally, improper searches were those associated with terms clearly not referring to biomedical research, such as 'car crash deaths' or 'martial arts'.

*To whom correspondence should be addressed. Tel: +39 10 5737 292; Fax: +39 10 5737 295; Email: giusy@ist.unige.it

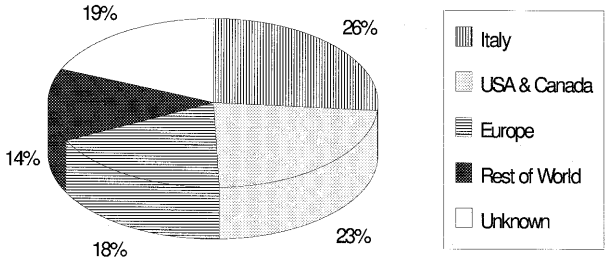

Figure 1. Percentages of connections by geographic area (January 1997–August 1997).

During the examination period, 554 searches were carried out by Internet-users. Analysis and classification of seed words indicated that there were 366 (corresponding to 66%) proper searches, 93 (16.8%) generic, 83 (15%) questionable and 12 (2.2%) improper. These results indicate that MPDB is being appropriately used and has reached the correct target, that is biomedical researchers interested in oligonucleotides.

Other interesting statistics relative to accesses to MPDB were also computed from the WAIS and Gopher log files; they regard the geographical distribution of the searches (Fig. 1), the number of searches and retrievals monthly carried out (Table 1) and the retrieval frequency for each oligonucleotide (Table 2).

Table 1. Number of searches carried out and of oligos' descriptions and complementary documents retrieved on a mothly base, from January 1997 to August 1997

Months	Retrieved oligos description	Executed researches	Complementary documents
January	143	97	84
February	108	87	72
March	12	31	21
April	49	65	87
May	177	94	54
June	24	61	55
July	79	68	47
August	90	58	47

Table 2. Distribution of single documents retrievals by data set

Frequency	From January to March 1997	From March to August 1997
One time	82.1%	89.5%
Two times	16.7%	8.3%
Three times	0.9%	1.8%
More than three times	0.3%	0.4%

The left column refers to the data set created in November 1996 that was available until end of March 1997. The right column refers to the following data set that was created at the end of March 1997 and left on line until end of August 1997.

Regarding the geographical distribution of the connections, we have considered four areas: Italy, the rest of Europe, USA and Canada, the rest of the World. A further category was created to group connections of unknown origin (only IP address available). Results of this analysis show that there is not a clearly prevailing area, although the European component is much higher than the North American one (44.8% including Italy versus 22.9%). Another significant consideration is that the ratio of unknown users is very high. This is mainly due to the lack of management of reverse IP addressing and to the tendency of not recording all personal computers linked to the Internet.

The data referring to the frequency of retrieved oligos show that there is not a polarization on only a few oligos and that all the oligonucleotides that are included in MPDB can be of the same relevance. In other words, there is not any special indications for extending the contents of MPDB towards a special research or application area, instead all areas seem to be equally relevant for the end users.

MPDB EVOLUTION

MPDB was first set up in 1992. Since then, many important changes have occurred in the field of networking and, in particular, in biologists' use of networks. Nowadays, personal computers, modems, networks and related software are available on each biologist's desk. And, even more significant, access to Internet services and archives is carried out through a common user-interface, the Web browser.

MPDB can make full profit from this modification. We are therefore developing a deep revision of its data structure and of data management procedures in order to take into account these changes. The main assumptions on which the new system will be based are that data management can now be easily carried out remotely by accessing a central database and that the knowledge and expertise on the use of oligonucleotides in different application fields is distributed between many researchers and many laboratories of different sizes.

Starting from these points, we are developing a new system in which researchers from external laboratories will be in charge of inserting and updating information on oligonucleotides that are used by them or other groups for application areas of their own special interest, by also carrying out surveys of relevant scientific literature. These 'curators' will be responsible as well for the accuracy and completeness of recorded information since they will be granted exclusive write and update rights on database records created by them.

The database itself, though, will be kept unique. Access to it will be only granted to curators, while end users will be able to search and retrieve information that will be kept in a separate hypertext, automatically extracted from the database as frequently as needed. All interactions with the new system will be carried out through a WWW server, thus enabling curators and end users to use Web browsers as the sole interface. The server will also allow for some graphical enhancements, such as the visualization of pattern sequences and positions by means of Java applets.

CONTACTS

Further information can be obtained by contacting: Dr M. Giuseppina Campi, Telematics Applications in Biotechnology, Biotechnology Department, National Institute for Cancer Research,

c/o Advanced Biotechnology Centre, Largo Rosanna Benzi, 10, I-16132 Genoa, Italy. Tel: +39 10 5737 292; Fax: +39 10 5737 295; Email: giusy@ist.unige.it

The relevant URLs are as follows:

BIOTECH Department WWW server:
http://www.biotech.ist.unige.it/
MPDB WWW home page:
http://www.biotech.ist.unige.it/interlab/mpdb.html
MPDB Gopher directory:
gopher://gopher.ist.unige.it/11/interlab/mpdb/
SRSWWW at the EBI in Cambridge (UK):
http://srs.ebi.ac.uk:5000/

ACKNOWLEDGEMENTS

The authors wish to thank Dr Massimo Romani for helpful scientific advice and Ms Paola Bianchi for secretarial help.

REFERENCES

1 Aresu,O., Parodi,B., Romano,P., Romani,M., Angelini,G., Manniello,M.A., Iannotta,B., Rondanina,G., Ruzzon,T. and Santi,L. (1992) *Nucleic Acids Res.*, **20** (supplement), 2009–2011.
2 Aresu,O., Campi,M.G., Romano,P., Parodi,B., Manniello,A., Thuroff,E., Molina,F., Saguato,F., Iannotta,B., Rondanina,G., Ruzzon,T. and Santi,L. (1994) *Nucleic Acids Res.*, **22**, 3474–3480.
3 Campi,M.G., Castoldi,M., Romano,P., Thüroff,E., Manniello,M.A., Iannotta,B., Rondanina,G., Ruzzon,T. and Santi,L. (1997) *Nucleic Acids Res.*, **25**, 92–95.