Molecular Probe Data Base: a database on synthetic oligonucleotides

Paolo Romano², Ottavia Aresu¹, Barbara Parodi¹, Assunta Manniello¹, Giuseppina Campi¹, Giovanna Angelini¹, Massimo Romani¹, Beatrice Iannotta¹, Gabriella Rondanina¹, Tiziana Ruzzon¹ and Leonardo Santi¹-²
¹Istituto Nazionale per la Ricerca sul Cancro, viale Benedetto XV 10 and ²Istituto di Oncologia Clinica e Sperimentale, Università degli Studi, viale Benedetto XV 10, 16132, Genova, Italy

ABSTRACT

The Molecular Probe Data Base (MPDB) was designed to collect and make information on synthetic oligonucleotides available on-line. This paper briefly describes its purpose, contents and structure, forms and mode of data distribution. Particular emphasis is given to recent data extension and system enhancements that have been carried out in order to simplify access to MPDB for unskilled users.

INTRODUCTION

The Molecular Probe Data Base (1) has been developed within the Interlab Project, a University-Industry joint project, funded by the Italian Ministry for University and Scientific and Technological Research, as part of the improvement of Italian research infrastructures. Access to databases on animal cell lines (Cell Line Data Base) (2) and HLA typed B lymphoblastoid cell lines (B Line Data Base) (3) is also provided in this project.

Recently, improvements were carried out both in the Cell Line Data Base and the Molecular Probe Data Base in order to extend data availability to a wider body of users and to make data access easier (4). This paper concentrates on the above mentioned modifications and on the data that have been included in MPDB during last year.

MPDB CONTENTS AND STRUCTURE

Data on oligonucleotides having a sequence of up to 100 nucleotides can be recorded in MPDB. Information on oligonucleotide identification, target gene, origin, technical aspects, applications and availability is included.

Identification is based on the name of the oligonucleotide, which is assigned by the data submitter, and on the nucleotide sequence. When the sequence refers to a coding region, the possibility is given to indicate the starting point of the corresponding peptide, and the aminoacid sequence, expressed by standard three-letter abbreviations (5), is then automatically generated by the computer.

Information on more than one target gene can be specified for a single oligonucleotide. Each gene is described on the basis of its name and code (taken from the HGM11 in the case of human genes (6)), map location, EMBL/GenBank sequence access number (7,8) and, for polymorphic genes, recognized allelic variants. Names of allelic variants are derived, when possible, from international workshop nomenclatures. This is the case, for example, of HLA gene alleles (9).

Technical data include melting temperature. When this information is not provided by the data submitter, values are automatically estimated by the system, according to Sambrook et al. (10).

Origin data include bibliographic references and details on laboratories that either submitted the data or synthesized the oligonucleotide. More than one application can be recorded for each oligonucleotide. These are taken from a vocabulary of controlled terms that includes the most frequent research, clinical and industrial applications. Finally, data on availability include oligonucleotide code in catalogues and the addresses of the distributors.

The database is being constantly updated and about 2000 oligonucleotides are presently described.

In detail, information on highly polymorphic human sequences has been included, with particular reference to the Major Histocompatibility Complex (130 oligonucleotides) and to chromosome specific PCR amplimers (about 1500 oligonucleotides). Nearly 150 oligonucleotides used in the diagnosis of genetic diseases are also present. Figure 1 shows the distribution of MPDB human chromosome specific oligonucleotides.

Moreover, artificial yeast chromosome specific oligonucleotides (about 50) and about 150 viral sequences, which can be employed for the diagnosis of human infectious diseases, are also included.

Data recorded in MPDB are organized in a quite complex structure, created according to data base relational theory. Its main purposes are: i) to optimize searches (both in speed and accuracy), ii) to reduce disk usage (avoiding repetition of data), iii) to improve user interface, and iv) to minimize errors (spelling, punctuation, etc...). In this context, a table has been created for each logical entity involved in MPDB, i.e. oligonucleotides, genes and laboratories, and for those attributes that can assume a particular value only within a set of predefined values, e.g. species, applications and allelic variants of a gene.
In the latter case, for each item of the tables (also referred to as controlled vocabularies) a mnemonic code is provided as well to facilitate and speed up searches and insertion of data.

**USER INTERFACE**

Particular care has been devoted to the realization of the user interface; this has led to menu-driven applications with easy to understand screens and valuable ‘on-line’ helps. Standard search and display procedures have been prepared and results of queries are available for possible downloads.

**Searches**

Developed applications give the user the opportunity to carry out searches on the basis of: i) the name and/or distribution code of the oligonucleotides, ii) their nucleotide sequence, iii) their target gene (including corresponding chromosome localization and specificities), and iv) their applications.

Mnemonic codes can often be used as an alternative to the plain textual definition of an item. Queries are made even easier by the availability of wild characters which can be used within the text to be searched as substitutes for either a single character or any sequence of characters.

**Reports**

Standard layouts for searching and displaying results have been prepared and are available in a format fit to transfer to a personal computer. Thus, at the end of a search, the user can ask for the report of all the information of interest. Data can be either ‘downloaded’ by capturing them while they are displayed, or requested via electronic mail.

Four different kinds of report can be obtained from MPDB: i) reports on all the characteristics of one specific oligonucleotide, ii) reports on relevant characteristics of all oligonucleotides, iii) lists of oligonucleotide names, indexed by some relevant characteristics, and iv) lists of items included in controlled vocabularies.

Oligonucleotides can be indexed on the basis of: i) application, ii) species and target gene, and iii) target gene and recognized specificities. Lists of controlled terms are available for species, target genes, specificities, applications, research laboratories and distributors.

**Recent enhancements**

A wider contextual help has been set up and a more effective on-line manual has been created as well. The latter can be reached and consulted by means of a menu-driven application and includes some advice and short statements about how to obtain specific results.

Direct integration of query specific reports into queries has also been added. In the previous version of MPDB, searching and reporting tasks were completely independent and there were no means of obtaining a report restricted to the results of a query. This separation has been eliminated by creating new applications that allow for the generation of reports limited to oligos retrieved during a specific search.

**MPDB FORMAT AND WAYS OF DISTRIBUTION**

Distribution of data has been considerably improved. MPDB is currently hosted in the Interlab Project Data Processing Center (IPPC), located at the National Institute for Cancer Research of Genoa, and can be reached via Internet. Data can be searched by means of two different softwares (one relational database and one information retrieval system).

MPDB was initially created by means of the relational database management system Oracle that is available on many hardwares (from personals to mainframes) and utilizes the standard query language SQL (Structured Query Language). A standard DEC VT100 terminal emulation is required in order to fully accomplish Oracle functionalities. No particular limitation exists on hardware used by the end users. In order to access MPDB data through a remote login session, an account must be requested to the Interlab Project User Service. All users recognized by MPDB through the login/password procedure will be directly prompted by Oracle menu-driven applications.

Oligos, a textual version of MPDB, can also be searched at IST. Detailed lists of all synthetic oligonucleotides included in MPDB are periodically created and indexed by means of a public domain software package, the Wide Area Information Servers (WAIS) system [11], that is distributed and supported by Thinking Machines Co., and that is rapidly gaining popularity in the biomedical research community.

WAIS is based on the so-called client-server philosophy: indexed data (documents) are physically resident at IST and are managed by a server software that can be queried from any Internet node having some WAIS client softwares installed. WAIS clients and servers only share a common protocol. Many different clients have been created, according to the specific features of the most widespread operating systems. Searches can be carried out by having brief descriptions (the 'source' files) of the databases of interest (see appendix for the Oligos source file).

Access to WAIS servers is also possible through the directory of servers, a WAIS database containing the 'source' files of all known WAIS servers.

A summary list of all oligonucleotides that are recorded in MPDB which includes, among the other data, oligonucleotide name, target gene code, and, when appropriate, recognized specificities, is also available by anonymous ftp, a common way of distributing documents and software based on free, but restricted, access to hosts connected to Internet. Lists of terms included in controlled vocabularies and indexes of oligos names grouped on the basis of applications and target gene/recognized specificity are also provided.

**Figure 1.** Distribution of MPDB human chromosome specific oligonucleotides.

ND = chromosome localization not determined.
ACKNOWLEDGEMENTS

The authors wish to thank Mr Tom Wiley for having revised the manuscript and Mrs Nadia Richetti for her secretarial help. This work was partially supported by a grant from the Italian Ministry for University and Scientific and Technological Research (DM 02081988).

CONTACT

Enquiries about MPDB and Oligos should be addressed to:

Interlab Project User Service
Servizio Biotecnologie
Istituto Nazionale per la Ricerca sul Cancro
viale Benedetto XV 10,
16132 Genova, ITALY

Tel. +39(10)3534511/2/3/4/5
Fax: +39(10)352888
E-mail: ipus@istge.ist.unige.it

REFERENCES


APPENDIX

In this appendix, the 'source' file of the Oligos WAIS database, that is the description of its contents and technical informations that are needed to search it, is reported.

:database-name 'Oligos'
:source-name '/wais/cldb/mpdb'
:cost 0.00
:cost-unit:free
:maintainer 'wais@istge.ist.unige.it'
:description 'Server created with WAIS release 8 b5 on Oct 22 1992
Server updated with WAIS release 8 b5 on Mar 10 1993
Keywords: biology, biotechnology, bioinformatics, synthetic oligonucleotides

At istge.ist.unige.it three databases devoted to availability of biological materials are maintained and updated, in the sphere of the Interlab Project.

Molecular Probe Data Base (MPDB) refers to oligonucleotides of up to 100 nucleotides. It includes data on ca. 1500 synthetic oligonucleotides.

Here is a sample record taken from MPDB:

11.10

Sequence:
5' GCAAGACACTCTAACATTG 3'

Target gene:
— Human F8C
   Coagulation factor VIIIc, procoagulant component (hemophilia A)
   Chromosome localization: Xq28
   Oligo localization: intron 7

Applications:
Genetic polymorphism—Hemophilia diagnosis—Molecular diagnosis of genetic diseases

Data received from:
— Istituto Nazionale per la Ricerca sul Cancro (BMOGE, Genova)
Internal probe for Factor VIII Polymorphism. Detects the product amplified by 11.6 and 11.2 amplimers.

Bibliography:

Target genes codes are taken from the Human Gene Mapping Library, when possible. Genbank/EMBL accession number are reported when available.

Complete data on a particular laboratory can be retrieved by searching MPDB with a term composed by the laboratory's code plus the fixed postfix LAB (e.g., searching BMOGELAB will retrieve complete data relative to the BMOGE laboratory).

Complete data on a particular catalogue can be retrieved by searching MPDB with a term composed by the catalogue's code plus the fixed postfix LAB (e.g., searching GENSETLAB will retrieve complete data relative to the GENSET catalogue).

MPDB version available via WAIS is also called Oligos.