A standardized format for handling data on plasmids, viruses and transposons: the PVT database format

M. Vicente, M. Aldea, A. Sacristán, C. Rohde, V. Weihs, M. Kracht, F. van Asma, E. Kampert, V. Hughes and C. Jones

The PVT format described here has been designed to store and retrieve genetic data on plasmids, viruses or transposons with special focus on their applications. Both naturally-occurring and engineered elements can be included in it. A variety of data can be accommodated in fields that are grouped in blocks: name and type of element, database administration, element administration, history, propagation, selection and host, biological properties, cloned insert and applications. The number of fields, now 157, can be expanded as required. Most properties can be described in simple logical fields. The format is organized to permit rapid searches and to facilitate communication between database and user; connection with culture and/or DNA collections is also envisaged and adequate fields for these tasks have been provided. The format allows cross-reference with that originated by the Microbial Information Network Europe (MINE) for computer storage and handling of bacterial or fungal strain data.

Key words: Database format, plasmids, transposons, viruses.

Plasmids, viruses and transposons are natural or constructed genetic elements widely used for genetic manipulations, biotechnology, and medical, environmental or basic research. The number of constructed elements has increased considerably in recent times. The adoption of a uniform format for handling genetic data on these elements would provide both a procedure to combine the information registered in different DNA, culture and data collections, and a valuable tool for scientists to obtain information on the choice of elements available for a particular purpose.

Databases, such as EMBL (Stoehr & Cameron 1991), GenBank (Burks et al. 1991), or VecBase (Pfeiffer & Gilbert 1988), containing the nucleotide sequence of many of these elements, are available; however the information stored in these sources is not specifically tailored for the use of the elements as tools for genetic manipulation, or other biotechnological and clinical applications.

Formats for bacterial, yeast, and fungal databases, which originated as the result of collaboration between European culture collections, have already been published and used (Gams et al. 1988; Stalpers et al. 1990). We describe here a format devised specifically for the description of plasmids, viruses, and transposable elements (the PVT format) emphasizing their usefulness for genetic manipulation and biotechnology. It is intended that both naturally-occurring and constructed elements are to be included in the database.

This PVT format was developed following the concept originated by MINE, the Microbial Information Network Europe, for the storage and handling of bacterial strain data (Stalpers et al. 1990). This will simplify cross-referencing between bacterial strains and the genetic elements they contain. The PVT format should be used to develop user-friendly management support programs, to encourage individual scientists or laboratories to keep their records so that they can be exported to central databases when needed.

Considering the foreseeable increase in the number
of available genetic elements, the format should be designed to accommodate new fields when dictated by circumstances; the PVT format facilitates the incorporation of new elements and their classification according to new properties.

**Software Tools**

The PVT format has been designed to be implemented using relational database programs. Of those commercially available, DataPerfect, dBASE IV and ORACLE have been used by the authors to construct databases conforming to the PVT format. The field ‘illustration’ has been implemented for use as input for programs for vector data management, such as CLONING (Aldea & Kushner 1988a, b).

**Description of the Format**

The unit used for constructing the PVT format is the field (Table 1). The field is formed by three essential elements: a name, a type assignment and a length. Each field contains additional information (line number and description) to help the user to understand the scientific structure of the format; this information is not part of the structure of the database.

The line number allows each field to be located within the format. The numbers are written to allow addition of other fields to the format whenever it becomes necessary. The name of the field is a brief mnemonic term formed by abbreviating the description; names should then be kept invariant by all the users of the database. A number forms part of the name whenever the field named forms part of a group of fields. Fields are grouped only when they define qualities of the same property, e.g. fields ET101 to ET199 (lines A.2.1 to A.2.11) are used to define the type of genetic element (ET stands for element type). The type assignment defines the kind of information (logical, or other type) accepted in the field. The description of the field indicates the kind of information to be stored in each field. Some fields have been marked ‘p’, ‘v’, or ‘t’ when they apply solely to plasmids, viruses, or transposable elements, respectively. Descriptions of most fields require no further explanation, comments on other fields are given below.

**Types and Lengths of Field**

For simplicity, most fields have been assigned a logical (log) type so that they can be easily filled by answering ‘yes’ or ‘no’. Besides the logical field, four other types have been used. In order of complexity they include the numeric field with n digits (num(n)), mostly for sizes, such as PACKMAX (line F.1.10) (see Table 1) to indicate the maximum packing capacity of viral cloning vectors. The alphanumeric fields with n characters (char(n)) are intended to contain short descriptions that do not fit into logical fields, e.g. NAME (line A.1.1), used to designate each element. Some data fields are provided to enter the date of relevant operations, both for the element, the database and the possible physical collection linked to the database, e.g. CHK (line C.1.5), used to annotate the last date on which the status of the element was checked in the collection. Finally, the memo fields are unlimited alphanumeric fields to be used as memoranda to contain lengthy descriptions, such as PROPAG (line E.1.4), in which a protocol for the propagation of the element can be indicated; the memo fields have been kept to a reasonable minimum number due to the heavy demands that they impose on the storage capacity of the hardware when some database programs are used.

**Addition of New Fields to the Format**

Several fields have been designed to include definitions of infrequent or new elements not covered by the present version of the format, e.g. VECAP998 (line H.5.1) is intended to be filled when the applications of a vector cannot be found in the logical fields that precede it. For these cases, an additional field is provided for specifying the property of the previous field, in this example VECAP999 (line H.5.2) will accommodate up to forty characters to describe the new application. Once the elements accumulated in one of these fields can be classified in one or more new fields they should be added to the format to satisfy the searching needs of the user.

**Cloning Sites**

Many elements to be included in the database require a functional description of their relevant restriction sites. For this purpose, a large alphanumeric field, RESCLON (line F.3.5.), has been included. To store the maximum amount of information in this space, the following recommendations are suggested: (i) use accepted abbreviations for the name of the enzyme that recognizes the site; (ii) a capital letter should be used in front of the name to indicate a functional description of each site; (iii) an ‘T’ defines a site in which insertion of a foreign fragment inactivates a genetic marker (the abbreviated name of the inactivated marker should follow between brackets), ‘U’ indicates that the restriction site is present only once (unique) within the element. ‘E’ designates a site in which the genetic structure of the element directs the expression of the inserted fragment and an ‘A’ is to be placed in front of site names that are absent from the element.

**Illustration**

The use of illustrations within a database has some difficulties, yet the user of a PVT-formatted database is likely to value most this kind of presentation because it provides information on an element at a glance. Moreover,