NEW ARCHITECTURES APPROPRIATE FOR LARGE 2-D GEL DATABASES

LEIGH ANDERSON
Large Scale Biology Corporation, 7503 Standish Place, Rockville, MD 20855 USA

ABSTRACT

There are at least two major classes of information arising from 2-D protein mapping studies: descriptive information about spots or sets of spots and quantitative (generally tabular) data describing protein abundances through a series of gels. The first type is called "**annotational**", since it takes the form of textual comments attached to entities on a map (in this case proteins). The second type is called "**quantitative**", in order to emphasize the numerical nature of the data. A third class of relevant databases, the "**connected**" databases, is composed of information resources assembled for reasons other than 2-D work, e.g., literature and sequence databases.

The two types of 2-D database, annotational and quantitative, are quite distinct and allow rather different opportunities. One is more in tune with the symbolic manipulation capabilities of artificial intelligence languages, and the other more in tune with conventional statistical analysis software packages. The one is primarily oriented towards text and symbols, the other towards numbers. Here we will attempt to illustrate the differences between these database types, consider some features desirable in an annotational database system and offer a few ideas applicable to their implementation.

1 INTRODUCTION

As large amounts of new data become available on biological macromolecules, the capacity of the research community to locate and use this data becomes a major issue. Two-dimensional electrophoresis of proteins, a technique capable of resolving and quantitating hundreds (sometimes thousands) of proteins in biological samples [1,2,3], makes it possible to develop a relatively complete picture of gene expression in cells and tissues and to accurately describe changes in gene expression associated with disease or experimental manipulation. Given a database of basic information concerning the proteins observed on 2-D gels, very sophisticated experiments can be undertaken, and new biological effects analyzed at the molecular level.

Data obtained by 2-D electrophoretic analysis of proteins is, however, very voluminous. Since it describes large numbers of proteins, often from a variety of experimental viewpoints, it is also often quite complex. For these reasons, the design and construction of databases generated from 2-D gel data represents a considerable technical challenge.

We have devoted considerable effort to the development of computer software systems for analyzing 2-D images and extracting quantitative data on the proteins observed [4,5,6,7,8,9]. This type of data, when treated by the methods of multivariate statistics, yields interesting insights into complex processes and allows an approach to the taxonomy of a range of biological systems from cells to species [10]. Precise quantitative data is also directly useful in searching for mutations at large numbers of loci [11,12].

However, there has so far been relatively little attention paid to the problems and opportunities associated with annotation databases in the context of 2-D mapping. We have suggested that the construction of such databases may involve the development of a "language" capable of representing basic facts of biochemistry and cell biology [13]. In thinking through the long term objectives of database generation it has become increasingly clear that one wants a system that not only stores biological characteristics of proteins but also allows certain inferences to be drawn from this information in a well-defined, automatic manner. If a protein is glycosylated, then it is likely to have been processed through the Golgi apparatus. This generalization is easily expressed as a "rule" concerning glycoproteins, or rather a rule associating certain symbols ("glycosylated", Golgi apparatus", passed through"), and need not be stored as a separate fact associated with each and every glycoprotein spot in the databases. Hence we believe a symbolic database constructed so as to allow storage and manipulation of information by symbolic procedures is required. Many of the concepts involved have been developed in the fields of artificial intelligence and expert systems.

2 GENERAL CONSIDERATIONS REGARDING ANNOTATION DATABASES

How can a single database system contain information on the physical properties, expression patterns, names, literature references and numerous other features of a large number of proteins, most known only as a spot on a 2-D gel and identified by 2-D map position? The problem is compounded by the fact that no one knows in advance all the sorts of data that should be recorded: thus provision must be made for flexibly adding any new type of data of interest to any user.

An annotation database ("ADB") can be viewed as analogous to a card file in which one maintains a series of notes on proteins (or on defined groups of them), keyed to a 2-D map by protein names, numbers or any other convenient nomenclatural system. The key feature is that the file must allow a user to record an unlimited variety of data: physical properties, literature references, hunches, associations with other proteins, etc. It is therefore very difficult to determine in advance all the types of data to be entered or to store them in some fixed structure where each separate kind of fact has its own pre-allocated space. One needs instead a system with great flexibility for storing and retrieving new kinds of information.

It is worth emphasizing the distinction between this concept and the classical idea of a quantitative database ("QDB") which is similar to a matrix of numbers. In a QDB there is generally a column for each gel (each analysis) and a row for each protein spot (each measurement), or vice versa. The values recorded are often the abundances of each protein on each gel (the integrated density of the corresponding spots on the gels), although other parameters can be examined (spot shape, peak height, etc.). In extracting answers from such data, one uses the techniques of multivariate statistics. Which spots show significant variation between gels? What changes are correlated with external variables? Is there a single overall pattern of change throughout the experiment, or a series of different effects? These questions require the application of sophisticated statistical techniques to voluminous and complex numerical data. Because of the wide range of application, the statistical techniques required are fairly well developed: what remains is the slow and laborious process of building large, high-quality datasets from well-designed experiments.

2.1 Annotation and Symbolic Computing It is important to recognize that a database of annotational information about proteins can be more than a collection of textual notes. Most existing

annotation databases have served as systems for filing simple text information in a way such that it could be retrieved through access to the spot pattern: a sort of electronic notebook. While more useful than a paper filing system, this approach limits the subsequent usefulness of the data.

Here we attempt to develop the concept of an annotational database in which the the information is to be entered primarily as symbolic data, meaning that the database records the association of particular symbolic properties with particular spots (or groups of spots). The value of this approach is that these symbols can then be used further by computer programs designed to perform what is now called symbolic computation: the manipulation of symbols rather than numbers. The fields of artificial intelligence and expert systems rely almost entirely on this type of computation, making use of a different sort of computer language (LISP or PROLOG) specialized for the manipulation of symbols rather than numbers (as is conventional with FORTRAN, Pascal or C). Modern database software systems also incorporate important elements of this approach.

2.2 Annotation Protein Index Architecture Designing a database is something that requires a considerable amount of thought. Not only does one want the most useful and economic design, but one wants to avoid structural mistakes whose correction will require extensive reworking of the data in the future. Since complexity grows naturally in any system, it is best to begin with simplicity based on a few general principles and elaborate later to cover special cases. As described below, we have begun with a series of desirable features and then developed some simple concepts that would allow us to express a rich set of protein mapping data in a simple database model: the relational model.

2.3 General philosophy Our approach in formulating an annotation database architecture is based on several basic notions:

- The architecture must be as simple as possible, so as to be implementable on a range of database software products. Speed and efficiency are at this stage less important than simplicity and generality due to the rapid evolution of computers toward higher speed and lower cost.

- The architecture must allow for future growth in functionality as well as growth in information content.

- The information to be stored should be reduced to as simple a form as possible in order to facilitate subsequent symbolic manipulation of properties. Ideally there should be just one record structure for all the factual data, regardless of type.

- All terms used in the database should also be defined in the database. This facilitates the introduction of new terms and common understanding of those in use.

- All important information should be tagged with the author and the date entered. This allows correct attribution of all discoveries (or errors) and allows reconstruction of a series of facts input over time. Our view is that proper attribution is a major requirement if multiple users are to cooperate in generation of a database.

- The system should, where possible, provide for the use of synonyms, thereby allowing several different names to be used for the same thing transparently. This feature, though perhaps inelegant, removes the problem of having to decide early on whose name to use for a given protein or property.

• A user should be able to access and add to the database using any of a variety of levels of computer sophistication. A database can achieve greatest use if it is accessible from a mainframe or large minicomputer, from a complete gel analysis graphics workstation or from a personal computer such as an IBM-PC. Thus the system itself should not depend directly on a computer graphical user interface (though this will certainly be most helpful to frequent users).

Given these general characteristics, an annotation database can be made widely available, easily accessed, and easy to add to. Perhaps the major advantage is that a variety of users may be able to share information in a simple format.

2.4 Symbolic properties, sets and hierarchical information As a means of achieving these goals we have elected to explore combining the concepts of spot characteristic (here called symbolic property) and spot set. A symbolic property is taken to define a set of spots (those that have the property, e.g. "glycosylated"), and a set of spots is named by its common property (e.g. the phosphoproteins are all marked "phosphorylated"). Thus users must create unique names for all the characteristics that a spot (or group of spots) might have; this is already largely done through the nomenclature of biochemistry. These names then become the names of the set of proteins having this property.

A natural extension of this idea allows the name of the set/property to be used in a manner similar to the number (or name) of a spot: additional properties can be attributed to the set as a unit. This makes the set defined by one property a member of a higher set defined by another property, creating a hierarchical tree structure of sets. Information can be searched for either upwards from a starting point (what sets does the starting spot or group belong to; what properties does it have) or downwards (what spots or groups are members of the set defined by the starting point).

These notions allow the creation of a database consisting of simple statements about spots and sets, constructed in such a way that any spot or set can be a member of several sets or hierarchies simultaneously. By the use of common but sophisticated database access tools, these hierarchies can be retrieved together with the properties attributed to the spots and sets.

2.5 Relational database systems Although computer database systems have been in use for decades, most have required of the user substantial computer expertise. More recently, however, the concept of the relational data model [14] has been advanced as a solution to the problem of simplifying database architecture and facilitating changes and updates to database design. In the relational model, all data is stored in tables of inherently simple structure. Each column of a table corresponds to a type of data and each row (or database record) corresponds to an instance (such as a 2-D spot) for which each of the types of data can be recorded. A key principal of the relational model is that each table should be as simple as possible and should not duplicate data in other tables unnecessarily. The alternative, primarily hierarchical-type database efficiently implements a hierarchical scheme, but this scheme is difficult to change; a defect in a situation where the structure of the knowledge is continuously updated.

An additional significant advantage of the relational model is the relative ease with which the entire database can be listed in tabular form, sorted in various ways to make a paper copy useful directly. This feature is important if non-computerized users are to be able to make any use of the data and to receive regular updates prepared in a painless manner.

3 EXAMPLES OF THE THREE DATABASE TYPES

In thinking about databases, concrete examples can be helpful. Here we provide several in hopes of reinforcing our arguments concerning the distinctions drawn above.

3.1 Quantitative Databases A typical quantitative experiment involves looking for statistically-significant differences among a group of individual samples. The table below presents data on 10 spots present in plasma protein patterns of four individuals, each run on five replicate gels, all gels normalized using a large set of spots. Each row shows (for a spot) master spot number (MSN), the number of replicate gels on which the protein was observed (N1, N2, N3, N4), the average abundance (integrated density) over the five gels of each individual (1-4), the coefficient of variation (standard deviation divided by mean, expressed as percent) for the five measurements of each individual (1-4), the statistical F-test result (a ratio of between-group to within-group variance) and an indication of the significance of the F-test result (** indicates difference significant at P<.001).

MSN	N 1	N 2	N 3	N 4	AVG 1	AVG 2	AVG 3	AVG 4	CV 1	CV 2	CV 3	CV 4	F	Sig
17	5	5	5	5	88496	72990	81395	82409	3	3	5	6	12	**
25	5	5	5	5	179411	199136	181596	199161	8	4	6	7	3	*
26	5	5	5	5	86519	87682	73159	71866	6	4	1	6	20	**
33	5	5	5	5	25165	24111	17406	32293	40	21	26	41	2	
34	5	5	5	5	8758	37485	41922	68040	23	27	27	40	3	*
35	5	5	5	5	72765	72930	61812	61749	4	4	8	5	12	**
50	5	5	5	5	50647	44345	51514	44686	5	3	2	4	16	**
53	5	5	5	5	79890	60238	82095	72877	2	31	3	9	4	*
54	5	5	5	5	68694	62873	61740	60028	10	12	3	12	1	
60	5	5	5	5	26727	40003	27416	46741	13	22	4	16	12	**

Five of the ten proteins shown exhibit very significant quantitative differences between individuals. Such proteins might form the object of further study in attempts to correlate specific protein markers with disease states (external variables).

3.2 An Example of the Use of a Connected Database: the 2-D Literature We have assembled a relational database of 2-D gel literature containing about 3,400 citations and abstracts. These have been indexed by an expert and the index terms used to make possible a subject-specific, rapid search. The SQL query language is used to retrieve the appropriate citations; such queries can be generated automatically through a graphical interface when a protein's name or characteristics are known. As an example, here are three of more than 50 papers having to do with 2-D electrophoresis of a specific protein: *tubulin* (abstracts are truncated here to economize on space).

Jun 1985 **Maytansine-resistant mutants of Chinese hamster ovary cells with an alteration in alpha-tubulin.**

Schibler MJ; Cabral FR 86001951 Can J Biochem Cell Biol; 63 (6) p503-10

Mutant clones of Chinese hamster ovary cells resistant to killing by the Vinca alkaloid maytansine have been isolated using a single-step procedure. These mutants are threefold more resistant to killing by the drug than the...

Nov 1985 **Retention of autoregulatory control of tubulin synthesis in cytoplasts: demonstration of a cytoplasmic mechanism that regulates the level of tubulin expression.**

Pittenger MF; Cleveland DW 86034172 J Cell Biol; 101 (5 Pt 1) p1941-52

Virtually all animal cells rapidly and specifically depress synthesis of new alpha- and beta-tubulin polypeptides in response to microtubule inhibitors that increase the pool of depolymerized subunits, or in response...

Jun 1986 **Effects of colchicine on cardiac cell function indicate possible role for membrane surface tubulin.**
Lampidis TJ; Trevorrow KW; Rubin RW 86220530 Exp Cell Res; 164 (2) p463-70
The effects of the tubulin-binding drug colchicine on cultured neonate cardiac cell function were investigated. Application of low doses of colchicine (but not lumicolchicine) caused an early reversible increase in...

• • •

Availability of such bibliographic information can substantially speed up the assimilation of earlier published 2-D work. It also makes possible the inclusion of direct connections between spots and references in the annotation database.

3.3 An Example of an Annotational Database Against the relatively simple structures of the quantitative and connected databases shown above, the annotation database described here requires a richer structure. Such a structure can be based on the following database relations (tables). Detailed specifications of the fields of annotation database relations will be contained in a separate publication.

3.3.1 The PROPERTY Relation The PROPERTY record is the fundamental data statement in the Annotational database. The record associates a named entity (eg. a spot or a set of spots) with a symbolic property. Symbolic properties can be general or specific, and can include physical properties (that the protein is glycosylated or contains no cysteine) or organizational information (such as the fact that the indicated spot is part of a larger named group). Each property must be defined in the Definition relation with sufficient clarity that users can be sure what it means and how the information was obtained.

Text and numeric value (and units) fields are included for storage of data additional to that carried by the property name itself (such as sequence or exact molecular mass derived from sequence).

Examples of Symbolic Properties

Symbolic properties are named using a brief but unique and informative word or phrase whose parts are connected by underscores so as to make one continuous string of characters. Examples of general property types and associated values include:

Glycosylated	Cytoplasmic	Variable_among_individuals
Phosphorylated	Mitochondrial	Methionine_starvation_diminished
Nuclear	Interferon_induced	Markers_for_mononucleosis
Acetylated	Short_half_life	Wheat_germ_agglutinin_binding
No_cysteine	Diminished_by_aging	Concanavalin_A_binding
Adenylated	Mitochondrial_matrix	Peroxisome_proliferator_induced
Proline_rich	Heat_shock_induced	Mitochondrial_encoded

An important feature of this approach is the possibility of expressing "tree" structures in a dynamic way, and of being able to add additional data to any level of this hierarchy.

3.3.2 The MAP Relation The MAP relation contains a list of all the spots in a particular 2-D pattern (whether a real or a "synthetic" pattern merged from a series of gels). Spots and sets of spots are identified in the Property relation by name (or number) and map, thereby eliminating confusion if two reference maps are incorporated in the same database.

This relation also contains the information necessary to generate an image resembling a real 2-D gel of the proteins in the database. Such a pattern may be used as the user interface to the database when appropriate computer graphics devices are available.

3.3.3 The PEOPLE Relation This relation identifies persons who contribute information to the database, and does so in sufficient detail to allow another user to communicate directly regarding detailed questions or suggestions. It also provides an automatic mechanism for generating the minimum list of persons to whom updates of the database should be sent.

3.3.4 The BIBLIO Relation A special relation is incorporated to handle references to the scientific literature, or indeed to any external document that can be described by a citation. The purpose is to make it easy to indicate the source of particular pieces of information. At this stage, the citations are not broken into discrete fields for individual authors, journal, volume number, etc., since these formats are variable. The citation key (Ref) can, however, be a pointer into one of the large literature databases (Medline or Biosis) if desired.

3.3.5 The DEFINITION Relation Definitions of terms are extremely important in a system designed for widespread use and continuous updating. Each person who enters data is responsible for defining the terms (especially the symbolic properties) that are introduced. In addition, a conjunction may be specified so that when information is retrieved it can be phrased in a grammatical way. Thus if the property is "glycosylated", the conjunction could be "is"; the sentence "haptoglobin is glycosylated" can be assembled from the named entity (haptoglobin), the conjunction and the property. If the property is "Hp_beta_chain", the conjunction might be "forms part of the": "Hp_beta_cleaved forms part of the Hp_beta_chain".

3.3.6 The SYNONYM Relation An important objective of an annotation database system is to allow different users to interact with the data on their own terms. In particular, this can be facilitated by allowing different users to use different words for the same spot(s) or properties. We propose therefore to add the capability to handle transparently the use of declared synonyms for any of the principal data terms in the database.

3.3.7 The CONTEXT Relation In order to allow the merging of related databases into a single homogeneous structure, a CONTEXT field is supported in several of the basic relations. The context referred to is the "experimental system" being described. If, for instance, a plasma protein database is to include data on humans, monkeys and rats, then three contexts could be defined, each with its own appropriate MAP. The relationships between the patterns could be expressed by creating a "pan-species" property for each cross-identified protein. Thus all three haptoglobin types (identified on the three species maps) would be assigned the property "All_haptoglobins", and would be joined at the root of the hierarchical tree.

3.3.8 The FUNCTION Relation In some cases, a specific numerical value can be computed from some other, known number for each spot. Examples are SDS-molecular mass and pI, given Y and X positions, respectively, and some information about the positions of standards. Thus it is useful to be able to provide a function that delivers these values when asked, rather than storing two records for every spot in the database. As for rules (see below), we do not limit the languages in which these functions are written. They will be implementation-specific, though in all cases an English explanation should also be provided, along with the relevant numeric parameters.

3.3.9 The RULE Relation Important applications for rules in a system of this type include situations where either a conclusion can be drawn by putting together statements in the database, or where a statement can be inferred from the absence of explicit data in the database.

In the first case, let us take an example: mitochondrial proteins. Imagine that we can say that a protein is a mitochondrial protein encoded by a nuclear gene if (a) it is present in purified mitochondria and (b) it is not encoded by the mitochondrial genome (ie. its synthesis is not stopped by treatment with chloramphenicol). Then assuming that the mitochondrial proteins and those encoded by the mitochondrial genome are identified in an annotational database, it should be easy to identify the nuclear-encoded ones by a simple logical manipulation of the database using the rule given above.

The other obvious case when a rule can be useful is that of the negative value of a recorded fact. It may be that in a particular 2-D pattern there are only a few glycoprotein spots, and so it is appropriate to record that these spots (or their groups, if defined) have a symbolic property representing glycosylation. It is also useful, however, to know which proteins are not glycoproteins. To add a record for each of the other proteins indicating this fact is inefficient, and instead a rule may be defined stating that a protein that does not have the property "glycosylated" is "unglycosylated". The later term can then be used in further symbolic manipulations.

We are not yet in a position to specify a single language in which all rules are written. Therefore the relation provides for the possibility that a rule is expressed in English (or French, etc.) for purposes of general comprehension, and also (in other records) in languages useful in various software systems.

4 DISCUSSION

The ability to exchange data between laboratories and between experiments will emerge as a major issue in the coming years. Our view is that the best means of achieving such exchange is the development of standard database architectures that can be implemented on a variety of machines under a variety of database software systems. Hence a general discussion of the different architectural approaches is needed.

In this paper, we have argued for a dissection of the databases problem into pieces that can be implemented more simply than can a single, all-encompassing, specialized database system. The notion of an annotation database described here can be initiated with personal computer-level software and information now kept in the form of labelled gel photographs. It can be extended to large databases resident on mainframes. As always, the test will be usefulness in practice.

5 REFERENCES

1) High resolution two-dimensional electrophoresis of proteins. O'Farrell, P. J. Biol. Chem. 250: 4007-4021, 1975.
2) Analytical techniques for cell fractions. XXI. Two-dimensional analysis of serum and tissue proteins: Multiple isoelectric focusing. Anderson, Norman G. and Anderson, N. Leigh. Anal. Biochem. 85: 331-340, 1978.
3) Analytical techniques for cell fractions. XXII. Two-dimensional analysis of serum and tissue proteins: Multiple gradient-slab electrophoresis. Anderson, N. Leigh and Anderson, Norman G. Anal. Biochem. 85: 341-354, 1978.
4) Estimation of two-dimensional electrophoretic spot intensities and positions by modeling. Taylor, J., Anderson, N.L., Coulter, B.P., Scandora, A.E., and Anderson, N.G. Electrophoresis '79, B. Radola, ed., W. de Gruyter, Berlin, pp 329-339, 1980.
5) A computerized system for matching and stretching two-dimensional gel patterns represented by parameter lists. Taylor, J., Anderson, N.L., and Anderson, N.G. Electrophoresis '81, Allen and Arnaud, eds., W. de Gruyter, Berlin, pp 383-400, 1981.

6) The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B.P., and Anderson, N. G. Clin. Chem. 27: 1807-1820, 1981.
7) Design and implementation of a prototype human protein index. Taylor, J., Anderson, N. L., Scandora, A. E., Jr., Willard, K. E., and Anderson, N. G. Clin. Chem. 28: 861-866, 1982.
8) Numerical measures of two-dimensional gel resolution and positional reproducibility. Taylor, John, Anderson, N. Leigh, and Anderson, Norman G. Electrophoresis 4: 338-345, 1983.
9) The Kepler™ workstation software system developed by Large Scale Biology Corp. for analysis of 2-D gel data.
10) Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. Anderson, N. Leigh, Hofmann, Jean-Paul, Gemmell, Anne, and Taylor, John Clin. Chem. 30: 2031-2036, 1984.
11) Suitability of two-dimensional electrophoretic protein mapping separations for quantitative detection of mutations. Taylor, John, Anderson, N. Leigh, Anderson, Norman G., Gemmell, Anne, Giometti, Carol S., Nance, Sharron L., and Tollaksen, Sandra L. Electrophoresis '86, M.J. Dunn, ed., Verlag Chemie GmbH, Weinheim, pp 583-587, 1986.
12) Detection of Heritable Mutations as Quantitative Changes in Protein Expression. Giometti, Carol S., Gemmell, M. Anne, Nance, Sharron L., Tollaksen, Sandra L., and Taylor, John. J. Biol Chem., 262: 12764-12767, 1987.
13) Some perspectives on two-dimensional protein mapping. Anderson, Leigh and Anderson, Norman Clin. Chem. 30: 1898-1905, 1984.
14) An Introduction to Database Systems, C. J. Date, Addison-Wesley, 1979.