A Database of Lipid Phase Transition Temperatures and Enthalpy Changes

MARTIN CAFFREY,* DENIS MOYNIHAN, and JACQUELINE HOGAN

Department of Chemistry, The Ohio State University, Columbus, Ohio 43210

Received August 14, 1990

The systematic study of the mesomorphic phase properties of synthetic and biologically derived lipids began some 30 years ago. In the past decade, interest in this area has grown enormously. As a result, there exists a wealth of information on lipid phase behavior, but unfortunately, these data have, until now, been scattered throughout the literature in a variety of books, proceedings, and journals. The data have recently been compiled in a centralized database with a view to providing ready access to the same and to the appropriate literature. The compilation facilitates review of what has thus far been accomplished and highlights what remains to be done in this active research area. As such, it represents a convenient summary of the existing data which, when evaluated, will enable us to identify where deficits exist in the data, to reveal the fundamental physicochemical principles upon which lipid phase behavior is based, and to understand more completely lipid phase relations in biological, reconstituted, and formulated systems. The compilation consists of a tabulation of all known mesomorphic and polymorphic phase transition temperatures and enthalpy changes for synthetic and biologically derived lipids in the dry and in the partially and fully hydrated states. Also included is the effect on these thermodynamic values of pH, and of salt and metal ion concentration and other additives such as proteins, drugs, etc. The methods used in making the measurements and the experimental conditions are reported. Bibliographic information includes complete literature referencing and list of authors. As of this writing, the database is current through June 1990 and contains 9500 records. Each record contains 28 fields. Here, we describe how the database originated, its scope and contents, data abstraction procedures, and issues relating to mesophase and lipid nomenclature, data analysis, and evaluation, and database maintenance and distribution.

INTRODUCTION

Lipids, also known as fats, constitute a diverse and important group of biomolecules. Lipids are liquid crystals, and their thermotropic, lyotropic, ionotropic, and barotropic behavior has been the subject of intensive study this past three decades. Work continues in this area at an increasing rate. The Lipid Thermodynamic Database (LIPIDAT) was initiated to collect in one, central depository all information on lipid mesomorphic and polymorphic transitions and miscibility. The data on the subject have been accumulating rapidly in the literature, and only episodic, less-than-comprehensive compendia have been published.[1] The data are presented in two broad categories in the literature, and the database project includes two foci to reflect this. Numerically reported phase transition temperatures and enthalpies comprise the first category, and phase diagrams the second. Described herein is the first half of the database, the tabulation of phase transition temperatures and enthalpies. The database is considered comprehensive for phospholipids, sphingolipids, and natural membrane extracts through June 1990. Studies of these lipids in the fully hydrated state form one file of the database; those in the partially hydrated state, including dry lipds, form the second. Lipid mixtures constitute the third file, while a fourth file contains bibliographic information, including authors, but not titles. To alert the reader to its existence, brief mention will be made of the phase diagram compilation.

The amount of work required to compile these data was considerable, but, at the same time, insignificant when compared to the collective toil of the many researchers over the years who originally acquired the data. From these efforts was assembled a vast body of information, much of which could have been considered, effectively lost, or, at best, retrievable only at great expense. More than 100 journals publish articles related to this topic. Few researchers could screen each of these for potentially relevant articles. Such an endeavor would actually prove counterproductive, as more than 50% of the information finds its voice in five or six journals, and some journals might publish no more than one relevant article per year. The lipid phase behavior research field and its close relatives, are large enough to warrant an institution responsible for screening each of those journals for data and for ensuring that these data are regularly and consistently recorded and made available.

Here is the double benefit of a database of this kind: not only are the data simply not lost, but their inclusion in an accessible database facilitates and encourages their frequent use. This easy access to once-lost data provides benefits to the educational, research, and industrial communities that many of its members have stated far outweigh the expense of the database's construction and maintenance. Examination of the database will provide workers in the field the means to assess the extent of knowledge, and to identify subjects of research that either have received an overabundance of attention or have been overlooked. Analyses of the data will be of a scope and comprehensiveness not possible with earlier tabulations or not feasible experimentally.

Origins. For the past two decades we have been working on various aspects of lipid research ranging from biosynthesis to the study of the physicochemical properties of lipids. In the recent past, time-resolved X-ray diffraction methods using synchrotron radiation have been developed that enable lipid phase diagrams to be constructed at unprecedented rates. The current phase data compilation began in earnest some six years ago in response to our frequent need to search the literature for information on lipid phase behavior and attempts to understand mesophase stability in the complex lipid mixtures of biological membranes.[2] The earliest version of the compendium was based on our own data and extensive files on the physicochemical properties of lipids. In an attempt to broaden the scope of the compilation, a letter was circulated to approximately 200 workers in the field worldwide requesting contributions to and suggestions for the database. The appeal elicited a 70% response, and the project was heavily endorsed by both academic and industrial researchers alike. It was obvious then that there was strong support in the community for the project. However, it continued at a "snail's pace" for

0095-2338/91/1631-0275$02.50/0

want of proper funding. The break came in 1988 when federal funding was received, which has enabled the project to progress to its present fruition. Traditionally, financial support for data compilations and databases has been difficult to come by. The sense is perhaps that the endeavor does not represent "science" in the true sense of the word and, as such, is not worthy of monetary support. Clearly, this view is limited since the immediate and long-term benefits of such a compendium are enormous.

Significance of the Compilation. The benefits of the compendium are many and varied. Firstly, it allows us to establish our present state of knowledge in this active field of research and points to further fruitful areas of study. It minimizes wasteful duplication of research and makes most efficient use of available data. The compilation in its present form provides immediate access to the actual data and to the appropriate literature. Having all the data together in one location provides us with a historical perspective on the field and allows us to gauge the influence of technological and materials innovations and new discoveries on the level of detail to be found in studies of this type. In evaluating the contents of the database, discrepancies in the literature will be highlighted and hopefully resolved. The compilation should bring with it an improvement in the quality and reliability of new data, a more standardized mesophase and lipid nomenclature (sorely lacking in this area) and the practice of reporting actual data and error bars and not just interpreted phase fields and boundaries as is often the case in the phase diagram area. Further, the compilation provides a summary of data which upon analysis will enable us to (a) distill from the data the underlying principles of lipid phase behavior, (b) relate this primary information to the situation that prevails in cellular and reconstituted membranes and in other lipid aggregates both in vivo and in vitro and (c) establish a plan of action to make good deficits in the data. Finally, the database will be used as a source of data for (a) directing the choice of lipid/lipid mixture, or additive with desired physicochemical properties or effects; (b) the selection of organisms with membrane or storage lipid profiles that satisfy particular growth environment, processing, and/or end-use requirements; (c) suggesting how a given lipid species might be chemically modified to express a particular physicochemical character; (d) use in the formulation of foods, feeds, and pharmaceuticals with desired organoleptic and miscibility properties; (e) testing theoretical models of lipid phase behavior and miscibility; and (f) the instruction of students in both the physical and life sciences.

How Database Records Were Obtained—The Data Sluice. The process by which the database was constructed can be viewed as a four-stage sluice down and within which the vast body of lipid phase behavior information, multifariously presented and variously relevant, is directed, abstracted, and refined to conform to the design and purpose of the database.

The First Stage: Reprint Procurement. Original, published articles serve as the raw data for the database. It is the eventual goal of the database to assign a quality index to each record within it. By abstracting data solely from published articles, the data can be said to have earned already at least the exacting quality demanded by peer review. The reprints from which were abstracted the database's approximately 9500 records were obtained through several methods, paramount among which were a direct solicitation for reprints from over 200 workers worldwide and an extensive on-line search of the Chemical Abstracts Service (CAS) database. The number of articles obtained through these means for the period in which the database is considered comprehensive (through June 1990) was approximately 2500 from over 250 different journals. Of those, about 50% was found to contain data appropriate for the database; 115 different journal titles were represented. The reprint procurement process is ongoing, with every edition of over 13 major journals, accounting for ca. 80% of the data in the database, screened for relevant articles. Literature-wide computer searches (CAS) are carried out on a quarterly basis also.

The Second Stage: Screen 1/File I. Original articles are screened for data, and any data abstracted are entered on the computer. During Screen 1 an abstracter scrutinizes a reprint, identifies data that satisfy the various fields of the database (see Figure 1), and transfers this information to a hand-written data sheet. The data are then entered on the computer into the file named File I, in the step of the same name (File 1). A printout of these data is made. As a reprint passes along each stage of the data sluice, its progress is recorded on a computer file called Reprint Control. The initials of the individual responsible for completing a step in the sluice and the processing date are recorded on this file. This allows tracking and physical location of reprints, as well as analysis of Data Sluice efficiencies and deficiencies. A recent addition to the data abstraction process has been the inclusion of Key Words in Reprint Control.

The Third Stage: Screen 2/File II. Screen 2 offers the original abstracter the opportunity to check the File I results for errors or omissions by comparison with the original source material. Corrections and additions are made to File I records, and these are then transferred to File II on the computer and printed.

The Fourth Stage: Screen 3/File III. A second abstracter performs Screen 3. This step exerts quality control on data abstraction by introducing an independent, critical viewpoint to the process. The Screen 3 abstracter performs a rigorous screening of the reprint. Any suggested changes to the File II data are discussed with the original abstracter, and final changes made. These records are then transferred to File III, the actual data depository. Screen 3 changes made to the data are checked for typographical errors. Reprint Control is updated to reflect that the reprint has completed its journey down the Data Sluice. The reprint itself is filed in the database office, and three backups of the computer files are made.

We estimate that on average it requires ca. 3 h (range: 1–4 h) to fully process (complete Stages 1–4) a single article, including initial procurement. Two-thirds of the time is devoted to bringing a paper through Stages 1 and 2. Even a paper that is deemed to have "No Data" passes through all four stages of the Data Sluice, thus ensuring that it has not been erroneously designated as such.

DESCRIPTION OF THE DATABASE

A. Hardware and Software. As of this writing the database is housed in a Macintosh II microcomputer with 8 Mbytes of RAM (4 Mbytes of RAM is sufficient to open the full database) and was constructed by using Microsoft Excel, version 2.2. The four files described in the Introduction (fully hydrated, low hydration, lipid mixtures, and bibliography) occupy just under 3 Mbytes of disk space when stored as Excel files. As described under Dissemination, the database will be made available in Macintosh- and PC-compatible and hardcopy forms.

An example of how entries are presented in the Excel version of the database is shown in Figure 1. Each transition temperature and associated information abstracted from the literature appears as a single record in the database. The transition temperature and ancillary data are presented in separate fields within each record.

The prototype of the IBM-compatible PC version of the database is illustrated in Figure 2. Three types of searches are possible in the latest edition of this version of the database. The first search option generates the complete set of records,

	Field Number									
	1	2	3	4	5	6	7	8	9	10
	ID #	File Type	Lipid Description chain, backbone	hdgrp	T	ΔH	Transition Type	Initial Phase	Final Phase	Method
Record #1	608	pl	DL-16:0/18:1c9	pc	11	n	gel - l.c.	B	A	FI(tParA)
Record #2	4808	sl	N-18:1c9-Sph	Gal	44.8	1.8	LC1 - ms.l.c.	1Bc2	(1A)	DSC

	Field Number								
	11	12	13	14	15	16	17	18	19
	Rate	Purity	Aqueous Phase	pH	Journal	vol	sp	fp	year
Record #1	2	Pure TLC	10mM K-phosphate	7.4	BBA	598	189	192	1980
Record #2	5	Pure TLC	70 wt% H2O	n	Biophys. J.	55	281	292	1989

	Field Number							
	20	21	22	23	24	25	26	27
	Footnotes	Author 1	Author 2	Author 3	Author 4	Author 5	Author 6	Author 7
Record #1	n	Lee, T.	Fitzgerald, V.					
Record #2	Incubated 24h at 49°C, cooled, immed. reheated.	Reed, R.	Shipley, G.					

Figure 1. Contents of two records showing details of the individual fields as they appear in the Excel version of the database. For clarity, only a limited number of fields per line are shown for each record in this example. In the actual database, each record exists as a single row containing the first 20 fields. Fields 20–27 are housed separately in the Bibliographic file. Because of the large variation in the way that authors names are listed in the different journals, we have chosen to include in the Bibliographic file the authors surname and first initials only.

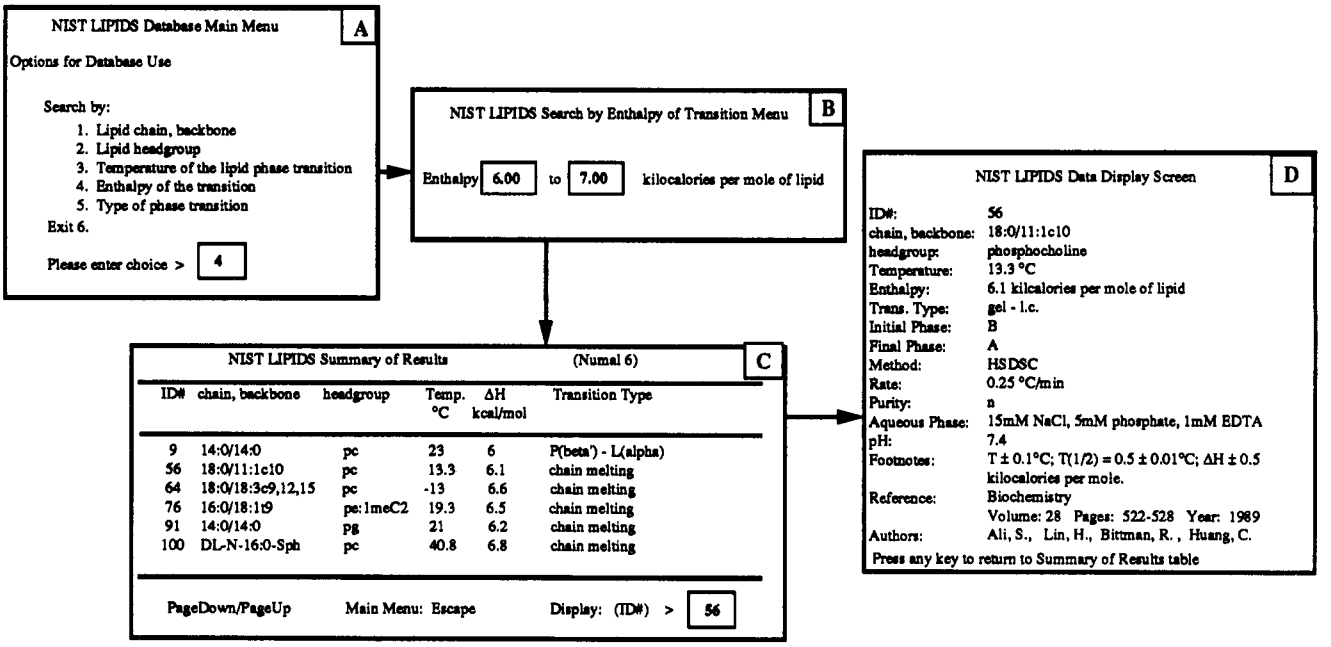

Figure 2. Structure and content of the IBM-compatible PC version (prototype edition) of the database is illustrated in this series of sequential screen prints. In this example a subset of the database containing 101 records is searched based on the transition enthalpy change criterion (option #4, panel A). In panel B the range of enthalpy change values over which the database is searched is chosen. In this case, the range is 6.00–7.00 kcal/mol. Following the search, a Summary of Results table (panel C) is generated containing all records in the database subset with enthalpy change values in the specified range. The final details in each record can be displayed individually (panel D) by selecting the appropriate ID number in panel C, in this case no. 56. As of this writing the database contains over 9500 individual records of the type shown in panel D.

of the form shown in Figure 2 (panel D), for a specified Lipid Species and Type, for example, all records for 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine. The second search option generates a summary table of all the records in the database satisfying a given single criterion such as a particular lipid species, a particular transition type, a particular temperature range, etc. The user then chooses the record from within the summary table to be displayed in the manner il-

lustrated in panels C and D of Figure 2. The last search option is a user-defined search. Here, the user can specify a search based on a combination of up to five criteria from 11 field decriptors. The results of the search are presented in a summary table from which particular records can be chosen for display. In this latest edition of the database, the use of a lightbar to highlight items for selection from within a panel has been implemented. Thus, the amount of typing while conducting a search is kept to a minimum. Another feature of this edition is the presentation of enthalpy change in units of both kcal/mol and kJ/mol and the provision of a calculated entropy change where appropriate. Future editions will incorporate graphics and statistical packages to facilitate data analysis.

B. Database Fields. A "field" is to a database what a column heading is to a table. These are the different elements that constitute a row in the database. Each row is called a "record". Each field will be described here to relate the scope of the database and to explain syntax and nomenclature with which users may be unfamiliar. Many different, acceptable means exist to relate the information for which the database was designed. For example, DLPC often appears as an abbreviation for a particular lipid in the literature. However, DLPC has been used variously to represent dilauroylphosphatidylcholine, dilinoleoylphosphatidylcholine, and dilinolenoylphosphatidylcholine. Enthalpy change is usually reported in units of kilocalories per mole, but also appear as either kiloJoules per mole or calories per gram. Such differences in nomenclature and other data are navigable within the context of a single paper, but not in a comprehensive database. The database consists of 28 fields. For ease of description, these will be treated under the following headings: The Lipids, The Data, The Ancillary Data, and The Bibliography.

1. The Lipids. The database fields to be addressed here describe the lipid hydrocarbon chain, backbone, and headgroup. However, before proceeding with a description of each, a comment on lipid nomenclature is in order. Vital to the utility of the database was the development of a lipid nomenclature scheme that would simultaneously reflect the great diversity of lipids while conforming to the database constraints of consistency and compactness. The nomenclature rules used are described briefly here. A more complete presentation of this topic will accompany the hardcopy version of the database. Most people familiar with lipids will recognize the majority of the lipids as they appear in the database with no knowledge of the rules, as they are derived primarily from the guidelines recommended by the IUPAC-IUB Commission on Biochemical Nomenclature[3-5] and have been implemented in the literature to varying degrees over the past several years. However, some excursions from this nomenclature "norm" were deemed necessary for the database in pursuit of rigorous uniformity.

The nomenclature scheme used in the database is as follows. To begin with, the lipid molecule is considered to be composed of three identifiable parts: (i) the apolar, hydrocarbon chain region; (ii) the polar headgroup region, and (iii) the backbone region to which the first two parts are attached. In the database they are described by two fields, one defines the chain and backbone region, while the other identifies the headgroup. The default structure is the L isomer with two, unbranched saturated hydrocarbon chains esterified to the C1 and C2 positions on a glycerol backbone with the headgroup covalently attached through a phosphodiester group at the C3 position of glycerol.

a. Chain and Backbone. 1. Chain Description and Their Modifications. For pure synthetic lipid preparations, all acyl and alkyl chain residues are fully specified, using a systematic nomenclature as follows. The two chain lengths, in units of carbon atoms (and with the first carbon of the chain defined to be that one bonded through an oxygen atom to the glycerol backbone) are given, each to the left of a colon (:). The two chain descriptors are separated from each other by a slash (/). Modifications to each chain are indicated to the right of the colon and are listed according to number, kind, and location. First, to the right of the colon appears the number of modifications on that chain. A zero (0) indicates that the chain is in the default configuration, with no modifications. Following the number of modifications, the modifications themselves are listed alphabetically. Following each modification is a number indicating the carbon atom position on the chain where the modification is located.

Modifications to lipid hydrocarbon chains are manifold. The database nomenclature system employs a strategy to reflect this diversity by describing changes from the default on an atom-by-atom basis along the chain. Modifications include, but are not limited to, position and type (ether, ester) of chain attachment to the glycerol, unsaturation at one or more sites along the chain, and presence of functional groups or heteroatoms. There are two commonly encountered systems of nomenclature for signifying double bonds. The first is to use the letters c or t to denote respectively, the cis or trans configuration of the double bond or bonds, and then to use the symbol Δ followed by superscripts to identify their positions. So, for example, a linolenoyl chain is described as 18:3ccc$\Delta^{9,12,15}$ The second most commonly used system is to describe the position of the double-bond counting carbon atoms starting at the terminal methyl group. Here, the format $(n - x)$ or $(\omega - x)$ is used where n or ω designates the terminal methyl carbon and x the position of the double bond relative to the methyl group. In the database we use a modified version of the former system, modified because of our desire for compactness and to avoid the use of subscripts, superscripts, Greek letters, etc. Thus, the linolenoyl chain is described simply by 18:3c9,12,15. Likewise, an elaidoyl chain is designated 18:1t9. If the configuration of the double bond is unspecified in the source literature then the letter e is used before positional information. If the configuration is known but the position is not, then the letter specifier is used without further qualification.

2. Backbone Description and Modification. Backbone modification refers to changes made to the default glycerol backbone. Rather than using a fully systematic nomenclature for this part of the molecule, an abbreviation denoting the structural unit considered to be the backbone is employed and is placed after the chain designation, conjoined with a hyphen. For example, the default abbreviation for glycerol is Gro. Thus, the full designation for the 1,2-dipalmitoyl-*sn*-glycero moiety is 16:0/16:0-Gro. However, because of the extremely common nature of this linkage the -Gro is dropped and is implied. Some lipids with backbones different from glycerol appear frequently enough in *Nature* and in the literature to earn their own generic name. The class known as sphingolipids (designated by the letters -Spd, -Sph, or -Spn, depending on the nature of the sphingoid base) and plasmalogens (-Psm) are good examples. There are occasionally lipids that appear in the literature that have normal headgroups and modified glycerol backbones. The modified glycerol is then represented in abbreviated form, all in capital letters, and is appended to the hydrocarbon chain description with a hyphen. For example, a deuterated glycerol backbone is designated 14:0/14:0-C(D)2-C(D)1-C(D)2. There also exist some more "exotic" backbones, for example, (2,3-dimyristoyl)cyclopentatriol, which is designated in the database as 2-14:0/3-14:0-CPENT, although at present these types of lipids are rarities in the literature. Lysolipids, i.e., monoacyl- or monoalkylglycero-

phospholipids are also considered as backbone-modified lipids and are denoted by the suffix "-lyso" attached to the chain designation. The unsubstituted hydroxyl group of the *sn*-glycerol backbone is specified at the beginning of the molecule. Thus, 1-palmitoyl-*sn*-glycero-3- is designated as 2-16:0-lyso.

b. Headgroup. This field identifies the polar or headgroup part of the lipid molecule. In the default condition the headgroup is assumed to be covalently bonded to the glycerol C3 via a phosphodiester linkage. The type of nomenclature employed in this field of the database is either simple structural unit, e.g., "pc" for a phosphocholine moiety, or generic, e.g., Cer for ceramide. Due to space considerations, most headgroups receive a two-letter designation, and the more complicated headgroups such as those encountered in the glycolipids are explained in an abbreviated form in the Footnotes field. Some modifications occur with sufficient frequency to warrant assignment of a new headgroup designation. The bulk of headgroup modifications are simply subtle changes effected upon a common headgroup (see, for example, record 76 in panel C of Figure 2).

After the glycerol backbone, the next most common backbone found in biological systems is that based on sphingosine. However, it must be stressed that many of the sphingolipids isolated from natural sources contain a variety of different sphingoid bases. The system currently in use in the database is as follows: unless the sphingolipid backbone is specifically identified as sphingosine, the letters "Spd" are attached via a hyphen to the name of the source of biologically derived lipid preparations and to synthetic lipid preparations, where necessary, to denote a generic sphingoid base.

2. The Data. The Data fields consist of the Temperature (T), Enthalpy Change (ΔH), and Entropy Change (ΔS) fields, and the Initial Phase, Final Phase, and Transition Type fields.

The Temperature field contains the phase transition temperature in °C. An "n" in this field, or any other of the database's fields, indicates that the information was not reported. As with all data in the database, transition temperatures come exclusively from published, original research articles, or abstracts. Manuscripts, theses, and reviews are not used. Transition temperatures are presented in many different forms in the source articles. The object of the database is to collect all of these in one, accessible place, uniformly. Consequently, some transition temperatures have been translated from the form in which they were published into the database's numerical form. Data for other fields were treated similarly, as necessary. The averages of temperatures presented as ranges are recorded in the Temperature field, and the published range is included in the Footnotes field (described below).

The initial data forms requiring translation were the presentation of transition temperatures in kelvin, in figures rather than in tables or the text, and as temperature ranges. For those temperatures presented in kelvin, 273.15 was subtracted to obtain °C, retaining the number of significant figures found in the original data. The Excel version of the database has temperature listed in °C, while the PC version provides a K, °C option. For data presented in figure form, transition temperatures (or enthalpies, pHs, etc.) were estimated from them, to a subjectively determined, conservative degree of precision. Such liberties taken in the interest of comprehensive data inclusion were properly footnoted. The figure number in the original paper from which the datum was taken is indicated in the footnote of the record.

The practice of estimating values from figures for the purpose of inclusion in the database might strike some as completely flawed. The guiding rationale behind this form of data abstraction has been this: an informed reader would come across a figure and estimate from it, to the best of their ability, the transition temperature, or whatever parameter the figure related. The reader would have an estimate of the value, along with the knowledge that it should be taken only as an estimate. Such estimates constitute 17% of the database records since much of the data are presented in figure form only. Each is footnoted accordingly. This is one of a number of reasons why users should consult the footnotes frequently while using the database.

The Enthalpy Change field was constructed by means identical with those used for the Temperature field. Values are in units of kilocalories per mole (kcal/mol), which enjoys widespread use despite its consideration by some as archaic. Thus, kiloJoules per mole (kJ/mol) were converted to kcal/mol through division by 4.184 kJ/kcal. When values were reported in calories per gram for compounds of known molecular weight, a conversion was made to kcal/mol. For those for which this calculation was not possible, as in the case of membrane extracts, a "(fn)" was placed in the Enthalpy field, and the reported value placed in the Footnote field. The entry (fn) is used throughout the database as a designator for "see Footnotes". The Excel version of the database has Enthalpy Change listed in units of kcal/mol while the PC version provides the option of kcal/mol or kJ/mol.

The PC version of the database contains a calculated value for entropy change ($\Delta S = \Delta H/T$) when both transition temperature and enthalpy change are reported. The ΔS units used are cal/(mol·K).

The Transition Type field contains text from the papers themselves describing the transition the lipid underwent at the recorded temperature and/or with the recorded enthalpy change. The exact wording of the authors has been recorded here as faithfully as possible, appropriately abbreviated to conserve space. This field was deemed necessary owing to the great variety of expressions and conventions used and constantly modified in the literature to relate the complex behavior of lipids. If authors describe a thermal event observed with a differential scanning calorimeter as the "main transition", then "main" is placed in the Transition Type field.

A systematic classification scheme for the phase transitions exhibited by lipids has been developed and will be described below. It is for this that the two fields Initial Phase and End Phase have been included in the database.

3. The Ancillary Data. The Ancillary Data fields contain the experimental details that were reported for the transition temperature and/or enthalpy determination. These fields are: Method, Rate, Purity, Aqueous Phase, pH, and Footnotes (Figures 1 and 2).

The Method field contains information on the device or procedure used in the determination. The items in this field appear in abbreviated form, and the key should be consulted for clarification. Some determinations are reported as having been performed with two or more methods. For these, the several methods are included in this one field, separated with a comma between each.

The Rate field contains the temperature scan speed at which the determination was made. As the phase transitions for which the database was designed are thermotropic, temperature change acts to trigger the phase change. Barotropic, lyotropic, ionotropic, and other significant transition classes have not been included in the current version of the database. The rate, when reported, is presented in the database in units of °C/min. Conversions were made during data abstraction as necessary, to convert from °C/h, for example. Many of the methods used to measure transition temperatures are performed statically; that is, the temperature is held constant, and the sample is allowed to equilibrate at that temperature before the next higher temperature measurement is made. For these, the Rate field contains an "s". Cooling scans are re-

corded as negative numbers, e.g., "−1" for a cooling scan of 1 °C/min. When authors report that scans were performed with a range of scan rates, the highest rate is taken, barring further specification for individual scans. In certain cases, the direction of the scan is reported but not the actual rate. Here, the "+" and "−" signs are used, respectively, to indicate heating and cooling at an unspecified rate.

The Purity field contains any results of analytical purity assessments. Sample purity information is vital to effective data evaluation, so the standards used in abstracting this field were especially stringent. Too often, authors state simply that "purity was checked" by TLC (thin-layer chromatography), for example. This might be viewed by many as sufficient, and as containing the implicit statement "and the purity so checked was deemed satisfactory". The database does not assume this. Sample purity was recorded as "Pure" when authors stated in the paper that purity was checked by a certain means and determined to be pure. For example, "Sample migrated to a single spot on a TLC plate" would warrant a "Pure" designation. Percent purities determined by analytical means are included in this field, such as "≥98% TLC". Multiple techniques, when used, are included, separated from each other by a comma, which, throughout the database represents the word "and".

Components of the sample's suspending medium are recorded in the Aqueous Phase field, along with their concentrations or relative amounts. Steroids and ions are included here, but not proteins, salts, and other additives that would have been included specifically as an experimental perturbant rather than a buffer ingredient.

Records with reported aqueous phases of less than 50 wt % of the total sample were placed in the "Low Hydration" file. Transition temperatures and/or enthalpies determined in nonaqueous media have been abstracted and recorded on an incomprehensive, not-yet-released file. For a list of the many database abbreviations necessitated by the wide variety of chemicals appearing in aqueous phases of this sort (over 1000 to date), the key must be consulted.

The Aqueous Phase field is generally taken verbatim from the Materials and Methods section of the source article and placed in the appropriate, abbreviated form. Some calculations are made to account for dilution factors or when components are reported only as their molar ratios to the lipid, for example. The pH field contains reported pH values. An "n" indicates that pH was not reported. The average of reported pH ranges are included and footnoted.

The Footnotes field is the space allotted for information that does not fall under the auspices of any other field. Descriptions of sample pretreatment crucial to the use of the record are included here. If the data were subjected to any significant transformation before inclusion in the database, mention is made here. By "significant" is meant the averaging of a reported range of values or the estimation of a value from a figure. Conversion from K to °C is not considered significant. The number of the figure from which the value was estimated is recorded in the Footnotes.

Details of molecules considered as additives to the pure lipid system are found here as well. Lipid/lipid mixtures and additives such as proteins are flagged in the headgroup field, with the type and amount described in full in the Footnotes. Also found in the Footnotes is information deemed valuable by the abstracter, such as data collected at pressures other than atmospheric, and expansions upon some abbreviated items in the hydrocarbon chain, backbone field, such as the scientific names of microorganisms.

4. The Bibliography. The Bibliography fields contain the journal title, abbreviated as per the key, the volume number, page numbers, and year of publication, as well as up to eight authors, listed in the order they appear on the original paper (Figures 1 and 2).

PROBLEMS ENCOUNTERED IN ESTABLISHING THE DATABASE

A. Nomenclature. One of the most challenging problems to be dealt with while establishing this database concerned the variation in nomenclature used to describe the lipids and the mesophases and polymorphs they form. For any compilation, a consistent nomenclature must be used. Difficulties arose in determining how to classify logically and simply the lipids and their phases and to recognize where in this scheme a particular lipid or phase cited in the literature actually belonged.

1. Lipid Phases and Polymorphs. The system devised by Luzzati some two decades ago[6] for naming the various lipid mesophases has endured well. However, the enormous growth in the number of new phases that have been identified in the recent past has meant that the Luzzati scheme is no longer adequate. Presently, the literature abounds with a confusion of Greek characters, subscripts, superscripts, symbols, and capitalized, lower case, and italicized letters. Since no generally accepted scheme exists and since the Luzzati system is limited, each new phase is assigned by its discoverer a new symbol or combination of symbols, characters, and/or letters, etc. Thus, with independent and simultaneous discovery and usage in different contexts emerges a plethora of nomenclatures. As an example, the bilayer phase with disordered acyl chains has been variously referred to as the L_α phase,[6] the neat phase,[7] phase D,[8] phase G,[9] the liquid-crystalline phase, the fluid-bilayer phase, the lamellar phase, the lamellar liquid-crystal phase, the smectic A-like phase, and others.

As part of this compilation project we have set about devising a nomenclature scheme for lipid phases and polymorphs (Figure 3). The purpose of the scheme is to simplify the naming process and the nomenclature itself. It is based on a logical structural hierarchy and has built-in flexibility such that new phases and/or modifications can be accommodated readily. It also dispenses with the need for cumbersome symbols, subscripts, superscripts, letters, and characters. One of the advantages of the scheme is illustrated in an example taken from Figure 3. Thus, "1Bh" denotes a phase which is periodic in one-dimension with ordered chains packed on a two-dimensional hexagonal lattice. In the Luzzati scheme[6] this corresponds to the lamellar gel phase without specifying whether the chains are tilted or not—which, incidentally, is not identified in the Luzzati scheme as a unique phase. In the new scheme, an additional specifier can be added to indicate whether the chains are tilted (1Bh1) or untilted (1Bh2), and so on. In the current version of the database we have included both the old (actually, the nomenclature used by the authors of the paper from which the data were obtained) and the new nomenclature since the user is likely to benefit more from the former, in the short term at least. Undoubtedly, the new nomenclature will be challenged and will meet with resistance. Regardless of whether or not it survives, a systematic and flexible lipid phase nomenclature which is comprehensive and universally accepted is sorely needed in this area. Our wish is that the proposed new scheme will stimulate interest and discussion in this important but neglected area.

2. Lipid Nomenclature. The nomenclature system implemented in the database to describe lipid structure has been described above under The Lipids. The system was developed based on our own needs to apply it systematically to the 800-plus different types of lipids in the database and for it to be simple and easily understood.

B. Data Procurement and Abstraction. There were many difficulties encountered in the course of establishing this da-

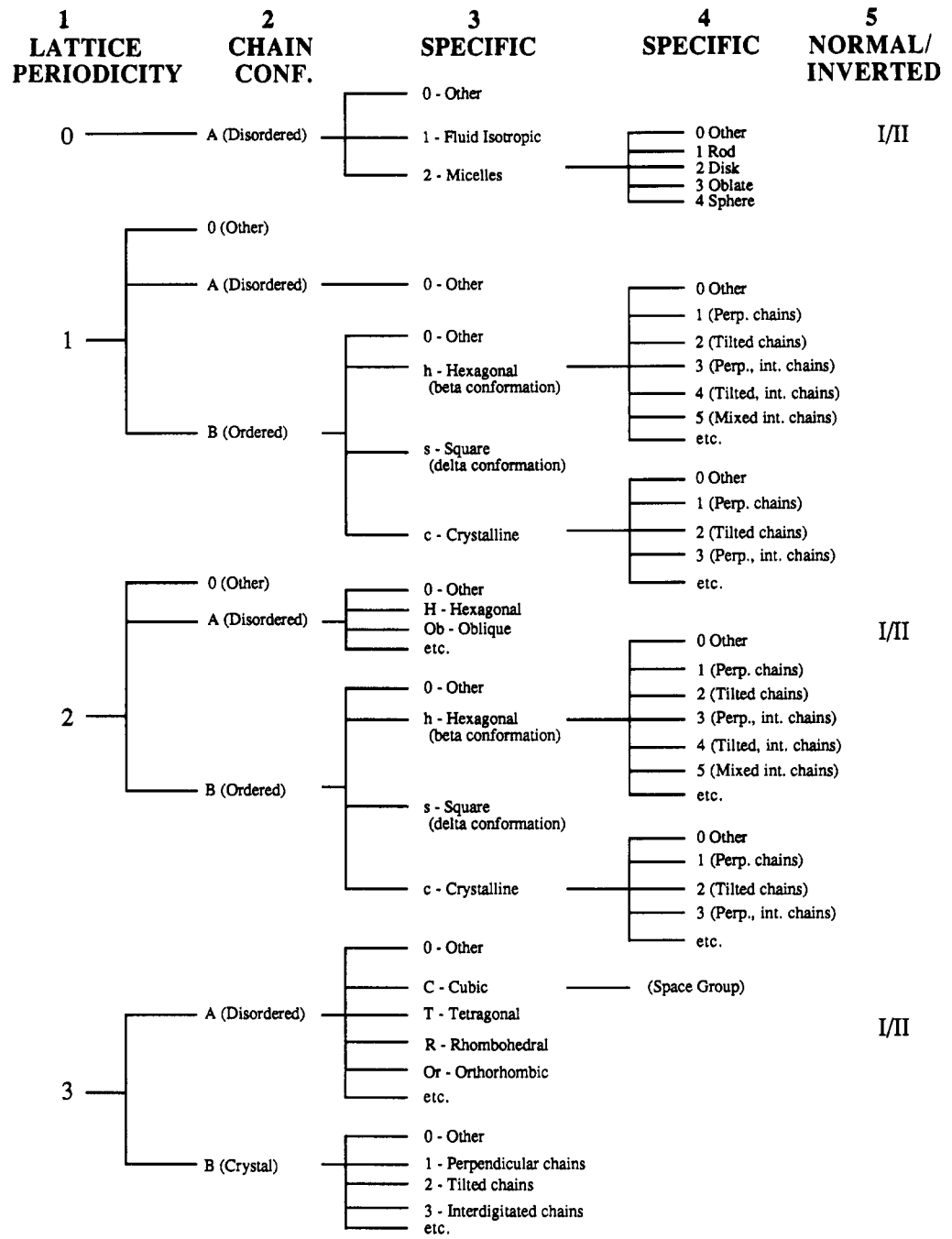

Figure 3. Proposed nomenclature scheme for lipid phases and polymorphs. In this scheme a maximum of five categories is needed to describe completely any phase. Due to space considerations, the words "perpendicular" and "interdigitated" have been abbreviated to Perp. and Int., respectively. A 0 (zero) is reserved for special cases when the characteristics of the phase do not seem to correspond to any of the other categories. It is proposed that for these phases, after a given period of time, a decision should be made by a panel of experts as to whether these cases require a designation denoting a new phase or if they can be accommodated within the existing categories.

tabase that had to do with ensuring that the database contained all the available literature on lipid phase transitions and with the actual process of data abstraction. The problem encountered when comprehensively searching the literature mainly involved what search terms or keywords to use. To simply use text for search terms was impractical since lipids have been described as phospholipid(s), lipid(s), bilayer(s), membrane(s), phosphatidylcholine, phosphocholine, etc. The search logic employed on CAS was to build a substructure for a particular lipid and conjoin this with the word "phase". This method produced the greatest number of articles meeting the criteria. However, it is possible that the search might miss more exotic lipids or articles that do not use the keyword "phase" as when "metastable behavior" is used as opposed to "phase behavior". In order to reduce this possibility, we have supplemented the search with other means such as manually screening major journals. Our task would be made easier, however, if consistent keywords are employed by researchers. For data abstraction, in certain cases, it proved an extremely demanding and time-consuming task simply to locate the actual datum itself within the text, a figure, or a legend. Thus, while extensive use of tables is not endorsed by journal editors, their provision greatly facilitates data abstraction. Oftentimes data are presented in the form of a figure wherein some phase-sensitive parameter is plotted versus temperature without reference to an actual transition temperature anywhere in the text. This then becomes the charge of the abstracter with the attendant difficulties of scale calibration, distance measurement, and transition temperature estimation. Needless to say, the abstraction process and data fidelity would be greatly enhanced if such information was gleaned from the data by the authors and documented clearly somewhere in the text of

the article.

Another difficulty arose in certain source articles when thermodynamic data and structural data were presented separately in the text. To piece together the two items can be difficult and time-consuming for the abstracter. This is particularly so when the phase behavior is complex as in systems exhibiting metastability and polymorphism. Combining the structural and thermodynamic data in a single scheme, of the type commonly found in papers originating from the Shipley group (see Reed and Shipley,[10] as an example), is of benefit both to the reader and to the abstracter.

FUTURE PLANS

The database has been in the works for several years and the product is soon to be released. A difficulty exists with a database of this type in that the data are constantly collecting in the literature. The future of the database has four main fronts: increase the scope, continue to collect data, disseminate what has been done, and evaluate the contents of the database.

A. Scope. The current database contains information on the thermotropic transitions of the phospholipids and the sphingolipids in the dry, and partially and fully hydrated states. The database is divided into four files as follows: (1) thermodynamic data and associated information on single-lipid species at less than 50 wt % water, (2) thermodynamic data and associated information on fully hydrated lipids (this file contains 88% of the database entries), (3) thermodynamic data and associated information on hydrated lipid mixtures, and (4) a bibliographic file. Later updates of the database will include separate databases on (1) the neutral lipids including the acylglycerols, steroids, and fatty acids, (2) lipids in combination with nonaqueous lyotropes/solvents, (3) isothermal (lyotropic, barotropic, ionotropic, pH-induced, etc.) transitions, (4) phase metastability and associated transitions, (5) data published in books, theses, etc., and (6) unpublished data. The latter is modeled on the Quid Pro Quo file soon to be available through the GenBank database and serves the purpose of alerting researchers to work in progress.

B. Updating. Ideally, the compilation should be updated continuously, and having the data available on-line will greatly facilitate this. An effective maintenance and updating protocol requires that the data be included in the compilation as close in time as possible to its date of publication. To this end, we propose that contributing authors to the relevant journals be requested to submit the data electronically and/or to complete a form which would be forwarded to the database, providing details and support information on the phase properties of lipids about which they intend to publish. This will mean that the compilation can be updated immediately with the new peer-reviewed data and that the information will be abstracted accurately for inclusion in the compilation since the authors themselves will perform the annotation.

We submit this in the spirit of a proposal and we encourage and welcome comments and suggestions. We have thought long and hard about the problem of data being "lost" in the literature and how best to approach the task of data consolidation and compilation maintenance and accuracy. Given the rate at which lipid phase data is being generated nowadays, we are of the view that for such a compendium to be effective, the data must be continually assessed and compiled immediately upon publication. We would hope that in time inclusion of data in the compilation would become an integral part of the publication process. Under no circumstances do we envision a compendium of this sort replacing or substituting for the original literature. Rather, it will serve in the true sense of a compendium: as a source and a sink for lipid phase data worldwide. Considering the value as well as the expense of an accurate and current database it seems likely that much of the responsibility for maintaining databases in the future will rest with the scientific community itself.

Parenthetically, we note that several journals have entered into an agreement with GenBank, the main U.S. gene sequence database and are rejecting papers that fail to present evidence of having previously submitted the sequence information to the database. It seems likely that prior electronic submission of data will become an integral and mandatory step in the publication process in this and related rapidly growing fields.

A future version of the database will be designed so that its contents can be examined in multiple, overlapping ways. Further, the data will be encoded with information necessary for reading the data based on a universally accepted format. Such data have been referred to as "self-describing".[11] This level of organization will render the data accessible to the largest possible audience and will facilitate the most thorough analysis and efficient use of the data. Plans are also underway to convert the database for thermodynamic data and the database for miscibility data from flat file structures to a relational structure.

C. Dissemination. The database will be available on disk for the PC, with no more supporting software than DOS required, from Standard Reference Data of the NIST. The authors have the database housed in a Macintosh II microcomputer and are currently using the Microsoft Excel (version 2.2) spreadsheet. It is likely that this version of the database will be released for general distribution also. No immediate plans exist to make the database available through an on-line host. As well, a hardcopy of the database will be published independently in a form that has yet to be determined. The hardcopy will contain the bulk of the work, the extent of phospholipid thermotropic phase transition data from, roughly, 1953 to June 1990. The literature will be screened annually, and the data compiled and released.

D. Data Evaluation. The first version of this compilation contains all of the data that have appeared in the literature to date without reference to the quality of the data. The next step must be to set about the demanding task of critically evaluating the data. This part of the project has not yet been undertaken. However, the evaluation is likely to be based upon criteria of the following type: sensitivity of the method used to make the measurement, instrument calibration, equilibration method, whether or not reversibility was tested, sample purity before and after measurement, scan rate used, details provided on suspending medium composition, thermal history of sample, data analysis method, accuracy and precision tests performed, and whether or not actual data were published (in tables, figures and text), or simply an interpretation of the data in the form of phase fields and boundaries, for example. The quality and reliability of the data will be indicated on a relative scale, the basis and exact details of which have yet to be resolved. However, each entry in the tabulation and phase diagram compilation will be assigned a quality symbol in the spirit of the rating system used by the Joint Committee on Powder Diffraction Standards (JCPDS) in the X-ray Powder Diffraction File. In particular cases, the quality of the data may not be apparent. For example, it may be that this same set of data is all that is available, in which case it would be included in the compilation with an "undetermined" quality rating.

Of course, the purpose of the evaluation is to identify the most accurate and precise data available for a given transition in a given system. Eventually, we hope to have available a version of the compilation which consists exclusively of critically evaluated data.

A panel of experts will be assembled consisting of researchers in the lipid field and in other areas with a focus on thermodynamic databases. Likely functions of the panel will

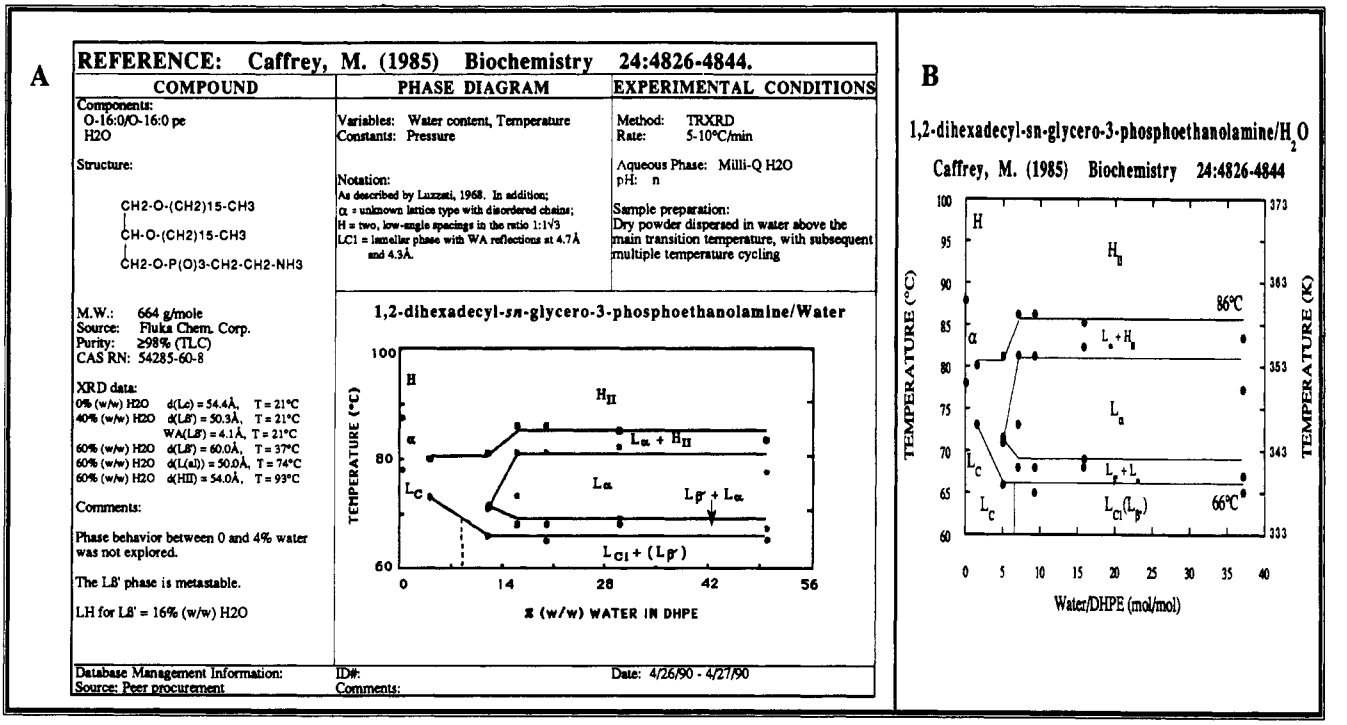

Figure 4. Proposed standard format for presenting binary lipid–water phase diagrams in the phase diagram compilation. The diagram in panel A represents a scanned image taken directly from the original article. The diagram in panel B represents a plot, in a standard format, of the experimental data points found in the original figure. These data were obtained by using a digitizing pad (Sigma Scan, Jandel Scientific, CA), and the coordinates of the data points along with the equation for the phase boundary lines are stored in a computer file (Kaleidagraph, Synergy Software) for easy retrieval. In panel B composition is shown in units of moles of water/mole of lipid. In the actual phase diagram database, a second plot is included showing composition in units of wt %.

be to establish the evaluation criteria, to advise on the evaluation process, and to approve the final ratings.

DATA ANALYSIS

Having the thermodynamic data and associated information assembled in compendium form will facilitate a thorough analysis of the data. Ultimately, we seek from such data an understanding of the principles of mesophase and polymorph stability and how phase stability and structure is related to sample composition and molecular structure. Analysis of the contents of the database will provide useful statistics such as the number of articles with data relevant to the database published per year, per journal, per author, etc. With a sufficiently large database it should be possible to evaluate such effects as method used to make the measurement, rate and direction of measurement, composition, etc., on the various thermodynamic quantities. Analysis will also highlight relationships between onset of appearance and frequency of a particular type of thermodynamic data in the database and the introduction of new and/or improved measurement and synthesis techniques and growth of interest in certain topics such as nonbilayer phases and the use of nonaqueous solvents and additives. A thorough analysis will eventually yield a compendium of reference data, the advantages of which include the identification of useful lipid standards, speed of future data retrieval and analysis, and a dramatic reduction in the volume of stored data. An analysis of the database and its organization will undoubtedly suggest ways in which the compendium can be restructured in as flexible a format as possible to facilitate future data processing and the identification of previously unseen relationships.

PHASE DIAGRAM DATABASE

We wish to alert the reader to the existence of a related compilation covering lipid/lipid and lipid/lyotrope miscibility. In this compendium, data are presented in the form of temperature–composition phase diagrams. The bulk of the material consists of lipid/water-phase diagrams and lipid/lipid-phase diagrams prepared dry and in the presence of excess aqueous phase. For the sake of completeness, the effects of a variety of small molecules, peptides, protein and nonaqueous solvents as well as of pressure on lipid phase behavior have been included. The lipid species presently in the phase-diagram compilation are the same as those considered relevant to the thermodynamic database, i.e., the phospholipids and sphingolipids. However, phase diagrams for the neutral lipids are also being collected for future editions of the phase-diagram compilation. Binary, multicomponent and theoretical-phase diagrams are included also. Monolayer-phase diagrams are not included in the current compilation although, again, as for the neutral lipids, any monolayer-phase diagram encountered in the literature is filed for future reference. Additional information provided with each phase diagram in the compilation includes methods (scan rate and direction), components, variables, details of sample preparation, composition and pH of suspending medium, equilibration methods, notation, comments, and molecular structure, formula weight, and CAS registry number of the lipid. Structural details on the various phases, when reported, will also be included with each phase diagram. A binary phase diagram taken from the compilation is shown in Figure 4. The authors welcome comments on any and all aspects of the Phase Diagram Database.

CONCLUSIONS

The great wealth of phospholipid, sphingolipid, and natural membrane thermotropic-phase transition temperature and enthalpy data has been collected in one depository. This task was prompted by the dissatisfaction of more than a few workers forced to depend on the episodic and unconcerted efforts of several workers to summarize the study of lipid phase transitions. The traditional difficulty in obtaining this information is not the only problem. The danger that the data could become "lost" in the literature exists as well. For quality

data to assume such a fate or to be unworthy of the Herculean effort required for retrieval calls to question the very purpose of obtaining the data to begin with. The database now exists to ensure that data are not lost, and that they are easily accessible and in a useful format.

ACKNOWLEDGMENT

The encouragement and support of R. G. Laughlin (The Procter and Gamble Company) and M. Chase, P. Fagan, and D. Bickham (National Institute of Standards and Technology [NIST]) throughout this project is gratefully acknowledged. Additional thanks are extended to all those who contributed comments, data, and reprints to the database. The project was funded in part through a grant for Critical Compilations of Physical, Chemical and Materials Data through NIST and a grant from The National Institutes of Health (Grant DK36849) to M.C.

REFERENCES AND NOTES

(1) Silvius, J. In *Lipid-Protein Interactions*; Jost, P., Griffith, G., Eds.; Wiley and Sons: New York, 1982; pp 239-281.
(2) Caffrey, M. In *Metabolism and Dry Organisms*; Leopold, A. C., Ed.; Cornell University Press: New York, 1986; pp 242-258.
(3) IUPAC-IUB Commission on Biochemical Nomenclature *J. Biol. Chem.* **1967**, *242*, 4845-4849.
(4) IUPAC-IUB Commission on Biochemical Nomenclature *Eur. J. Biochem.* **1977**, *79*, 1-9.
(5) IUPAC-IUB Commission on Biochemical Nomenclature *Eur. J. Biochem.* **1977**, *79*, 11-21.
(6) Luzzati, V. In *Biological Membranes, Physical Fact and Function*; Chapman, D., Ed.; Academic Press: New York, 1968; Vol. 1, pp 71-123.
(7) Fontell, K.; Mandell, L.; Ekwall, P. *Acta Chem. Scand.* **1968**, *22*, 3209-3223.
(8) Ekwall, P. *Advances in Liquid Crystallography*; Brown, G. H., Ed.; Academic Press: New York, 1971; Vol. 1, Chapter 1.
(9) Windsor, P. A. *Chem. Rev.* **1968**, *68*, 1-40.
(10) Reed, R. A.; Shipley, G. G. *Biochim. Biophys. Acta* **1987**, *896*, 153-164.
(11) Waldrop, M. M. *Science* **1990**, *248*, 674-675.

Structure Searches in Patent Literature: A Comparison Study between IDC GREMAS and Derwent Chemical Code

KARL HEINZ FRANZREB, PIA HORNBACH, CLAUDIA PAHDE, GOTTFRIED PLOSS, and JÜRGEN SANDER*

Hoechst Aktiengesellschaft, D-6230 Frankfurt/Main 80, Federal Republic of Germany

Received October 24, 1990

Patent searches by fragment codes in the Derwent WPI/L and IDC GREMAS Files as well as by closed substructures in the CAS Registry and the Derwent WPIM Files are compared. The most relevant answers are found in the IDC GREMAS File; by the Derwent fragment code the IDC result is almost reached. On the contrary, the search results obtained from the CAS Registry File are so incomplete that they are ruled out as an alternative. Derwent WPIM File as a successor to Derwent Chemical Code cannot serve as a replacement for the GREMAS File at the present time because of lack of completeness. Higher costs at IDC GREMAS compared with Derwent WPI/L for database use contrast with lower costs for relevance checking. On the whole, the IDC GREMAS File offers the best cost-benefit ratio for Hoechst AG.

INTRODUCTION

Particularly when investigating a patent status, it is important for industry to be able to have full information provided with as few irrelevant answers as possible. The Internationale Dokumentationsgesellschaft für Chemie mbH (IDC) was founded by chemical industries with this objective. Its task is to create and update files of information in the areas of primary interest to the shareholders; this also includes the maintenance and further development of the input and retrieval systems.

The GREMAS code is used for encoding full structures and generic structures in the field of organic chemistry. An efficient retrieval method developed for searching in this GREMAS File enables us to search with a high recall and precision.[1]

The IDC File is directed particularly to the needs of the shareholders and accordingly only covers the patent areas important to them. On the contrary, the WPI/L[7] File of Derwent covers all patent areas.[2] The GREMAS code on which the IDC Files are based is more precise than the chemical code used by Derwent for encoding structures; as a result IDC requires more time and effort for indexing but consequently yields more precise search results. This qualitative statement is known and has also been described in the literature.[3]

The aim of this study was to make a cost-benefit analysis between GREMAS and WPI/L searches. In addition, searches in the Registry and CA Files of the Chemical Abstracts Service (CAS) were included. On the one hand, we were concerned to make a quantitative analysis as to whether the higher indexing costs for the IDC could be compensated by cost savings in personnel for the relevance checking of a lower IDC proportion of false drops. On the other hand, we were concerned to make a value statement regarding the IDC File in comparison to the WPI/L and/or CAS Files on the basis of the number of relevant hits found.

Derwent will be suspending the indexing by chemical code in favor of the topological encoding by Markush DARC. The WPIM File based on Markush DARC was therefore also taken into account in this test; however, only for the small period of time hitherto available from Derwent week 01/87.

FILE CONTENTS

In principle the period from 1963 onward may be used for comparing the IDC GREMAS and the Derwent WPI/L Files. Since 1970, however, the CPI abstracts from Derwent have been used as a common basis for both files. They are indexed by Derwent using the chemical code and in parallel by IDC using the GREMAS code.

In the CA File from CAS also included in the test, patent literature is indexed in addition to journal literature. Searching of these patents using structure formulas can however only be made using specific patent examples filed in the Registry. Generic parts such as, for example, alkyl, are permitted for the structures of the query formulation in CAS, but the CAS Registry contains only specific compounds. The references

0095-2338/91/1631-0284$02.50/0