[Frontiers in Bioscience 2, a31-36, November 1, 1997]

FUNCTIONAL BIOINFORMATICS: THE CELLULAR RESPONSE DATABASE

James Sorace[1,2,3], Kip Canfield[1], Steven Russell[1]

[1] Department of Information Systems University of Maryland Baltimore County, [2]The Department of Pathology and Laboratory Service Baltimore VA Medical Center, Baltimore MD.[3] Department of Pathology, University of Maryland at Baltimore School of Medicine

TABLE OF CONTENTS

1. ABSTRACT

Biological Scientists function in an increasingly data rich environment. The emerging field of bioinformatics is attempting to insure that this flow of information can be structured to support the generation of significant biological hypothesis and ultimately new knowledge. To date, most of the current databases have focused on protein and nucleic acid sequence information as the principle type of data stored for further interpretation. In this paper, we describe the Cellular Response Database. This database stores functional information regarding the changes of cellular gene expression associated with various stimuli, and supports queries linking cell types, expressed genes, and inducers. The database is designed to support information-intensive queries to aid in the determination of biological function, and is flexible enough to allow the storage of a broad range of experimental data such as cytotoxicity data, immunoassays of target gene protein expression, and others.

2. INTRODUCTION

Biological scientists function in an increasingly data rich environment. With the rapid advancements in molecular biology, particularly at the level of characterizing large sets of gene products, the roles of factors influencing cellular gene expression and regulation will become increasingly important. The teratogenic drug thalidomide is an excellent example of this level of complexity. Depending on the dose, carrier and cell lines tested, various changes in lipopolysaccharide (LPS) induced tumor necrosis factor-alpha (TNF-alpha) gene regulation have been reported. These include inhibition (1,2,3,4,), as well as upregulation_ (5). Another example can be found in the use of cellular cytotoxicity data and cellular activity data to search for candidate anticancer drugs as outlined in a recent publication (6). In this strategy the cytotoxicity profile of several thousand compounds in 60 cell lines was compared against a panel of 113 biochemical properties of the cell lines. This allowed previously unrecognized relationships between drug sensitivity and cellular biochemical activities to be established. An improved information infrastructure will be required to manage the experimental data that is beginning to be produced in these settings.

A necessary prerequisite for any type of enterprise-wide information system implementation, is the development of detailed specifications and data models that determine what information is to be stored and what types of queries the database will support. The importance of this early phase of planning cannot be overemphasized. If useful information is not entered, or if it is poorly stored, the ability of the system to support user needs, will be compromised. Difficulties in querying the current biological databases have already been noted as a concern of the biotechnology and pharmaceutical industries (7). In this paper, we describe the ***Cellular Response Database*** **(CRD;***http://130.85.105.176:8080/assaydb/*). The purpose of this database and its associated web site is to develop the prototype database and collaborative research tools required to store and retrieve the functional data outlined above. The database is unique in that it treats the cell population, the target gene of interest, and the agents influencing gene expression equally, allowing querying on traits of any of these entities. This is important as genetic responses can vary widely depending on cell type, state of differentiation, and on the biological agent.

3. METHODS

This project required three distinct phases of development. First, it was necessary to conceptually model

Received 4/3/97 Accepted 10/24/97
Send correspondence to: James Sorace MD, Chief Blood Bank and Hematology Laboratories, Baltimore VA Medical Center, 10 N. Greene St., Baltimore MD 21201, Tel: (410)-605-7000 Ext. 5335, Fax: (410)-605-7911, E-mail: jmsorace@ ix.netcom.com

Table 1: Examples of Agent Notation

EXPERIMENT	NUMBER OF AGENTS	TEST AGENT	CONTROL AGENT	TARGET GENE/PROTEIN	TEST CELL POPULATION
The investigator wishes to determine if the production of TNF-alpha in human PBMC that is induced by cytokine A is inhibited by drug B.	2 (cytokine A, Drug B)	Drug B	None	TNF-alpha	Human PBMC.
As in Experiment 1, except that drug B is not water soluble. Instead it is dissolved in DMSO, a chemical that is known to alter gene expression in some circumstances. The experimenter controls for this by adding an equal concentration of DMSO to the control group. However because it is present in both the experimental and control groups DMSO is not a control agent.	3 (Cytokine A, Drug B, DMSO)	Drug B	None	TNF-alpha	Human PBMC
The investigator has determined that IFN-gamma induces the expression of a transcription factor. Using an anti-sense oligonucleotide A directed against the transcription factor, the investigator wishes to see if IFN-gamma induction of IL-12 is down regulated in the murine RAW 264.7 macrophage cell line. As a control an oligonucleotide B of identical composition but differing sequence to A will be used. Since oligonucleotide B is found only in the control group it is the control agent.	3 (IFN-gamma, oligonucleotide A, oligonucleotide B)	Oligonucleotide A	Oligonucleotide B	IL-12	murine RAW 264.7 macrophage cell line
The investigator wishes to determine if an anti-Met oncogene monoclonal antibody (Mab 1), inhibits hepatocyte growth factor induced MAP kinase phosphorylation in NIH 3T3 cells. As a control the investigator will use a monoclonal antibody directed against an irrelevant antigen (Mab 2).	3 (Mab 1, Mab 2, hepatocyte growth factor).	Mab 1	Mab 2	MAP Kinase	NIH 3T3 Cells

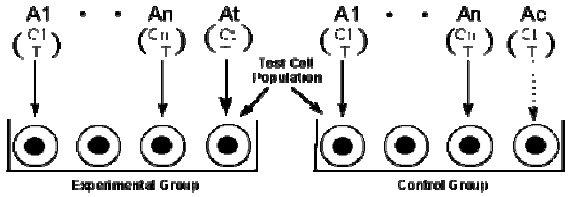

Figure 1: Experimental notation for the Cellular Response Database. Agents A1 to An are added to identical test cell populations at time (T) and concentrations (Cx). The test agent At (bold arrow) is added only to the experimental group. The control agent Ac (dashed arrow) may be added to the control group. A biological response between the groups is then measured.

the type of experimental design used in this field. Many cellular response experiments can be modeled as in figure 1.

Three types of biological entities are involved, the test cell population, the target gene/protein whose activity is assayed, and agents that are tested to see if they alter the target gene's expression or activity. These agents include protein molecules such as cytokines, sterol hormones or drugs. Their common property is that, at a specific time point in the experiment, the investigator adds them at a predetermined concentration. It must be remembered that various combinations of agents may be used within any experiment, and that dose response and kinetic data may be presented. However, appropriate experimental design allows comparisons between two groups of test cells that differ only by treatment with one agent. In figure 1, this is shown with a bold arrow and it represents the ***test agent***. In the thalidomide example noted above, LPS is present in both the experimental and control groups. Thalidomide is present only in the experimental group and is defined as the test agent. In other experimental designs, in addition to the test agent, a control agent must also be defined. In figure 1, this is shown with a dashed arrow. The control agent is added only to the control population of cells. In the thalidomide example noted above, a biologically inert thalidomide analog can be used as a control agent. In antisense experiments, the control agent may be an oligonucleotide with different sequence that is identical in composition to the test agent. Alternatively, if the test agent is an antibody, it would be an antibody of the same isotype with differing specificity. An experiment must always have one test agent. However, it may have several other agents, one of which may be a control agent. Table 1 gives several examples of how this notation is used.

The next phase of development consisted of several rounds of data modeling based on the experimental model discussed above. This resulted in the relational database schema presented in figure 2. This schema enables queries to be performed on test cell populations, the target protein, and the agents that are being tested. We have attempted to store important information associated with each of the experimental entities. For example, the target cell population may be a cell line (cloned or polyclonal), or primary culture (e.g. murine peritoneal exudate cells). The data model also supports searching by species, cell name, cell type, ATCC number, or organ of origin. Additional relevant data such as cellular concentration or culture conditions can be stored in a memo field in the database, or linked information in an electronic manuscript. Secondly, the gene/ protein target

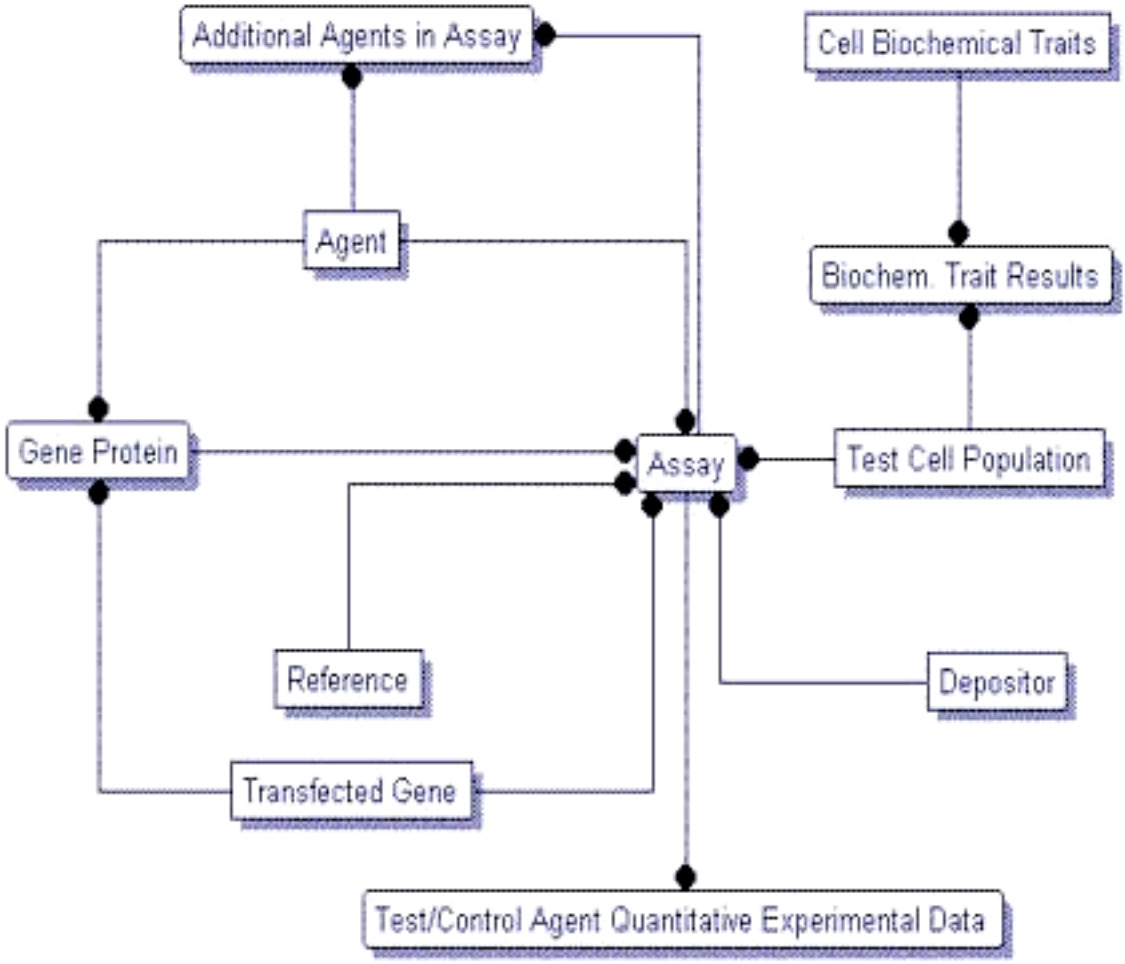

Figure 2: The data model for the Cellular Response Database: The most recent data model for this database is displayed.

can be referenced by name, species or GenBank number thus linking the target gene of interest to current sequence databases. The gene/protein target may be detected experimentally in several ways. In the thalidomide examples noted above this has included enzyme linked immunosorbent assay (ELISA) measurements of protein target (TNF-alpha), Northern blot measurement of the mRNA, or bioassay (L929 cell cytotoxicity). This broad range of possible detection systems, and the lack of numerical quantitation of many of them (e.g. bands on a Northern blot), is one of the major challenges for the database designer. In our implementation, we have adopted several approaches. First, a qualitative description of the change is entered (table 2) along with the type of detection assay used (ELISA, cytotoxicity, Northern blotting). Secondly, quantitative values and their units can also be defined and entered.

In some cases, assays may be run in which a specific gene/protein target is not measured, but a general biological property is. For example, considerable public domain data exist on the growth inhibition of tumor cell lines (6). The CRD can handle this type of data without modification by defining the gene/protein target as "none" (a null default value), and the assay type as GI50 (for growth inhibition 50%). Finally, a figure and caption can be linked. Also, as noted above, additional information can be stored in text fields or obtained if necessary by linking to a manuscript. Next, a Microsoft Access database has been developed and populated with data derived from several literature references. Several queries of this database are accessible through a CGI interface.

4. RESULTS OF TRIAL QUERIES

In order to demonstrate the utility of the data model, data from several of the thalidomide references noted above (1,2), and from a set of studies describing a family of interferon (IFN) induced GTPase like molecules (8,9,10), have been entered into the prototype CRD. The GTPase family currently consists of four published sequences, which show significant sequence homology with one another (8,9,10,11,12). Their expression is known to be modulated by a wide variety of agents including LPS, and IFNs. For some members of this family, patterns of inhibition or lack of induction by cytokines or drugs have also been reported. Based on these data, we have developed several queries that include:

Table 2: Patterns of target gene response

PATTERN #	DESCRIPTION
1	Up regulated, not detectable before treatment
2	Up regulated, but detectable before treatment
3	Down regulated
4	Basal level of expression unchanged
5	Not detected before or after treatment
6	Variable depending on dose

The database stores the qualitative interpretation of the response pattern in the Assay Table. In addition, numerical data can also be stored by the database.

1. Find the pattern(s) of response and manuscript reference for a given test agent. An example of this query, using thalidomide as the test agent, is shown in table 3, Panel A. Source data can be found in figure 2 and table 1 of reference 1, and figure 2 of reference 2.

2. Find the target genes up regulated by a cytokine, and display the test cell population. An example of this query is shown in table 3. Panel B, for IFN-gamma. Source data may be found in the first figure of references 8,9 and 10.

3. Find conditions in which an agent down-regulates the expression of a gene. The data from this query is shown in table 3, Panel C. This illustrates that LPS inhibits the ability of IFN-gamma to induce the MG21 GTPase. Source data may be found in figure 5 of reference 9.

4. List all other agents which have been assayed in combination with a given test agent. table 3, Panel D, lists all agents and assays that are linked to thalidomide as a test agent. Source data may be found in tables 1 and 2 of reference 1 (assay numbers 1-5, and 33), and figure 2A of reference 2 (assay number 6).

5. DISCUSSION

Our effort centers upon developing common data models for the storing and retrieving of data in biology and medicine. In addition to the current data model, we have proposed a data model for a Molecular Diagnostics Laboratory Information System intended for the clinical use (13). It is particularly important that these databases support queries based on the biological or medical context in which an experiment occurs. The undertaking of this type of data modeling is necessary if the concepts of intelligent agents, distributed collaboration, data mining and a shared intelligence (14,15) are to be applied to practice.

If widely implemented, a database like the one proposed above would have many advantages in generating new biological knowledge. First, comprehensive queries are rapidly displayed and more easily interpreted. Secondly, networks of control and regulatory pathways can begin to be investigated. It is increasingly recognized that the next phase of biological advances will involve the exploration of complicated networks of biological control (16). Once a critical mass of data has been entered it will be potentially possible to search for physiologically relevant pathways, and produce biologically relevant hypothesis for further experimental study. Finally, such an approach to data management would reduce redundant experimentation, and allow researchers with conflicting data to be aware of these results earlier. However, the development of shared data models, such as the one outlined above, is an absolute prerequisite before this level of enterprise informatics can occur.

Many difficulties associated with the development of an electronic database repository will need to be addressed. Several of these issues are outlined below:

1. Lack of standard data models and a formal syntax for the description of experimental data. Perhaps, the greatest challenge and opportunity in the field of bioinformatics is to design data models and logical notations that will support the future generation of knowledge. The data model proposed here and its notation (e.g. agent, test agent, control agent) represent one approach to some of these issues. These concepts can be extended. Some users may wish to use transfected genes as agents that modify the test cell population's target gene expression. For example, does the transfection of epidermal growth factor into fibroblasts increase cellular c-myc expression? The current database contains tables to support this particular possibility (figure 2). While these tables have not been fully implemented, they represent an explicit example of one way the database might be upgraded. However, no data model is entirely inclusive in that it can capture every type of data generated by an experimenter. There always will remain a need for the data to be annotated through the use of memo fields, and for links to be established to journal like entities.

2. Peer review: Assuring the entry of high quality data is critical for database success. Database entry would occur in 2 stages. First, the investigator would submit the paper to a participating journal for review, and its associated database would be placed in a provisional CRD database. This review would focus on the scientific merit of the paper. The reviewers would now be able to query and reformat the submitted data directly, as well as search the published database for related information. Once accepted for publication by the journal, the data entered in the provisional database would be transferred to the fully on-line version.

It is important to recognize that the paradigms governing the scale of biological research are also changing. Software like the CRD may enable large collaborative groups, using good laboratory practice protocols, to establish high quality databases. These groups may divide along the test agent used (chemokines or IFN) or assay types (protein or mRNA expression). By using a common database the compilation and quality control of such an effort can be improved. Regardless of whether databases are populated by individual submissions

Table 3: Outputs of example queries*

PANEL A

Agent.Name	Gene_Protein.Nam	Species	Pattern of Response	First Author	Journal	Volume	Page
Thalidomide	TNF-Alpha	Human	UP regulated, but present before stimulation	Makonkaweyoon S	Proc Natl Acad Sci, USA	90	5974
Thalidomide	HIV-1 Reverse Transcritptase Activity	Human	Down regulated	Makonkaweyoon S	Proc Natl Acad Sci, USA	90	5974
Thalidomide	TNF-Alpha	Human	Down regulated	Peterson PK	J Infect Dis	172	1137
Thalidomide	HIV-1 Reverse Transcritptase Activity	Human	Variable depending on conditions	Makonkaweyoon S	Proc Natl Acad Sci, USA	90	5974

PANEL B

Agent.Name	Agent Role	Assay_Counter	Cell_Name	Cell_Type	Gene_Protein.Name
IFN- gamma	TEST	14	RAW 264.7	Macrophage	LRG-47
IFN-gamma	TEST	20	BCL1	B-cell lymphoma	LRG-47
IFN-gamma	TEST	25	RAW 264.7	Macrophage	iGTP
IFN-gamma	TEST	28	RAW 264.7	Macrophage	iGTP
IFN-gamma	TEST	30	Peritoneal Cells	Macrophage	Mg21
IFN-gamma	TEST	31	Peritoneal Cells	Macrophage	Mg21

PANEL C

Agent.Name	Exposure (hour)	Concentration (microgram/ml)	Additional Agents By Assay	Concentration	Units
LPS	6	10	IFN-gamma	1000	Units/ml

Gene_Protein.Name	Cell._Name	Cell_Type
Mg21	Peritoneal Cells	Macrophage

PANEL D

Test agent	Assay_Counter	Additional Agent
Thalidomide	1	Phorbol 12-myristat 13-acetate
Thalidomide	2	GM-CSF
Thalidomide	2	LPS
Thalidomide	3	GM-CSF
Thalidomide	3	IL-6
Thalidomide	4	IL-6
Thalidomide	4	LPS
Thalidomide	5	IL-6
Thalidomide	5	IL-3
Thalidomide	6	LPS
Thalidomide	33	Phorbol 12-myristat 13-acetate

*: See text for details. The amount of information displayed per query has been limited for clarity.

or by large group efforts, issues regarding review and quality assurance will be the subjects of considerable future debate.

3. Data entry and presentation: Careful consideration will need to be given to interface design to assure that data can be entered quickly and accurately. If the laboratory were to use a database with similar data structures to the depositories, submission could be largely automated. Conversely, once the data is entered, the development of useful query and presentation formats will be crucial to support end users.

4. Archiving of historical data: Data that has already been generated will not be archived in an electronic database. It is possible that a highly selected subset of such publications may be retrospectively archived. However, this criticism overlooks the fact that the rate of data generation will only increase as biological research continues to advance, inundating the current information infrastructure and rendering it obsolete.

In this report, we have outlined one possible way to advance bioinformatics. However, the biological community should be given the opportunity to influence the types of databases that need to be developed and supported by collaborative data entry. In an effort to initiate a dialogue, the interested reader is invited to complete a survey associated with the CRD web site. We will publish the results of this survey to help guide the future database design. The interested reader is invited to further participate in the development of this database by visiting the prototype CRD web site. Currently, investigators can deposit their data provided the work has been accepted or published by a peer-reviewed journal.

6. REFERENCES

1. Makonkaweyoon S, Limonson-Pobre RNR, Scauf V, & Kaplan G: Thalidomide inhibits the replication of human immunodeficency virus type 1. *Proc Natl Acad Sci USA* 90, 5974-5978 (1993). PMID: 8327469

2. Peterson PK, Hu S, Sheng WS, Kravitz FH, Molitor TW, Chatterjee D, & Chao CC.: Thalidomide inhibits tumor necrosis factor-alpha production by lipopolysaccharide and lipoarabinomannan-stimulated Human microglial cells. *J Infec Dis* 172, 1137-40 (1995). PMID: 7561198

3. Moreia AL, Sampaio EP, Zumidzinas A, Frindt P, Smith KA, & Kaplan G: Thalidomide exerts its inhibitory action on tumor necrosis factor alpha by enhancing mRNA degradation. *J Exp Med* 177, 1675-1680 (1993). PMID: 8496685

4. Shannon EJ, Sandoval F, & Krahenbul JL: Hydrolysis of thalidomide abrogates its ability to enhance mononuclear cell synthesis of IL-2 as well as its ability to suppress the synthesis of TNF-alpha. *Immunopharamacol Immunotoxicol 36, 9-15 (1997).* PMID: 9129992

5. Shannon EJ, & Sandoval F: Thalidomide can be either agonistic or antagonistic to LPS evoked synthesis of TNF-alpha by mononuclear cells. *Immunopharamacol Immunotoxicol* 18, 59-72 (1996). PMID: 8683039

6. Weinstein JN, Myers TG, O'connor PM, Friend SH, Fornace AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwin JK, vasn Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, & Paul KD: An informational intensive approach to the molecular pharmacology of cancer. Science 275, 343-349 (1997). PMID: 8994024

7. Williams N: Drug firms move to link databases. Science 277, 902 (1997). PMID: 9281071

8. Sorace JM, Johnson RJ, Howard DL, & Drysdale BE: Identification of an Endotoxin and IFN-inducible cDNA: possible identification of a novel protein family. *J Leukoc Biol* 58, 477-84 (1995). PMID: 7561525

9. Lafuse WF, Brown D, Castle L, & Zwilling BS: Cloning and characterization of a novel cDNA that is IFN-gamma-induced in mouse peritoneal macrophages and encodes a putative GTP-binding protein. *J Leukoc Biol* 57, 477-83 (1995). PMID: 7884320

10. Taylor GA, Jeffers M, Largespada DA, Jenkins NA, Copeland NG, & Vande Woude GF: Identification of a novel GTPase, the inducibly expressed GTPase, that accumulates in response to interferon gamma. *J Bio Chem* 271, 20399-20405 (1996). PMID: 8702776

11. Gilly M, & Wall R: The IRG-47 gene is IFN-gamma induced in B cells and encodes a protein with GTP-binding motifs. *J Immunol* 148, 3275-81 (1992). PMID: 1578148

12. Carlow DA, Marth J, Clark-Lewis I, & Teh HS: Isolation of a gene encoding a developmentally regulated T cell-specific protein with a guanine nucleotide triphosphate-binding motif. *J Immunol* 154, 1724-34 (1995). PMID: 7836757

13. Sorace JM, Ritondo M, &Canfield K: Information engineering for molecular diagnostics. In: Proceedings of the 18th Annual Symposium on Computer Application in Medical Care: 293-297, (1994). PMID: 7949937

14. M. Wooldridge, JP Muller, & M Tambe eds. Lecture Notes in Artificial Intellegence 1037: Intellegent Agents II, agents, theories, architectures, and languages. Proceedings of IJCAI'95 Worhshop (ATAL) Montreal Canada, August 1995. Springer-Verlag 1996.

15. Zare RN. Knowledge and distributed intellegence, *Science* 275, 1047 (1997).

16. Cohen J. The Genomics Gamble. *Science* 275, 767-772 (1997). PMID: 9036536