Signature: Med Sci Monit, 2003; 9(8): MT89-95
PMID: 12942040

Received: 2003.03.14
Accepted: 2003.07.15
Published: 2003.08.27

Authors' Contribution:
A Study Design
B Data Collection
C Statistical Analysis
D Data Interpretation
E Manuscript Preparation
F Literature Search
G Funds Collection

From transcriptomics to bibliomics

Bertrand Henri Rihn[1], Solveig Vidal[2], Claude Nemurat[1], Sébastien Vachenc[1], Steve Mohr[1], Florian Mazur[2], Philippe Houdry[2], Francoise Grandjean[1], Sophie Visvikis[3], Jacques Ducloy[2]

[1] Institut National de Recherche et de Sécurité (INRS), Vandoeuvre-les-Nancy, France
[2] Institut de l'Information Scientifique et Technique (INIST-CNRS), Vandoeuvre-les-Nancy, France
[3] Equipe 4 INSERM U525, Faculté de Pharmacie, Nancy, France

Source of support: none.

MT

Summary

Background: Current biological investigations tend to operate with genomes, instead of genes as during the last century. It is possible to compare entire genomes, transcriptomes or proteomes, using alphanumeric data corresponding to the differential expression levels of thousands of genes. What remains difficult is to link array results to factual or bibliographical data and retrieve information that is highly structured and - in Shannon's sense - rare.

Material/Methods: We have developed a tool, Documentation and Information LIBrary (DILIB), that enables us to retrieve, organize and analyze huge amounts of data available on the Internet and related to microarray experiments. DILIB can link hundreds of differentially expressed genes – through their Single Identifier or GenBank accession number – to hundreds of Medline records, and can retrieve, analyze, and compare automatically thousands of non-trivial descriptors related to gene clusters.

Results: As exemplified with frequency comparison of MEdical Subject Headings and Registry Number descriptors, we reanalyzed the involvement of 'integrin', 'interleukin' and 'CD Antigens' in mesotheliomas. Thus, DILIB allowed us to: (i) associate literature to expressed genes, (ii) link functional transcriptomes in various experiments, (iii) associate specific descriptors to experiments, (iv) define new research areas, and eventually (v) find new functions for co-expressed genes.

Conclusions: We propose a new concept, 'bibliomics', representing a subset of high quality and rare information, retrieved and organized by systematic literature-searching tools from existing databases, and related to a subset of genes functioning together in '-omic' sciences.

key words: **microarray • data mining • investigation server • systematic bibliometry • metathesaurus • bioinformatics • transcriptomics • genomics • proteomics • bibliomics • mesothelioma**

Full-text PDF: http://www.MedSciMonit.com/pub/vol_9/no_8/3543.pdf
Word count: 3187
Tables: 4
Figures: 2
References: 24

Author's address: Bertrand Henri Rihn, Equipe 4 INSERM U525, Faculté de Pharmacie, 30 Rue Lionnois 54000 Nancy, France, email: Bertrand.Rihn@pharma.uhp-nancy.fr

BACKGROUND

Along with analyses of the expression of several thousand genes using microarray technology, there is a parallel need for retrieval and analysis of text-derived information from existing databases [1,2]. To achieve this task, a high priority has been given by various bioinformatic teams to find fast, cheap and time-saving methods. In microarray experiments, usually only 3% of the analyzed genes display a differential expression level. Thus for the estimated human genome, composed of approx. 25,000 genes, a given pathological or physiological state may involve approx. 1,000 differentially expressed genes [3]. It is also a major challenge to achieve algorithms for other warehoused data, such as macromolecule sequences and genetic modifications at the genome level in cancer [4].

At present there is a huge amount of underused information in natural language stored in existing biological databases. Craven and Kumlien [5] developed methods for the automatic analysis of non-coded sentences; these methods, based on the knowledge of semantic rules, offer a promising alternative to handcoding information and extraction routines. By extracting data *via* 'relational learning', these authors found information on localization of proteins at the subcellular level. The corresponding algorithm can be used for semantic as well as syntactic tasks. One possible avenue of development of such tools is their use in relational training with Unified Medical Language Subject (UMLS), a highly structured vocabulary used for MeSH indexing in Medline® (*http://www.ncbi.nlm.nih.gov/*). Because of this unmanageable set of data sources and formats spread throughout the Internet, a SQL-like query language has been proposed to link the outputs of Genbank and Genecards databases. Therefore, natural language queries for syntactic and semantic heterogeneities were translated into a computational language that makes it possible to search for and retrieve relevant data [6]. This use of this method can be exemplified by searching for a target DNA sequence for a new drug or by distributing data source access through XML (eXtended Markup Language) outputs. Intelligent information systems for processing natural language have also been shown to work in noun automatic analysis, as exemplified by the work of Ono et al [7]. The latter performed an automated analysis of all descriptors related to protein interaction, e.g. of keywords like 'association', 'binding', or 'complex'. The collection of proteins involved in interaction(s) allowed them to build a protein interaction database. Shatkay et al. [8] pointed out the need for bioinformatic software to work in a biological environment to allow better surveying of relevant literature. They developed a method to retrieve automatically data on functional relationships and specific literature, in order to produce a short summary for every single gene.

There was no example in the literature, however, of automatic retrieval, analysis and association processing of literature related to clusters of genes differentially expressed and identified by microarray experiments. Here we present the Documentation and Information LIBrary (DILIB), a software platform, consisting in a set of engineering programs (see http: //dilib.inist.fr/Demos.html). At present, DILIB makes it possible to create bibliographic investigation servers and perform automatic analysis and processing of hundreds of Medline® records. These records correspond to differentially expressed genes from 7,000- [9] and 10,000-gene [10] microarray experiments. The automatic single or associated terms [Title], [Abstract], [Registry Number] and [MEdical Subject Heading] related to records of differentially expressed genes make it possible to perform bibliometric studies and easily compare descriptors and themes of over- and underexpressed gene series. Both the comparison and association studies developed here allowed us to:

(i) find biologically relevant data that were hidden in microarray experiments;
(ii) gain new insights into the physiopathology of mesothelioma;
(iii) define hypothesis-driven research areas.

MATERIAL AND METHODS

DILIB (Document and Information LIBrary) is a prototype workbench, developed by the Centre National de Recherche Scientifique (CNRS), Institut National de Recherche en Informatique et Automatique de Lorraine (INRIA) and Institut de l'Information Scientifique et Technique (INIST), for document engineering. DILIB works under the C language and allows for automatic processing and integration of data retrieved from various databases. It is basically composed of two kinds of tools:

(i) functions and commands for handling SGML trees and records;
(ii) instruments to build Information Retrieval Systems and generate so-called 'investigation servers', such as the ones we built for the present study.

DILIB works under the Unix operating system. Using XML derived from SGML, as defined by ISO-8879 and described by Jolibois et al. [11], DILIB allows heterogeneous records to be collected and standardized in tagged fields. Records are stored in an HFD repository (Hierarchical File for Documentation). This method allows the the DILIB platform to store as many as 10^6 records in basic HFD configuration. Other items, such as inverted files, whose records are also coded into SGML form, improve easy access to related records, e.g. through records with similar keys (keywords or authors).

This function has been used to develop most of the bibliometric modules. From an inverted file, it is possible to build association files, where the frequency of co-occurrence of keywords is stored. Finally, this association file is used to build a cluster file, which is also a set of SGML documents, one per cluster. A cluster forms a group of 'internal associations', describing the relations existing between descriptors inside the cluster, which represent a scientific 'theme' according to our nomenclature. On the other hand, associations that link several clusters are called 'external associations.'

Two cell lines used in a previous study – Met-5A pleural cells (CRL-9444) and MSTO-211H mesothelioma cells

Table 1. Word sorting in "Abstract", "Medical Subject Heading", "Registry Number" and "Title" indexes.

Gene corpus	Index	Words before sorting	Words after sorting	% rejected words
Overexpressed	Abstract	672	359	47%
	MeSH	828	540	35%
	RN	508	465	8%
	Title	885	494	45%
Underexpressed	Abstract	618	408	34%
	MeSH	774	580	25%
	RN	455	431	5%
	Title	818	566	31%

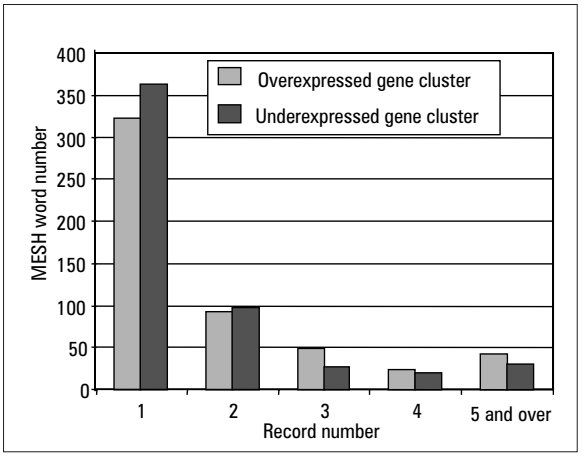

Figure 1. Distribution of MeSH descriptors by number of records in over- and underexpressed clusters.

(CRL-2081, American Type Culture Collection, Manassas, Virginia, USA) – were also used in the present study. Target cDNA from both cells was hybridized on Incyte Pharmaceuticals™ arrays containing 6969 probes with sequences complementary to 3962 human genes and 3007 human expressed sequence tags (ESTs) (Human Unigem V™, Genome Systems Inc, St. Louis, Missouri, USA). A typical experiment comparing their expression level showed 4% differentially expressed genes [9]. The entire set of screened genes can be downloaded at *http://www1.inrs.fr/INRS-PUB/inrs01.nsf/inrs01_search_view_view/C3350637CB90F92CC1256CDF00618CF9/$FILE/visu.html?OpenElement*(to download click on ‹‹Analyse de cellules humaines en culture››). Thus 242 genes were overexpressed in mesothelioma cells (here called 'overexpressed genes' in mesothelioma cells) and 257 were underexpressed in mesothelial cells (here called 'underexpressed genes' in mesothelioma cells). The same number of 'Second Identifiers' [SI], corresponding to Genbank© accession numbers, were queried in Pubmed© (*http://www4.ncbi.nlm.nih.gov/entrez/query.fcgi*). However, only 200 and 202 Medline© records were retrieved, representing records related to overexpressed genes of mesothelioma and mesothelial cells, respectively. Inversely, these records were related to underexpressed genes of mesothelial and mesothelioma cells, respectively. These differences were due to ESTs (Expressed Sequence Tags) which do not belong to any Medline© record.

Similarly, the cDNA of two mesothelioma tumors was compared to pleural cells, Incyte Pharmaceuticals™ arrays (Human UnigemVII™, Genome Systems Inc), which enable the screening of 10,200 elements, including 9,984 known human genes and expressed sequence tags and 216 internal controls [10]. In this experiment, 71 and 139 genes were respectively over- and underexpressed in tumor specimens, compared to pleural cells that corresponded respectively to 304 and 394 Pubmed references. This entire set of genes may be downloaded at *http://www1.inrs.fr/INRS-PUB/inrs01.nsf/inrs01_search_view_view/C3350637CB90F92CC1256CDF00618CF9/$FILE/visu.html?OpenElement*(to download click on ‹‹Analyse de tumeurs humaines››).

The Medline© fields kept in the DILIB database were [AB, AD, AU, ID, MH, RN, SI, TA, TI]. Two bibliographic record corpuses extracted from Medline© and respectively related to over- and underexpressed gene clusters were loaded as two separate databases in two integrated 'investigation servers', the so-called 'Cell transcriptome' (http://dilib.inist.fr/dps/sdv/Genome/Server/EN.Genome.index.html) and 'Tumor transcriptome' (http://dilib.inist.fr/dps/sdv/transcriptome/Server/EN.Genome.index.html) for the cell and tumor comparison experiments, respectively. For each transcriptome, four indexes were made: MH (MeSH descriptors), TI (title descriptors), ABS (abstract descriptors), and RN (registry number), which contained all descriptors of their respective Medline© fields, namely MH, TI, AB and RN of the retrieved Medline© records. All MeSH descriptors (MEdical Subject Heading, [MH] field) were retrieved from Medline©, but not all were kept in DILIB index tables. The non-significant words were pooled in a 'rejection' vocabulary table. These words were not relevant or too general, e.g. '*Rodentia*', software', support US government', cDNA'. However, descriptors related to organs, organelles, chemicals, biological processes, pathways, functions, pathologies, etc. were kept and indexed using the DILIB functions. The same selection was done with trivial vocabulary in titles and abstracts. The MeSH and Registry Number [RN] descriptors appearing in both databases (under- and overexpressed genes) were labeled with '(c)' (for common words), making the comparison more easy. For the sake of example, the index and rejected tables of the cell transcriptome servers are shown at http://dilib.inist.fr/dps/sdv/Genome/Server/PagesINRS/EN.Index1.html. A function of both investigation servers enabled us to determine the frequency of a given keyword in MeSH, title, and abstract, as well as in Registry Number descriptors. The most cited descriptors and frequency of co-occurrence of two given words can also be measured and compared in both corpuses of the investigation server, which can be consulted at the above-mentioned websites.

RESULTS

Table 1 shows the number of records before and after manual selection of words from Medline©: ABS, MH, TI

Table 2. Pubmed© publication numbers and ID (PMID) for "Integrin*" and "Mesothelial" or "Mesothelioma"*.

Integrin Type	Mesothelioma	Mesothelial
Integrin without ID	10783328; 9789201; 8681612	7909596
β1	9013837; 9212227; 9872600; 9378538	9212227; 11266240
β3	–	9212227; 7560092; 10427122; 11266240
β4	9013837; 10427122	7560092
β5	10427122; 7536280	7560092
α3	9013837; 9872600; 9378538	11266240
α2	9378538	11266240
α6	10427122; 9378538; 9013837	11266240
α1	9013837	
α5	9013837; 10427122	10427122; 11266240
αv	9013837; 7536280; 10427122	11266240
Total Pubmed© records	24	15

* as retrieved by 20 December 2001

Table 3. Pubmed© records for "Interleukin*"queries.

Query*	"Interleukin and Mesothelial or Pleural"	"Interleukin and Mesothelioma"
Interleukin	18	15
Interleukin 1	15	4
Interleukin 2	2	–
Interleukin 4	1	–
Interleukin 7	–	1
Interleukin 8	–	5
Interleukin 12	–	–
Interleukin 15	–	–

* as retrieved by 20 December 2001

Table 4. List of inner associations for the RN cluster "0 (integrin *alpha6*) – 0 (*Antigens*, CD)".

0 (integrin *alpha6*) – 0 (Antigens, CD)(c)
60-92-4 (Cyclic AMP)(c) – 362-74-3 (Bucladesine)(c)
362-74-3 (Bucladesine)(c) – 0 (CLA-1 protein)
0 (Platelet Membrane Glycoproteins) – 0 (CLA-1 protein)
0 (Platelet Membrane Glycoproteins) – 0 (Antigens, CD34)
0 (Platelet Membrane Glycoproteins) – 0 (Antigens, CD)(c)
0 (CLA-1 protein) – 0 (Antigens, CD34)
0 (CLA-1 protein) – 0 (Antigens, CD)(c)
0 (Antigens, CD34) – 0 (Antigens, CD)(c)

(c) – the RN descriptor is common for both over- and underexpressed gene databases;

0 – a generic protein without CAS number

and RN descriptors of the cell transcriptome servers. The rejected words, stored in the respective anti-dictionaries, varied from 5 to 47% and corresponded to generic words, such as 'DNA', homology', and genus characterizations, such as '*Homo sapiens*','*Rodentia*'. These anti-dictionaries can be applied to any new experiment, as displayed at http: //dilib.inist.fr/dps/sdv/Genome/Server/PagesINRS/EN.Index1.html. Common descriptors for both corpuses in the investigation server were also labeled with a '(c)' The RN descriptors are usually very informative, as they represent the names of proteins which are coded by over- and underexpressed genes. This may explain the relative low level of rejected descriptors in this series (5 to 8%). The number of associated Medline© records for a given descriptor varied mainly from 1 to 4 (Figure 1). In our example', protein conformation' referred to 16 records and 'DNA binding' to 30 records in over- and underexpressed gene clusters, respectively, as retrieved from both MeSH fields. The power of automatic bibliography retrieval with DILIB tools is indicated by the four examples elaborated below.

In the cell transcriptome servers, the descriptor 'integrin*' appeared approximately 4.4 times more frequently in the overexpressed series of descriptors (regardless of whether they are ABS, MH, TI, or RN descriptors) as compared to the underexpressed one. For example 'integrin' was cited 8 and 2 times in record titles of over- and underexpressed genes, respectively. When a Boolean search 'integrin' and 'mesothelioma' limited to the abstract field was performed in Medline©, the result was 24. This number is to be compared with 15 references with co-occurrence of 'mesothelial OR pleural AND integrin*' (Table 2). Thus the increased occurrence of the 'integrin' descriptor in the overexpressed series seems to be indicative of a biological fact. Indeed, it turned out that integrin α6, β4 subunits and αv-β5 were expressed in malignant mesothelioma as compared to a single integrin isoform (β3 subunit) which was found in normal mesothelial cells [12].

An analysis of the 'interleukin*' query in MeSH indexes of the cell transcriptome servers is also of interest. Although no interleukin encoding gene (of the 27 that were screened in the cell transcriptome experiment) was differentially expressed, the 'interleukin' MeSH descriptors were associated in both the indexes of underexpressed and overexpressed genes. Automatic retrieval of documentation with DILIB made it possible to associate the overexpressed gene series to interleukin-1, -7, and -8, and the underexpressed to interleukin-1b, -2, -4, -12, and -15. Moreover, when these descriptors were retrieved from the tumor transcriptome servers, only inter-

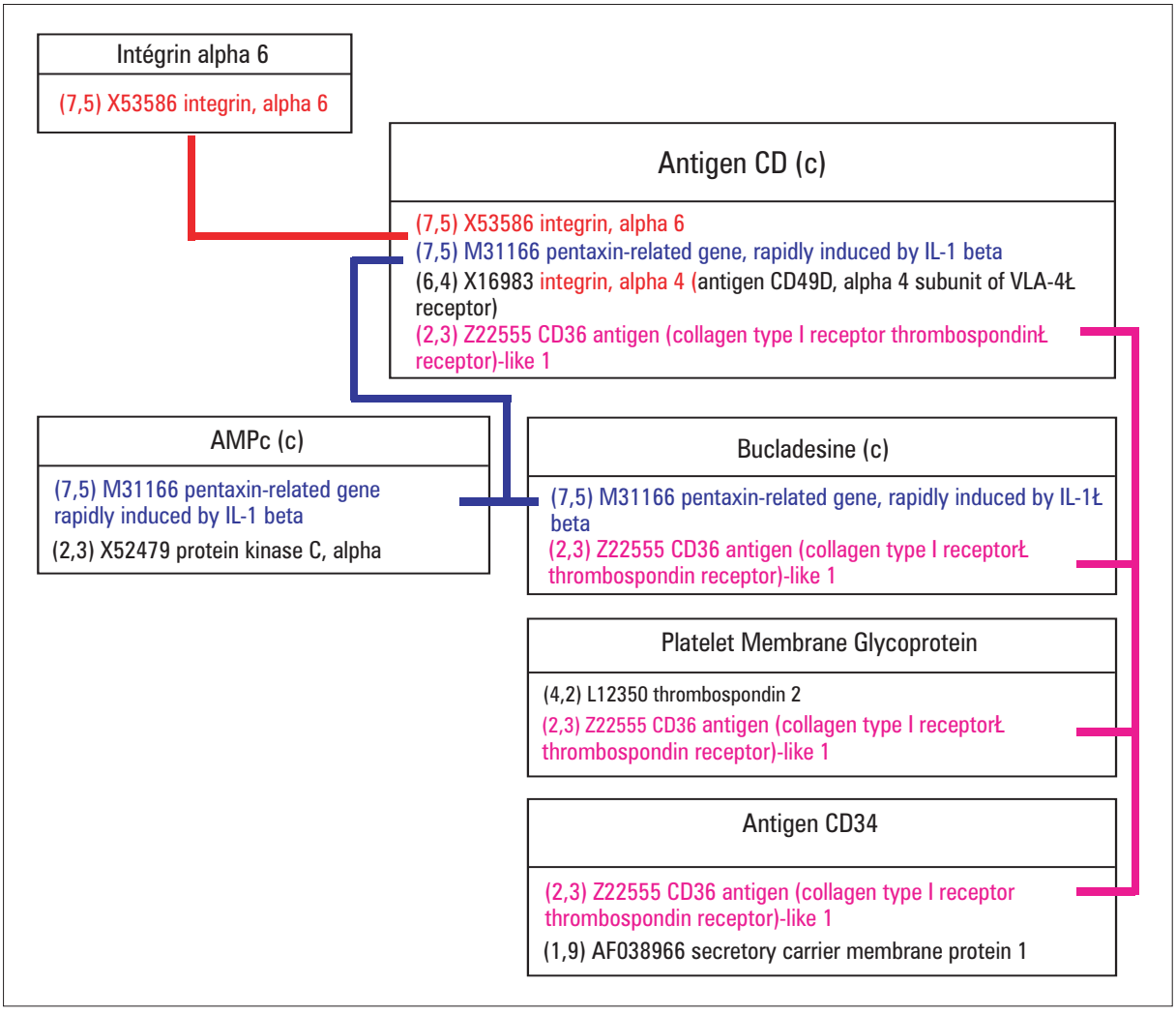

Figure 2. How gene cluster are related to the RN cluster; e.g. by '0 (integrin alpha6) - 0 (Antigens, CD)'. The number before the Genbank ID (Single Identifier [SI] in Medline(taxonomy) of a given gene corresponds to the level of overexpression [9]. Colored gene ID indicates a primary association of 3 genes (X53586, M31166 and Z22555). For example 'Antigen CD' is a common Registry Number (RN) of X53596 and M31166, displaying a primary association. 'Bucladesin' is a common RN of M31166 and Z22555, therefore showing a primary association. However, the latest (Z22555) was linked through the 'Platelet Membrane Glycoprotein' registry number by a secondary association represented by the gene ID in boldface type. Indeed X16983, X52479, L12350 and AF038966 were each linked through a secondary association with a primary associated gene (a 'colored'-typed gene) in 4/5 displayed RN associations. Therefore all of these genes appeared more strongly linked, as they were all overexpressed. It should be interesting in future experiments to see if such subsets of gene functions and their products vary together.

leukin-1 was associated with the overexpressed gene series, and both interleukin-2 and -4 were found in the MeSH descriptors of the underexpressed series. Furthermore, a Medline© Boolean keyword search, namely 'Mesothelial OR Pleural AND Interleukin*', performed on 12/20/01, showed 18 references associated with these keywords, of which fifteen concerned interleukin-1, two interleukin-2 and one interleukin-4 (Table 3). Moreover, in 9 Medline© records that associated 'interleukin* AND mesothelioma', 4 concerned interleukin-1, one pertained to interleukin-7, and five dealt with interleukin-8. Thus, in a certain way, relevant biological data may be found with the automatic information retrieval, given that interleukin-8 was recently recognized as an autocrine factor in mesothelioma cells [13]. An additional advantage could be rapid identification of some new research areas, e.g. the role of interleukin-12 and -15 in mesothelial cells, as both keywords were retrieved from the underexpressed gene database of the investigation server and were never described in mesothelial cells at the time of query.

Another example is provided by occurrence analysis of cluster differentiation antigens, called 'Antigen CD' in the cell transcriptome servers. The mesothelial cells were associated with CD 4, CD 55 and CD 59 according to bibliometric retrieval of the microarray underexpressed genes. In contrast, mesothelioma cells were associated with antigens CD 9, 34, 36, 44 and 95. A further Medline© search displayed 24, 14 and 1 records for respectively

'Mesothelioma AND CD34', Mesothelioma AND CD44' and 'Mesothelioma AND CD95'. Only one record was returned by the association of 'Mesothelial AND CD4'. Rapid analysis of the role of CD95 (i.e. Fas ligand) in mesothelioma showed that in 6 malignant pleural mesothelioma lines, either Fas or Fas ligand was expressed at the protein level [14]. Such associations, evidenced by RN automatic descriptor analysis, are thus supported by relevant literature from the Medline© database.

A powerful application of DILIB automatic keyword clustering and inner associations of themes was given by association studies of RN (Registry Number). The RN cluster '0(Integrin alpha 6)-0(Antigens, CD)' retrieved from the overexpressed gene server provided 7 descriptors and 9 inner associations (Table 4), finally corresponding to 7 overexpressed genes, which were linked by DILIB as indicated in Figure 2. As a matter of fact, this inner association directly linked 3 overexpressed genes, and through a secondary association, 4 others (Figure 2). Thus these genes, strongly linked by the literature searching-tool developed here, may be analyzed together in further designed experiments on mesothelioma, and may eventually lead to the discovery of previous unsuspected relations. This DILIB function allowed for a rapid survey of related literature of gene networks created by DILIB.

Discussion

Andrade and Bork [15] developed a tool which enables association of keywords related to diseases from electronically published literature described in the Online Mendelian Inheritance in Man (OMIM) database. The OMIM keywords have been categorized and organized in ontology fields, so that knowledge can be acquired or discovered. Previously, software was developed that allowed automatic protein annotations by ranking and scoring vocabulary related to proteins and selected in Medline© [16]. Such applications can be of use when researchers try to compare microarray data for hundreds or thousands of genes, differentially expressed and stored in *ad hoc* 'warehouses' [17], as we did in our study.

Jenssen et al. [18] described a tool for automatic literature extraction using a protein co-citation index. Their hypothesis was that a 'biological meaningful relationship' between two genes exists if they were co-cited in the MeSH descriptors. The automatic analysis of protein co-citation allowed them to establish a virtual gene network called Pubgene (*http: //www.pubgene.uio.no*) and to rank genes according to biological processes. Parts of these gene networks have been validated by microarray results. Similarly, semantic associations achieved by DILIB functions also allowed us to define gene clusters related through MeSH and RN descriptors in order to find some functional relationships and complex interplay, as shown in Figure 2.

The question that arises in the present post-genomic era is how to weigh biological relevance and measure the strength of the gene clusters issuing from microarray experiments. To make microarray powerful [19], it is absolutely necessary to link them to the most complete set of biological facts established by previous experimentation, by using more 'classical' methods (cloning and discovery of gene new functions, protein expression analysis, RT-PCR, and so forth) and retrieved from the most complete databases [20]. Comparing microarray data as soon as they are standardized and annotated, possibly using XML language, would be achievable using software such as that described by Hayes [21]. As array results are purely numeric, the corresponding genes should be linked to databases such as Medline©, to make the array results more informative. Such work has been performed by Masys et al. [22]. They used, as we did, controlled terminology, e.g. MeSH descriptors and RN keywords, which is known to reflect biomedical papers, to the greatest possible extent, by short and concise descriptions. Retrieved MeSH descriptors were translated into parent concept identifiers, according to the classification of the UMLS metathesaurus. The authors validated their method with the microarray data obtained by Golub et al. [23]. The DILIB platform described in this paper also allowed us to extract MeSH descriptors, RN, keywords and abstract words for over- and under-expressed genes and to compare them. We recently improved the DILIB tools by using the parent concept hierarchies given by the UMLS metathesaurus applied to the cell transcriptome servers, as shown at http: //dilib.inist.fr/dps/sdv/GenomeAP/Server/EN.genome.index.html.

Another example of the power of computational analysis can be found in the work of Tsoka and Ouzounis [24], who developed such a tool to query databases (e.g. Swissprot©) and retrieve the distribution of Enzyme Classification (EC) number along a given metabolic pathway. They aimed eventually to describe by such queries all possible cellular functions of a given cell. As shown in our study, all EC numbers were easily retrievable through the RN number of encoded enzymes.

Conclusion

After semantic comparison of words, associations and theme clusters, and elimination of the common descriptors of both under- and overexpressed gene series, we isolated some frequent descriptors in each series and queried them in Pubmed©. The above mentioned examples clearly showed that information retrieved through DILIB and associated to over- and underexpressed genes, or any other chosen cluster of genes, may be related to relevant biological facts and may be faster evidenced through screening of hundreds or thousands of Medline© records associated with those genes. Finally, in this work, we showed how DILIB allowed us to *(i)* link microarray data to existing bibliographic data, *(ii)* compare descriptors associated to over- and underexpressed genes, *(iii)* 'weight' clusters of genes differentially expressed, and *(iv)* link gene descriptor associations to biological facts previously evidenced by other groups working on mesothelioma. Hence we propose to introduce the concept of 'bibliomics', corresponding to the use of systematic searching strategies that enable biological facts evidenced by genomic, transcriptomic,

proteomic or metabolomic studies to be linked, as quickly and efficiently as possible, to existing scientific literature and databases. One future task will be to link Genecards©, Locuslink© or other genomic databases existing on the Internet through DILIB functions, whatever their warehouse file format, to data obtained by microarray experiments.

Acknowledgments

The authors are indebted to Alain Zasadzinski, Aude Nedelcot, Cécilia Fabry, and David Moulin for their initial work and improvements in DILIB, and to Gérard Keith and Marcel Rubio for manuscript reading. This publication is dedicated to Laure Anaïs Marie, for her courage.

REFERENCES:

1. Alfred J: Mining the bibliome. Nature Reviews Genetics, 2001; 2: 401
2. Brush M: Making sense of microchip array data. The Scientist, 2001; 15: 25
3. Mohr S, Leikauf GD, Keith G et al: Microarrays as cancer keys: an array of possibilities. J Clin Oncol, 2002; 20: 3165-75
4. Paton NW, Khan SA, Hayes A et al: Conceptual modelling of genomic information. Bioinformatics, 2000; 16: 548-57
5. Craven M, Kumlien J: Constructing biological knowledge bases by extracting information from text sources. In Proceedings of the AAAI Conference On Intelligent Systems In Molecular Biology, 1999; 77-86
6. Eckman BA, Kosky AS, Laroco LA Jr: Extending traditional query-based integration approaches for functional characterization of post-genomic data. Bioinformatics, 2001; 17: 587-601
7. Ono T, Hishigaki H, Tanigami A et al: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics, 2001; 17: 155-61
8. Shatkay H, Edwards S, Wilbur WJ et al: Genes, themes and microarrays: using information retrieval for large-scale gene analysis. In: Proceedings of the AAAI Conference On Intelligent Systems In Molecular Biology, 2000; 8: 317-28
9. Rihn B, Mohr S, McDowel SA et al: Differential gene expression in mesothelioma. FEBS Letters, 2000; 480: 95-100
10. Mohr S, Galateau-Salle F, Keith G et al: Cell protection, resistance and invasiveness of two malignant mesotheliomas as assessed by 10k-microarrays. Submitted for publication, 2003
11. Jolibois S, Nauer E, Chouaniere D et al: La gestion informatisée de corpus bibliographiques: adaptation des normes et formats documentaires. Bulletin des Bibliothéques de France, 2000; 45: 98-108
12. Giuffrida A, Vianale G, Di Muzio M et al: Modulation of integrin expression on mesotheliomas: the role of different histotypes in invasiveness. Int J Oncol, 1999; 15: 437-42
13. Galffy G, Mohammed KA, Dowling PA et al: Interleukin 8: an autocrine growth factor for malignant mesothelioma. Cancer Res, 1999; 59: 367-71
14. Stewart JH 4th, Nguyen DM, Chen GA et al: Induction of apoptosis in malignant pleural mesothelioma cells by activation of the Fas (Apo-1/CD95) death-signal pathway. Thorac Cardiovasc Surg, 2002; 123: 295-302
15. Andrade MA, Bork P: Automated extraction of information in molecular biology. FEBS Letters, 2000; 476: 12-7
16. Andrade MA, Valencia A: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatics, 1998; 14: 600-7
17. Bassett DE, Eisen MB, Boguski MS: Gene expression informatics - it's all in your mine. Nature Genetics, 1999; 21: 51-5
18. Jenssen TK, Laegreid A, Komorowski J et al: A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, 2001; 28: 21-8
19. Masys DR: Linking microarray data to the literature. Nature Genetics, 2001; 28: 9-10
20. Gaasterland T, Bekiranov S: Making the most of microarray data. Nature Genetics, 2000; 24: 204-6
21. Hayes A: The second international meeting on microarray: data standards, annotations, ontologies and databases. Yeast, 2000; 17: 238-240
22. Masys DR, Welsh JB, Fink JL et al: Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics, 2001; 14: 319-326
23. Golub TR, Slonim DK, Tamayo P et al: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999; 286: 531-7
24. Tsoka S, Ouzounis CA: Recent developments and future directions in computational genomics. FEBS Letters, 2000; 480: 42-8

Index Copernicus

Global Scientific Information Systems for Scientists by Scientists

TM
INDEX COPERNICUS INTERNATIONAL

www.IndexCopernicus.com

- EVALUATION & BENCHMARKING
- PROFILED INFORMATION
- NETWORKING & COOPERATION
- VIRTUAL RESEARCH GROUPS
- GRANTS
- PATENTS
- CLINICAL TRIALS
- JOBS
- STRATEGIC & FINANCIAL DECISIONS

Index Copernicus integrates

IC Scientists

Effective search tool for collaborators worldwide. Provides easy global networking for scientists. C.V.'s and dossiers on selected scientists available. Increase your professional visibility.

IC Virtual Research Groups [VRC]

Web-based complete research environment which enables researchers to work on one project from distant locations. VRG provides:

- customizable and individually self-tailored electronic research protocols and data capture tools,
- statistical analysis and report creation tools,
- profiled information on literature, publications, grants and patents related to the research project,
- administration tools.

IC Journal Master List

Scientific literature database, including abstracts, full text, and journal ranking. Instructions for authors available from selected journals.

IC Patents

Provides information on patent registration process, patent offices and other legal issues. Provides links to companies that may want to license or purchase a patent.

IC Grant Awareness

Need grant assistance? Step-by-step information on how to apply for a grant. Provides a list of grant institutions and their requirements.

IC Conferences

Effective search tool for worldwide medical conferences and local meetings.

IC Lab & Clinical Trial Register

Provides list of on-going laboratory or clinical trials, including research summaries and calls for co-investigators.