Journal of Informetrics xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays

Andrei Mogoutov[a], Alberto Cambrosio[b,*], Peter Keating[c], Philippe Mustar[d]

[a] *Aguidel, Paris, France*
[b] *Department of Social Studies of Medicine, McGill University, Montreal, Canada*
[c] *Département d'Histoire, Université du Québec à Montréal, Montreal, Canada*
[d] *Centre de Sociologie de l'Innovation, École Nationale Supérieure des Mines, Paris, France*

ARTICLE INFO

Article history:
Received 26 May 2008
Accepted 30 June 2008

Keywords:
Microarrays
Biomedical innovation
Triple-helix
University–industry–government relations
Publications
Patents
Research projects
Text-mining
Network analysis
Visualization

ABSTRACT

Using the example of microarrays, one of the constitutive technologies of post-genomic biomedicine, this paper introduces a method for analyzing publications, patents and research grants as proxies for "triple-helix interfaces" between university, industry and government activities. Our method creates bridges that allow one to move seamlessly between publication, patent and research project databases that use different fields and formats, and contain different information. These links do not require pre-defined categories in order to search for correspondences between sub-topics or research areas in the three databases. Finally, our results are not restricted to quantitative information but, rather, allow one to carry out qualitative investigations of the content of research activities. Our approach draws on a combination of text-mining and network analysis/mapping software packages.

1. Introduction

This paper is part of a larger project focusing on the development of one of the emergent technologies of contemporary biomedicine: microarrays, also referred to as biochips, gene chips and so on. A single DNA microarray contains thousands of short DNA sequences arrayed on a solid surface. Compared to previous molecular genetic approaches, a microarray experiment involves the simultaneous analysis of many hundreds or thousands of genes, as opposed to single genes, thus making microarrays a tool of choice of the post-genomic era. As a result, the number of articles based on microarray technology has grown exponentially during the last decade, from less than 200 in 1999 to over 6000 in 2005, for a cumulative total of approximately 30,000 articles in 6 years. These articles are not confined to experimental reports: an increasing number of publications describe clinical applications of microarrays (e.g., Harris & Horning, 2006; Perkel, 2005; Quackenbush, 2006; Simon & Wang, 2006; see Keating & Cambrosio, 2004 for an initial analysis). The growth of commercial activities is equally impressive: the annual compounded growth rate of the microarray market between 1999 and 2004 has been estimated at 63% (Constans, 2003), and the 2006 revenues of the two major companies in the field, Affymetrix and Illumina, were,

* Corresponding author.
E-mail addresses: andrei@aguidel.com (A. Mogoutov), alberto.cambrosio@mcgill.ca (A. Cambrosio), keating.peter@uqam.ca (P. Keating), philippe.mustar@ensmp.fr (P. Mustar).

1751-1577/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2008.06.005

respectively, US$ 355 million and US$ 184 million.[1] The number of patents in the domain numbers in the thousands. In short, microarray technology is not only growing rapidly but it is presently being translated into the scientific, clinical and commercial domains.

As a field, microarrays instantiate our claim that in order to chart and understand the dynamics of the development of biomedicine one must follow the (in no way unidirectional) circulation of innovations between laboratories, clinical settings and biotechnology companies (in particular, biotech start-ups and spin-offs). Choices made in academic laboratories often play a key role in the definition of new biomedical platforms that shape novel clinical approaches in the life sciences (Keating & Cambrosio, 2003). Key-users and lead-users play a central role in the development and formatting of products (Oudshoorn & Pinch, 2003; von Hippel, 1976, 1986, 1989), for instance, by establishing conventions, standards and other forms of regulations that make possible the routine use of a given technology. In the field of microarrays this is done, for instance, both by creating hybrid (academic-industrial) bottom-up institutions such as MGED (Microarray Gene Expression Data Society) and through state-sponsored initiatives such as the FDA MicroArray Quality Control (MACQ) Project (see Cambrosio, Keating, Bourret, Mustar, & Rogers, in press; Rogers & Cambrosio, 2007). In turn, the survival of firms that commercialize the technology depends on the evolution of research practices in academic settings (Mustar, 1997, 1998, 2001). Finally, the development of microarray technology has led to increasingly porous boundaries between (public and private) clinical, laboratory and commercial activities, and thus to the establishment of hybrid institutional forms; this is especially true of bioinformatics, a new line of work that is essential to the pursuit of microarray experiments (Fransman, 2001; McMeekin, Harvey, & Gee, 2004; Powell, Koput, & SmithDoerr, 1996).

A detailed rendering of the processes we have just mentioned can only be provided by thorough ethnographic investigations, but the latter can in turn be informed by the semi-quantitative analysis of the interfaces between university, industry and government as represented by proxies such as publications, patents and state-funded research projects. To gain an understanding of their interaction, these three sources of information should be analyzed as a whole and not as separate entities. Salerno, Landoni, and Verganti (2006) provided the initial impetus for this paper. Noting that publications and patent data have often been used as proxies to examine the two-way traffic between the laboratory and the world of commerce, they championed an extension of this approach to a new source of data, namely research projects listed in the databases of grant agencies. Investigating nanotechnology, they used keywords to select sub-topics within this broad domain and compared the differential pattern of emergence of each sub-topic by juxtaposing curves expressing the number of projects and patents awarded by the US National Science Foundation and by the US Patent Office during the last 10–20 years. We decided to expand on this approach by adding publications to projects and patents, and by devising a method to explore the "triple-helix interfaces" that emerge between these three sources of data. Our method sought to fulfill three requirements:

- A first, basic requirement was to create "bridges" that would allow us to move seamlessly between the three data sets: publications, patents and research projects. The challenge here resides in the fact that the different databases use different fields and formats, and the information they contain is structured and indexed differently.
- Secondly, we sought to avoid using pre-established categories, derived either from existing research typologies or from standardized nomenclatures and keywords, to define sub-topics or research areas. Since one of the properties of innovations is, by definition, to chart novel areas and to redefine boundaries between domains, one should "let the data speak for themselves" when identifying research topics, rather than forcing novelty into old pigeonholes.
- A final requirement was to obtain results that are not limited to a quantitative assessment of the state of the field but, rather, allow one to undertake a qualitative inspection of the content of the research and commercial activities in a given domain.

To fulfill these requirements, we tested a combination of two approaches. The first uses text-mining software to extract relevant concepts from titles and abstracts (as contrasted, for instance, with keywords); the second uses network analysis/mapping software, to replace statistical indicators with the analysis of semantic, citation, institutional and authorship networks. While scientometric approaches that have tried to go beyond mere productivity measurements and assess the content of scientific publications have traditionally focused on the development of more sophisticated statistical methods for analyzing citation or keyword clusters, we believe that future advances in this domain will depend on the use of the increasingly sophisticated *linguistic* tools that are now available for the analysis of unstructured, natural language documents. This article is primarily a methodological note, rather than a substantive analysis of the microarray domain, and demonstrates an approach to the semi-quantitative exploration of socio-technical interfaces in a rapidly expanding biomedical domain.

2. Databases and software packages

The first step in any procedure such as the one described in this article is to gather the relevant data. This involves two choices: first, the choice of databases from which to retrieve source data for further treatment, and, second, the choice of

[1] According to the Datamonitor Company Profiles: Document Code DCC06282-9FEA-4BD4-A49B-DCB7E4C2E4F4 (Affymetrix/April 2007) and A8DD7B13-69B1-4126-A04C-71F7920B1876 (Illumina/May 2007). According to a Frost and Sullivan report, the total 2005 DNA microarray revenue in the United States was US$ 446 million (Flanagan, 2007, p. 1, 42, 43).

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

a method for retrieving those data. We discuss below the choice of databases but, before doing that, we comment of the choice of a method for querying them. To be fully consistent with our second requirement that calls for avoiding the use of pre-established categories, we should have chosen a text-mining approach rather than resorting to a query based on existing classification systems, such as the MeSH keyword system built into the *PubMed* database. Text-mining allows, in principle, to retrieve references that are missed by a keyword query because of, for instance, sloppy indexing practices or a reference's deviant profile. In order to do so, however, one should text mine the whole database that, in the case of PubMed, contains millions of references: a very impractical solution.[2] We therefore opted for a compromise solution, namely to use a keyword query to retrieve a (large) set of source data for subsequent analysis via text-mining. Our choice presupposes, in principle, the availability of a robust keyword indexing system such as the one used by PM. As discussed, however, by Mougoutov and Kahane (2007), in the case of databases with incomplete or weakly structured keyword systems, and in the case of emergent domains that have not yet been captured by a stabilized set of keywords, it is possible to refine keyword queries by using a scalable, evolutionary query design.

2.1. Publication data

Data concerning microarray publications can be obtained from several sources, including the following two major bibliographic databases: *MEDLINE*/*PubMed* [PM] of the U.S. National Library of Medicine and *Web of Science* [WoS] from Thomson Scientific. As just mentioned, PM has a robust, hierarchical keyword system (the *MeSH* controlled vocabulary) that allows for data retrieval with a very good sensitivity/specificity ratio. Major drawbacks of PM include the fact that it is limited to biomedical journals, it provides only the address of first authors and does not contain a list of the references cited by individual articles. This latter characteristic is, of course, the major advantage of WoS, which also dispenses with the two other shortcomings of PM by listing the addresses of all the authors of a given article and by including publications from all scientific fields. Unfortunately, the WoS does not have the equivalent of the *MeSH* indexing system and thus WoS queries generate more noise. One way of combining the advantages of each database is to use a matching algorithm to automatically extract references elicited with a more specific PM query from a broader WoS database (Cambrosio, Cotterau, Popowycz, Mogoutov, & Vichnevskaia, 2007): a PM-WoS matching algorithm is available in the network analysis software *ReseauLu* (see below). In the present case, preliminary investigations showed that, in spite of the fact that PM also covers bioinformatics and bioengineering publications, the technological, computational and statistical aspects of microarrays were better represented in WoS. Given the purpose of our project and the fact that subsequent text-mining and network mapping treatments eliminate the noise (low frequency concepts are, for instance, eliminated), we opted for a WoS database. By way of comparison, a PM query[3] and a WoS query,[4] both performed in January 2007, resulted, respectively, in 27,519 and 34,553 (29,242 if we limit to research articles) references.

2.2. Research project data

While publication databases contain references catalogued at an international level, research project databases are mostly limited to national data. Given the major role played by US scientists and clinicians in the microarray field, two obvious choices for databases were the CRISP (Computer Retrieval of Information on Scientific Projects) database of the US National Institutes of Health (NIH)[5] and the database maintained by the US National Science Foundation (NSF).[6] Mowery (2001) claims that more than 70% of the federally funded academic R&D is supported by the NIH, with a much smaller share supported by the NSF. Moreover, given the largely biomedical nature of microarray technology, it is hardly surprising that the NSF database listed (December 2006) only 600 microarray-related projects, as compared to the 13,954 projects[7] harvested from the CRISP database.[8] Maintained by the NIH Office of Extramural Research, CRISP includes projects funded by all the institutes of the

[2] Moreover, while in the case of, say, the US Patent Office database, one can argue that the database contains the entire set of US patents, databases such as *PubMed* or *Web of Science* are anyway not comprehensive, insofar as they only list references from a selected (albeit large) number of journals.

[3] We used the following query (Microarray Analysis) OR ((Nucleic Acid Hybridization) AND (Microarray*)) OR (cDNA Microarray*) OR (cDNA Array*) OR (DNA Microarray*) OR (DNA Chip*) OR (DNA Microchip*) OR (DNA Array*) OR (Gene Expression Chip*) OR (Gene Chip*) OR (Gene Expression Microarray Analysis) OR (Oligonucleotide Array*) OR (Oligonucleotide Microarray*) OR ((Gene Expression Profiling) AND (Microarray*)) OR (cDNA Microarray*) OR (cDNA Array*) OR (DNA Microarray*) OR (DNA Chip*) OR (DNA Microchip*) OR (DNA Array*) OR (Gene Expression Chip*) OR (Gene Chip*) OR (Gene Expression Microarray Analysis) OR (Oligonucleotide Array*) OR (Oligonucleotide Microarray*).

[4] We used the following query: "Microarray*" OR "Oligonucleotide Array Sequence Analysis" OR "Microarray Analysis" OR "cDNA Microarray*ÖR "cDNA Array*ÖR "DNA Microarray*ÖR "DNA Chip*ÖR "DNA Microchip*ÖR "DNA Array*ÖR "Gene Expression Chip*ÖR "Gene Chip*ÖR "Gene Expression Microarray AnalysisÖR "Oligonucleotide Array*ÖR "Oligonucleotide Microarray*ÖR "Gene Expression Profiling.

[5] http://crisp.cit.nih.gov/.

[6] http://www.nsf.gov/awardsearch/.

[7] CRISP has its own system of Thesaurus Terms that includes the keyword "Microarray Technology". The number of retrieved references rises to 16,974 if one includes subprojects (i.e., distinct research projects within large grants). Many thanks to Elliot and Rediet Berhane from the NIH who provided us with explanations concerning the CRISP database and, most importantly, with an Excel version of the data.

[8] On the European side, the CORDIS database (http://cordis.europa.eu/) that lists projects funded by the European Community yielded only approximately 70 microarray-related projects, which is not surprising given that for the 2002–2006 period CORDIS lists a total of 234 projects for the "Life sciences, genomics and biotechnology for health" Thematic Priority. European researchers have, of course, alternative (national) funding sources.

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

NIH (National Cancer Institute, National Institute of Mental Health, etc.) and by additional agencies such as the Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDCP).

2.3. Patent data

Patent information can be obtained from national patent offices, such as the US Patent Office or from the *Derwent Innovation Index* database curated by Thomson Scientific, that contains patents from all national and international patent offices. The design of queries for patent data is notoriously tricky given the idiosyncratic indexing system and the specialized vocabulary used by these documents. In the present case, after using a string of keywords derived from MeSH terms and drawing on personal acquaintance with the literature,[9] we compiled and parsed the resulting set of over 8000 patents in order to discard documents that simply listed microarrays in passing: by keeping only patents identified by keywords in the title and the abstract claims, and rejecting those patents in which the relevant keywords only occurred in the extended abstract, we obtained a final yield of 6480 patents.

2.4. Text-mining software

Keywords such as MeSH terms are often used to explore the content of scientific texts. While readily available, they have several drawbacks: terms absent from a text are routinely added by indexers to its list of keywords, new biomedical terms are added to the MeSH Thesaurus only after a certain delay and the retrospective indexing of articles is far from consistent; most importantly, because of their standard nature, MeSH keywords are ill-suited to map distinctive and emerging research patterns. Text-mining technologies, although they burden the analysis with additional treatments, provide an optimal solution by extracting from publications the single term and multi-term expressions, such as "gene expression profile" actually used by the authors of a given text. Space limits prevent us from discussing here the basic principles of text-mining: for a general introduction readers can refer to Feldman and Sanger (2007) as well as, for the specific case of biomedical texts, to Ananiadou and McNaught (2006). We note, however, the difference between text-mining tools based exclusively on statistical approaches and Natural Language Processing (NLP) tools that resort to linguistics-based algorithms, since considerations related to this distinction have guided our choice of a software program. Briefly put, most statistics-based systems simply count the number of times terms occur and calculate their statistical proximity to related terms – these tools, therefore, work best with large samples – whereas NLP tools equipped with internal, hard-coded dictionaries use a sequence of morphological, syntactic, semantic, pragmatic and statistical treatments in order to assign a part-of-speech category (noun, verb, adjective, . . .) to terms, examine relations between terms, solve ambiguity issues and select candidate concepts. Concept extraction is improved by using external, domain-specific dictionaries (e.g., genomics), synonym dictionaries, and user-defined dictionaries to exclude or include certain concepts. NLP-based tools, in other words, can recognize and extract compound words, phrases and idioms that would typically be treated as individual words by other products, thus dramatically increasing the overall accuracy of the system.[10]

Faced with the choice of several text-mining software programs, including some available from academic institutions, we opted for a commercial package (*SPSS LexiQuest Mine* and *Text-Mining Builder* [henceforth LQM])[11] because commercial products, in addition to the fact that they come with full technical support and documentation, offer standard interfaces and stable versions that, in principle, should improve inter-researcher comparability. LQM is a linguistics-based, Natural Language Processing software program. Thus, using the NLP and statistical algorithms we just mentioned, LQM can identify multi-term concepts that carry semantic meaning (e.g., "breast cancer patients" rather than the constituent single terms) as well as equivalence classes (e.g., "cancer of the breast" = "breast cancer"). Other features include the possibility of extracting terms only when they are part of multi-term expressions,[12] and of easily replacing abbreviations (a major issue in biomedical texts; see Chang & Schutze, 2006) via its linguistic-resource customization interface that, moreover, comes equipped with a series of dedicated dictionaries, for instance, a genomics, a gene ontology and a MeSH dictionary. Other dictionaries are available for domains other than biomedicine or can be imported, if needed.

[9] The search parameters included the Chemical, the Electrical and Electronic, and the Engineering databases. We used the following query: TS = (Microarray*) OR TS = (micro_array*) OR TS = (oligodeoxyribonucleotide_microchip*) OR TS = (oligodeoxyribonucleotide_micro_chip*) OR TS = (oligonucleotide_array*) OR TS = (oligonucleotide_microchip*) OR TS = (oligonucleotide_micro_chip*) OR TS = (cDNA_microarray*) OR TS = (cDNA_microchip*) OR TS = (cDNA_micro_chip*) OR TS = (cDNA_micro_array*) OR TS = (cDNA_array*) OR TS = (DNA_microarray*) OR TS = (DNA_microchip*) OR TS = (DNA_micro_chip*) OR TS = (DNA_micro_array*) OR TS = (DNA_Chip*) OR TS = (DNA_array*) OR TS = (gene_expression_chip*) OR TS = (gene_expression_microchip*) OR TS = (gene_expression_micro_chip*) OR TS = (expression_chip*) OR TS = (gene_chip*) OR TS = (oligonucleotide_array*) OR TS = (oligonucleotide_microarray*) OR TS = (oligonucleotide_micro_array*) OR TS = (nucleic_acid_array*) OR TS = (nucleic_acid_microarray*) OR TS = (ribonucleic_acid_array*) OR TS = (ribonucleic_acid_microarray*) OR TS = (ribonucleic_acid_microchip*) OR TS = (ribonucleic_acid_micro_chip*) OR TS = (nucleic_acid_microchip*) OR TS = (nucleic_acid_micro_chip*).

[10] Having used for a previous paper (Bourret et al., 2006) a statistical/neural network-based text mining software program, in preparation for the present work we reanalyzed those data using NLP software as described in this article and we obtained vastly superior results.

[11] http://www.spss.com/lexiquest/lexiquest_mine.htm.

[12] While the keywords that have been used to query a database will occur in a large percentage of documents and co-occur with a majority of the extracted concepts, thus carrying little differential meaning, complex expressions that contain those terms are meaningful and should be extracted: "cancer" is too generic in a cancer database, but "breast cancer", "colon cancer" or "breast cancer epidemiology" carry important information.

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

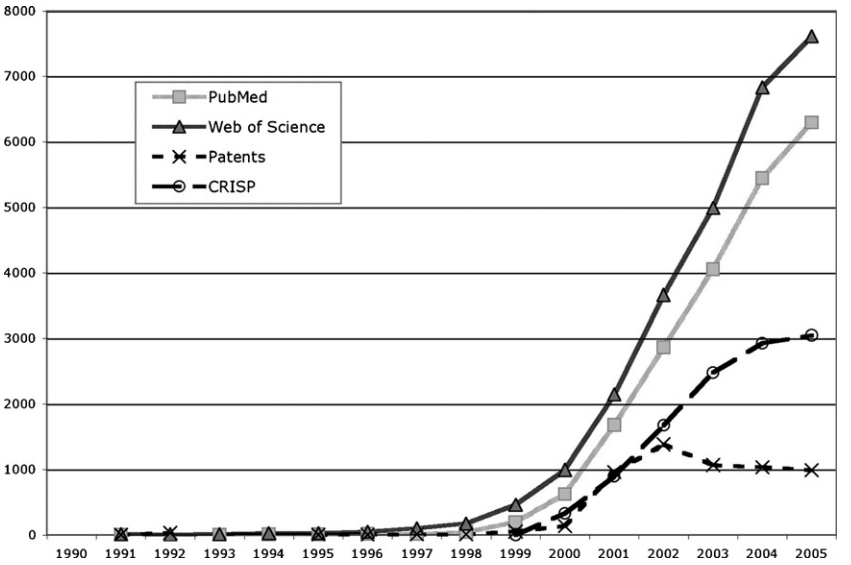

Fig. 1. Number of publications (source: PubMed and Web of Science), patents (source: Derwent Innovation Index) and NIH research grants (source: CRISP database) by year in the field of microarrays.

2.5. Network analysis software

To map and analyze the semantic network of concepts extracted by LQM as well as other kinds of relevant networks (co-authorship, inter-citation, etc.) we used *ReseauLu* [RL], a network analysis software designed specifically for the treatment and mapping of complex, heterogeneous relational data.[13] RL uses a dynamic node placement algorithm that mobilizes several approaches, including the simulation of the interaction between geometric objects submitted to random forces, the non-linear projection on a plane of a multidimensional structure and a "spring" model for the optimisation of the distance between objects. The final configuration of the networks results from a three step optimisation process: (i) global initial positioning of the object vis-à-vis all the other nodes in the space; (ii) micro-optimisation of the positioning of the nodes vis-à-vis the other nodes to which it is directly connected ("network neighbours"); and (iii) meso-optimisation of groups of highly connected objects ("clusters"). The optimization process depends on explicit rules defining symmetry properties, structural equivalence of points inside the structure, centrality and "between-ness" of nodes. The resulting map has no meaningful axes. With the most recent version of RL (RLX2) the results of queries from several databases, including PM, WoS and the output of LQM treatments, can be automatically imported and parsed to generate a large number of predefined maps that include both homogeneous maps (e.g., co-authorship, journal-to-journal citation, and institutional collaboration maps) and heterogeneous maps (e.g., authors and journals, and countries and keywords maps), whose parameters can be further refined, for instance, to display only the more specific links or to change the number of nodes displayed on a map.

3. Methodology and results

3.1. Characterizing the field of microarrays

Prior to developing a method for bridging the databases described in the previous section, we used bibliometric indicators and network analysis to gain a better understanding of the characteristics of the microarray domain. Fig. 1 shows the growth of projects, publications and patents retrieved by the queries described in the previous section. Although articles describing microarray technology and experiments first appeared in the early 1990s, the growth of the field did not begin in earnest until the end of the decade. It then proceeded at a hectic pace. The initial take-off period occurred simultaneously for publications, projects and patents. Given that research budgets cannot grow at the same pace and rate as publications, the growth of research projects included in the CRISP database appears to have subsequently reached a plateau. As for patents, as we will explain below, a restriction of the criteria for obtaining patents in this domain accounts for the peculiar pattern of the curve after the initial growth period.

Fig. 2, a co-authorship network of the most cited authors in the field, shows that the domain is characterized by the presence of a few leading actors that are strategically situated between clusters of co-authors (the size of the nodes is proportional to the number of co-signed articles). Table 1 confirms this analysis by showing, for instance, that two authors

[13] The latest version is ReseauLuX2; http://www.aguidel.com; on *ReseauLu*, see also Bourret et al. (2006) and Cambrosio et al. (2006, 2007).

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

Fig. 2. Co-authorship network of the most cited authors in the field of microarrays (data source: Web of Science; network analysis and visualization: ReseauLu X2).

– Brown and Botstein – who occupy a strategic central position in Fig. 2, dominate the field in terms of the number of publications and the number of citations they received; most importantly, a large number of their publications figure among the most cited papers in their domains. The other authors further down the list also occupy significant topological real estate as measured by these same parameters. As a result of this concentration around a few key actors and of the simultaneous

Table 1
The most visible authors in the domain of microarrays (Source: Web of Science; data analysis: ReseauLu X2; NbPub = total number of publications; NbCit = total number of citations; NbCitMax = number of citations received by an author's most-cited article; Top 0.01% [0.1% cum; 1% cum] = number of an author's articles in the top 0.01% [0.1%; 1%] most-cited articles in the corresponding field)

Name	NbPub	Nbcit	NbCitMax	Top 0.01%	Top 0.1% cum	Top 1% cum	Top 10% cum
Brown, P.O.	116	31793	4231	17	33	71	96
Botstein, D.	77	21524	4231	15	27	49	67
Tibshirani, R.	45	8363	2339	14	20	27	32
Golub, T.R.	43	9405	2561	9	16	28	35
Staudt, L.M.	52	6946	2339	8	17	30	40
Shinozaki, K.	42	1721	328	8	15	23	34
Seki, M.	37	1692	328	8	15	20	32
Simon, R.	76	3916	732	7	17	32	54
Pinkel, D.	49	2813	693	6	8	25	37
van de Rijn, M.	38	5432	1455	6	12	24	34
Rosenwald, A.	47	5122	2339	6	15	24	32
Chinnaiyan, A.M.	57	2849	549	5	11	27	40
Davis, R.W.	45	8259	3223	5	11	20	31
Li, C.	49	2643	918	5	8	16	32
Albertson, D.G.	39	2547	693	4	5	18	29
Gerstein, M.	38	2380	685	4	6	14	23
Rubin, M.A.	55	2603	549	3	6	18	40
Zhang, L.	84	1497	797	3	6	17	29
Quackenbush, J.	43	1748	385	3	8	16	28

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

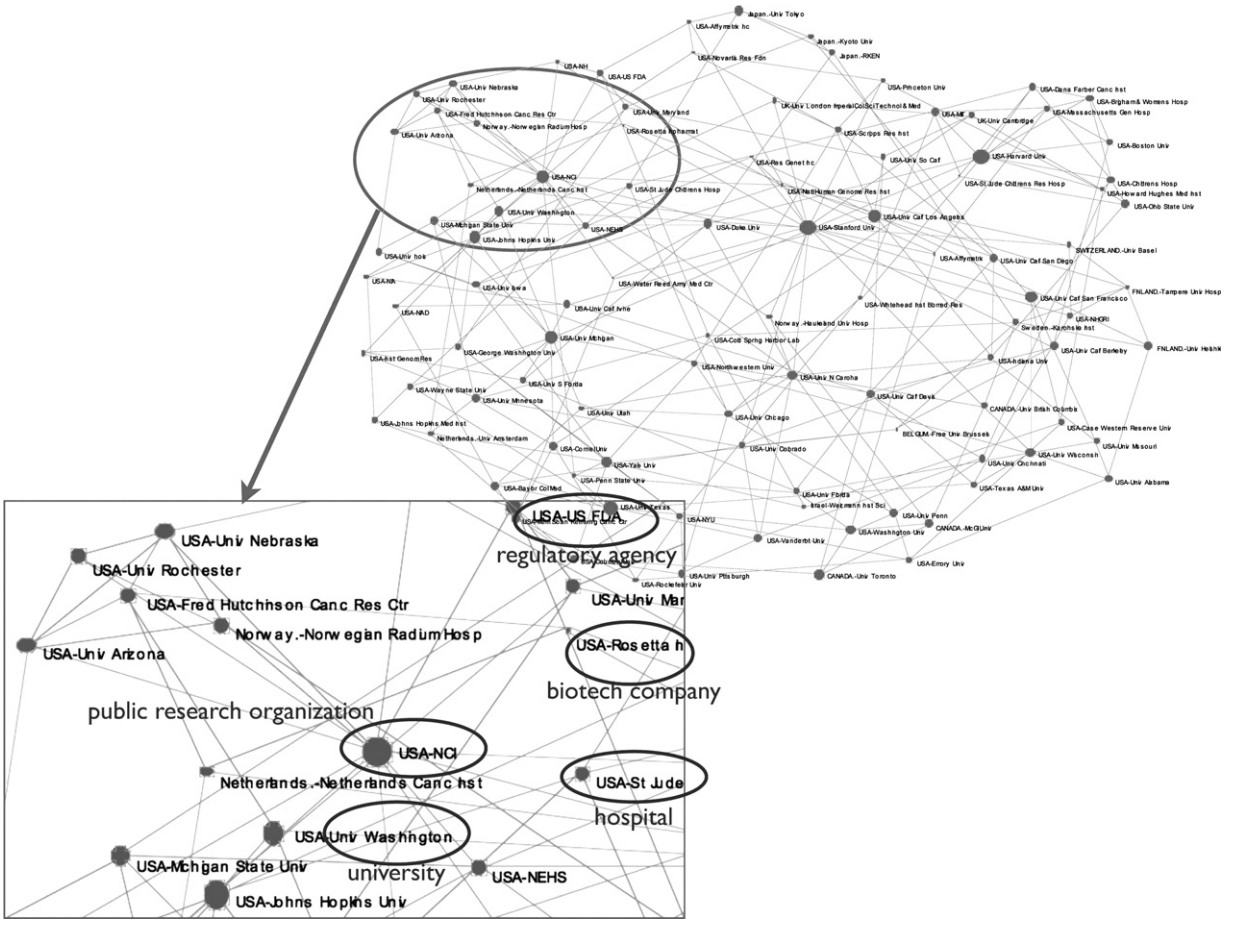

Fig. 3. Institutional collaborative network in the field of microarrays (data source: Web of Science; network analysis and visualization: ReseauLu X2; see text for explanations).

rapid growth of the research and industrial aspects of the field, we should expect to find close connections between our three databases.

Fig. 3 displays the network of institutional collaborations: two institutions are linked if they appear in the addresses of co-authors. Whereas the map (not showed) displaying all collaborations amounts to a single cloud of closely interconnected, individually undistinguishable links, Fig. 3 only shows the more specific links. The latter are calculated via a RL algorithm that uses the cosine measure of proximity to select, compile and map the n (in the present case: four) nearest nodes to each node. As can be seen in the blow-up window, collaborations in this field include academic institutions (e.g., University of Washington), public research organizations (e.g., the NCI), clinical institutions (e.g., St. Jude hospital), commercial institutions (e.g., Rosetta Inpharmatics) and state regulatory agencies (e.g., the FDA). Here again, because of this intertwining of different kinds of institutions we should expect to find close connections between our three databases.

3.2. Building bridges between databases

Patents, and thus patent databases, contain non-patent references; conversely, publications, and thus publication databases, contain references to patents. But the stated goal of our methodology is to link patents and publications *across*, rather than *within*, databases. Moreover, our third kind of source data, research projects, cannot be searched for references to patents and publications. Finally, we are interested in building database bridges that focus primarily on the *content* of projects.

One relatively simplistic way of performing this task – we mention it here only for the sake of comprehensiveness – is to begin with the CRISP database of NIH-funded projects, since grants are categorized by the specific NIH institute awarding them. Fig. 4 shows that the highest share of awards is provided by six institutes corresponding to the following medical domains: cancer, allergy and infectious diseases, general medical sciences, diabetes and digestive and kidney diseases, heart lung and blood, and neurological disorders and stroke, the remaining grants being distributed among 17 other institutes. This is, admittedly, a very coarse subdivision of the fields of application of microarray research, but it highlights, for instance,

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

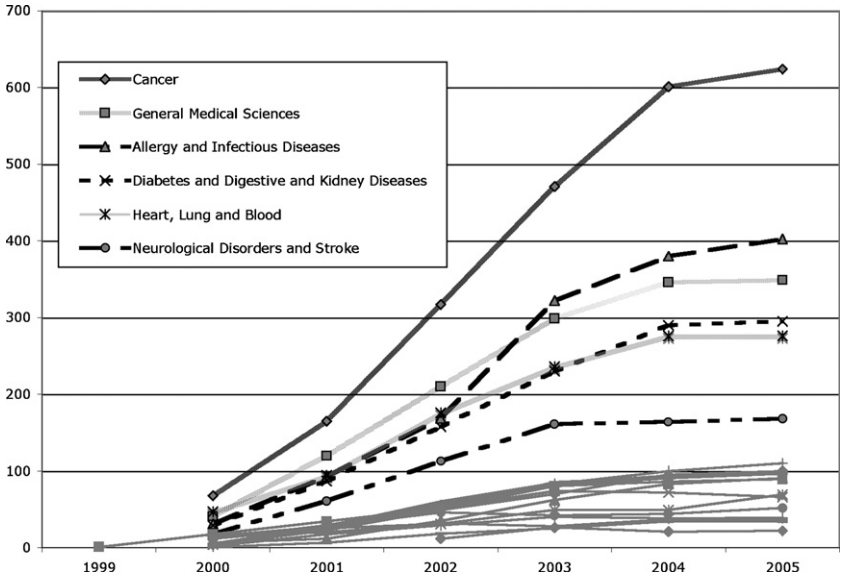

Fig. 4. Number of NIH research grants by year and institute (source: CRISP database).

to the central role played by oncology at the biomedical research front, at least as far as genomic technologies are concerned. This categorization can also be used to establish interfaces between different databases by establishing a list of authors of the grant proposals and then matching their names with those of the authors and inventors listed in the publication and the patent databases. Fig. 5 illustrates the results of this kind of matching: the *X*-axis corresponds to the number of CRISP projects for each domain and the *Y*-axis to the number of publications. The cancer domain has proportionally more publications than would have been predicted by the number of grants awarded in this domain; the opposite is true, for instance, for diabetes and digestive and kidney diseases. Similar diagrams (not shown) can be easily produced relating the number of CRISP projects to the number of citations received by publications as well as to the number of patents. In this latter case, 8.45% of the authors listed in the CRISP database also appear in the patent database, a number that is consistent with the order of magnitude mentioned by other authors for the percentage of academic scientists who file patents (Forti et al., 2007).

This approach to building bridges between databases, in addition to operating at a highly aggregate level, is not consistent with our requirement to avoid using pre-defined categories when delineating subtopics. We thus devised and tested the following approach:

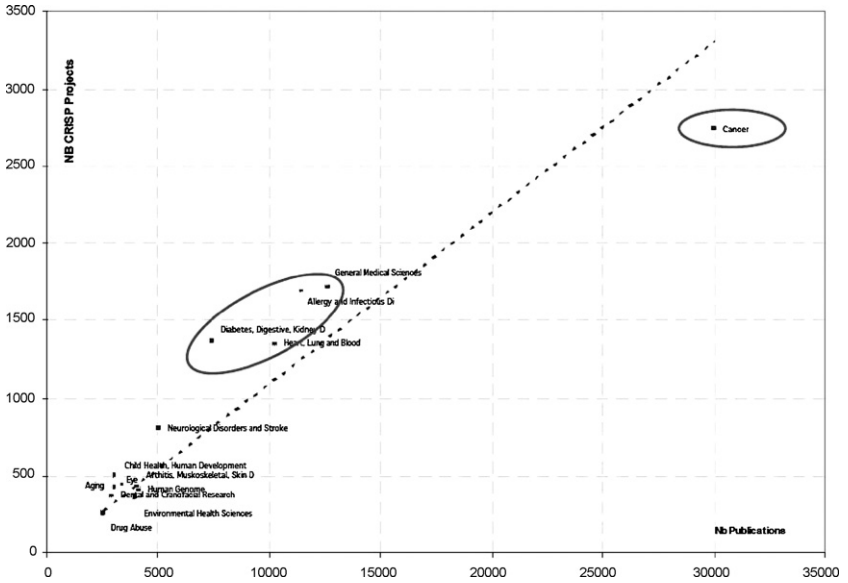

Fig. 5. Number of NIH research grants vs. number of publications in institutionally defined domains.

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

(a) First, we selected the largest and arguably the most comprehensive of the three databases, namely WoS, and text-mined titles and abstracts with LQM to extract a list of relevant concepts.
(b) The resulting list of concepts being quite extensive, we selected the most specific concepts by applying a Chi-square filter to the table of co-occurring concepts.[14]
(c) Using RL, we produced a semantic network of these concepts and automatically identified clusters of co-occurring concepts using a fuzzy *K*-means clustering algorithm.
(d) The list of concepts corresponding to each cluster was manually inspected in order to assign a topical sub-domain name to the cluster.
(e) In order to map these domains to the two other databases, we text-mined them and each document in those databases was characterized by a list of concepts extracted from it.
(f) Finally, a matching algorithm was used to assign patents and projects to one of the subtopics: the algorithm functions by picking the best match between the concepts defining each sub-domain and the concepts characterizing each document.

Concerning step (a), an alternative approach would have been to put all the documents (projects, publications and patents) in a same, general set, text mine this combined set and proceed with the subsequent steps (b–f) in order to define sub-domains and assign documents to them. A drawback of this approach is that patents use a very different vocabulary from publications and projects, in particular by resorting to legal terminology. As a result of this terminological heterogeneity across databases, the resulting set of concepts contains a lot of "noise" that can make the identification of techno-scientific sub-domains more difficult.

Concerning step (b), LQM retrieved 55,168 individual concepts occurring at least three times in the WoS database. Each individual concept occurs, of course, in several documents: a total of 860,962 concepts were found in the 29,242 documents, i.e., on average, 29 concepts in each document. All these concepts are meaningful ones: they do not include "stop words" and other terms that carry no semantic information – they are automatically discarded by LQM – nor terms such as "cancer", "tumor", "cell" or "microarrays" that we listed in a user-defined exclude dictionary since they were too generic for our present purpose. However, we instructed LQM to retrieve concepts such as "breast cancer", "beta cell apoptosis" or "epidermal peripheral-nerve sheath tumor" that include these generic single terms. The distribution of concepts in the database follows Zipf's law (Zipf, 1949): of the 55,168 individual concepts, only 22 occurred more than 1000 times, 103 occurred at least 500 times and 13,093 at least 10 times. The small set of most frequent concepts (the "trivial zone" of a Zipf curve) and the long queue of very low occurrence concepts (the "noise zone") are likely to carry little or less distinctive value: thus, after selecting the top 500 most frequent concepts, we retained only the top 300 concepts with the highest Chi-square value (the "information zone").

Concerning step (c), the "fuzzy" *K*-means algorithm built into RL allows for the possibility for a given concept to belong simultaneously to two different clusters. While the *K*-means approach is well established, it has a few drawbacks, including the fact that the number of clusters must be specified in advance: one tinkers with the number of clusters until they look sufficiently homogeneous, consistent and specific. We are presently testing an alternative approach based on a random-walk, Monte Carlo simulation algorithm (Burlatsky, Oshanin, & Mogoutov, 1990; Burlatsky, Oshanin, & Mogoutov, 1992; Oshanin, Mogoutov, & Burlatsky, 1990) that calculates the optimal number of clusters and assigns concepts to those clusters by iteratively computing the probabilities for each concept of belonging to a given cluster. In other words, in addition to its network visualization modules, RL contains a set of data analysis modules for purposes such as selecting the more specific links or automatic clustering according to different algorithms.

Finally, concerning step (f), the matching algorithm works by comparing the list of concepts defining a given sub-domain to the list of concepts retrieved from each document: if the concepts characterizing a given document belong to different sub-domains, the "best match" is defined as the sub-domain with the highest number of shared concepts. A threshold can of course be defined for the minimal number of concepts necessary for a document to be declared a match.

Fig. 6 shows the semantic network map as well as the list of sub-domains corresponding to the 12 clusters identified by the clustering algorithms. It will be noticed that we have four cancer-related clusters, namely three sub-domains related to the mechanisms involved in the development of and spread of tumors (apoptosis, metastasis and angiogenesis) and a fourth sub-domain concerning clinical aspects of cancer treatment. Among the other topics, we should mention gene expression and single nucleotide polymorphisms (SNPs), that correspond to the two main focal points of microarray research and microarray applications on both the academic and the biotech side: for instance, the commercial leader in the field, Affymetrix, has traditionally made most if its revenues on its RNA (gene expression) products, but more recently DNA genotyping (SNPs) has increased its share of total sales (Petrone, 2007).

The matching algorithm was able to retrieve and categorize, respectively, 93 and 69% of the references in the project and the patent databases. The relatively low matching percentage of patents can be accounted for by the previously mentioned idiosyncratic terminology used in patents. Fig. 7 is a diagram comparing the percentage of documents corresponding to two selected sub-domains in the three databases: SNPs represent a larger share of patents than of publications or projects,

[14] The co-occurrence table compiles the frequency of co-occurring concepts within the same titles and abstracts. In order to select the most specific combinations of concepts, we used an index of specificity (I) that measures the discrepancy between the observed frequency (O) of concept co-occurrence and its expected value (X). The following formula was used to calculate the normalized difference between observed and expected values: $I = (O - X)^2/X$.

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

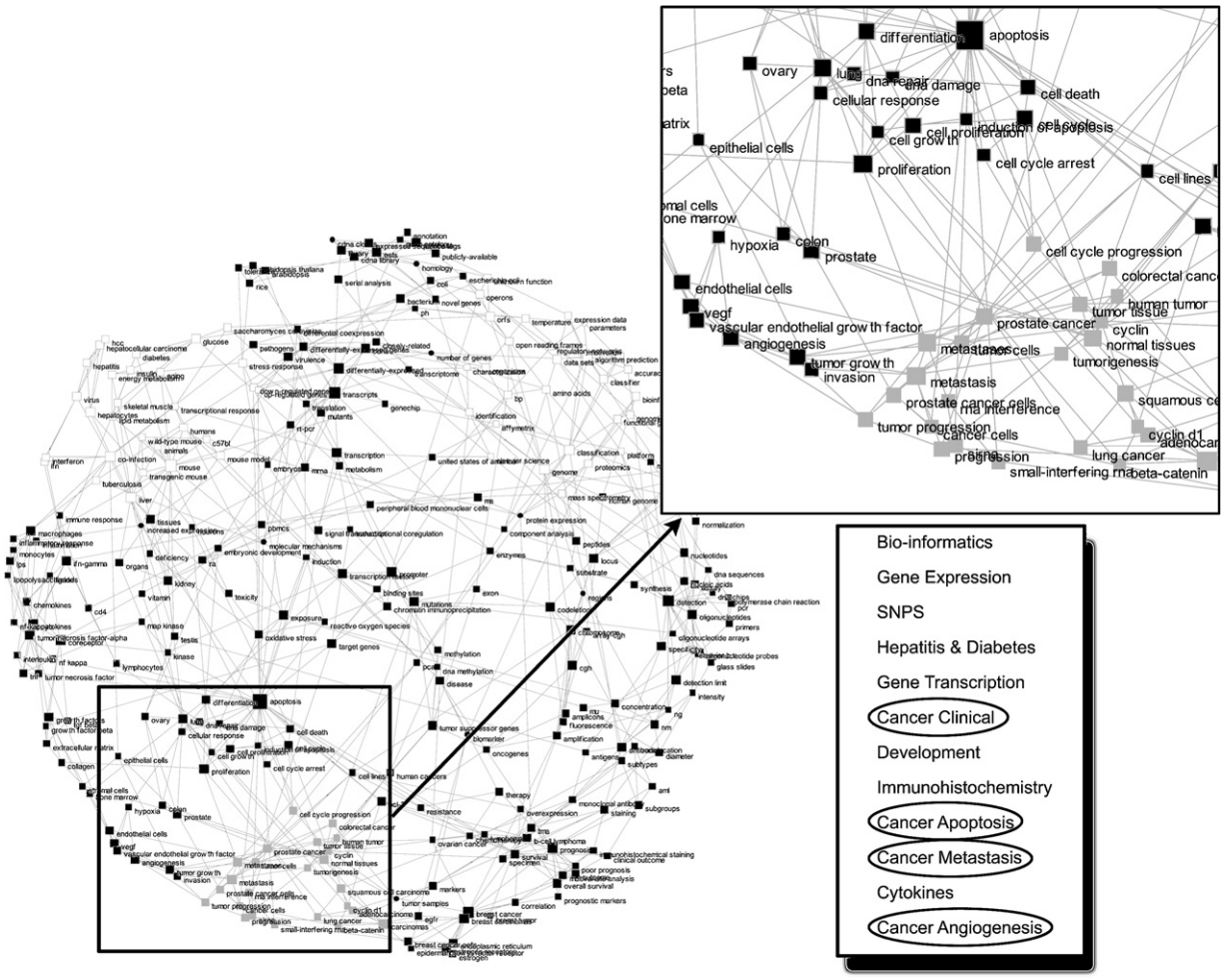

Fig. 6. Co-occurrence network of concepts extracted from the Web of Science microarray database and thematic clusters (text-mining: SPSS Lexiquest Mine; data analysis and visualization: ReseauLu X2).

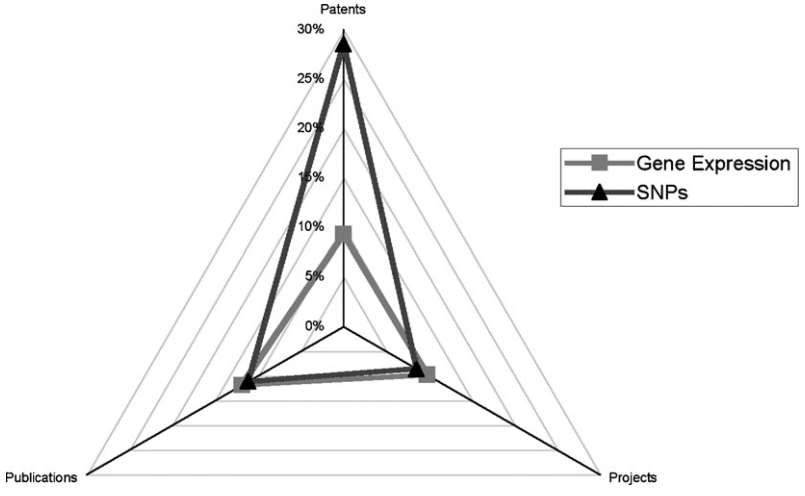

Fig. 7. Comparison of the relative percentage of patents, publications and research projects retrieved from the patents, publications and research projects databases in two selected sub-domains (see text for explanations).

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

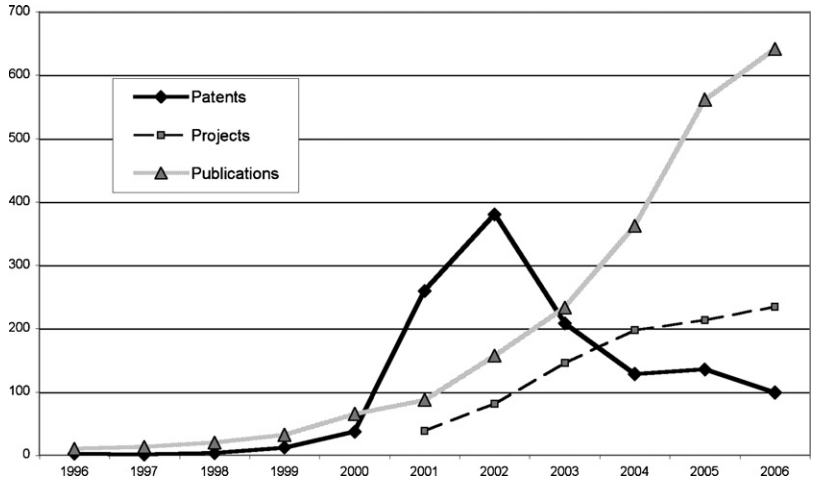

Fig. 8. Number of publications, patents and research projects by year in the SNPs sub-domain (see text for explanations).

whereas the situation is more balanced for gene expression. Fig. 8 shows an example of the growth of publications, patents and projects in a specific sub-domain, in this case SNPs[15]: we notice a rapid growth of publications and a less rapid increase in the number of grants, but also a somewhat surprising peak of patents in 2002, followed by a decline. The shape of the microarray/SNPs patent curve can be accounted for by the fact that while the rapid growth of the microarray technology put thousands of SNPs on the market, subsequent interventions by the US Patent Office restricted the criteria for obtaining patents in this domain[16]; moreover, the development of public databases and of the HapMap project (an international partnership officially started in late 2002 to find genes associated with human disease and response to pharmaceuticals) provided open access to SNPs information.[17]

Traditionally, in order to explore, say, the connection between industry and academic or public research, analysts have looked at citations to scientific articles within patents, i.e., within a single type of document. Such approaches can be extended by linking non-patent references in patents to publications and then combining data from both sources (Verbeek et al., 2002, 2003). The present article provides "proof of principle" of a novel way of looking at triple-helix interfaces. Our method makes it possible to look at such connections across types of documents by: (a) taking into account and starting simultaneously from different kinds of databases, (b) including research project data (and potentially other kinds of source data) in addition to patents and publications, and (c) pursuing a fine-grained exploration of the *content* of these documents via text-mining instead of relying on statistical compilations of pre-established categories. Indeed, our results allow for the qualitative inspection of the content of the research and commercial activities, insofar as they are not limited to numbers but consist of tables containing all the relevant information (references, authors, institutions, text-mined concepts) related to each sub-domain. One can thus easily produce maps displaying, for instance, semantic, institutional or co-authorship networks (or combinations thereof) with the important, additional bonus that these networks do not correspond to a single type of source (publications, patents or projects) but simultaneously deploy all three types of sources.

Ongoing work is devoted to exploring the new opportunities provided by our approach. As mentioned in the previous paragraph, a first step will consist in producing substantive results in order to empirically test the value of the methodological approach presented in this article with regards to specific research questions and in relation to specific domains. A second, related step is external validation by experts in the field. While well-established text-mining algorithms such as those of LQM can be trusted to extract meaningful concepts from documents, i.e., concepts that correspond to their content, the categorization of sub-domains by clustering is a more delicate procedure and expert validation is a necessary step. Given the fact, however, that multiple descriptions of a given field are possible, this kind of validation is better pursued in relation to specific research questions and results.

4. Conclusion

Our method for the exploration of "triple-helix" connections could be of practical value for at least three different purposes: as a tool for technological foresight (for firms and public policy analysts), as a method for the strategic assessment of public programs, and, last but not least, as a way to better understand the diversity of the processes partaking in the construction of new science-based technologies.

[15] On the SNPs literature until 2001, see Coronini et al. (2003).
[16] For an early discussion of these issues, see Flattmann and Kaplan (2001).
[17] Many thanks to Bertrand Jordan (Marseille-Nice Genopole) for suggesting this explanation.

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

With regards to the first purpose, several countries (e.g., UK, Japan and France) have initiated technology foresight exercises to identify future key technologies as part of the priority setting activities of policy makers and firms (Georghiou, 2003). These foresight exercises involve periods of reflection, consultation and networking and are based mainly on qualitative methods. To provide material for discussion, however, they also often mobilize quantitative data about scientific articles and patents since these data can provide useful insights concerning the emergence of new scientific and technological fields. As argued by Salerno et al. (2006), the analysis of the content of research projects is a useful addition, insofar as research projects will lead to the publication of papers and/or the granting of patents and they thus qualify as a complementary source of information to understand innovation processes.

With respects to the second purpose, one of the major questions raised by the economy of technical change concerns the role played by scientific research in the innovation process. To understand the relationships between academic research and technology, scholars have analyzed references to scientific articles in patents and have devised ways to map the combined dynamics of scientific and technological activities (e.g., Moed, Glanzel, & Schmoch, 2004). Our relational analysis of projects, publications and patents adds a triple-helix perspective to this classical issue, and could actually lead to a re-examination of the relations between university, government and industry, by focusing, for instance, on the role played by public programs in the emergence of new technological fields (Callon, Larédo, & Mustar, 1997). The use of contract databases for evaluation purposes is not new. Carter (1974) and Narin (1983) have assessed the NIH research policies by examining 747 research projects and 51 research programs financed by the NIH in 1967. As noted by Mauguin (1997), "one of the results of these studies was to show the existence of a correlation between bibliometric indicators of research productivity and non-bibliometric quantitative indicators, including financial data concerning contracts". Back then, the goal of these evaluations was to calculate "the number of publications produced per dollar invested as a measure of the economic effectiveness of research programs financed by the NIH". This mechanical view of the relationship between public funding and science has been abandoned. Today, contract database analysis has become a tool for the strategic management of public research programs.

Finally, our way of analyzing "triple-helix interfaces" foregrounds the multiple university, industry and government configurations that lead to the differential development of technologies. Different theoretical frameworks, often corresponding to different specialties (the sociology of innovation, R&D economics, etc.) have described specific aspects of the evolution of specific techno-scientific fields, but a focus on interactions and interfaces should prompt researchers to merge the insights from these different perspectives in order to develop a more comprehensive approach. Far from being a mere technological "trick", the combination of text-mining and network analysis is a first step towards the more ambitious program of mapping the specific institutional and conceptual arrangements that underlie the development of a given technology and that call upon a heterogeneous set of items, such as public programs, equipments and techniques, research materials and results, scientific entities, teams, companies, and skills, to mention just a few.

Acknowledgements

Research for this paper was made possible by grants from the Canadian Institutes for Health Research (CIHR MOP-64372), the Fonds Québécois de la Recherche sur la Société et la Culture (FQRSC ER-95786), and the Social Sciences and Humanities Research Council of Canada (SSHRC 410-2002-1453).

References

Ananiadou, S., & McNaught, J. (Eds.). (2006). *Text mining for biology and biomedicine*. Boston: Artech House.
Bourret, P., Mogoutov, A., Julian-Reynier, C., & Cambrosio, A. (2006). A New Clinical Collective for French Cancer Genetics: A Heterogeneous Mapping Analysis. *Science, Technology, & Human Values*, *31*, 431–464.
Burlatsky, S. F., Oshanin, G. S., & Mogoutov, A. V. (1990). Direct Energy-Transfer in Polymer Systems. *Physical Review Letters*, *65*, 3205–3208.
Burlatsky, S. F., Oshanin, G. S., Mogoutov, A. V., et al. (1992). Directed Walk in a One-Dimensional Lattice Gas. *Physics Letters A*, *166*, 230–234.
Callon, M., Larédo, P., & Mustar, P. (Eds.). (1997). *The strategic management of research and technology*. Paris: Economica International.
Cambrosio, A., Cotterau, P., Popowycz, S., Mogoutov, A., & Vichnevskaia, T. (2007). Analyse des réseux hétérogènes: Le projet ReseauLu. In C. Brossaud & B. Reber (Eds.), *Humanités numériques* (pp. 165–180). Paris: Hermès Science.
Cambrosio, A., Keating, P., Bourret, P., Mustar, P., & Rogers, S. (in press). Genomic platforms and hybrids. In P. Atkinson, P. Glasner, & M. Lock (Eds.), *Handbook of genetics and society: Mapping the new genomic era*. London: Routledge.
Cambrosio, A., Keating, P., Mercier, S., Lewison, G., & Mogoutov, A. (2006). Mapping the Emergence and Development of Translational Cancer Research. *European Journal of Cancer*, *42*, 3140–3148.
Carter, G. M. (1974). *Peer review, citations and biomedical research policy: NIH grants to medical school faculty*. Report for the Health Resources Administration and the Office of the Assistant Secretary for Planning and Evaluation of the Department of Health, Education and Welfare, R-1583-HEW, December. Washington, DC: HEW.
Chang, J., & Schutze, H. (2006). Abbreviations in biomedical text. In S. Ananiadou & J. McNaught (Eds.), *Text mining for biology and biomedicine* (pp. 99–119). Boston: Artech House.
Constans, A. (2003). The state of the microarray. *The Scientist*, *17*(3), 34.
Coronini, R., de Looze, M.-A., Puger, P., Bley, G., & Ramani, S. V. (2003). Decoding the literature on genetic variation. *Nature Biotechnology*, *21*, 21–29.
Feldman, R., & Sanger, J. (2007). *The text mining handbook. Advanced approaches in analyzing unstructured data*. Cambridge, UK: Cambridge University Press.
Flanagan, N. (2007). DNA chips provide key to human genome analysis. *GEN*, *27*(11)
Flattmann, G. J., & Kaplan, J. M. (2001). Patenting expressed sequence tags and and single nucleotide polymorphisms. *Nature Biotechnology*, *19*, 777–779.
Forti, E., Franzoni, C., & Sobrero, M. (2007). The effect of patenting on the networks and connections of academic scientists. *Project IRIS working papers series* [available at: http://www.iris.unibo.it/Dload/WP/piWPS_0001-2007.pdf].
Fransman, M. (2001). Designing dolly: Interactions between economics, technology and science and the evolution of hybrid institutions. *Research Policy*, *30*, 263–273.
Georghiou, L. (2003). Foresight: Concept and practice as a tool for decision making. In *Paper presented at the technology foresight summit*

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005

Harris, N. L., & Horning, S. J. (2006). Burkitt's lymphoma: The message from microarrays. *New England Journal of Medicine*, *254*, 2495–2498.
Keating, P., & Cambrosio, A. (2003). *Biomedical platforms. Realigning the normal and the pathological in late-twentieth-century medicine*. Cambridge, MA: MIT Press.
Keating, P., & Cambrosio, A. (2004). Signs, markers, profiles, and signatures: Clinical hematology meets the new genetics (1980–2000). *New Genetics and Society*, *23*, 15–45.
Mauguin, P. (1997). Contract databases as tools for characterizing technological programmes. In M. Callon, P. Larédo, & P. Mustar (Eds.), *The strategic management of research and technology* (pp. 117–131). Paris: Economica International.
McMeekin, A., Harvey, M., & Gee, S. (2004). Emergent bioinformatics and newly distributed innovation processes. In M. McKelvey, J. Laage-Hellman, & A. Rickne (Eds.), *The economic dynamics of modern biotechnology* (pp. 35–261). Cheltenham: Edward Elgar Publisher.
Moed, H. F., Glanzel, W., & Schmoch, U. (Eds.). (2004). *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*. Dordrecht: Kluwer.
Mougoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy*, *36*, 893–903.
Mowery, D. C. (2001). The United States national innovation system after the cold war. In P. Larédo & P. Mustar (Eds.), *Research and innovation policies in the new global economy* (pp. 15–46). Cheltenham, UK: Edward Elgar.
Mustar, P. (1997). How French academics create high tech companies: Conditions of success and failure of this form of relation between science and market. *Science and Public Policy*, *24*(1), 37–43.
Mustar, P. (1998). Partnerships, configurations and dynamics in the creation and development of SMEs by researchers. *Industry and Higher Education*, 217–221.
Mustar, P. (2001). Spin-offs from public research: Trends and outlook. *STI (Science, Technology, Industry)*, *26*, 165–172.
Narin, F. (1983). *Concordance between subjective and bibliometric indicators of the nature and quality of performed biomedical research*. Program Evaluation Report for the Office of Program, Planning and Evaluation, NIH, April. Washington, DC: NIH.
Oshanin, G. S., Mogoutov, A. V., & Burlatsky, S. F. (1990). A 2-dimensional model of trapping reactions with Gaussian coils. *Physics Letters A*, *149*, 55–59.
Oudshoorn, N., & Pinch, T. (Eds.). (2003). *How users matter: The co-construction of users and technology*. Cambridge, MA: MIT Press.
Perkel, J. M. (2005). Medicine gets personal. With new regulations and new diagnostics, pharmacogenetics comes to the clinic. *The Scientists*, *19*(8), 34.
Petrone, J. (2007). As gene expression market tapers off, SNP genotyping becomes Affy's sales driver. *BioArrayNews*, *7*(31).
Powell, W. W., Koput, K. W., & SmithDoerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, *41*, 116–145.
Quackenbush, J. (2006). Microarray analysis and tumor classification. *New England Journal of Medicine*, *354*, 2463–2472.
Rogers, S., & Cambrosio, A. (2007). Making a new technology work: The standardization and regulation of microarrays. *Yale Journal of Biology and Medicine*, *80*, 165–178.
Salerno, M., Landoni, P., & Verganti, R. (2006). The role of funded projects content analysis in early stage disciplines exploration: The case of nanotechnology. In *Paper presented at the SPRU 40th anniversary conference—the future of science, technology and innovation policy*.
Simon, R., & Wang, S. J. (2006). Use of genomic signatures in therapeutics development in oncology and other diseases. *Pharmacogenomics Journal*, *6*, 166–173.
Verbeek, A., Debackere, K., & Luwel, M. (2003). Science cited in patents: A geographic "flow" analysis of bibliographic citation patterns in patents. *Scientometrics*, *58*, 241–263.
Verbeek, A., Debackere, K., Luwel, M., Andries, P., Zimmermann, E., & Deleus, F. (2002). Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics*, *54*, 399–420.
von Hippel, E. (1976). The dominant role of users in the scientific instrument innovation process. *Research Policy*, *5*, 212–239.
von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science*, *32*, 791–805.
von Hippel, E. (1989). New product ideas from 'lead users'. *Research Management*, *32*(3), 24–27.
Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Cambridge, MA: Addison Wesley.

Please cite this article in press as: Mogoutov, A., et al. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics* (2008), doi:10.1016/j.joi.2008.06.005