First Monday, Volume 12, Number 6 — 4 June 2007

PEER-REVIEWED JOURNAL ON THE INTERNET

A Simple Method to Improve Life Sciences Patent Searches Using the Cyberinfrastructure at the National Institutes of Health

by Kyle Jensen, Chen Jinan, and Fiona Murray

Contents

Introduction

In the life sciences, an immense publicly–funded cyberinfrastructure contributes significantly to the rapid accumulation of knowledge and innovation. Much of the traditional public infrastructure consists of physical materials such as journal articles, data repositories, and repositories for biological and chemical samples. However in recent years, the U.S. National Institutes of Health and in particular the NIH’s National Center for Biotechnology Information generated a cyberinfrastructure in the form of electronic databases for journals, gene sequences, protein structures, and other experimental data. (The NCBI’s cyberinfrastructure is described in detail elsewhere [1].) The NCBI’s cyberinfrastructure is notable because it is both extensive and highly interconnected across diverse types of data. That is, scientists that contribute an individual “unit” to one of these resources have their data standardized and cross–indexed against other databases, thereby facilitating discoveries that may only be illuminated in the context of other, related scientific resources. For example, consider the Pubmed entry shown in Figure 1. In addition to bibliographical information, the entry is tagged with numerous words and phrases that were manually added by an independent scientist reading the manuscript. The largest set of these tags is called the “MESH headings,” referring to the NIH’s “MEdical Subject Headings”: a hierarchical ontology developed to classify the biomedical literature.

Figure 1: Pubmed entry for Chaudhuri, *et al.* [3]. As the figure shows, the Pubmed entry contains both basic bibliographical information such as the journal and author names, and “value added” information that was manually appended by researchers at NIH. This value–added information includes the MESH headings and a list of compounds relevant to the manuscript. The MESH headings are a logical breakdown of the major themes of the manuscript and can be useful “hooks” for querying Pubmed.

1: J Biol Chem. 1994 Mar 18;269(11):7835-8.

Expression of the Duffy antigen in K562 cells. Evidence that it is the human erythrocyte chemokine receptor.

Chaudhuri A, Zbrzezna V, Polyakova J, Pogo AO, Hesselgesser J, Horuk R.

New York Blood Center, New York 10021.

The human malarial parasite Plasmodium vivax invades erythrocytes by binding to a cell surface protein identified as the Duffy blood group antigen. The molecular properties of the Duffy antigen, which was recently cloned, are very similar to those of a chemokine binding protein known as the human erythrocyte chemokine receptor. This has led to the suggestion that these two molecules are the same protein. To further investigate the suspected double identity of the Duffy antigen we have transfected it into a human erythroleukemic cell line, K562. Cells stably expressing the Duffy antigen were isolated and used to characterize the protein. K562 cells transfected with the Duffy antigen displayed specific 125I-melanoma growth-stimulating activity (MGSA) binding while mock transfected cells did not. Comparison of 125I-MGSA binding to the Duffy antigen and the human erythrocyte chemokine receptor showed that the specific 125I-MGSA binding to both proteins was displaced by excess unlabeled MGSA, interleukin-8, RANTES, monocyte chemotactic peptide-1, and platelet factor 4, but not by macrophage inflammatory protein-1 alpha or -1 beta. Scatchard analysis of competition binding studies with these unlabeled chemokines revealed high affinity binding to the Duffy antigen with KD binding values of 24 +/- 4.9, 20 +/- 4.7, 41.9 +/- 12.8, and 33.9 +/- 7 nM for MGSA, interleukin-8, RANTES, and monocyte chemotactic peptide-1, respectively. A monoclonal antibody, Fy6, to the Duffy antigen inhibited 125I-MGSA binding to K562 cells expressing the Duffy antigen. Cell membranes from K562 cells permanently expressing the Duffy antigen were chemically cross-linked with 125I-MGSA. SDS-polyacrylamide gel electrophoresis analysis of the cross-linked products showed covalent incorporation of radiolabeled MGSA into a protein of molecular mass 47 kDa, and cross-linking was inhibited in the presence of unlabeled MGSA.

These studies provide evidence that the Duffy blood group antigen is the same protein as the human erythrocyte chemokine receptor.

MeSH Terms:
Amino Acid Sequence
Antibodies, Monoclonal/pharmacology
Cell Line
Chemokines, CXC*
Chemotactic Factors/metabolism*
Duffy Blood-Group System/biosynthesis
Duffy Blood-Group System/metabolism*
Erythrocytes/metabolism*
Growth Substances/metabolism*
Humans
Intercellular Signaling Peptides and Proteins*
Kinetics
Leukemia, Myeloid, Chronic
Molecular Sequence Data
Neoplasm Proteins/metabolism
Plasmids
Receptors, Cytokine/biosynthesis
Receptors, Cytokine/drug effects
Receptors, Cytokine/metabolism*
Restriction Mapping
Transfection
Tumor Cells, Cultured

Substances:
Antibodies, Monoclonal
CXCL1 protein, human
Chemokines, CXC
Chemotactic Factors
Duffy Blood-Group System
Growth Substances
Intercellular Signaling Peptides and Proteins
Neoplasm Proteins
Receptors, Cytokine
melanoma growth stimulating activity receptor

PMID: 8132497 [PubMed - indexed for MEDLINE]

In essence, the NCBI performs an “information upgrade” on the public literature and other life sciences resources, thereby supporting downstream knowledge accumulation and innovation. This is analogous to the classic public policy rational for the patent system, wherein dissemination of patent disclosures is key to promoting innovation. However, by comparison, the cyberinfrastructure for the “private knowledge” in the patent system is quite poor. Here we describe a very simple step for upgrading the information content of the life sciences patent literature by leveraging the public cyberinfrastructure developed by the NCBI. Specifically, we show how to link specific patents broadly into the diverse NCBI cyberinfrastructure using citation analysis.

Methods

Using software written in Python programming language, we extracted all citations to non–patent literature from issued U.S. patents dating back to 1976. From the list of citations, we extracted those citations which were matched by a series of regular expressions capturing possible permutations of the names of journals that are indexed by Pubmed. (These possible permutations, including abbreviations, were downloaded from NCBI’s Web site. The regular expressions, in general, captured abbreviations and variations in whitespace or other nonalphanumeric characters.) This yielded a set of roughly 1.2 million citations.

Using software designed by the authors, we parsed each of these citations for recognizable features including author names, year of publication, volume and number of publication, and statistically unlikely words in the title (see Table 1). These features, in turn, were used to make queries that were submitted to NCBI via the e–utils interface (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html). In order that we not overload the NCBI servers, these queries were submitted once every three seconds. As of the writing of this manuscript, roughly 150,000 of the 1.2 million citations have been sent to Pubmed via queries. For each of these queries we recorded the patent number–Pubmed number pair if it existed. In the following sections, all results refer to this subset of 150,000 citations unless otherwise noted.

Results and discussion

Roughly one–third of the queries failed to find any Pubmed number for a given citation, meaning that these citations may not be indexed by Pubmed, or that they are indexed by Pubmed but that correct features cannot be discerned from the citation in such a way that the Pubmed number can be discovered. For the remaining two–thirds of citations, we were able to find patent–Pubmed pairings. Most of the Pubmed indexed articles are cited only once; however, there are small number of articles that are cited tens or hundreds of times. That is, the distribution of citations is highly skewed. Many of the highly cited articles seem to describe methods. For example, the most highly cited article is Pubmed number 2440339: “Single–step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction,” which, as the title implies, describes a widely used method for isolating RNA [2]. In the 150,000 citations that we have parsed so far, this article is cited over 160 times.

How are these patent number-Pubmed number pairs useful? We feel they are principally useful for discovering prior art without resorting to searching the patent literature explicitly using fulltext searches via the USPTO’s Web site (http://www.uspto.gov/). For example, researchers commonly use citations of other patents as a method for discovering prior art. That is, say that we’ve discovered a handful patents that are pertinent to our new invention — it is likely that other patents that these patents cite are also pertinent. Or, that patents which cite this handful are pertinent. Very few tools exist that allow the same kind of citation analysis to be done using citations to the scientific literature instead of to the patent literature. The reason for this is that scientific citations vary widely in formatting (see Table 1). Here, we are able to unambiguously resolve a variety of formats to a single Pubmed number. This allows researchers to easily navigate a network of scientific citations just as they would patent citations. For example, we are able to find the set of all other patents that cite any one of the scientific articles also cited by a handful of pertinent patents.

Table 1: The seven unique citation formats for Chan, *et al.* [4] found in the patent literature. As the table shows, there is no standard format enforced by the USPTO for citing scientific literature. Some citations like the authors names, others lack title, and many use abbreviations. These formats are all relatively readable by humans; however, can be difficult when designing software to recognize arbitrary formats.

1) Chan et al. Identification of a New Class of Et Selective Endothelin Antagonists by Pharmacophore Directed Screening, Biochemical and Biophysical Research Communications, May 30, 1994, vol. 201, No. 1, pp. 228-234.
2) Chan et al., "Identification of a New Class of ET.sub.A Selective Endothelin Antagonists by Pharmacophore Directed Screening", Biochemical and Biophysical Research Communications, vol. 201, No. 1, May 30, 1994, pp. 228-234.
3) Chan et al., Biochem. Biophys. Res. Commun., 201(1) 228-34 (1994).
4) Chan et al., Biochemical and Biophysical Research Communications, vol. 201, No. 1, May 30, 1994, pp. 228-234.
5) Identification of New Class of ET.sub.A Selective Endothlin Antagonists by Pharmacophore Directed Screening, Biochemical and Biophysical Research Communications, May 30,1994, vol. 201, No. 1, pp. 228-234.
6) Identification of a New Class of ET.sub.A Selective Endothelin Antagonists by Pharmacophore Directed Screening, Biochemical and Biophysical Research Communications, May 30, 1994, vol. 201, No. 1, pp. 228-234.
7) Identification of a New Class of ET.sub.A Selective Endothelin Antagonists by Pharmacophore Directed Screening, Biochemical and Biophysical Research Communications, May 30, 1994, vol. 201, No. 1, pp. 228-234. Chan et al.

In addition to the application illustrated in the previous paragraph, there is a more sophisticated usage scenario for these patent number–Pubmed number pairings that leverages the cyberinfrastructure developed at NCBI. For example, consider a hypothetical situation in which we have developed a novel method of treating diabetes using a variant of testosterone. In order to search for prior art in the patent literature using traditional fulltext searches, we would probably start by searching for both “diabetes” and “testosterone.” We would then expand our search, replacing “testosterone” with the names of other known variants. An orthogonal and much more broad approach can be achieved using our database of patent number–Pubmed number pairs. For example, by searching Pubmed for articles that contain the MESH headings “Diabetes Mellitus” and “Androstanes” (androstane is the name of the family of steroids to which testosterone belongs) we find roughly 600 scientific articles each with a unique Pubmed number. By cross–indexing these articles with our patent number–Pubmed number database, we discover seven patents that cite at least one member of the 600. For example, patent 5654313: “Method for modifying or regulating the glucose metabolism of an animal or human subject,” which describes a method for treating diabetes by regulating levels prolactin and dopamine. The patent includes claims drawn towards arbitrary antagonists of prolactin and dopamine. Notably, although it is not mentioned in the patent, dopamine and prolactin levels are affected by steroids including testosterone. Therefore, it is possible that the claims in the ’313 patent could limit the scope of the claims we might file, or indicate to us a possible licensing arrangement. In any case, the ’313 patent is most certainly prior art in this area and cannot be found using the fulltext USPTO search strategy described above.

The limited examples above are only a small fraction of the many ways in which NCBI’s cyberinfrastructure can be leveraged to enable very sophisticated prior art searches. We feel that this database will be a useful resource for both scientists and law practitioners and we plan to complete this work in the next few months then make these data freely available.

About the authors

Kyle Jensen is the Director of Information and Analysis at the Public Intellectual Property Resource for Agriculture and a Research Affiliate at the Harvard–MIT Health Sciences and Technology program. He studies the role of intellectual property in the life science and agriculture sectors, focusing on the United States and the People’s Republic of China. Kyle received his Ph.D. from MIT and B.S. from the University of Illinois at Urbana–Champaign, both in chemical engineering.

Chen Jinan is a doctoral candidate at the Chinese Academy of Sciences' Institute of Biophysics. He studies biological networks and, in particular, computational methods for predicting protein–protein interactions. He has a B.S. degree in Atmospheric Sciences from Nanjing University.

Fiona Murray is the Class of 1922 Career Development Professor and Associate Professor Management of Technology Innovation & Entrepreneurship at the MIT Sloan School of Management.
E–mail: fmurray [at] mit [dot] edu

Notes

[1] *Nucleic Acids Res.* 2007 Jan;35 (Database issue):D5–12. Epub 2006 Dec 14.

[2] *Anal Biochem.* 1987 Apr;162(1):156–9.

[3] *J Biol Chem.* 1994 Mar 18;269(11):7835–8.

[4] *Biochem Biophys Res Commun.* 1994 May 30;201(1):228–34.

Contents Index

A simple method to improve life sciences patent searches using the cyberinfrastructure at the National Institutes of Health by Kyle Jensen, Chen Jinan, and Fiona Murray
First Monday, volume 12, number 6 (June 2007),
URL: http://firstmonday.org/issues/issue12_6/jensen/index.html