World Patent Information 31 (2009) 36–47

Contents lists available at ScienceDirect

World Patent Information

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

Searching biotechnology information: A case study

Luca Falciola *

RiboVax Biotechnologies SA, 12 Avenue Morgines, 1213 Petit-Lancy, Switzerland

ARTICLE INFO

Keywords:
Patents
Patent classification
Antibody
Internet
Search engines
Time factors
Literature
Biotechnology
Text Mining
Databases

ABSTRACT

The elaboration of strategies for the effective search of biotechnology information is a challenging task. In fact, the large amount of data in the public domain on biotechnology products and technologies is scattered among many databases and provided in different formats of document. This situation can make particularly difficult the identification, the extraction, and the aggregation of the information that are needed for performing detailed patent or scientific analyses.

The article presents a case study presenting different text-based approaches for searching and analyzing biotechnology information in patent and scientific literature using a series of exemplary searches on antibodies, a class of biological products having broad scientific and commercial interest.

The results show the complexity of defining how biotechnology information is actually searchable through the variety of available resources and to what extent. Some major factors that should be taken into consideration when searches are performed for evaluating scientific/patent trends, selecting documents potentially relevant for patentability, or identifying valuable technical information.

1. Introduction

1.1. Defining and searching biotechnology information

Modern biotechnology provides a large variety of products and technologies broadly applicable in scientific, medical, and industrial environments. Biotechnology is also a model for studies of social and economic research due to the high degree of innovation and the important commercial perspectives [1,2]. Previous studies combined quantitative and qualitative analysis of scientific and patent activity in biotechnology for determining their impact on innovation [3] and as well as on science policies [4].

According to the National Center for Biotechnology Information (NCBI):

> "Biotechnology is the body of knowledge related to the use of organisms, cells or cell-derived constituents for the purpose of developing products which are technically, scientifically and clinically useful." [5]

This definition acknowledges the many opportunities, mainly due to genetic engineering, for providing a large variety of products and technologies. The search and the analysis of the information on these subjects is a more and more challenging task. In fact, the high number of public and private entities that are active in biotechnology, together with the decreasing cost and time that is needed for generating and analyzing biological data, lead to the steady growth in the amount and in the complexity of biotechnology information that is available in the public domain. Moreover, such information can be:

- Provided in different (non-)textual formats (articles, biological sequences, patent documents, tables summarizing and comparing biological data, images of biological samples, graphics representing experiments, etc.).
- Scattered among many types of publications and databases.
- Published directly through Internet.

These factors increase considerably the efforts for comparing and aggregating the information at the scope of evaluating patents, products and technologies, or of interpreting trends in patent and scientific production.

1.2. Biotechnology information and biotechnology inventions

The search and the analysis of biotechnology information is a major element in the proceedings for establishing the patentability of biotechnology inventions which, according to the EPO Guidelines for Examination, are:

> "Inventions which concern a product consisting of or containing biological material or a process by means of which biological material is produced, processed or used". [6]

* Present address: Genfit SA, 885 av Eugène Avinée 59120 Loos, France. Tel.: +33 3 20 16 40 00; fax: +33 3 20 16 40 01.
E-mail address: luca.falciola@genfit.com

0172-2190/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2008.05.006

The Trilateral Search Guidebook in Biotechnology, a document that has been recently published by the Trilateral Project EPO-JPO-USPTO [7], provides some indications on how the information on such products, uses, and processes can be searched for establishing the patentability of biotechnological inventions. Exemplary searches are herein presented, giving a sufficiently broad overview of techniques, of major databases, as well as of keywords and search criteria that are applied at EPO, JPO and USPTO. This guidebook is particularly helpful since both EPO and PCT guidelines give limited technical information on search issues for biotechnology related subject-matters.

Patent Offices have defined specific examination criteria for biotechnology inventions. The Trilateral web site provides a comparative study on requirement for disclosure and claims for biotechnology, including legal and technical details on the patentability criteria at the three Patent Offices [8]. This document can be compared with the specific Examination Guidelines that are issued by USPTO [9] and other authorities, such as CIPO [10] or UKIPO [11].

1.3. Searching biotechnology information: the example of antibodies

Biotechnology information is associated with a large and heterogeneous panel of subject-matter: protein and DNA sequences, medical uses, cells, biochemical analysis, microbiological processes, etc. How are all these technical aspects are actually captured and searchable in patent and scientific resources?

The present article suggests how a similar question can be addressed by providing an overview of text-based information resources that are searched using different "biotextology" [12] techniques. In particular, quantitative analyses of documents and/or database records were performed by extracting information using specific criteria related to antibodies, a specific class of glycoproteins that circulate in blood and biological fluids and represent the principal effectors of mammal immune system.

Antibodies have very distinctive molecular and biological features that have been studied, modified, and exploited for many different uses, as summarized in reviews [13,14]. Due to their strong binding to organic or inorganic targets (defined as antigens) and the great structural and functional adaptability, antibodies are possibly the molecular tools most extensively used in modern biology for both medical and non-medical applications having scientific and, at least in some cases, major commercial importance [15,16].

Structurally, antibodies are multimeric protein complexes formed by heavy and light chains that are held together by disulfide and non-covalent bonds (Fig. 1). Amongst the different type of antibodies, those best characterized and widely used are monoclonal antibodies, which are clonally derived from a single cell and produced using cell biology and recombinant DNA technologies. The high sequence diversity of monoclonal antibodies is responsible of the large number of distinct antigens that can be bound by them *in vivo* and/or *in vitro*.

The complexity and the amount of information and specific technical language increased dramatically in the last 30 years for defining all the new antibody-based products and technologies. The main area of interest when searching information on antibodies can be summarized and put in relationship (Fig. 1). An antibody is a biological product characterized by a structure and a format that provide an activity potentially useful for (non-)medical applications. Distinct R&D technologies allow generating and characterizing antibodies having a specific activity for the desired uses (for example, biological assays for selecting, within a library, the antibodies that bind a pathogenic agent).

General aspects of the search and analysis of antibody information have been described in the Trilateral Search Guidebook on Biotechnology [7], as well as in materials prepared by information providers on biotechnology in general [17] or more specifically on antibodies [18]. Compared to these references, the present study intends to identify some general issues when searching biotechnology information by analyzing the results of identical search strategies in different databases and/or platforms for text-based searches.

This article does not take into consideration the search of biological sequences (i.e. DNA and protein sequences) which deserve specific considerations. Some general guidance on main databases and search techniques in connection to specific scopes has been provided in previous articles [19,20] and in the Trilateral Search Guidebook on Biotechnology [7]. Most recently, a study has critically compared the content of databases that allow searching for biological sequences disclosed within patent documents [21].

2. Methodologies

Patent databases were accessed directly through their specific websites as in the case of PATENTSCOPE (http://www.wipo.int/pctdb/en/), EPO (http://www.epo.org) and USPTO (http://www.uspto.gov) databases. Commercial databases were accessed through internet for MICROPATENT (http://www.micropat.com) or through STN on the Web platform (http://stnweb.fiz-karlsruhe.de/html/english/) for WPINDEX. The literature databases were accessed through STN on the Web platform (http://stnweb.fiz-karlsruhe.de/html/english/) in the case of MEDLINE, EMBASE, BIOSIS, and SCISEARCH, or directly from Internet as in the case of HIGHWIRE (http://highwire.stanford.edu/). Other relevant websites are indicated in the text.

The search criteria applied in each analysis are indicated in the description of the figure and/or in the text. Such criteria were assembled using the operators, the field restriction, and the format appropriate for the specific database. Searches performed with truncated terms are indicated with an asterisk (e.g. antibod*), while terms used as a phrase are indicated within " " (e.g. "monoclonal antibody"). The graphs were elaborated using Microsoft Office suite.

3. Searching for antibody information in patent literature

3.1. Patent protection & disclosures related to antibodies

The task of drafting patent applications that claim inventions arising from biological research is complicated by the steady pace with which biotechnology evolves and knowledge is accumulated. In fact, a quick search in main databases shows that, weekly, several thousands of articles and several hundreds of patent documents are published in this area. The amount, the heterogeneity, and the complexity of such disclosures can make particularly difficult the task of defining correctly prior art at a specific date, and consequently patentable subject-matters, becoming a source of frustration [22].

On their side, Patent Offices are presented with the problem of preparing meaningful Search Reports as basis for the examination. Many patent applications in the field of biotechnology are considered as "complex applications" since they contain large sets of claims and contemplate multiple possibilities or sophisticated technical parameters [23]. This inherent complexity seems to have a direct effect on patent proceedings, since biotechnology-related patent applications are those combining the highest number of claims and of pages with the slowest patent prosecution [24].

The highly competitive field of antibody research and discovery is no exception, given also that the technical progress now provides large varieties of antibody formats, applications, and activities [13–15]. As summarized in recent presentations [25,26], the case law in both Europe and USA demonstrates how applicants are now required to consider attentively enablement and inventive