World Patent Information 43 (2015) 12–24

Contents lists available at ScienceDirect

World Patent Information

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

Public disclosure of biological sequences in global patent practice

Osmat A. Jefferson [a, b, *], Deniz Köllhofer [a, b], Prabha Ajjikuttira [a, b], Richard A. Jefferson [a, b]

[a] Queensland University of Technology, Brisbane, QLD 4000, Australia
[b] Cambia, P.O Box 3200, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 5 January 2015
Received in revised form
20 July 2015
Accepted 23 August 2015
Available online 22 October 2015

Keywords:
Patent
Biological patent
Patent sequence
Patent office
Sequence listings
Patent sequence data
Patent sequence download
PatSeq tools
Patent disclosure

ABSTRACT

Biological sequences are an important part of global patenting, with unique challenges for their effective and equitable use in practice and in policy. Because their function can only be determined with computer-aided technology, the form in which sequences are disclosed matters greatly. Similarly, the scope of patent rights sought and granted requires computer readable data and tools for comparison. Critically, the primary data provided to the national patent offices and thence to the public, must be comprehensive, standardized, timely and meaningful. It is not yet. The proposed global Patent Sequence (PatSeq) Data platform can enable national and regional jurisdictions meet the desired standards.

1. Introduction

In the traditional working of the patent system, an inventor secures governmental rights to exclude others from making, using, or selling his/her invention for a limited time in exchange for publicly disclosing the full details of the invention - what is called 'the teachings'. The teachings derived from the disclosure and the practice of an invention enable the public to use the invention through licensing, to use the invention freely without license outside the jurisdiction, scope and timeframe of protection, build upon the invention through research and development, improve upon it, or design around it to advance scientific and technological capabilities and ultimately to benefit society.

In the contemporary use of patents to secure rights over genetic material, the quality of these teachings has come under public scrutiny and the role of patent offices in the disclosure process has been challenged [1,2].

Within patent documents, genetic sequences have been viewed both legally and practically as either chemical compounds or as information-encoding elements, and within the context of patent eligibility or infringement issues, their structure and function value has gained more importance as various jurisdictions — including the United States and Europe - attempt to balance competing interests either in favor of the inventors, as the case in Europe, or the public, as the case in USA [3].

As genetic sequences are made up of combinations of four bases — designated as A, C, G, and T (U), in the case of DNA (RNA) — or 20 amino acids each with different chemical properties - designated with single or triple letter codes - in the case of protein, they can only be interpreted using specialized computer software tools. Such tools clarify the structure, function and similarity of any sequence relevant to other sequences. Therefore, during the disclosure process, the applicant, the patent office, and upon publication, the public should be able to access the disclosed sequence data and use the computer tools to interrogate it within the context of all known sequence listings to interpret, understand, and value their combined effect on biological innovations. While some patent offices claim to have internal computer-mediated searching, analysis and visual tools to interpret the contextual value or meaning of patent sequences, public access is still lacking. Moreover, creating patent landscapes that can integrate sequence information with global patent rights and disclosures remain expensive, slow and

* Corresponding author. Cambia/QUT, P.O. Box 3200, Canberra ACT 2601, Australia.
E-mail address: Osmat@cambia.org (O.A. Jefferson).

http://dx.doi.org/10.1016/j.wpi.2015.08.005

cumbersome to the public and to those professionals who cannot afford the costly services of commercial providers.

Rules for handling of sequences in patent prosecution, implemented by United States Patent and Trademark Office (USPTO) and other major patent offices in the 1990s, required the submission of any sequence (nucleotide or peptide) disclosed in any national or foreign application [4]. At that time, the disclosure standard format, known as "Sequence Listing", was simple and file submissions were accepted either electronically or on paper [5]. As sequence disclosures grew exponentially over time, more legal rulings were introduced regarding submissions and with respect to compliance with standard formats. While the major offices recommended international standards such as ST. 25 [6], for the disclosure of sequence listings in the submitted patent applications, the submitted file formats remained flexible until recently (Table 1). Full compliance with ST. 25 and the inclusion of the associated metadata such as the origin of the sequence, its length and type, function, and other markup in a computer readable format [7], were actually achieved in only a few offices; variations in the readability of file formats of disclosed sequence data and in its accurate matching when transferred to public databases persist [8,9]. For example, from 2001 until 2007, most international applications did not comply with ST. 25 text format rules and the disclosed sequences were in tiff or pdf files and contained NON ASCII binary data (Table 1, "Format of published sequences" category at WIPO in 2007).

2. Availability of published patent sequences to the public

Each of the major patent offices adopted a strategy regarding the publication and provision of sequence listings to the public. Table 1, column "Format of published sequence listings" depicts the practice adopted over time by USPTO, World Intellectual Property Office (WIPO), and European Patent Office (EPO). Throughout the past 25 years, variations have existed among these offices. For example, the published sequence data from US patent documents is available for bulk downloads under various file formats, however USPTO does not offer a sequence search facility to interrogate the data. The office passes its published data to the National Center for Biotechnology Information (NCBI) [10]. This center provides a comprehensive public sequence search facility, BLAST, allowing contextual interrogation of sequence data and a world class database, GenBank [11], hosting nucleotide and peptide sequences from primarily large sequencing projects and individual labs as well as over 12 million sequences from US granted patents since 1982.

In an effort to enable access and interrogation of larger sets of patent sequence data, NCBI and two other major public databases providers, the European Bioinformatics Institute (EMBL-EBI) [12], and the DNA Databank of Japan (DDBJ) [13] initiated an informal collaboration in the early 1990s; The International Nucleotide Sequence Database Collaboration (INSDC) [14], to exchange nucleotide (DNA or RNA) -*not protein*-sequences, including those disclosed in patents [15], and allow public access and interrogation of the data.

Similarly, the European Patent Office releases their published sequence listings mainly from published patent applications to EMBL-EBI that incorporates into ENA [16] database within the patent data class (PAT). The sequences are served to the public along with other received sequence listings from partner institutions. The EMBL-EBI databases also provide access to protein-based sequences in the Universal Protein Resource (Uniprot) [17]. Unlike NCBI, EMBL-EBI parses the received sequence listings and extracts associated metadata before serving it in the ENA database [18]. Furthermore, EMBL-EBI provides non-redundant sequence databases based on patent sequences stored in ENA and protein databases. The non-redundant databases are created at two levels and contain additional annotation, patent family information and links to patent literature [19,20].

Sequence listings disclosed in published patent documents from Japan Patent Office (JPO) and Korean Intellectual Property Office (KIPO) are shared through DDBJ, which is administered by the Center for Information Biology of the National Institute of Genetics in Japan. The Databank includes the nucleotide-based sequence listings from patent documents published in Japan and Korea since 1997 [21]. In 2010, two amendments were introduced into this database. First, the NCBI taxonomy ID was added to each sequence listing based on the original organism declared for that sequence in the patent application and the newly revised entries for nucleotides and proteins were released in May 2010 with a scheduled update once per year [22]. The second amendment included the release of protein sequence listings from JPO and KIPO for ftp downloading and later the availability of a sequence similarity search facility for protein sequence listings from USPTO, EPO, JPO, and KIPO [23].

Other public databases that provide access to and search facility of yet smaller collections of published patent sequences include Patome@Korea database serving nucleotide and protein patent sequences provided by the Korean Intellectual Property Office (KIPO) [24] from 2004 to 2008 and maintained by the Korean Bioinformation Center (KOBIC). Similarly, NASDAP, a semi-public Chinese database, provided free sequence search services to explore Chinese gene patents (applications and grants from 1999–Feb 2006), but it seems it is no longer available in our latest search of May 2015. The database covered 123,218 sequence listings from 8563 Chinese patents acquired from State Intellectual Property Office as hard copies or images [25].

3. Why do we need a global and transparent patent sequence dataset?

As NCBI, EMBL-EBI, and DDBJ decide which sequence listing data to include in their databases and what sequence search facility to provide on what data and when, accurate and comprehensive accounting of published sequence data as disclosed in patents is then hard to achieve. Upon reviewing the maze of the available patent sequences from the public or commercial sources, Andree et al. (2008) reported that each public database has still a unique dataset and for any comprehensive searching and analysis, users may need to access and use several databases [26].

Moreover, Cambia's 2011 survey of patent offices reveals that over the past twenty years [27], there has been progress in harmonizing sequence filing rules but sharing that knowledge in a meaningful way and at a global level with the public has lagged, as has ensuring compliance with these rules both by applicants and internally. An optimally functioning patent office embracing such a public disclosure responsibility would meet certain standards.

Biological inventions often disclose biological sequences, such as DNA or proteins or portions of them, which may or may not be claimed, and their teaching value depends on obtaining a clear understanding of the nature and function, clear differentiation between what is disclosed and what is claimed, and how such sequences are used in follow-on inventions, and in innovations (products and services) by whom, and where in the world. For example, zooming on GALNT18 gene in PatSeq Analyzer [28] reveals that a 15 mer portion in the 3′ end region (GGTTGGTGTGGTTGG) can be/has been used in several patent documents in different contexts. Table 2 lists the issued patents that reference that sequence in the claims and under various SEQ IDs. The table also depicts the corresponding claim referencing the SEQ ID, the claim category based on the use of that SEQ ID within each patent, the applicant name, and the filing date.