

Title

Patenting gene sequences in plant biotechnology

Authors

Koen Jonkers, Institute of Public Goods and Policies (CSIC-IPP), Institute for Public Goods and Policies (IPP), Consejo Superior de Investigaciones Científicas (CSIC), Albasanz 26-28, 28037, Madrid, Spain.

E-mail: koen.jonkers@cchs.csic.es

Catalina Martínez, Institute of Public Goods and Policies (CSIC-IPP), Institute for Public Goods and Policies (IPP), Consejo Superior de Investigaciones Científicas (CSIC), Albasanz 26-28, 28037, Madrid, Spain.

E-mail: koen.jonkers@cchs.csic.es

Keywords

Patent, gene, plant.

I. BACKGROUND AND OBJECTIVES

A large increase in the patenting of plant sequences has occurred in the wake of whole genome sequencing programs and other major advances in molecular biology and molecular genetics. Previous work (Jonkers 2010) showed a concentration of the global plant molecular life science research agenda around certain model organisms and economically important crops. This article explores whether patenting also mainly occurs around the plants that are used as model organisms and those used as major commercial crops – or if applied research and technological development for which patents can be used as an indicator have a different focus, namely other organisms that are relevant for medical and industrial applications.

A model organism is a species that has been widely studied, usually because it is easy to maintain and breed in a laboratory setting and because it has particular experimental advantages (Twyman 2002). Apart from *Arabidopsis thaliana*, maize, tobacco, tomato, pea, barley and rice have also been used as plant model organisms, but they have been replaced in terms of importance, in terms of number of articles and the citations received by them, by *Arabidopsis thaliana* in the past two decades (Jonkers 2010). *A. thaliana*, or thale cress, has a relatively small genome, a small size, and a short generation time, features which facilitate experimental laboratory research. *A. thaliana* was the first plant whose genome was sequenced and published in 2000. In the same year a ten year multi-national functional genomics program started. For reviews on the use of *A. thaliana* as model organism see among many others: Meinke *et al.* (1998) and Koornneef and Meinke (2010).

The objective of this paper is to study the evolution of patenting of plant genetic sequences by making use of the GBPAT section of GENBANK, a collection of patented sequences in patents hosted by the National Center for Biotechnology Information (NCBI) of the National Institutes of Health of the United States. For the purposes of this study, we have created a new database by linking information from GENBANK to the European Patent Office (EPO) Worldwide Patent Statistics Database, PATSTAT, and using additional information from PATENTLENS to improve the classification of organisms, as described below.

We explore the global (EPO/USPTO) trends in plant sequence patenting: more specifically the paper shows the share of plant *versus* other sequence patents, main source organisms, region of inventors, trends in the number of sequences patented, the number of patents made at EPO, USPTO as well as the number of patent-families (Martinez 2011), the (public/private) nature of applicant, etc. It aims to answer questions such as the following:

- What is the share of model organisms and major commercial crops in all patents?
- Are most patent applications filed by US applicants?
- Are business applicants specialized in specific types of organisms?
- What can the analysis of patent families tell us about applicant strategies?
- Are there significant differences if the analysis is limited to granted patents only?

Genetic patenting is an area of heated debate in scientific, legal and economic fora, especially as regards the human genome, but increasingly also with respect to plant biotechnology given the large economic stakes involved (Louwaars *et al.*, 2009; Baillie & Connett Porceddu, 2012).

Economic scholars have analysed competition issues related to patents on genetically modified crops, mainly relying on US data (Harhoff *et al.* 2001; Conti *et al.* 2003; Regibeau and Rockett 2003, 2005) and the value of European plant biotechnology patents (Schneider 2011). But, to our knowledge, this is the first time that a study provides a broad overview of the current state and recent changes in the patenting of plant gene sequences across organisms and jurisdictions and tries to delve into the different issues involved.

II. METHODOLOGY

We downloaded the patent sequences files from NCBI GENBANK flat file release 183.0 (February 15 2011 and April 15 2011)¹ which were used to create a relational database (for other studies using GENBANK see: Jensen and Murray 2005 on the human genome; Arrieta *et al.* 2010 and Arnaud-Haoud *et al.* 2011 on marine species). GENBANK has 19,189,921 different gene sequence entries and 424,238 patent publication numbers from different patent offices. The analysis presented here is limited to patents filed in EPO and USPTO and relies on information about these patents from the PATSTAT database, version of October 2010.

A potential obstacle to the analyses of interest is that 74% of the identified patent applications, corresponding to 36% of the sequences in GBPAT of GENBANK have not been classified according to their source organism, because the USPTO does not classify the genes in GENBANK by source. This problem has been addressed through the following strategy. First in PATSTAT all patents with the IPC code C12N 15/29 which refers to: GENES ENCODING PLANT PROTEINS were identified.² These were matched with the USPTO gene sequence patents to isolate a sample of plant sequence patents. A source organism has been manually assigned to each of the 1323 identified sequences in the identified patents for the years 1990, 1995, 2000 and 2005.³ The approach taken in doing so was the following. In PATENTLENS (Cambia 2012) the list of DNA/RNA sequences mentioned in each of the patents was identified. The likely source organism of these sequences can be found through a nucleotide BLAST search (Altschul *et al.* 1990; NCBI 2012).

“The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to

¹ <ftp://ftp.ncbi.nih.gov>

² This is a conservative estimate as there may also be some plant sequence patents that are classified under other IPC codes. The share of C12n 15/29 EPO patents of all GENBANK EPO plantae patents is 293/581=50%, so it is assumed that in our first estimate approximately 50% of all gene plant USPTO patents is missed by only looking at those with this IPC code. The following IPC codes were also considered but discarded for not being of direct relevance: A01H; A01N; B01J; C07H; G11B; C07K; C12N; C12N 15/29; A21D; C08L; C12Q; A23G; A61K; C12M; A23L; G06F; C08B; C12R; A23K; C12P; A01K; G01N; A23D; C11B; A61P; A23B; H01J.

³ Only very few patents were found for the years 1990 and 2000. A probable explanation for this is that: in 1990 there was still relatively little plant sequence patenting activity. In 2000 there is likely to have been more patenting activity than in 1995, but this was not visible because it was only in 2001 that the USPTO started publishing patent applications instead of only granted patents.

infer functional and evolutionary relationships between sequences as well as help identify members of gene families (NCBI, 2012)”.

By running such BLAST search in the entire NCBI database of genetic sequences a likely source organism can be identified.

III. EXPECTED RESULTS

First of all, we expect to achieve a better understanding of the geography of invention and ownership of patents on plant gene sequences of different types of organisms, in particular as regards differences between United States v Europe, businesses v academic institutions, and types of plants.

Second, we will perform a more in depth analysis of patenting related to Arabidopsis gene sequences, given its character of model organism and increasing importance in plant biotechnology scientific research. Model organisms are ‘research tools’ in plant biotechnology, so the patenting of Arabidopsis gene sequences raises relevant issues for science, technology and innovation policy related to sequential innovation, pre-emptive patenting, research exemptions, etc.

Although European-based research on Arabidopsis has grown considerably in recent years, US-based companies are expected to have applied for most of the patents involving Arabidopsis genome and applications. Also, it is expected that most of these patents have been applied for at the USPTO, and most of those which have been applied for at EPO had their priority in previously filed US patents with US firms as applicants.

The distance between US and European research on Arabidopsis is narrowing over time, with Europe catching up, but the gap on patenting is expected to have remained stable or to have even grown over time. To explore this expectation this paper explores how the counts of USPTO and EPO plant sequence patent applications and patent families with applications in either bureau, as well as the geographical distribution of inventors and applicants in each office have changed over time. Another element of interest is the public/private nature of these applicants in the respective countries.

Research on Arabidopsis in Europe is expected to have mainly been carried out by public research institutions and universities, whereas on the US side the involvement of private companies is expected to have been much more important since the beginning (also in the context of the genome sequencing programs). This, together with the time advantage provided by more permissive domestic patenting policy in the US, might lead to dominance of US companies in the market for GM crops in the future and have implications for the development of the European seed industry. If this is the case this study offers a more detailed analysis of a European paradox in a field in which Europe is strong in scientific terms.

The database that was created for this project can be used for in depth analyses of the patenting of genetic sequencing in other fields.

Acknowledgements

Stéphane Maraut built a relational database from the GENBANK flat files and provided IT advice throughout the project. Laura Manzano Baena and Maria Eugenia Espinosa Calvo helped with the codification of USPTO patents. We also wish to acknowledge funding from CSO2011-2350 and CSO2009-10845.

REFERENCES

Altschul, S; Gish, W; Miller, W; Myers, E; Lipman, D (October 1990). "Basic local alignment search tool". *Journal of Molecular Biology* 215 (3): 403–410.

Arnaud-Haond, S., Arrieta, JM, Duarte, CM., 2011, Global Genetic Resources - Marine Biodiversity and Gene Patents, *Science*, Vol. 331 no. 6024 pp. 1521-1522

Arrieta JM, Arnaud-Haond S, Duarte CM. What lies underneath: Conserving the oceans' genetic resources. *PNAS*. 2010;107(43):18318 -18324

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GENBANK. *Nucleic Acids Res.* 36 (Database issue), D25-30 (2008)

Baillie, B., Connett Porceddu, M., 2012, Patent Landscape of the Arabidopsis genome and related plant species, <http://www.patentlens.net/daisy/ArabidopsisGenome/1662.html>

Cambia, 2012, patentlens: a free public resource for patent system navigation worldwide, available at: www.patentlens.net

Conti M, Régibeau P, Rockett K, 2003, How Basic is (Patented) University Research? The Case of GM Crops. Working Paper University of Essex, Department of Economics, Economics Discussion Papers 558.

Harhoff D, Régibeau P, Rockett K, 2001, Some simple economics of GM food, *Economic Policy*, 16, 33, 263-299.

Jensen K, Murray F, 2005, Intellectual Property Landscape of the Human Genome, *Science*, 310, 239.

Jonkers, K., 2010, Models and orphans, concentration of the plant molecular life science research agenda, *Scientometrics*, 83 (1): 167-179

Koornneef M, Meinke D., 2010, The development of Arabidopsis as a model plant. *Plant J.* 61(6):909-21.

Louwaars, N., Dons, H., van Overwalle, G., Raven, H., Arundel, A., Eaton, D., Nelis, A., Breeding business, the future of plant breeding in the light of developments in patent rights and plant breeder's rights, CGN Report 2009-14, Centre for Genetic Resources, the Netherlands.

Martinez, C., 2011, Patent families, when do different definitions really matter, *Scientometrics*, 86: 39-63.

Meinke, DW., Cherry, JM., Dean, C., Rounsley, SD., Koornneef, M., 1998, Arabidopsis thaliana: A Model Plant for Genome Analysis, *Science*, Vol. 282: 662-682

NCBI, 2012, BLAST - Basic Local Alignment Search Tool, available at: <http://blast.ncbi.nlm.nih.gov/>

Régibeau P, Rockett K, 2003, Are More Important Patents Approved More Slowly and Should They Be? Working Paper University of Essex, Department of Economics, Economics Discussion Papers 556.

Régibeau P, Rockett K, 2006, Competition, regulation, and intellectual property management in genetically modified foods: evidence from survey data, in *Knowledge accumulation and industry evolution. The case of pharma biotech*, Eds. M. Mazucatto and G. Dosi, Ch. 12: 346-377, Cambridge University Press.

Schneider C, 2011, The battle for patent rights in plant biotechnology: evidence from opposition filings, *Journal of Technology Transfer*, 36: 565–579.

Twyman, R., 2002, What are model organisms? Available at: http://genome.wellcome.ac.uk/doc_WTD020803.html