The discovery of 'introns': analysis of the science-technology interface

J.J. (Jos) Winnink[*, **, ***], Robert J.W. Tijssen[*, **] and Anthony F.J. van Raan[*]

[*]{winninkjj, tijssen, vanraan}@cwts.leidenuniv.nl
Leiden University, Centre for Science and Technology Studies (CWTS),
Wassenaarseweg 62A, 2333 AL Leiden (the Netherlands)

[**] Leiden University Dual PhD Centre The Hague, P.O. Box 13228,
2501 EE The Hague (the Netherlands)

[***]NL Patent Office, P.O. Box 10366, 2501 HJ Den Haag (the Netherlands)

Abstract

The study addresses visible clues and empirical data that can be used to discover early stage breakthroughs in science that evolve into new technological developments. We analyse bibliographical information from scholarly publications, patent publications, and links between patent and scholarly publications. Our overall goal is to develop an analytical methodology for pinpointing the stage in which fundamental discoveries occur. This particular case study focuses on the discovery of 'introns' in chromosomes. We discuss two breakthrough events related to this discovery, classify them according to the 'Cha-Cha-Cha' theory, and show that the second breakthrough reflects a 'phase shift' where the scientific discovery moves into the first stages of technological development.

Introduction

Our longitudinal bibliometric analyses of 'breakthrough processes' aims at obtaining better insight into general patterns that characterise 'revolutionary' R&D dynamics (Winnink & Tijssen, 2011; Winnink 2012). We use the 'Cha-Cha-Cha' theory (Koshland, 2009) to classify breakthrough discoveries into distinctly different types. A short description of the 'Cha-Cha-Cha' theory is given in the next paragraph. The subject of this case study is the discovery of so-called 'introns'[i]. We categorise this discovery as a 'Challenge' because it can be characterised as posing a challenge that required a major 'paradigm shifting' adaptation of previous theories to explain new experimental observations (rather than solving an 'obvious' scientific problem).

The underpinning scientific discoveries were published in 1977 (see for instance Gelinas and Roberts, 1977; Sharp, 1993; Roberts, 1993) and revealed that chromosomes comprise of a mosaic of 'exons', used to form new copies of genes, and of 'introns' that appeared to be non-functional 'interspaces'. These discoveries proved the assumption 'chromosomes for prokaryotic[ii] and eukaryotic[iii] organisms are similar in structure' to be wrong. The 1993 Nobel Prize in Physiology or Medicine was awarded to Philip A. Sharp and Richard J. Roberts for their pioneering work on 'introns'. Sharp (1993) describes in his Nobel Prize lecture the 'intron' problem as follows:

> *'By the late 70s the physical structure of a gene was firmly established from work in bacteria. The sequences of the gene, the RNA and the protein were colinearly organized and expressed. Since the science of genetics suggested that genes in eukaryotic organisms behaved similarly to those of prokaryotic organisms, it was naturally assumed that this bacterial gene structure was universal. It followed that if the gene structure was the same, then the mechanisms of regulation were probably very similar, and thus what was true of a bacterium would be true of an*

elephant. However, many descriptive biochemical aspects of the genetic material and its expression in cells with nuclei suggested that the simple molecular biology of gene expression in bacteria might not be universal'

The two main research questions we address in this study relate to how these revolutionary breakthroughs can be identified and monitored over time:

- "Can such a breakthrough, involving a paradigm shift, be identified and characterized in terms of bibliometric variables and indicators?"
- "Can one identify, bibliometrically, the stage in the development process where scientific knowledge is being used for work on science-based technologies?"

Conceptual framework

Our analytical approach builds on 'Cha-Cha-Cha'=theory by Koshland (2009), who's work on discoveries

'In looking back on centuries of scientific discoveries, however, a pattern emerges which suggests that they fall into three categories —Charge, Challenge, and Chance— that combine into a "Cha-Cha-Cha" Theory of Scientific Discovery.',

resulted in the following general typology that differentiates discoveries based on their nature:

'"Charge" discoveries solve problems that are quite obvious — cure heart disease, understand the movement of stars in the sky — but in which the way to solve the problem is not so clear.

"Challenge" discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time. The discoverer perceives that a new concept or a new theory is required to pull all the phenomena into one coherent whole.

"Chance" discoveries are those that are often called serendipitous[iv] and which Louis Pasteur felt favoured "the prepared mind." In this category are the instances of a chance event that the ready mind recognizes as important and then explains to other scientists.'

Charge discoveries can be considered 'normal science' in Kuhn's (1962) terminology. Challenge and Chance discoveries can be seen, within this particular context, as 'revolutionary science'.

Further work by Hollingsworth (2008) defines a 'major breakthrough' or a 'discovery' as:

'a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem.'

which is essentially equivalent to Koshland's 'Challenge' discovery, whereas 'small advances' can be seen as a 'Charge'. Hollingsworth's definition does not address 'Chance' discoveries as they are serendipitous and therefore have no precursory research directly linked to the discovery.

This discovery-oriented approach of describing knowledge creation dynamics differs from philosophy of science 'systems level' approaches, such as Kuhn's (1962) distinction between 'normal' and 'revolutionary' science, the latter being characterized by 'paradigm shifts'. Paradigm shifts start as a 'localized' diversion of an existing theoretical framework to explain observations that conflict with the theory. The consequence is that from a relatively small nucleus the new theory gains support of scholars resulting in growing numbers of publications. Paradigm shifts evolve gradually (Isnard and Zeeman, 1976) caused by the fact that scientists are reluctant to switch theories and concepts. During a paradigm shift

increasing numbers of unique authors, institutions, journals, journal categories, and countries, where institutions are located, on publications related to the new paradigm can be observed.

Related studies, such as Andersen, Barker, and Chen (2006) focus on the cognitive structure of scientific revolutions and the changes in concepts during paradigm shifts. Worall (2003) discusses the various views on theory changes in scientific progress, whereas Isnard and Zeeman (1976) use catastrophe theory to explain the mechanisms behind the switching of views within societies.

Methodology

The various types of discoveries should manifest themselves by different types of events that can be externally detected through 'bibliographical signals'. The solution for a long-standing scientific problem, for instance producing free standing graphene (Winnink, 2012), can lead to an impulse-like sudden increase in the number of research publications. To select relevant scholarly publications we used the topic 'intron' in the on-line version of Thomson-Reuters Web-of-Science database (WoS). This search resulted in 31408 publications, articles and conference proceedings, for the period 1980-2012.

Patent publications were gathered by ***s***earching the EPO Worldwide Patent Statistical Database (PATSTAT), October 2011 version, for the terms "intron" and "non coding DNA" in titles and abstracts this resulted in 1,284 patent publications in 677 patent families for the period 1984–2012. We group patent publications describing the same invention in 'simple patent families' to prevent double counting[v].

The data set contains intron-related publications from 1980 (Figure 1) onwards. The figure shows a very significant increase for scholarly publications in 1990-1991. To link scientific discoveries and technology we use the citing-cited links between patent-publications and scholarly publications. These links are called non-patent-literature references (NPL).

Results

General trends

Figure 1 shows, on a logarithmic scale, the trends for the numbers of scholarly publications and patent applications. The sudden increase in the number of scholarly publications from 1990 to 1991 stands out. We show that there are two breakthroughs of different type involved. One is the discovery of 'introns' in 1977, a 'challenge' breakthrough. The second is a 'charge' breakthrough marked by the sudden increase in intron-related scholarly publications from 1990 to 1991. The figure also depicts a continuous increase in patent applications from 1990 onwards until reaching a plateau level in 1998. The number of scholarly publications reaches an output plateau level at about the same time.

Figure 1 Trends in scholarly publications and patent applications

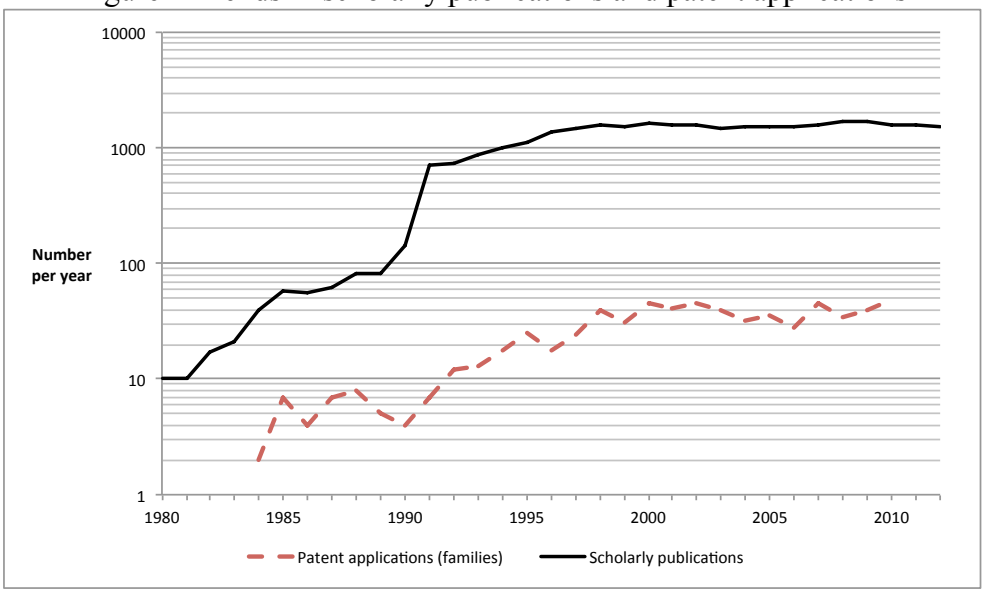
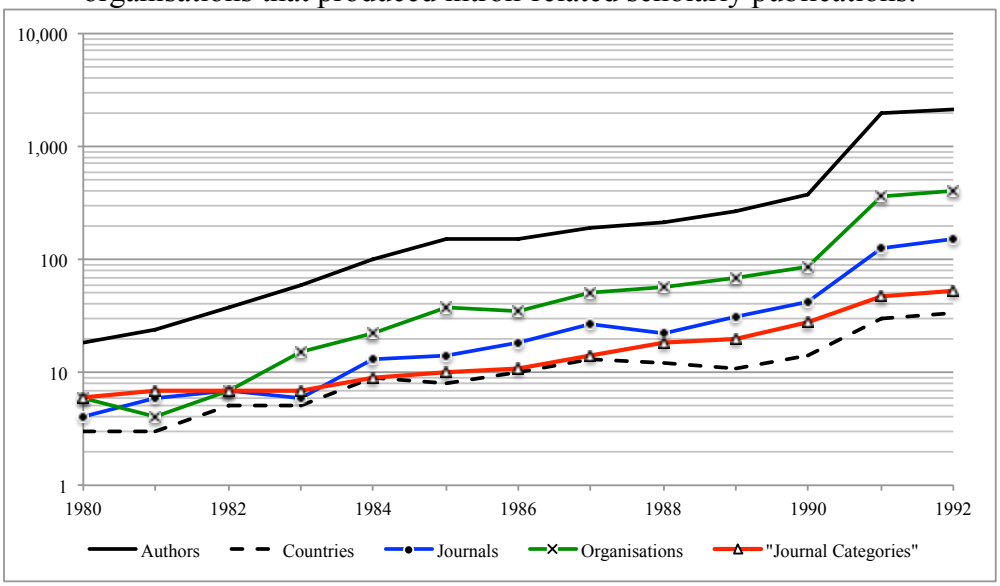


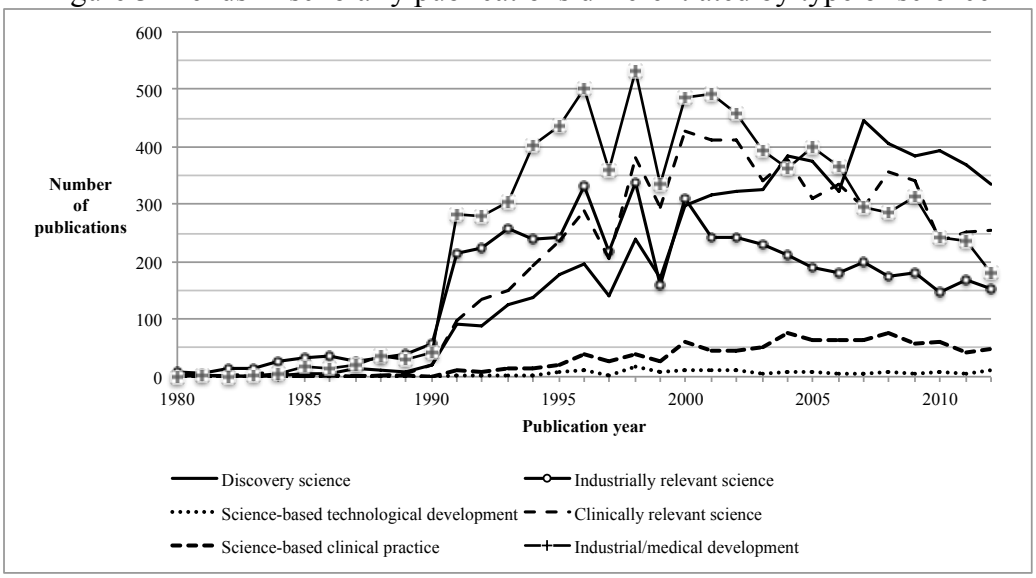
Figure 1 also reveals a low number of patent applications before 1992 indicating that intron related scientific knowledge was not yet used extensively in the development of new technologies. Figure 2 exhibits annual trends for the numbers of unique authors, journals, journal categories, organisations, and the number of countries where organisations that are 'active' in the intron-field at a certain moment are located; we presume that an actor that became active stays active. All these variables show a simultaneous and steady increase from 1980 onwards indicating that the theoretical concepts that form the basis for the selected scholarly publications spreads among the scientific community. These variables show the same 'phase shift' from 1990 to 1991 as in Figure 1.

Figure 2 Frequency counts of unique authors, journals, journal categories, countries, and organisations that produced intron-related scholarly publications.



In an attempt to explain the sudden increase 1990-1991 we classified the scholarly publications[vi] into six different categories, each encompassing a different type of science according to the cognitive-institutional environment in which the research was done (Tijssen, 2010). We could classify 85% of the publications in the document set. Figure 3 shows the results. The sudden increase is especially marked in 'application-oriented' categories 'Industrial/medical development', 'Clinically relevant science' and 'Industrially relevant science'. Our checks of the *Web of Science* indicate that this increase is not caused by new journals that were introduced into this bibliographical database. Although our inspection of the author affiliate addresses indicates that new institutional contributors (universities, research institutes, or other organisations) did enter the intron-field in 1991, the rise of publication output occurs mainly within incumbent institutes, especially the US universities.

Figure 3 Trends in scholarly publications differentiated by type of science



We constructed two subsets of scholarly publications; one for 1990 and one for 1991. Using the VOSviewer[vii] we created two maps (Figures 4 and 5) showing the spatial configuration of keywords and phrases as they occurred in the titles and abstracts[viii] of publications. Each of these content identifiers in these maps occurred in at least three publications. The same clustering and mapping algorithms were used to assure that the visible differences in the configuration reflect changes in content and relational structures.

In general, maps using terms from relevant documents to illustrate the evolution of a research area in science look for successive years quite similar. The differences between the two maps reflect considerable changes from 1990 to 1991 (Figure 3). Close inspection of the keywords and phrases in the maps shows a shift towards application of scientific knowledge for technological development.

Figure 4 Term map for scholarly publications published in 1990

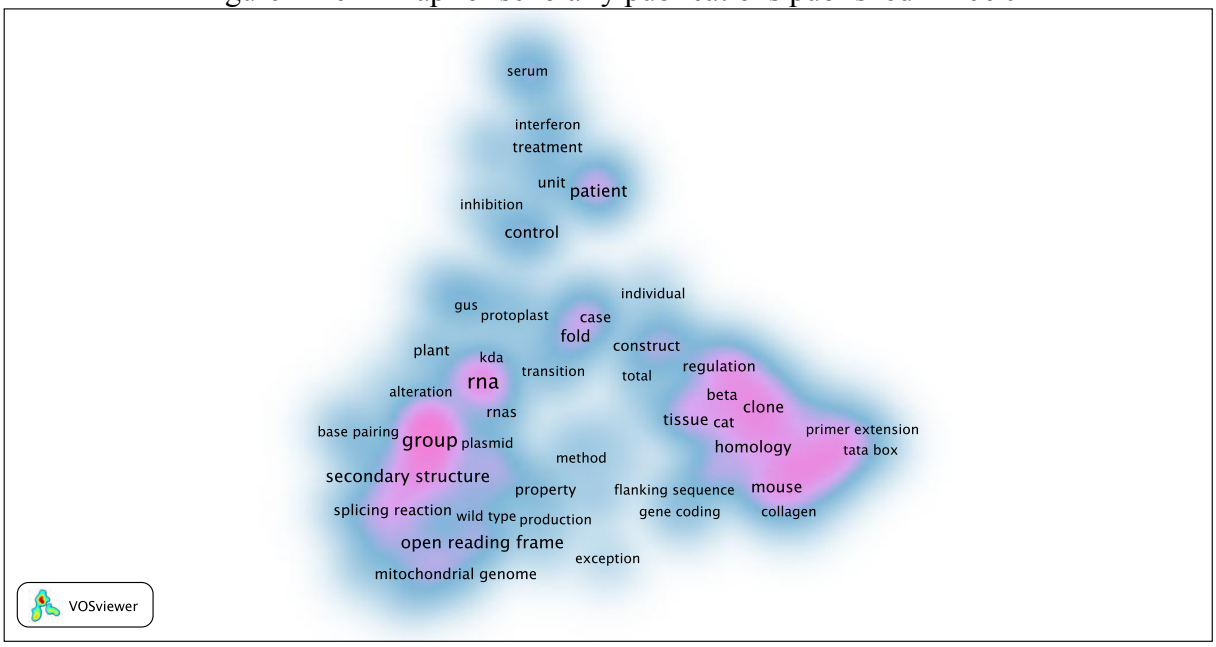
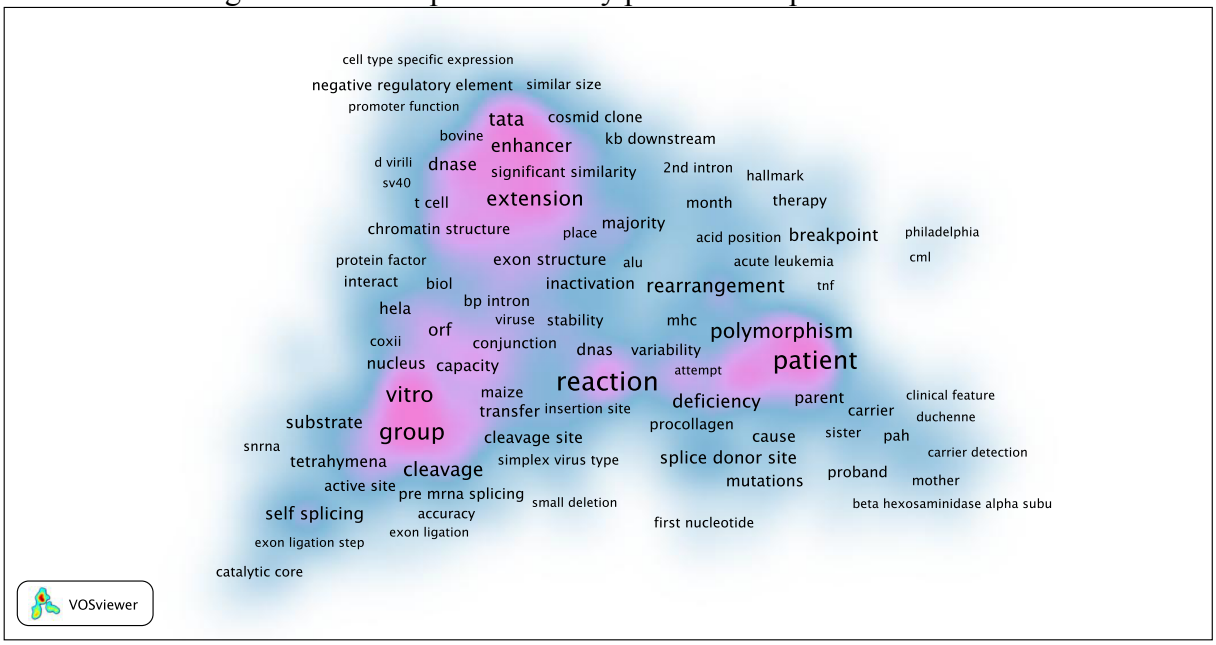


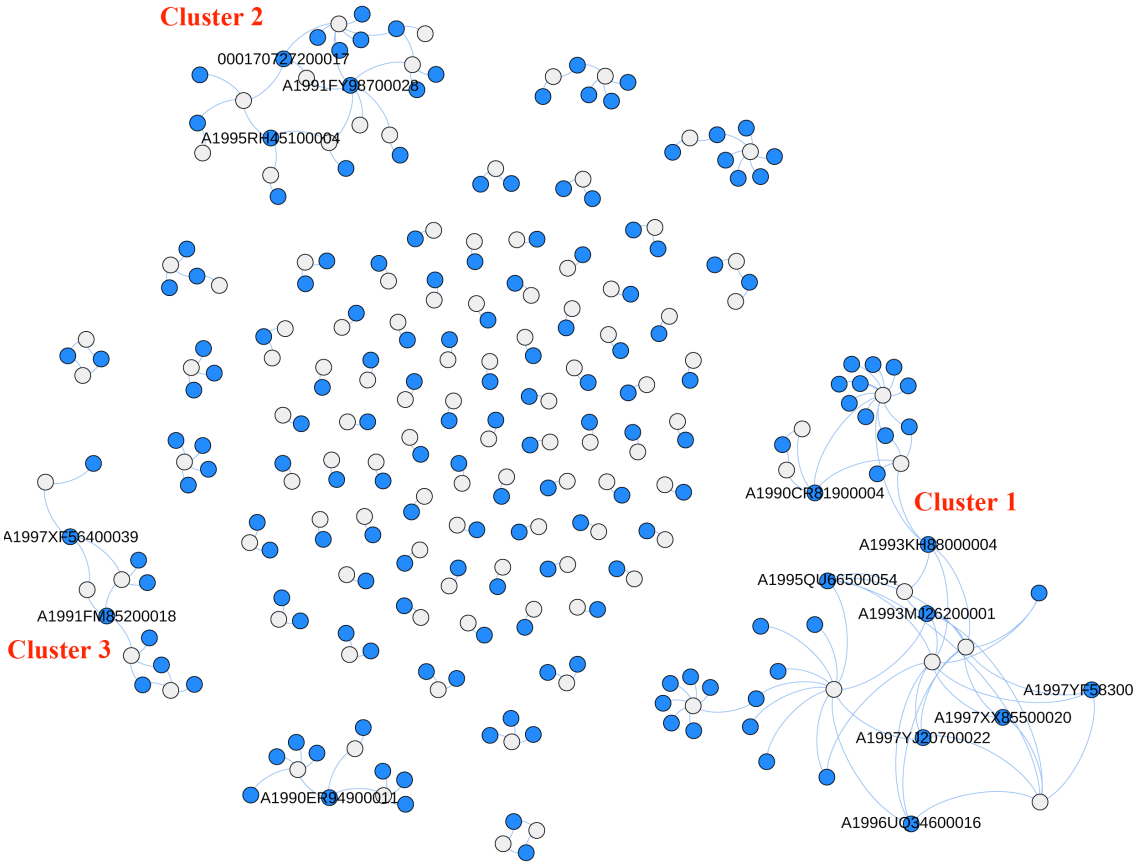
Figure 5 Term map for scholarly publications published in 1991



We find that 175 scholarly publications are cited within 112 patent families as non-patent literature (NPL) citations. These NPL citing-cited links are direct and visible connections between published science and patented technology. The network map in Figure 6 displays these NPL citing-cited links (citing-cited links between patents are omitted in order to highlight the science-technology interface) where the blue bubbles represent publications, while the white ones represent patent families. The graph shows a large number of dyads, mainly consisting of relations between a single publication and patent. These are the multitude of 'duplets' that occupy the centre. This patent-publication network contains three major patent-publication clusters. To place these clusters on a time-line we use the earliest application date of a patent family as application year for the whole patent family. For

publications cited three or more times (maximum is 6 citations) as NPL the identification numbers from the Web of Science (UT's) are shown.

Figure 6 Patent-publication network according to non-patent-reference connections (1980 - 2010)



We used the patent classification codes assigned to the patent publications to determine the main 'technological' topic of a cluster. Based on the same scientific knowledge base (i.e. 'introns' science) the technologies represented in these three clusters diversify into three distinct technologies. 'Recombinant DNA[ix]-techniques' is the topic of Cluster 1 and the patent applications were applied for in the period 1992-2002. Cited research publications in this cluster were published between 1988 and 2005[x]. Cluster 2 focuses on 'DNA vectors'[xi], patent applications in this cluster were applied for from 1998 until 2010, where the NPL-cited publications were published between 1991 and 2004. Cluster 3 covers the area of 'Techniques to modify DNA to express or suppress genes'. The earliest patent application in this cluster is from 1995 and the most recent from 2005. The cited publications were published from 1986 until 2001.

Concluding remarks

A paradigm shift in science starts as a 'localized phenomenon', sometimes a discovery or breakthrough, the effects of which gradually spread throughout relevant research fields. This knowledge dissemination process reveals as simultaneous increases in numbers of unique

authors, journals, journals categories, research organisations and the spread of different countries where those organisations are located. Figure1 shows that the number of patent applications is low prior to 1992, indicating that scientific knowledge shortly after the breakthrough in 1977 is not sufficiently mature for applications in new technologies. It also illustrates that the discovery in 1977 was a 'Challenge' breakthrough.

The sudden increase in the quantity of scholarly publications in 1990-1991 marks the stage in which intron-related science becomes mature in the sense of relevance for new technological applications; the increase is mainly visible in application-orientated scientific areas and marked by application oriented terms. The citing-cited network between scholarly publications and patents reveals three major clusters of publications. Patent applications in these three clusters start to appear after 1991, thus signalling a direct link between science and technology. The three clusters present technological differentiation based on the same scientific knowledge.

The observations we made support the two main research questions we addressed in this study. We are able to identify and characterize a breakthrough, involving a paradigm shift, in terms of bibliometric variables and indicators, and identified, bibliometrically, the stage in the development process where scientific knowledge is being used for work on science-based technologies. We furthermore conclude that the link between scientific knowledge and patented technology is established after a 'Charge' discovery successive to a 'Challenge' discovery.

Further research will focus on transformative developments in intron-related research since the 'challenge' breakthrough in 1977 that led to the 'charge' breakthrough in 1990-1991.

References

Andersen, H., Barker, P., and Chen, X. (2006). *The cognitive structure of scientific revolutions*. Cambridge University Press

Van Andel, P. (1994). Anatomy of the unsought finding. serendipity: Origin, history, domains, traditions, appearances, patterns and programmability. *The British Journal for the Philosophy of Science,* 45(2): 631–648.

Hollingsworth, J. R. (2008). Scientific discoveries: An institutionalist and path- dependent perspective. In Hannaway, C. (Ed.), *Biomedicine in the Twentieth Century: Practices, Policies, and Politics,* 317–353. National Institute of Health.

Isnard, C. A. and Zeeman, E. C. (1976). Some models from catastrophe theory in the social sciences. In Collins, L. (Ed.), *The Use of models in the social sciences*, International Behavioral and Social Sciences Ser.: Classics from the Tavistock Press, 44–100. Tavistock Publications.

Koshland, D.E. (2007). The Cha-Cha-Cha theory of scientific discovery. *Science* 317(5839): 761–762.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* The University of Chicago Press.

Gelinas, R.E. and Roberts, R.J. (1977) One predominant 5'-undecanucleotide in adenovirus 2 late messenger RNAs. *Cell* 11, 533-544.

Roberts, R. J. (1993). Nobel lecture: 'Split genes and RNA splicing', December 8, 1993. (http://www.nobelprize.org/nobel_prizes/medicine/laureates/1993/roberts-lecture.pdf)

Sharp, P. A. (1993). Nobel lecture: 'An amazing distortion in DNA induced by a methyltransferase', December 8, 1993. (http://www.nobelprize.org/nobel_prizes/medicine/laureates/1993/sharp-lecture.pdf)

Tijssen, R. J. W. (2010). Discarding the 'basic science/applied science' dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology* 61(9): 1842–1852.

Winnink, J. J. and Tijssen, R. J. W. (2011). R&D dynamics in the development of HIV/AIDS drugs. In Noyons, E., Ngulube, P., and Leta, J., editors, *Proceedings of the 13th International Conference of the Insternational Society for Scientometrics & Informatics (ISSI 2011),* volume II, 855–860.

Winnink, J. J. (2012). Searching for structural shifts in science: Graphene R&D before and after Novoselov et al. (2004). In Archambault, E., Gingras, Y., and Larivière, V., editors*, Proceedings of the 17th International Conference on Science and Technology Indicators (STI 2012)*, volume 2, 837–846.

Worrall, J. (2003). Normal Science and Dogmatism, Paradigms and Progress: Kuhn 'versus' Popper and Lakatos. In Thomas Nickles (Ed.) *Thomas Kuhn* (pp. 65–100). Contemporary Philosophy in Focus. Cambridge University Press.

---

i The word intron is derived from '**intr**agenic regi**on**'

ii Prokaryotic organisms are microscopic single-celled organisms, e.g. bacteria and cyanobacteria, that have neither a distinct nucleus with a membrane nor other specialized organelles.

iii Eukaryotic organisms are organisms consisting of a cell or cells in which the genetic material is DNA in the form of chromosomes contained within a distinct nucleus. Eukaryotes include all living organisms other than the eubacteria and archae.

iv See for instance Van Andel (1994))

v Because patents rights are national rights and patent procedures have several distinct phases patenting is a complicated process, in which several publications of the same invention can, and normally do, co-exist.

vi This classification is developed at CWTS and linked with the in-house version of the Web-of-Science database (TR-CWTS WoS). Linking with the on-line version of the WoS is based on 'Accession Numbers'.

vii For information on the VOSviewer see http://vosviewer.com

viii Before 1991 the Web-of-Science contained abstracts for only a fraction of the publications. We managed to obtain 138 abstracts for the 143 1990-publications, and 696 abstracts for the 703 1991-publications using additional sources.

ix DNA that has been formed artificially by combining constituents from different organisms

x Due to the complexity of the patenting process documents published after the filing of a patent application can show as cited non-patent-publication

xi An autonomously replicating DNA molecule into which foreign DNA fragments are inserted and then propagated in a host cell