7th Global Text Mining Conference
Atlanta, October 2017

EXTENDED ABSTRACT

Tracking the Emergence of Synthetic Biology

Philip Shapira pshapira@mbs.ac.uk
University of Manchester / Georgia Institute of Technology
Seokbeom Kwon skwon61@gatech.edu
Georgia Institute of Technology
Jan Youtie jan.youtie@innovate.gatech.edu
Georgia Institute of Technology

Overview. Synthetic biology is an emerging domain that combines biological and engineering concepts and which has seen rapid growth in research, innovation, and policy interest in recent years. This research presentation will discuss an effort to delineate this emerging domain using a newly constructed bibliometric definition of synthetic biology. The approach is dimensioned from a core set of papers in synthetic biology, using procedures to obtain benchmark synthetic biology publication records, extract keywords from these benchmark records, and refine the keywords, supplemented with articles published in synthetic biology dedicated journals. The search strategy is compared with other recent bibliometric approaches to define synthetic biology, using a common source of publication data for the period from 2000 to 2015. The research details the rapid growth and international spread of research in synthetic biology in recent years, demonstrates that diverse research disciplines are contributing to the multidisciplinary development of synthetic biology research, and visualize this by profiling synthetic biology research on the map of science. Further shown is the roles of a relatively concentrated set of research sponsors in funding the growth and trajectories of synthetic biology. In addition to discussing these analyses, the presentation notes limitations and explores lines for further work including refining the approach by further applications of machine learning and by adapting for analyses of patent landscapes in synthetic biology.

Approach. The paper puts forward a new bibliometric approach to delineating synthetic biology. We recognize the broad notion that synthetic biology involves the design and engineering of biological components and systems at the genetic level. We also acknowledge that there is significant debate about details that affect the operationalization of a bibliometric definition of synthetic biology. We thus tread carefully through these debates, realizing that they are not yet resolved, to put forth a pragmatic strategy for creating a bibliometric definition of synthetic biology. There is relatively little work so far available on the bibliometric definition of synthetic biology, and the definitions published to date are either too narrow or too expansive. We seek to contribute by refining an approach that better captures the complex scope of synthetic biology. We employ a multi-stage method, drawing from two publication indices (Web of Science and PubMed). The approach is used to identify scientific papers published in the synthetic biology domain and to trace patterns of emergence including international spread, funding, and disciplinary contributions.

Method: Our technique for developing a bibliometric definition of synthetic biology starts with a corpus of synthetic biology benchmark papers from which we build out procedures to capture other papers that can be considered in synthetic biology domain. We aim to include papers that are clearly acknowledged as synthetic biology as well as papers that should be included as part of the synthetic biology domain, even though they may not explicitly use "synthetic biology" in their title, abstract or key words. We also seek to exclude papers that are in related or other fields but which are not using the concepts, methods, or sources that are associated with synthetic biology. A four-step procedure is pursued (see Table 1). First, we gather a set of benchmark synthetic biology publication records. Second, we extract additional keywords from the abstracts of these benchmark record abstract by using Natural Language Processing (NLP). Third, we test and refine these keywords, and also delineate exclusion terms (Table 2). Finally, we include papers published in dedicated synthetic biology outlets.

Table 1 Overview of search strategy procedure

Step	Search strategy description and sub-steps
1	Retrieve benchmark records
	1.1. Download publication records searched by "synthetic biology" as benchmark records
	1.2. Retrieve abstracts from the benchmark records
2	Extract keywords with keyword (co-)occurrence pattern and add keywords from prior studies
	2.1. Extract candidate keywords from abstracts of the benchmark records
	2.2. Keep high frequency keywords, drop low-frequency keywords
	2.3. Combine keywords according to the keyword co-occurrence pattern
	2.4. Add suggested keywords from prior studies
3	Keyword screening by noise ratio test and face validation
	3.1. Measure noise ratio of each keyword
	3.2. Select keywords that have low noise ratio
	3.3. Extract exclusion terms by manually checking the abstract and title of the search records
	3.4. Download the publication records searched by the constructed keywords set from the Web of Science
	3.5. Download publication records searched by MeSH= "synthetic biology" from MedLine
4	Synthetic biology journal and special issue inclusions
	4.1. Search for the synthetic biology journals including special issues
	4.2. Download all the records of the published articles in the journal and special issues

Table 2 Collated keyword terms for synthetic biology search strategy (in Web of Science)

Search strategy – synthetic biology inclusion and exclusion terms
(((TS = ("synthetic biolog" OR "synthetic dna" OR "synthetic genom*" OR "synthetic *nucleotide" OR "synthetic promoter" OR "synthetic gene* cluster") NOT TS=("photosynthe*")) OR (TS = ("synthetic mammalian gene*" AND "mammalian cell") NOT TS="photosynthe*") OR (TS="synthetic gene*" NOT TS=("synthetic gener*" OR "photosynthe*")) OR (TS = ("artificial gene* network") NOT TS="gener*") OR (TS = ("artificial cell") NOT TS = ("cell* telephone" OR "cell* phone" OR "cell* culture" OR "logic cell*" or "fuel cell*" or "battery cell*" or "load-cell*" or "geo-synthetic cell*" or "memory cell*" or "cellular network" or "ram cell*" or "rom cell*" or "maximum cell*" OR "electrochemical cell*" OR "solar cell*")) OR (TS = ("synthetic cell") NOT TS = ("cell* telephone" OR "cell* phone" OR "cell* culture" OR "logic cell*" or "fuel cell*" or "battery cell*" or "load-cell*" or "geo-synthetic cell*" or "memory cell*" or "cellular network" or "ram cell*" or "rom cell*" or "maximum cell*" OR "electrochemical cell*" OR "solar cell*" OR "photosynthe*")) OR (TS = ("artificial nucleic acid*" OR "artificial *nucleotide")) OR (TS = ("bio brick" or "bioBrick" or "bio-brick"))))*

Note: This definition is applied directly into the advanced search feature of the Web of Science. It does not incorporate the additional synthetic biology journal search strategies described in the paper

We test and compare our approach against three other search strategies for synthetic biology (Oldham et al, 2012; Raimbault et al, 2013; and Hu and Rousseau, 2015). As detailed in the paper, our approach performs well in terms of both precision and recall.

Scientific Disciplines of Synthetic Biology. Synthetic biology is frequently described as an interdisciplinary research domain with contributions from biology, engineering, chemistry, computer science and other disciplines. Yet, there are also debates about the fields and specialties that underpin synthetic biology. Such discussions are important not only for definitional purposes but also because they suggest different trajectories for the emergence of synthetic biology.

To offer further insights on the nature of the disciplines that are contributing to synthetic biology, we draw on our synthetic biology publications dataset to analyze the subject categories associated with these records. Each publication is assigned to at least one of the more than 250 subject categories designated by the Web of Science based on citation patterns and judgement. These results are presented and discussed. Then, in an extension of the analysis, we layer the synthetic biology publication dataset onto a base map of science (Porter and Rafols 2009; Rafols et al. 2010). We draw on an enhanced overlay science base map and the clustering of subject categories into 18 macro-disciplines constructed with 2015 WoS journal data by Carley et al. (2015) and visualized with VOSviewer (van Eck and Waltman 2016). (Fig 1). We observe concentrations in three clusters. The largest cluster (by total papers for the 2000 to 2015 period) is "biochemistry, and molecular and cell biology." This is followed by "chemistry" (which includes the "biochemical research methods" category) and "biotechnology" (which includes "plant sciences"). A parallel analysis of publications in the leading macro-disciplines of synthetic biology over time indicates a redistribution of relative emphasis among the top three clusters, with chemistry being a major driver of change. This suggests growing interest in industrial biotechnology and biochemistry aspects of synthetic biology.

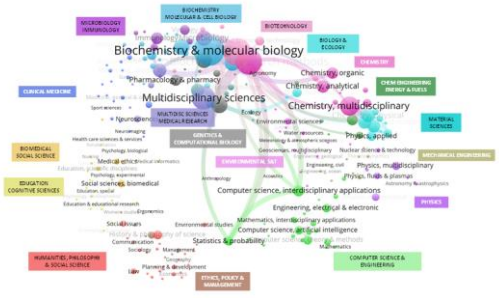

Fig 1| Profile of synthetic biology research by clusters and subjects, arrayed on the map of science.
Source: Synthetic biology articles (from WoS SCI-EXPANDED and SSCI), 2000-2015 (N=5,806). Map of science method from Carley et al. (2016), using VOSviewer (Van Eck & Waltman, 2016), with customization of 2015 WoS 18-category macro-discipline labels for synthetic biology.

Acknowledgements. This work was supported by the Biotechnology and Biological Sciences Research Council [Grant Number BB/M017702/1] (Manchester Synthetic Biology Research Centre for Fine and Speciality Chemicals) and by the National Science Foundation [Grant No. 0937591] (Center for Nanotechnology in Society CNS-ASU).

Keywords: Emerging Technology, Synthetic Biology, Bibliometric Analysis, Search Strategy, Map of Science, Research Sponsors

Fit with Conference Call: C. Translating analyses to useful intelligence: Informative indicators and compelling visualizations.