DISCUSSION PAPER No.197

bioRxiv に着目したプレプリントの分析

Analysis of preprints on bioRxiv

2021 年 8 月

文部科学省　科学技術・学術政策研究所

林 和弘　　小柴 等

【執筆者】

林 和弘　　文部科学省科学技術・学術政策研究所　データ解析政策研究室長

小柴 等　　文部科学省科学技術・学術政策研究所　データ解析政策研究室　上席研究官

【Authors】

HAYASHI Kazuhiro　Research Unit for Data Application,
National Institute of Science and Technology Policy (NISTEP), MEXT

KOSHIBA Hitoshi　Research Unit for Data Application,
National Institute of Science and Technology Policy (NISTEP), MEXT

本報告書の引用を行う際には、以下を参考に出典を明記願います。
Please specify reference as the following example when citing this paper.

林 和弘, 小柴 等 「bioRxiv に着目したプレプリントの分析」, *NISTEP DISCUSSION PAPER*, No.197, 文部科学省科学技術・学術政策研究所.
DOI: https://doi.org/10.15108/dp197

HAYASHI Kazuhiro, KOSHIBA Hitoshi "Analysis of preprints on bioRxiv," *NISTEP DISCUSSION PAPER*, No.197, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/dp197

bioRxiv に着目したプレプリントの分析

文部科学省　科学技術・学術政策研究所
林 和弘，小柴 等

要旨

定量的なデータに裏打ちされたエビデンスに基づく科学技術政策形成が求められる中，学術ジャーナルに掲載される原著論文の量（論文数）と被引用数に基づく質に関する調査研究を補完することを目的に，原著論文の草稿であるプレプリントに着目した試行分析を行った。

arXiv の調査に引き続き，2010 年代に入って生物学系で進展しているプレプリントサーバである bioRxiv に着目し，原著論文との関係，プレプリントの引用などの観点から，bioRxiv の特徴および分野別特性を分析した。

その結果，arXiv に比較して分野間の差はほとんど見られないことがわかった．原著論文となった割合は 4 割程度であり，オープンアクセス誌に掲載されているものが多いこと等が分かった．国・地域別の特徴も見られており,プレプリントサーバごとの特性を踏まえた分析と，政策への展開が望まれる．

Analysis of preprints on bioRxiv

HAYASHI Kazuhiro, KOSHIBA Hitoshi
National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

Due to the growing demand for the formation of science and technology policy based on evidence backed by quantitative data, we conducted a trial analysis focusing on preprints, which are the drafts of original papers, in order to supplement the research on the quantity (number of papers) and quality based on the number of citations of original papers published in academic journals.

Following the survey on arXiv, we focused on bioRxiv, a preprint server that has made progress in biology since the beginning of the 2010s, and analyzed the characteristics of bioRxiv from the viewpoint of the relationship with the original papers and the citation of preprints.

As a result, we found that there was little difference between the fields compared to the analytics of arXiv. The percentage of original papers afterwards was about 40%, and many of them were published in open access journals. The characteristics by country/region were also observed, and it is well considered that the analysis for policy implementation should be based on the characteristics of each preprint server.

目次

本文

1　序論

我が国の科学技術イノベーション（以下，STI とする）政策立案において，エビデンスに基づく政策立案 (以下，EBPM：Evidence-based Policy Making) 機能の強化が求められている．「第 6 期科学技術・イノベーション基本計画 (令和 3 年 3 月 26 日閣議決定)」（以下，第 6 期基本計画という）においても，科学技術・イノベーション行政において，客観的な証拠に基づく政策立案を行う EBPM を徹底することとされ，内閣府などにおいてもエビデンス等を整備する関連する取組が進められてきた．

その中でも研究力については，計量書誌学や科学計量学を基礎とした学術ジャーナルの査読を通った原著論文（以下，原著論文とする）に着目した定量的な分析と，政策づくりへの反映が試みられている．例えば，科学技術・学術政策研究所（以下 NISTEP とする）においては，サイエンスマップ，国別ベンチマーキング，大学ベンチマーキング，などの調査分析を行い，その内容が STI 政策づくりの一助となっている．一方，原著論文の分析においては，研究成果が生まれてから，査読・編集・出版を経て公開されるまでのタイムラグが含まれるため，全体の傾向（trends）をレビューすることには向いているが，新興領域を早く押さえることは構造的に難しい．また，原著論文においては原則それぞれの領域で確立された科学的判断基準によって査読が行われるため，学際領域や融合領域，あるいは全く新しい概念の論文が不利になりやすい．

ここで，データジャーナルなども含めた研究プロセスにおける研究データの共有や公開に着目して，より早期の把握を行うことも考えられる．こうした動向はオープンサイエンスの一環として近年大きな注目を浴びており，その将来性は大きく期待されるものの，原著論文とは違って，データの粒度，形式，流通形態が標準化されていないために，現状では研究力を測る分析に原著論文と同じレベルで適用することが難しい．

そこで本報では，論文原稿の草稿であり，査読による選別もされていない「プレプリント」に着目し，その分析を行うことで，原著論文を基とした研究力の分析に対して相補的に新しい知見を得ることができるのではないかと考えた[1)]．

本調査研究は，プレプリントサーバに登載されたプレプリントの分析について，生物学系の分野での活用が進んでいる bioRxiv に着目して試行的に行うと共に，科学技術政策への示唆を得ることを目的に行った．

[1)] 第 6 期基本計画においても，“論文のオープンアクセス化や研究成果の迅速な公開の場の一つとしてのプレプリントの活用も一層加速しており，研究データの公開・共有を含め，オープンサイエンス等の世界的な知の共有を目指した研究成果のオープン化が進みつつある” との認識が示されている．

2 bioRxiv

プレプリントそのものや，プレプリントサーバについては，すでに様々な文献で説明されている [1, 3] が，ここでもまず，プレプリントやプレプリントサーバについて簡単に概要を確認する．

プレプリントとは，主に査読付きジャーナルに投稿する前の草稿原稿のことを指す．したがってプレプリントは論文の体裁は満たしているものの，査読済みでも出版されたものでもないという位置づけである[2)]．プレプリントを研究者仲間に事前に共有して意見を求めることは従来より分野を問わず広く行われる情報共有活動であった [2] が，1990 年代に入って Web が登場すると，このプレプリントを Web に掲載して誰でも読めるようにするプレプリントサーバが物理系分野で登場し，新しい知見の迅速な共有とより多くのフィードバックを得ることができるようになった．

現在，プレプリントサーバにも様々なものが存在する [2] が，その代表例としては arXiv（アーカイブ）が挙げられる [1, 3]．arXiv はプレプリントサーバの嚆矢で，1991 年にロスアラモス国立研究所の Paul Ginsparg 氏によって開設されたものである．現在は米国コーネル大学を中心として世界各国の協力のもとに運営され，特に物理・数学・情報系の分野でメジャーかつ最も歴史が長いプレプリントサーバとなっている．

arXiv 以外に著名なプレプリントサーバとしては，本稿で対象とする bioRxiv をはじめ，SSRN や medRxiv などが挙げられる[3)]．

bioRxiv は 2013 年から，非営利の研究教育機関であるコールドスプリングハーバーラボラトリーによって運営されている生物学系の研究分野で著名なプレプリントサーバである[4)]．arXiv と違って 2013 年スタートと比較的新しいため，記事のユニーク ID として DOI(Digital Object Identifier) を採用していたり，記事へのコメント機能，Twitter や Facebook でシェアするための機能やそれらの件数を表示する機能などが提供されている．その他，2019 年にスタートした医療系のプレプリントサーバである medRxiv とは運営母体や，運用システムがほぼ同一という特徴もある．2020 年 4 月からはテキストおよびデータマイニング（TDM）を目的に API を通じた記事データの取得や，バルクデータの提供もスタートしており，arXiv の様に様々な分析を容易に行える環境が整っている．また，投稿件数やジャーナル出版状況などに関する年・月単位での統計情報の提供も行っている[5)]．

しかしながら，bioRxiv 上の論文に関して引用件数やジャーナル出版までの期間などについては示されておらず，arXiv についてそれらの広範な調査を行った文献 [1] などと比較することが難しい項目も存在する．そこで，本論文では bioRxiv に投稿された論文について文献 [1] と同様の分析を試み，bioRxiv の状況を明らかにするとともに，研究分野の動向把握等への活用可能性など科学技術・学術政策への寄与について検討する．

[2)] 公開することで，より多くの読者の目に触れることになり，これにより読者全員が査読者であるという見方もあるが，一般的にジャーナル論文でいう意味での査読とは異なる．

[3)] これらを含め様々なプレプリントサーバをある程度まとめたリストとしては [3] に詳しい．

[4)] bioRxiv はライフサイエンス系のすべての投稿を受け入れるとしている．

[5)] `https://api.biorxiv.org/reporting/home`

3　分析の手法

分析の手法について以下にまとめる．

まず bioRxiv の論文書誌データについては，bioRxiv が提供する API やバルクデータセット[6] を通じて収集する．

書誌データの項目としては例えば以下が挙げられる．

- DOI
- タイトル
- 概要
- 著者名
- 投稿先分野
- 初版投稿日

この他にも項目が存在するが，詳細は前述した API の記事に譲る．

ここで，AI2 [7] が提供する Semantic Scholar [8] では，bioRxiv の引用文献情報（任意の bioRxiv の論文を引用している文献の情報）を得ることができる．そこで，Semantic Scholar の API [9] を通じて，bioRxiv の論文ごとに被引用先のデータ（タイトル，雑誌名，DOI 等）も収集する．なお，被引用数は分析実施時点におけるものであるため，将来に向かって数が増えていく可能性が高い．また，bioRxiv 搭載から Semantic Scholar に搭載されるまでのタイムラグも存在する．結果の分析においては，これらの点について若干の注意を要する．

ところで，bioRxiv の記事がジャーナルなどに投稿された場合，個別記事ページにそれらの情報は掲載されるものの，前述の API 等で取得可能な情報にそれらは含まれない．一方，bioRxiv の記事の DOI を Crossref[10]が提供する Crossref REST API [11] で検索すると，関連項目として最終的に出版されたジャーナルの DOI を得ることができる．さらに，当該ジャーナルの DOI を検索すれば雑誌名や公開日などの情報を得ることもできる．そこで Crossref REST API を通じて，掲載雑誌名等のデータも収集する．

以上の収集したデータについて，分野，期間，ジャーナル DOI の有無，などの軸で計量することで，bioRxiv の状況を明らかにするとともに，研究分野の動向把握等への活用可能性について検討する．

---

[6] https://www.biorxiv.org/tdm
[7] Allen Institute for AI
[8] https://www.semanticscholar.org/
[9] http://api.semanticscholar.org/
[10] https://www.crossref.org/
[11] https://github.com/CrossRef/rest-api-doc

4 結果

4.1 データ件数等

2021 年 4 月 17 日時点で収集可能なものを全収集し，以下の通りとなった．

データ総数 117,293 件
期間 2013 年 11 月 7 日 〜 2021 年 4 月 17 日

4.2 bioRxiv の分野について

bioRxiv には独自の分野割りがなされおており，現在までの累積では表 1 に示す 27 分野 (Subject Area) が存在する．また，まれに分野が記載されていないものがあり，それを含めると 28 分野となる．

bioRxiv では分野は基本的に 1 つのみを選択する形式のため，28 分野のどれかひとつに所属することになる．

ただし，2013 年時点には論文が存在しなかった分野（Animal Behavior And Cognition, Clinical Trials, Epidemiology, Paleontology, Pathology, Pharmacology And Toxicology, Physiology, Scientific Communication And Education），2021 年時点で論文が存在しない分野（Clinical Trials, Epidemiology）も見られる．

これらについては，分野は存在していた・いるが投稿がなかった，分野自体が存在していなかった，という少なくとも 2 通りの解釈ができる．2021 年時点で論文が存在しない分野（Clinical Trials, Epidemiology）については，姉妹版の medRxiv の方がより適切と思われるため，分野がなくなった可能性が高いと推測される．

表 1　bioRxiv における分野分類

分野分類 (Subject Area)	**和訳**
Animal Behavior And Cognition	動物の行動と認知
Biochemistry	生化学
Bioengineering	バイオエンジニアリング
Bioinformatics	バイオインフォマティクス
Biophysics	生物物理学
Cancer Biology	癌生物学
Cell Biology	細胞生物学
Clinical Trials	臨床試験
Developmental Biology	発達生物学
Ecology	生態学
Epidemiology	疫学
Evolutionary Biology	進化生物学
Genetics	遺伝学
Genomics	ゲノミクス
Immunology	免疫学
Microbiology	微生物学
Molecular Biology	分子生物学
Neuroscience	神経科学
Paleontology	古生物学
Pathology	病理学
Pharmacology And Toxicology	薬理学・毒性学
Physiology	生理学
Plant Biology	植物生物学
Scientific Communication And Education	科学的コミュニケーションと教育
Synthetic Biology	合成生物学
Systems Biology	システム生物学
Zoology	動物学

4.3　データ登録数推移

図 1 および図 2 に，2013 年からの年ごとの論文投稿数を示す．順調に数が伸びているため累積に見えるが，実際にはその年ごとの投稿数である．

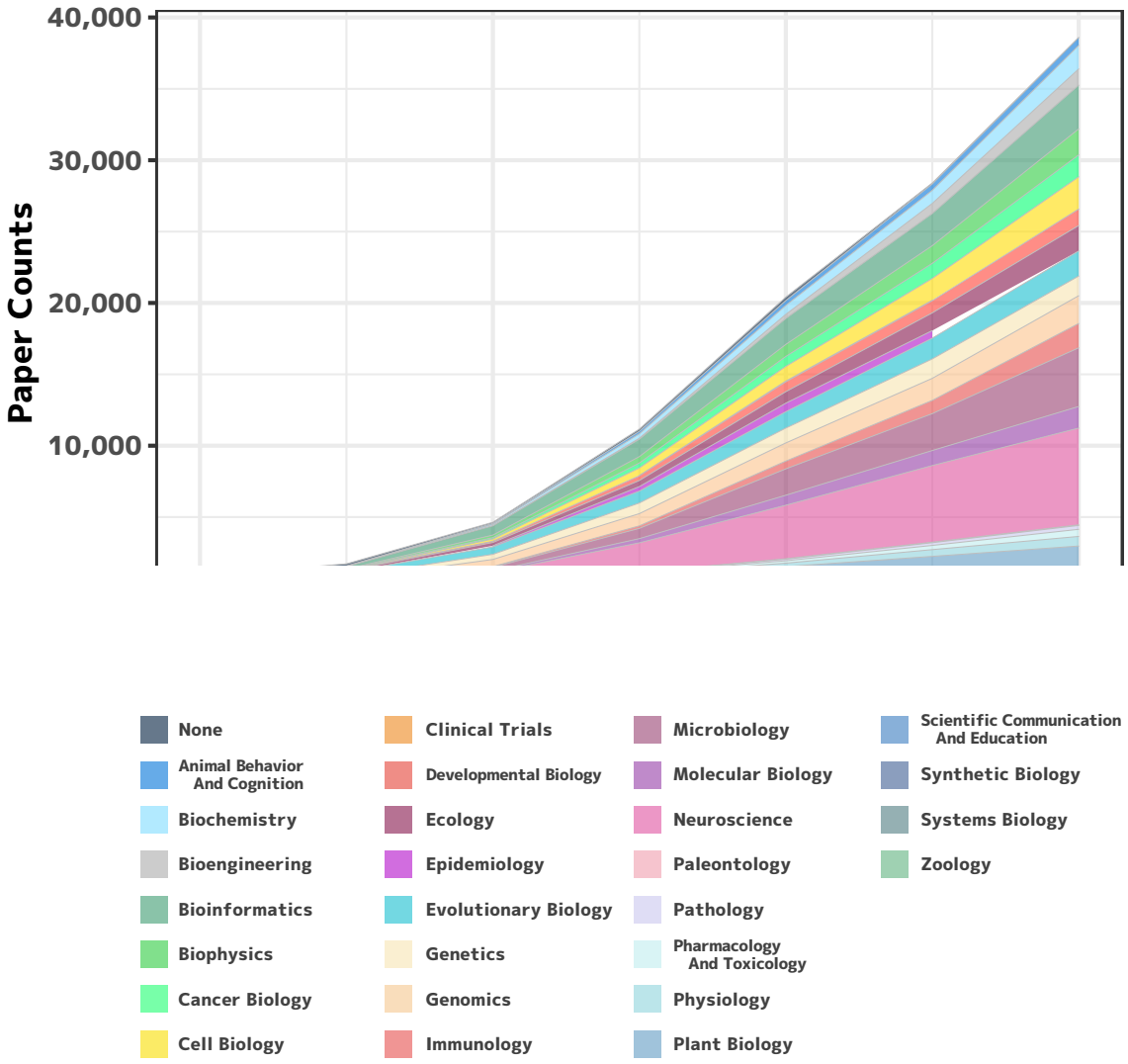
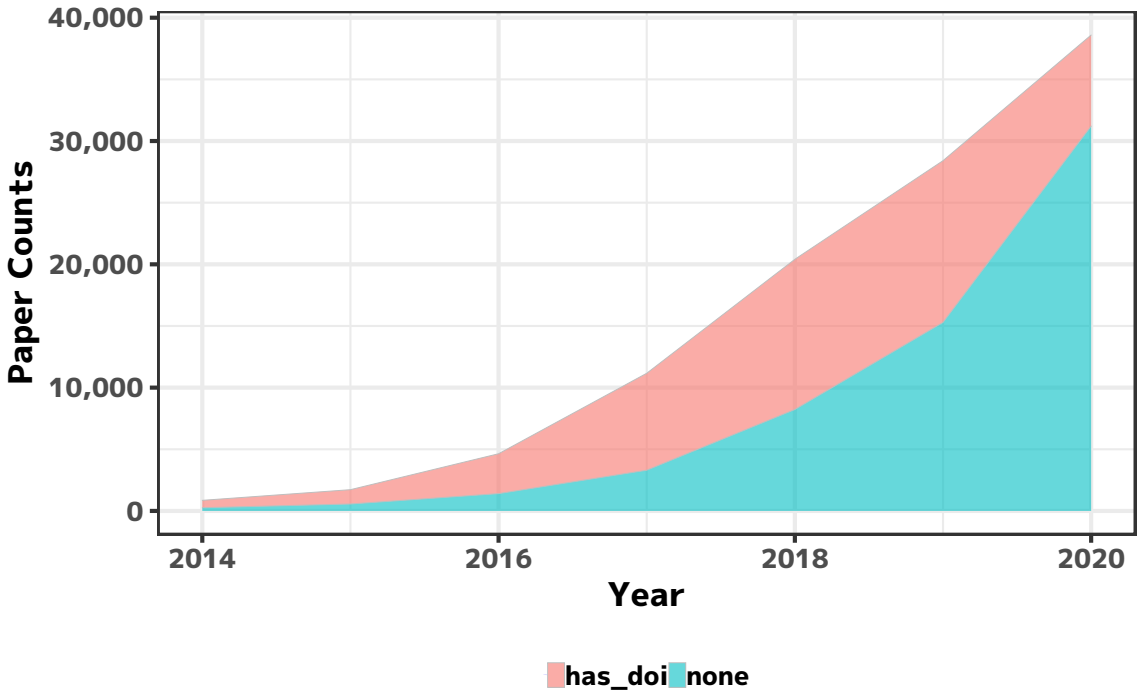


図 1　年・分野別の投稿数推移

図 1 をみると，全体の投稿数自体が順調に増加しており，2017 年から毎年 1 万件程度投稿件数を伸ばし，2020 年は年間おおよそ 4 万件の投稿を受け付けている．分野別では Neuroscience の投稿数が特に増加している傾向が読み取れる．また，Microbiology についても順調に数を伸ばしている傾向が読み取れる．

同じデータをジャーナル DOI(Digital Object Identifier) の有無で塗り分けた図 2 を見ると，2019 年に傾向が変化する様子が読み取れる．ここから，単純には bioRxiv 投稿後 1 年の辺りでジャーナ



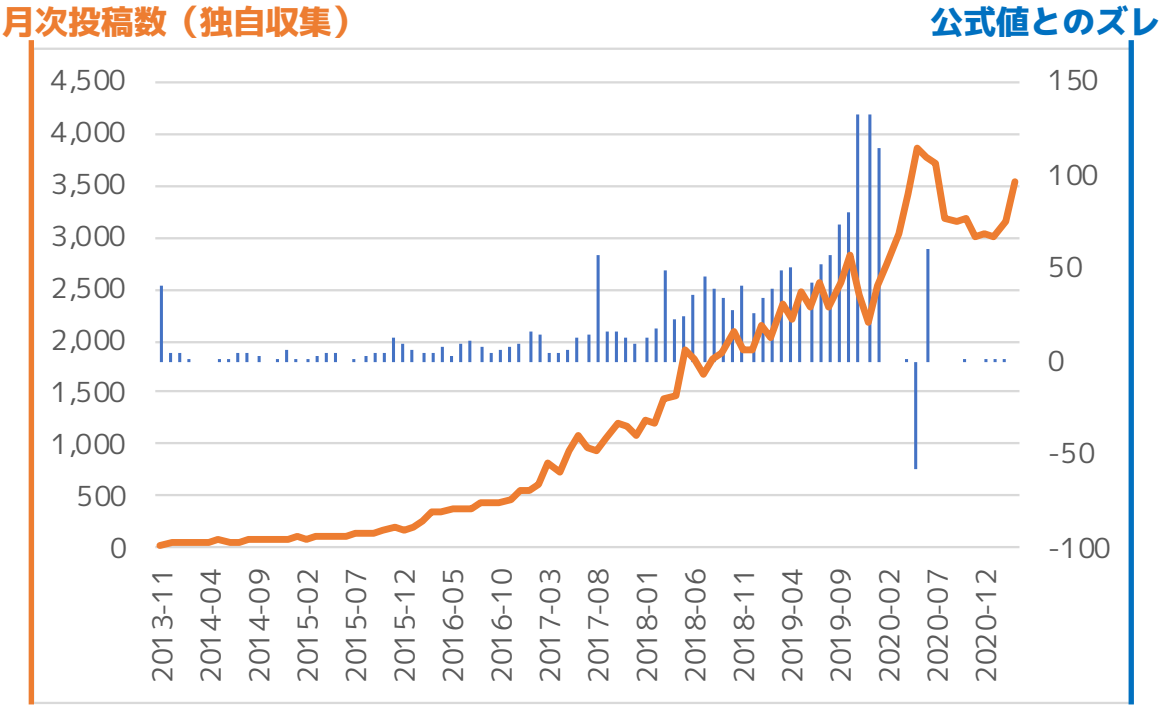


図 3　公式値とのズレの程度

なお，bioRxiv の公式値[12)]と，今回収集したもののズレの程度について図 3 にまとめた．割合で考えると大きくはなく，特に 2021 年に入ってからは差は少ないものの，月によっては独自収集分が公式値に比べて数百件程度少ないケースが多い．また，2020 年 3 月には独自収集分の方が多いという状況も生じている．誤差の範囲とみることもできるが，公式の集計値と差があることについて一定の留意を要する．

[12)] https://api.biorxiv.org/reporting/home

4.4　ジャーナル DOI 中の Award 情報

前節でジャーナル DOI が付与された bioRxiv 論文数の推移を確認した．ところで，Crossref を通じて得ることのできる DOI 情報の中には Award（研究助成）の情報も含まれる．

そこで，ジャーナル DOI のうち Award 情報が付与されたものの件数および，その中に “Japan” の文字列を含むものが何件程度あるかを調査した．

結果を以下に示す．

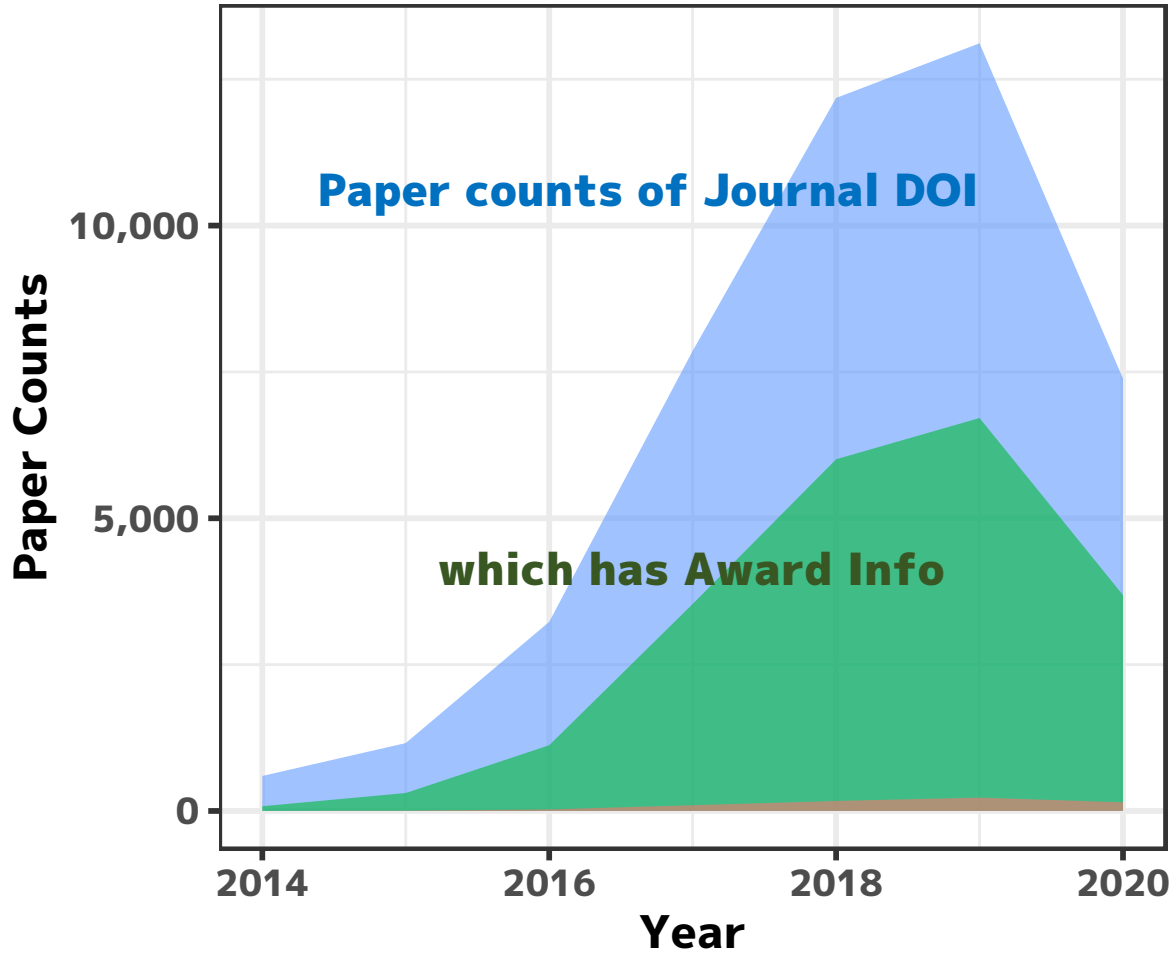


図 4　DOI 付き論文のうち Award 情報を有するものの数

図 4 をみると，Award 情報は 2016 年頃を境に急激に増加しているように見受けられる．

次に図 5 をみると我が国からの研究助成を得て実施されたと推測される論文も 2016 年あたりから観察されはじめ，2019 年には 200 件程度検出されている．

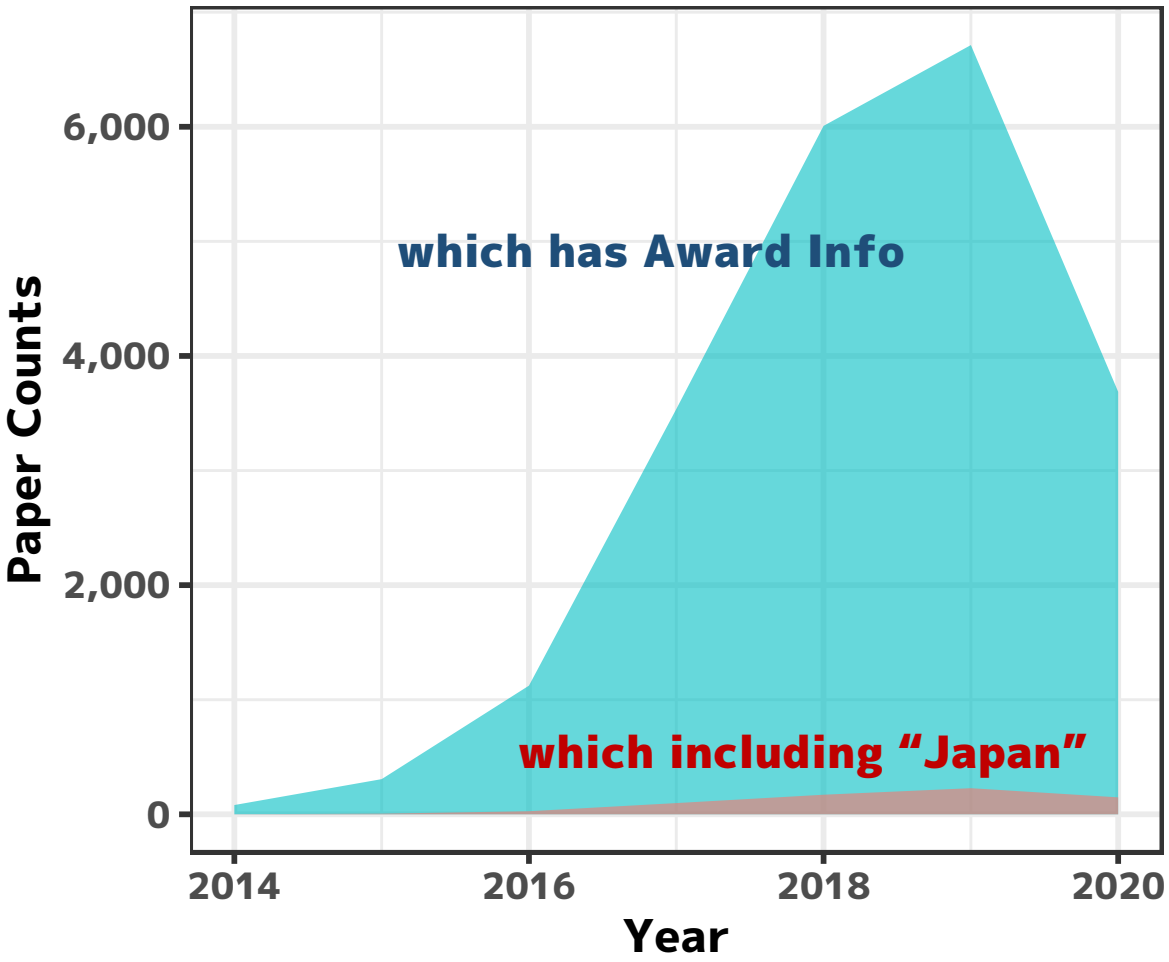


図 5　Award 情報に “Japan” の文字を含むものの数

4.5 分野と DOI の関係性

先に述べたとおり，bioRxiv の論文にジャーナル DOI が付与されるということは基本的には何らかの雑誌に掲載されたこととほぼ同義と捉えられる．したがって，分野毎にどの程度ジャーナル DOI が付与されているかを観察すれば分野毎の論文誌採択率に相当するものを推定できる．

ところで，ここまで未定義のまま “ジャーナル DOI” という語を用いてきたが，これは，bioRxiv 論文の DOI そのものについて Crossref の API を通じて情報を取得した際，その属性値に含まれる "relation" 中の "is-preprint-of" 要素に設定されている DOI を意味する．

ここではこのジャーナル DOI について分野単位での付与率を算出した．結果を図 6 に示す．

図 6 では図 2 の傾向を参考に，期間 2016 年 10 月頭から 2020 年 9 月末投稿の 5 年分をに限定した上でジャーナル DOI 付与率を示している．

図 6 を見ると分野毎のジャーナル DOI の付与率はほぼ同様で概ね 4 割程度と目される．

ここで，既に述べたとおりジャーナル DOI が付与されているということは何らかの雑誌に掲載されていることを意味する．他方，**DOI が付与されない場合**には以下のような複数の状態が考えられる．

1. 論文誌に投稿中・査読中
2. 論文誌に投稿したがリジェクトされた
3. そもそも bioRxiv への掲載に留めており，論文誌に投稿していない
4. 論文誌に採録されたが論文誌側に DOI がない
5. 論文誌以外（例えば会議録，書籍）のメディアで発行されたが DOI が付与されていない
6. 論文誌や他のメディアに採録され DOI も付与されたが，bioRxiv の情報更新を忘れている

したがって，図 6 に示した結果からジャーナル DOI 付与率が低いことのみをもって，ある分野の採択率が厳しい，ある分野では論文誌に投稿しない，といった推定を行うことは一定の留保を要する．

続いて，bioRxiv に登録されてから DOI の公開日までの期間を分野毎に調べた．結果を図 7 に示す．

図 7 は図 6 とは期間を変え，bioRxiv 開始時点の 2013 年から 2020 年 9 月末投稿までの 8 年分を採用した．これは期間の算出に際して十分な幅の “窓” を設ける，すなわち DOI が付与されるまでの期間が長かったプレプリントもできるだけ拾うことを意図したものである．また，最新の 2021 年 4 月から半年以上の期間を空けることで採録に要する期間の影響をある程度軽減しようと試みたものである．期間が異なることから図 6 の結果との単純な比較は困難である点に留意を要する．

その上で図 7 をみると ここでも分野による差はほとんどないことが観察できる．平均的には概ね半年（6 ヶ月）長い分野でも 8 ヶ月もあればジャーナル DOI が付与されているようである．

最後に，全期間を通じて具体的にどういった雑誌に掲載されているかを調査した．ここでは 2016 年 10 月〜2020 年 9 月末まで投稿分の 5 年分を対象として集計した．結果を表 2 に示す．

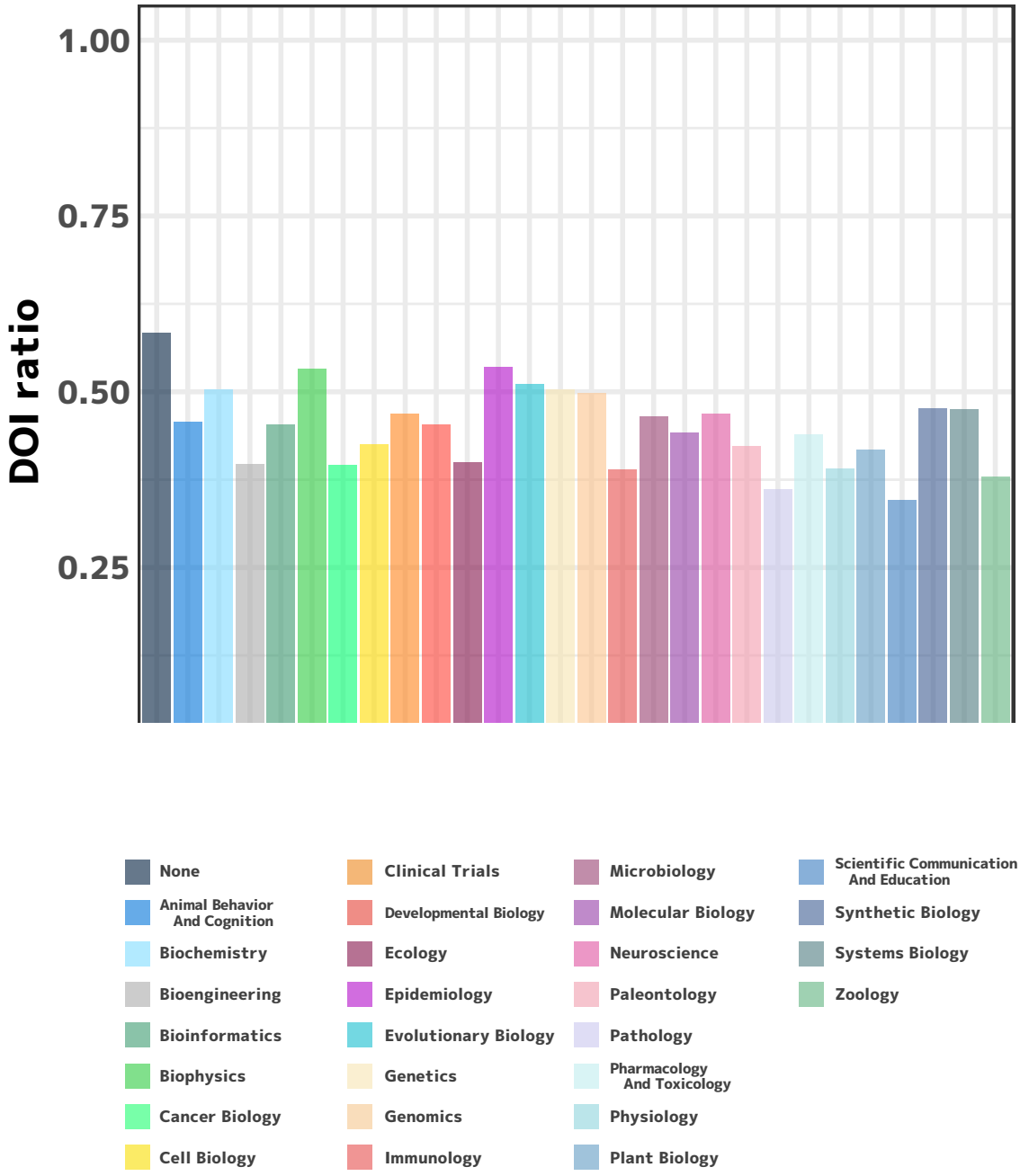


図 6　分野毎の DOI 付与率

表 2 の右列に印がないものは基本的にオープンアクセスの雑誌である．*1，2 も オープンアクセスとカウントすると，上位 20 件のほとんどがオープンアクセス誌と言える．

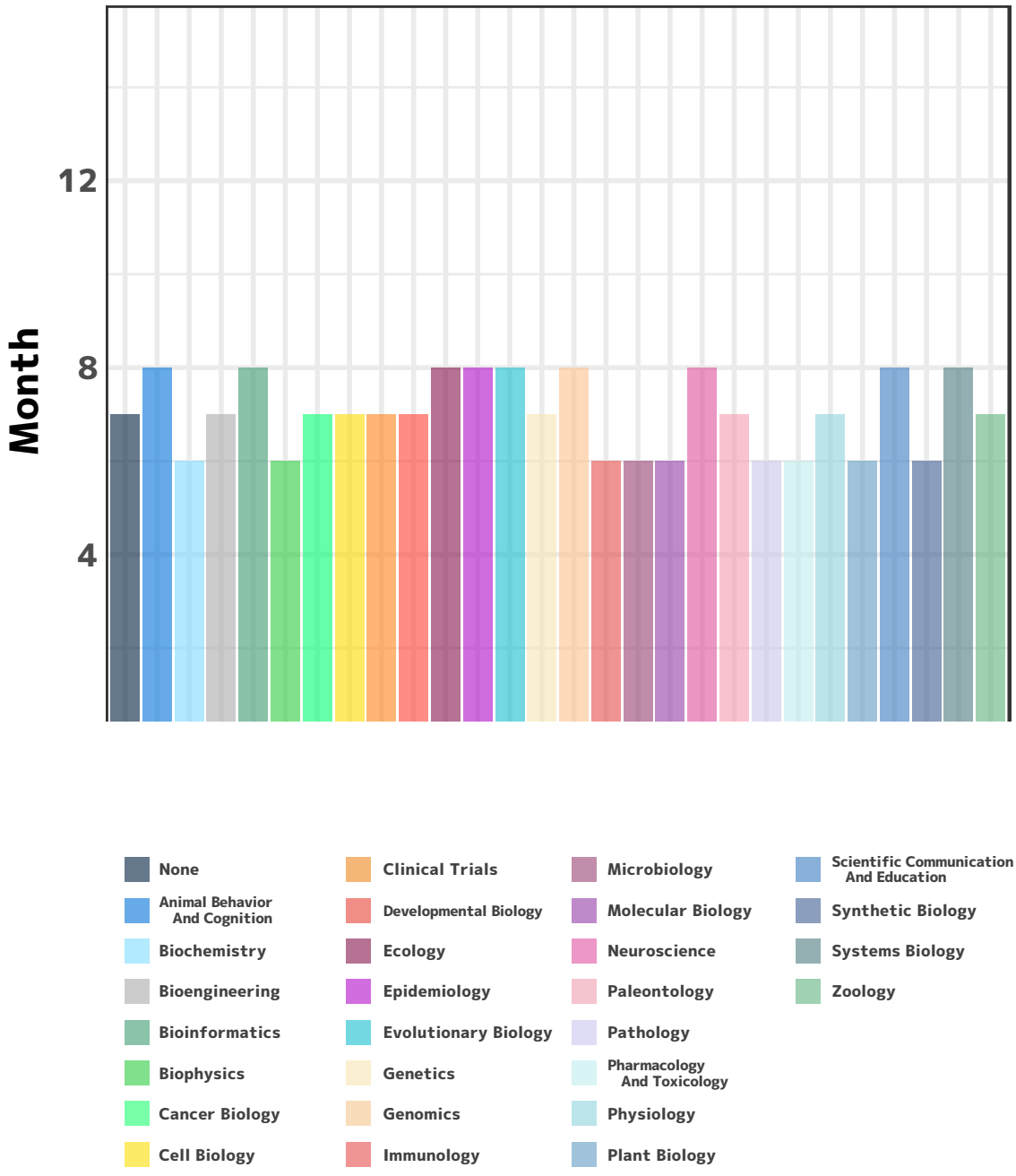


図 7　分野毎の DOI 付与までの期間

表 2　投稿先雑誌名 Top20

	Journal	Count	
1	PLoS ONE	2,613	
2	eLife	2,142	
3	Scientific Reports	2,072	
4	Nature Communications	1,727	
5	Proceedings of the National Academy of Sciences	1,130	*1
6	Bioinformatics	936	*2
7	PLoS Computational Biology	891	
8	PLoS Genetics	627	
9	Nucleic Acids Research	625	
10	NeuroImage	621	*3
11	G3 Genes|Genomes|Genetics	521	
12	The Journal of Neuroscience	464	*3
13	Cell Reports	452	
14	Genetics	437	*2
15	Genome Biology	383	
16	BMC Genomics	349	
17	Biophysical Journal	341	*2
18	BMC Bioinformatics	340	
19	Molecular Biology and Evolution	334	*3
20	mBio	324	

*1　6 ヶ月後にオープンアクセス

*2　ハイブリッド・オープンアクセス

*3　従来型のジャーナル誌

4.6　分野と被引用の関係性

ここまで bioRxiv に投稿された論文がどの程度の期間を経てどのような雑誌に掲載されているかを分析した.

続いて bioRxiv に掲載された論文が その後の論文にどのような影響を与えたか，具体的には bioRxiv に掲載された個々の論文がどの程度引用されたか，について見る．結果を図 8 に示す.

図 8 は 2016 年 10 月～2020 年 9 月末まで投稿分の 5 年間に bioRxiv に投稿された論文を対象に被引用情報を取得し，分野毎に平均引用回数を示したものである.

図 8 をみると，ここでは分野毎に異なる傾向も観察され，たとえば Genomics は比較的引用回数が多い傾向が読み取れる.

関連して分野に関係なく同期間での引用数 Top25 を表 3 に示す.

表 3　高被引用論文 Top25

	DOI	Post Date	Category	Title	Cite
1	10.1101/080333	2016-10-12	Neuroscience	Genetic, transcriptome, proteomic and epidemiologi...	741
2	10.1101/099192	2017-01-09	Genetics	Watching the clock for 25 years in FlyClockbase: V...	587
3	10.1101/203943	2017-10-16	Neuroscience	Degeneracy in hippocampal physiology and plasticit...	573
4	10.1101/535005	2019-02-01	Ecology	GIFT – A Global Inventory of Floras and Traits for...	562
5	10.1101/310763	2018-04-30	Epidemiology	MicroCOSM: a model of social and structural driver...	558
6	10.1101/2020.03.23.003384	2020-03-23	Genetics	Rat models of human diseases and related phenotype...	487
7	10.1101/2020.03.23.003392	2020-03-23	Genetics	Rat models of human diseases and related phenotype...	487
8	10.1101/833988	2019-11-07	Neuroscience	Arc Regulates Transcription of Genes for Plasticit...	485
9	10.1101/425488	2018-09-24	Ecology	Complex responses of global insect pests to climat...	453
10	10.1101/503334	2018-12-26	Ecology	Data paper: FoRAGE (Functional Responses from Arou...	445
11	10.1101/142760	2017-05-28	Bioinformatics	Opportunities and obstacles for deep learning in b...	432
12	10.1101/307652	2018-04-28	Neuroscience	Mapping molecular datasets back to the brain regio...	412
13	10.1101/405688	2018-08-31	Animal Behavior and Cognition	The evolution of infanticide by females in mammals	406
14	10.1101/152264	2017-06-22	Bioinformatics	Informatics for Cancer Immunotherapy	404
15	10.1101/2020.07.14.202085	2020-07-14	Neuroscience	10 years of EPOC: A scoping review of Emotiv's por...	398
16	10.1101/2020.03.15.992479	2020-03-17	Neuroscience	Williams syndrome and autism: dissimilar socio-cog...	382
17	10.1101/107268	2017-02-09	Evolutionary Biology	Between semelparity and iteroparity: empirical evi...	377
18	10.1101/2020.02.05.935189	2020-02-05	Developmental Biology	Foxg1 Organizes Cephalic Ectoderm to Repress Mandi...	372
19	10.1101/166785	2017-07-21	Neuroscience	Computational Foundations of Natural Intelligence	364
20	10.1101/2020.06.04.133199	2020-06-05	Cancer Biology	Use of signals of positive and negative selection ...	353
21	10.1101/142034	2017-05-28	Bioinformatics	Reverse-engineering biological networks from large...	352
22	10.1101/203307	2017-10-14	Evolutionary Biology	The genomics of local Adaptation in trees: Are we ...	349
23	10.1101/237586	2017-12-21	Genomics	Firefly genomes illuminate the origin and evolutio...	343
24	10.1101/577320	2019-03-14	Cell Biology	The model of local axon homeostasis - explaining t...	342
25	10.1101/583625	2019-03-21	Ecology	Functional diversification enabled grassy biomes t...	334

表 3 をみると上位は Neuroscience, Genetics など特定分野に偏っている傾向が見られる.

次に被引用数 0 件から 100 件までの頻度分布を図 9 に示す.

計量書誌学では「ロトカの法則」など「べき分布」に従うようなデータが見られるが，ここでも「べき分布」に類する分布形状が観察でき，多くの論文は被引用が 0 件である一方表 3 に示したように多くの引用を稼ぐ論文も存在することが分かる.

この引用頻度と分野の関係について，別途図 10 にまとめた.

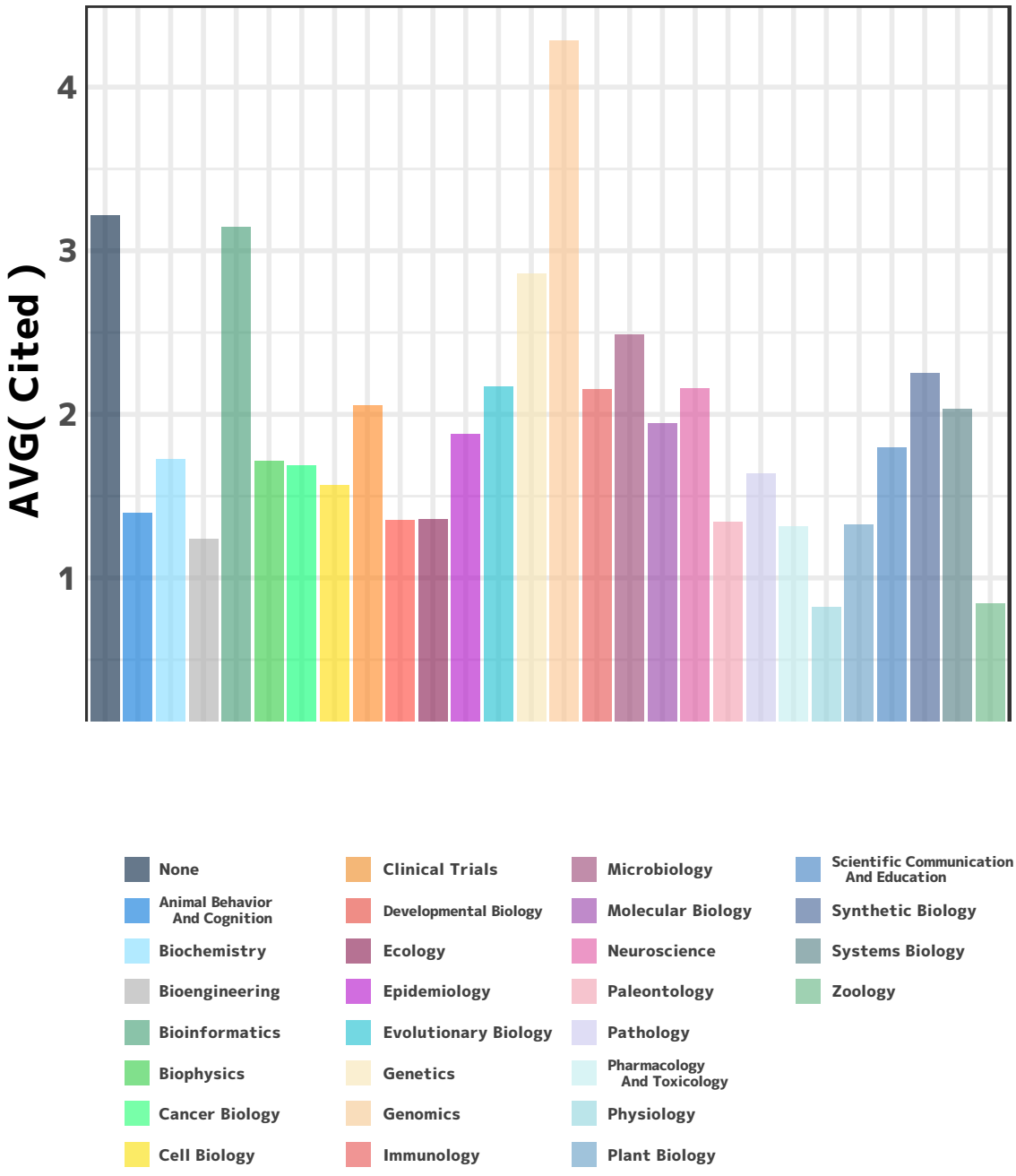


図 8　分野と被引用の関係性

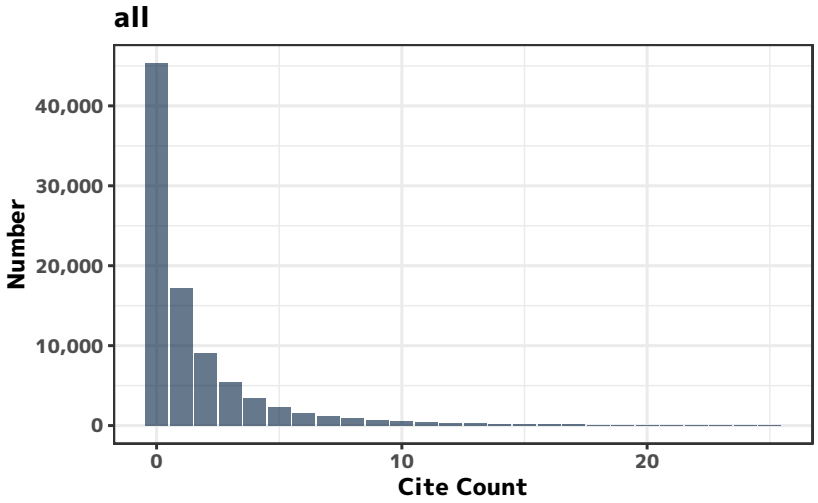


図 9　被引用件数と頻度

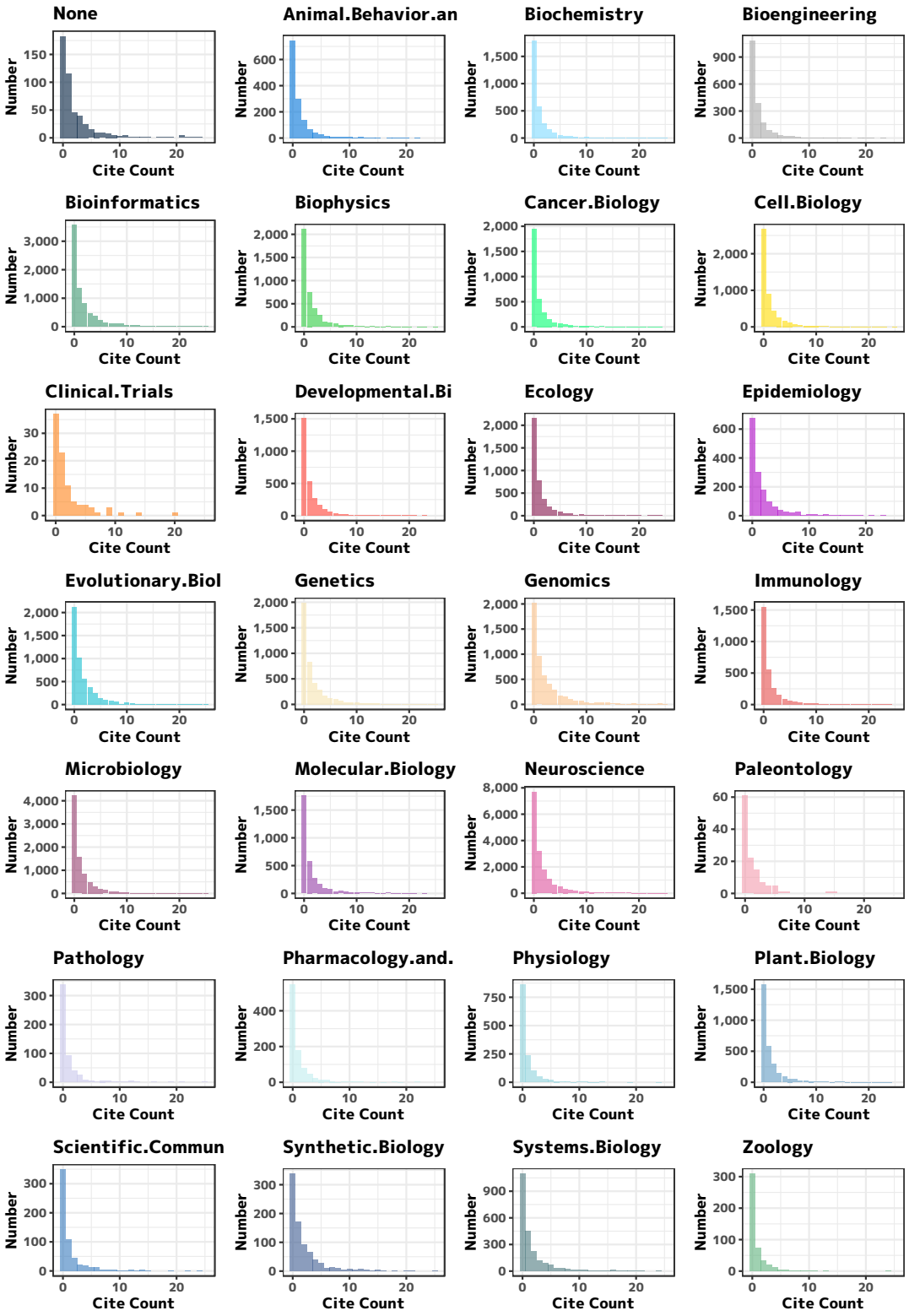


図 10　分野と被引用の関係性

4.7　国・地域との関係性

bioRxiv の書誌情報も arXiv と同様に投稿者所属機関の国・地域情報は必ずしも整備されていない．より正確には国・地域の入力欄はあり，それらのデータも取得可能だが，記載されていないケースも多く表記揺れも多い（例えば，米国について US,USA,United States, などの記載があったり，CA, IL, NY, New York の様に州名を書いているなど）ため使いづらい．

そこで，文献 [4] と同様に著者の連絡先メールアドレスを基に国・地域を推定し，国・地域との関係性分析を試みた．

ここでは文献 [4] と同様，筆頭メールアドレス（最初に出現するメールアドレス．基本的には 第 1 著者 や Corresponding Author に該当するが，第 1 著者がメールアドレスを記入していなく，第 2 著者以降の著者がメールアドレスを記入しているケースもあり，必ずしも第 1 著者/Corresponding Author に限らない.）をベースに，そのトップレベルドメインや管理者所属国籍で振り分けを行っている．Gmail や Hotmail などは不明 (Unknown) に分類し，その他の.edu, .com, .org などは管理者所属国籍を調べて割り当てる（たとえば，126.com は China，london.edu は UK など）．

結果を表 4 に示す．

表 4　国・地域の分布（上位 40 件）

	Region	Count		Region	Count
1	US	37,678	21	Korea	535
2	Unknown	27,217	22	Finland	525
3	UK	9,226	23	Austria	503
4	Germany	5,828	24	Singapore	460
5	China	4,843	25	Portugal	344
6	France	3,976	26	New Zealand	321
7	Canada	3,601	27	Poland	314
8	Australia	2,508	28	Mexico	292
9	**Japan**	**2,218**	29	Taiwan	285
10	Switzerland	2,106	30	Argentina	283
11	Netherlands	1,842	31	Russia	272
12	India	1,751	32	Czechia	268
13	Spain	1,305	33	Hong Kong	261
14	Sweden	1,242	34	EU	258
15	Italy	1,103	35	Ireland	246
16	Israel	965	36	Hungary	182
17	Brazil	859	37	Chile	161
18	Denmark	822	38	Turkey	158
19	Belgium	750	39	Iran	148
20	Norway	583	40	South Africa	135

表 4 の通り，不明（Unknown）が全体で 2 位とそれなりのボリュームを占めており，分析等に際

して留意を要する[13)]．

これらの地域ごとの投稿数に関する時系列での変化を表 5,6 に示す．

また，上位 15 地域の全期間における分野比率について図 11 から 25 に示す．

加えて，上位 15 地域の分野構成比を基に，多次元尺度法で 2 次元でまとめたものを図 26 に示す．図は無次元で距離が類似度に相当し，近くにあるものは類似する分野構成を，遠くにあるものは異なる分野構成を有すると言える．

図 26 からは，上位 15 の国・地域において，中国とインド，日本，イタリア，不明 (Unknown) は他とは異なる分野構成比を有しそうなことが伺える．

たとえば，中国とインドは Neuroscience の割合が特に小さい．日本は Cell Biology の割合が大きい．イタリアは Neuroscience の割合が特に大きい．

このように，bioRxiv 内の分野においても，国・地域ごとに特色が分かれていそうなことが観察された．

[13)] もともとの国名記入欄ベースでも空欄が 23,991 件存在し，2 位を占める．これを見る限り国名ベースの方がわずかながらカバー率が広い．ただし，前述通り国名でないものが記入されている場合などもある他，例えば文献 [4] などの既報と比較する際に同じ手法で処理しておく方が比較容易性の面で望ましく，ここでも既報と同じ手法を採用した．なお，国名ベースでかつ簡易に名寄せした場合，1 位は米国で約 2.9 万，3 位は英国で約 1 万，4 位はドイツで約 7 千である．

表 5　国・地域の投稿件数推移（上位 10 件，2014-2017 年）

Y	M	Total	US	Unknown	UK	Germany	China	France	Canada	Australia	Japan	Switzerland
2014	1	40	1	29	2	0	2	0	1	0	0	2
2014	2	49	12	27	2	0	0	1	1	0	0	1
2014	3	44	13	22	2	0	0	0	2	2	1	1
2014	4	55	11	30	4	4	3	2	0	0	0	0
2014	5	61	18	34	4	1	0	0	1	0	1	0
2014	6	71	15	39	5	0	2	0	3	1	0	0
2014	7	55	10	31	3	2	0	0	2	1	0	1
2014	8	75	17	37	4	1	1	1	0	2	0	1
2014	9	78	21	43	6	2	1	0	1	0	0	2
2014	10	82	11	54	5	3	1	1	0	0	0	2
2014	11	71	12	45	2	2	1	2	2	1	1	1
2014	12	81	17	44	8	2	1	1	0	2	1	0
2015	1	95	20	51	5	4	0	0	0	2	0	1
2015	2	83	7	48	8	3	1	1	3	2	1	0
2015	3	109	19	63	9	4	1	1	1	1	0	0
2015	4	111	17	65	8	5	0	2	2	1	0	2
2015	5	97	22	55	1	4	0	2	2	2	0	1
2015	6	112	20	68	5	2	1	0	3	1	0	1
2015	7	138	34	71	9	3	4	3	5	0	0	1
2015	8	134	34	73	4	4	5	0	2	2	3	2
2015	9	135	26	74	8	6	4	2	3	1	1	0
2015	10	177	54	77	9	6	1	2	3	1	2	3
2015	11	186	41	88	8	4	7	6	7	1	2	3
2015	12	156	27	80	13	4	2	6	2	0	1	5
2016	1	165	34	91	6	8	4	2	3	1	1	1
2016	2	231	45	113	15	8	3	6	6	3	1	3
2016	3	326	76	164	20	11	5	4	13	6	0	2
2016	4	295	78	135	21	7	2	9	8	3	6	5
2016	5	348	93	172	19	5	7	7	9	2	2	2
2016	6	325	78	176	18	8	4	4	8	3	0	2
2016	7	319	71	161	22	6	2	6	6	8	0	5
2016	8	394	99	195	24	7	6	4	7	8	2	1
2016	9	393	78	205	29	13	6	9	5	7	1	9
2016	10	395	90	203	18	11	7	8	7	7	2	2
2016	11	411	77	204	19	18	8	8	13	10	5	7
2016	12	471	89	242	26	14	9	8	16	11	4	4
2017	1	517	110	246	36	14	9	14	12	6	3	6
2017	2	522	112	248	39	15	12	10	11	5	5	5
2017	3	698	147	348	44	24	15	11	22	5	7	13
2017	4	663	155	329	34	28	19	10	14	7	6	7
2017	5	807	200	375	57	26	17	15	15	15	7	10
2017	6	951	253	452	56	22	23	15	14	9	6	12
2017	7	886	210	425	61	23	19	23	27	13	7	7
2017	8	904	222	426	64	18	23	15	20	18	9	9
2017	9	985	244	490	58	18	10	17	21	6	9	13
2017	10	1,076	263	504	64	37	22	21	24	15	12	9
2017	11	1,113	238	541	85	34	18	20	18	15	11	18
2017	12	987	205	455	75	30	18	28	18	18	15	11

表 6　国・地域の投稿件数推移（上位 10 件，2018-2021 年）

Y	M	Total	US	Unknown	UK	Germany	China	France	Canada	Australia	Japan	Switzerland
2018	1	1,172	298	524	73	44	31	28	16	17	9	18
2018	2	1,126	257	518	70	32	27	28	25	24	11	15
2018	3	1,380	423	376	113	62	45	42	41	36	16	28
2018	4	1,450	346	659	97	35	43	49	28	22	16	18
2018	5	1,799	465	723	98	59	63	39	45	29	36	22
2018	6	1,748	390	805	108	55	71	33	36	22	25	20
2018	7	1,601	409	664	96	57	63	46	29	21	15	19
2018	8	1,744	369	753	132	62	70	19	42	21	29	23
2018	9	1,747	427	763	91	50	59	44	27	29	18	34
2018	10	1,952	448	870	126	69	54	52	35	34	33	24
2018	11	1,841	410	822	109	73	61	48	37	24	19	28
2018	12	1,860	396	812	126	56	71	44	30	33	20	29
2019	1	2,079	483	914	109	66	79	51	47	31	24	24
2019	2	1,888	440	813	107	78	62	44	40	27	27	30
2019	3	2,265	518	884	184	92	60	65	55	38	33	37
2019	4	2,157	765	317	228	124	114	68	76	38	47	37
2019	5	2,397	919	238	227	142	78	83	92	64	67	35
2019	6	2,287	924	263	208	125	96	77	64	44	34	46
2019	7	2,519	908	266	244	149	114	113	96	60	49	57
2019	8	2,368	940	248	189	125	113	81	80	68	47	49
2019	9	2,550	971	248	234	126	115	105	105	69	42	44
2019	10	2,793	1,100	273	260	180	112	117	98	52	52	58
2019	11	2,372	843	270	245	151	108	98	84	62	43	40
2019	12	2,282	826	265	188	142	114	115	77	50	49	57
2020	1	2,628	989	278	204	144	158	126	84	66	50	52
2020	2	2,717	963	295	245	168	164	83	109	80	48	62
2020	3	3,078	1,083	305	210	186	226	132	119	77	74	61
2020	4	3,365	1,194	351	311	188	211	131	118	67	80	57
2020	5	3,903	1,500	362	375	200	167	162	120	126	112	74
2020	6	3,732	1,397	396	355	217	147	136	117	100	74	77
2020	7	3,771	1,332	395	326	239	194	153	124	82	102	91
2020	8	3,297	1,277	349	269	192	156	88	122	93	81	59
2020	9	3,412	1,304	343	288	203	164	118	134	93	83	53
2020	10	3,450	1,300	395	254	189	156	133	123	94	86	67
2020	11	3,301	1,093	404	279	192	158	157	128	82	86	72
2020	12	3,438	1,190	410	272	226	162	143	116	81	85	72
2021	1	3,580	1,374	394	267	187	162	141	128	74	83	82
2021	2	3,767	1,343	387	284	246	171	147	152	101	96	80
2021	3	4,401	1,486	456	385	306	206	213	156	105	116	114
Total		115,694	36,846	26,956	9,030	5,708	4,731	3,859	3,518	2,460	2,152	2,058

* Total は　表 5,6 に示した 2014 年 1 月頭から 2021 年 3 月末までの範囲での総計.
従って表 4 より少ない値を示す.

US

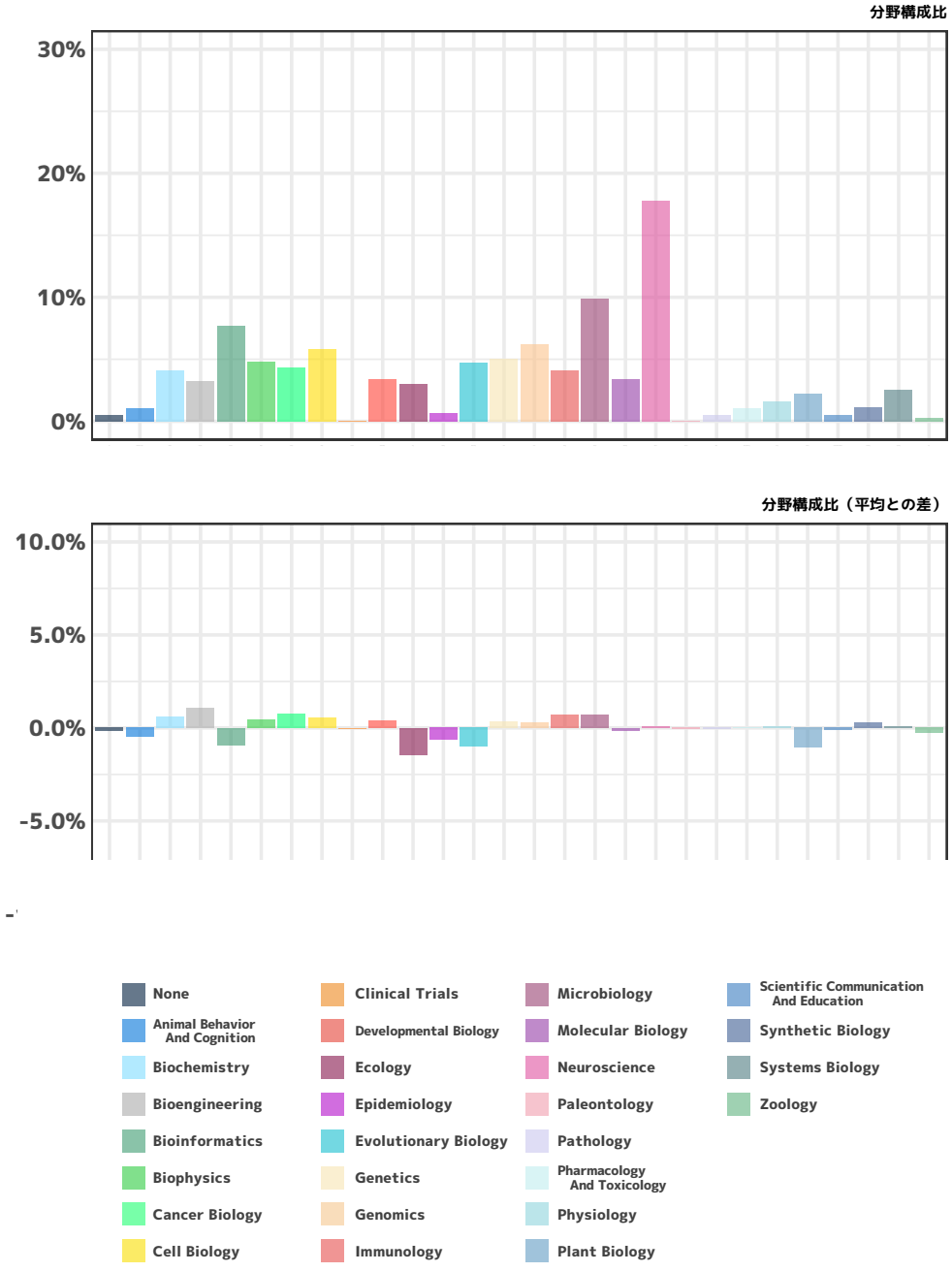


図 11　分野比率（US）

Unknown

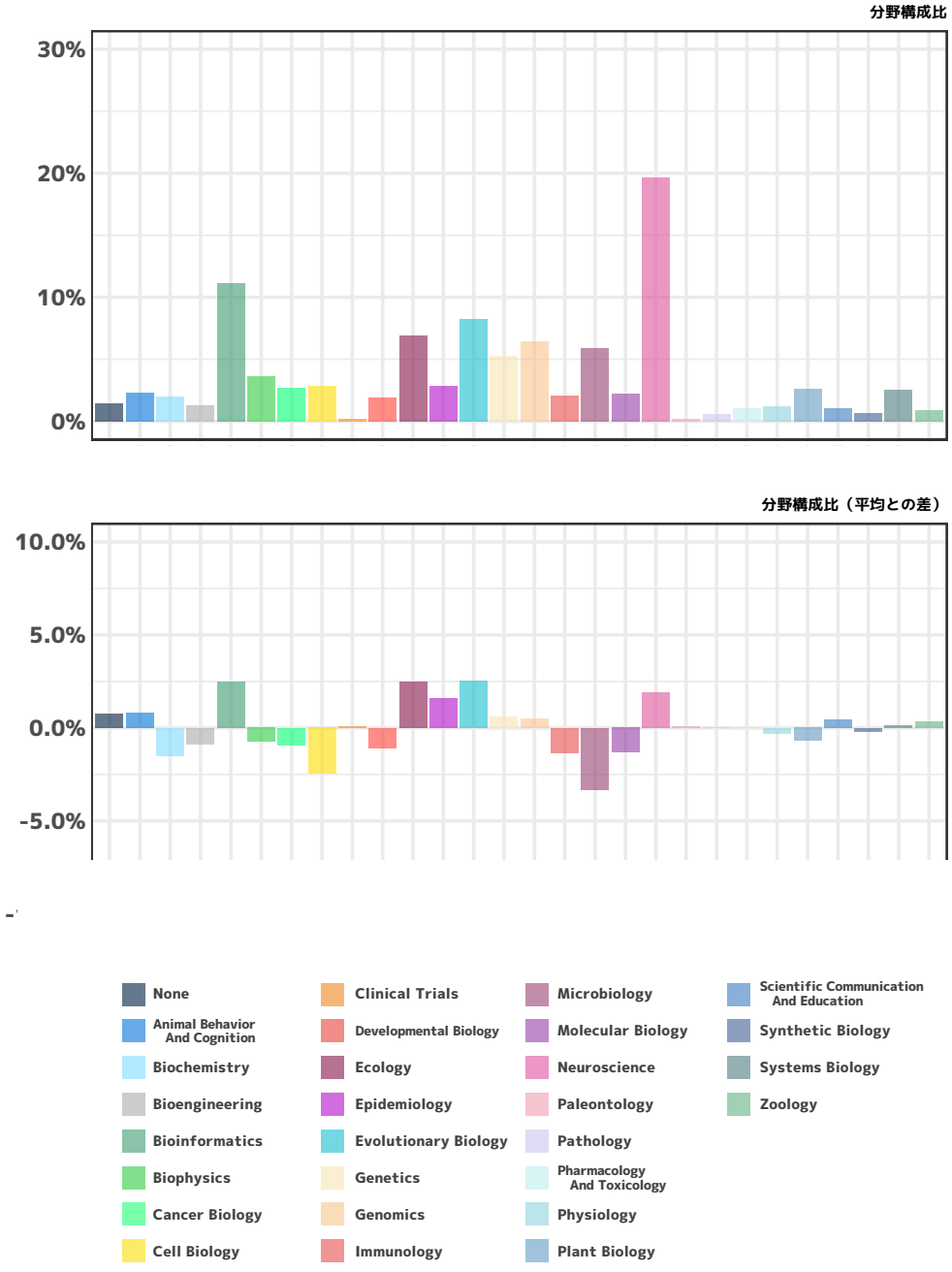


図 12　分野比率（Unknown）

UK

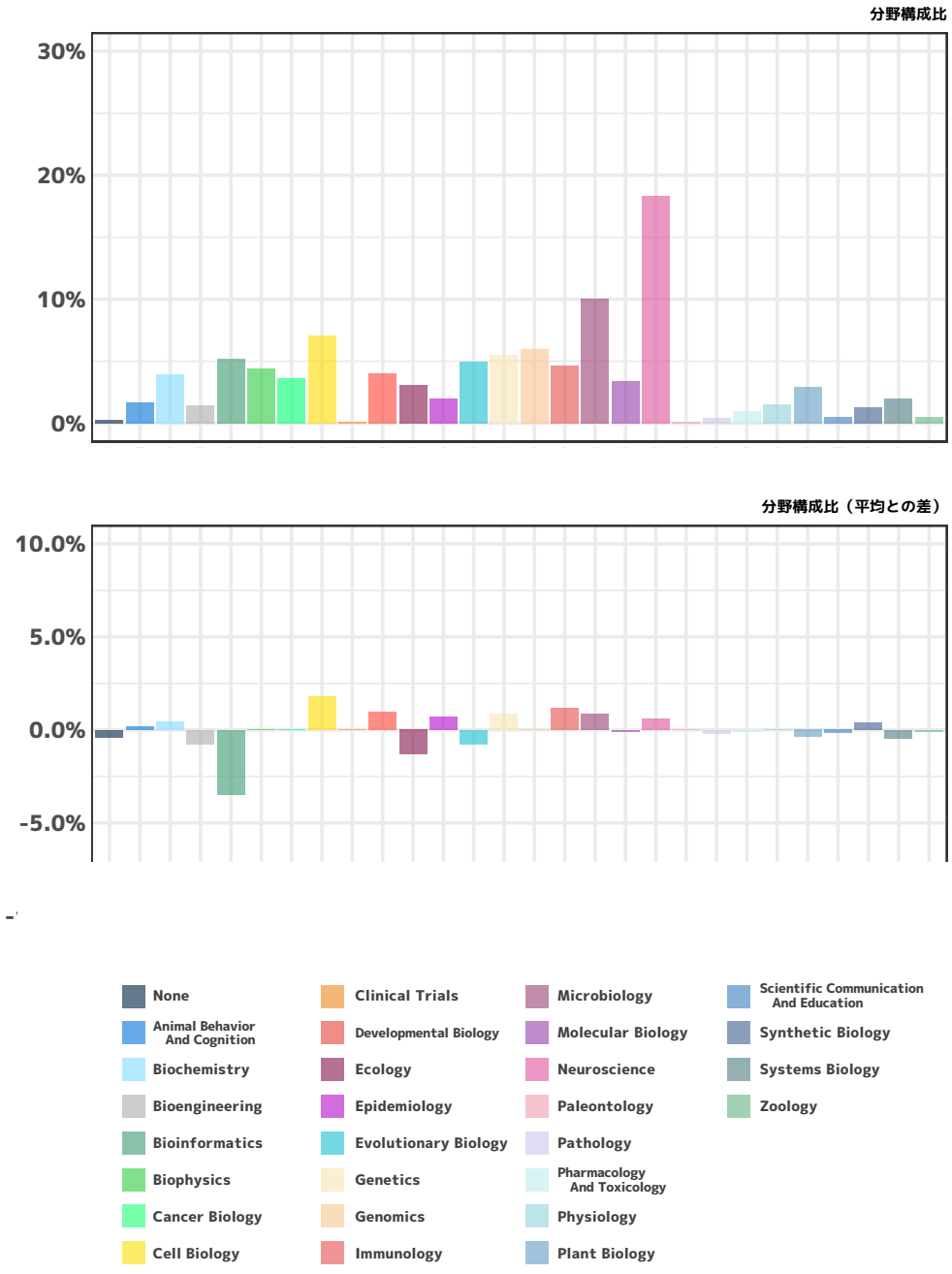


図 13　分野比率（UK）

Germany

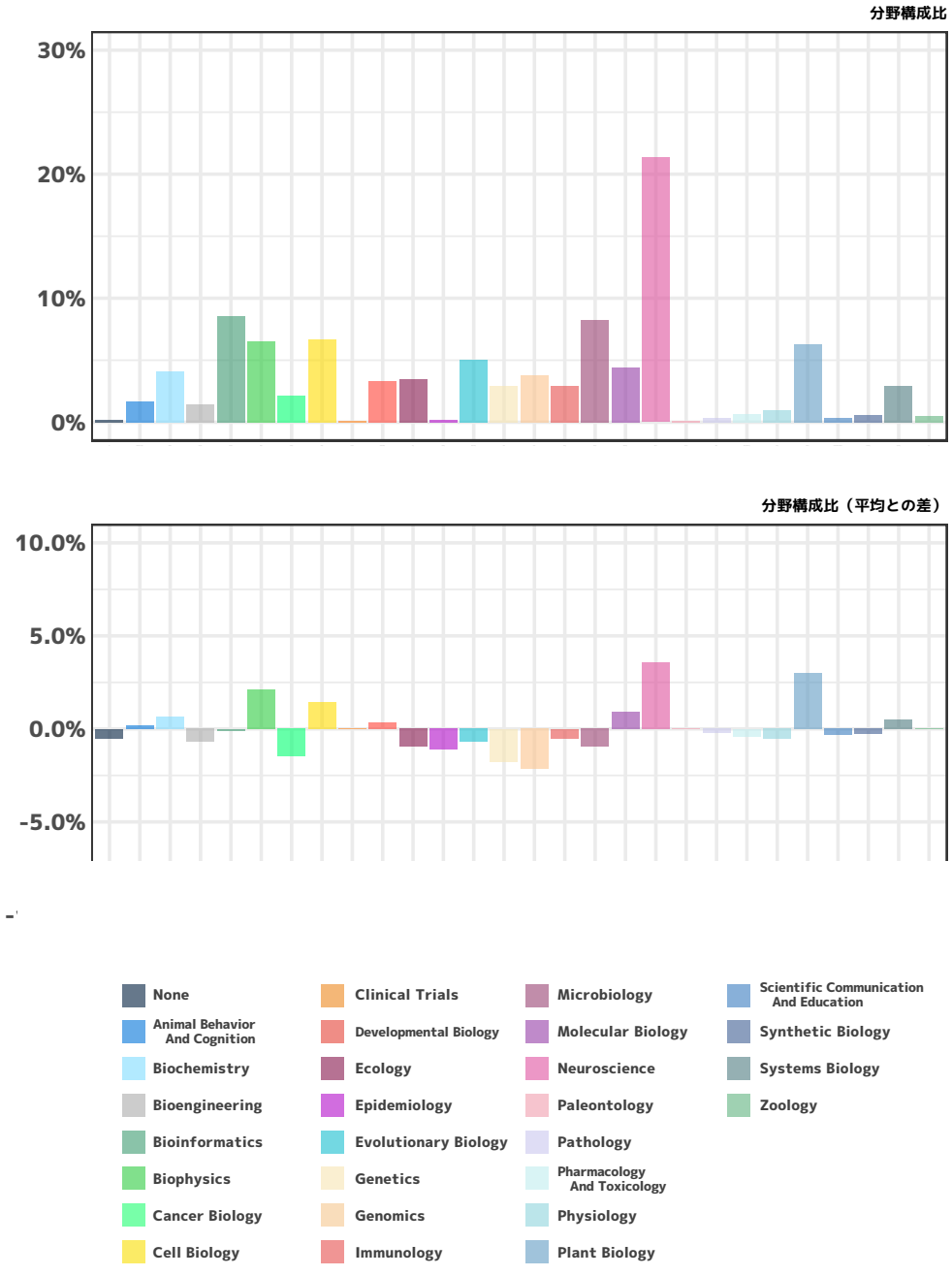


図 14　分野比率（Germany）

China

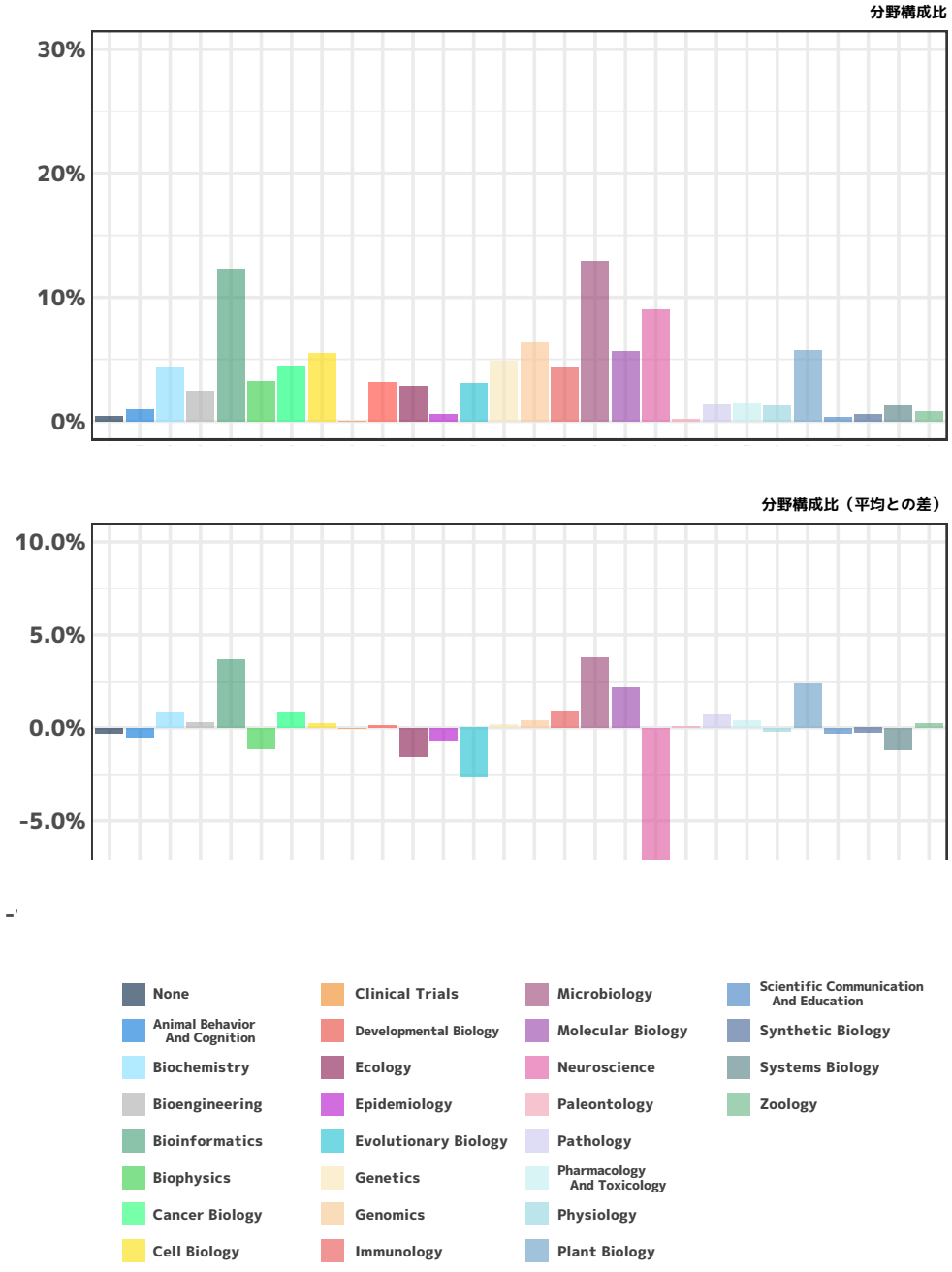


図 15　分野比率（China）

France

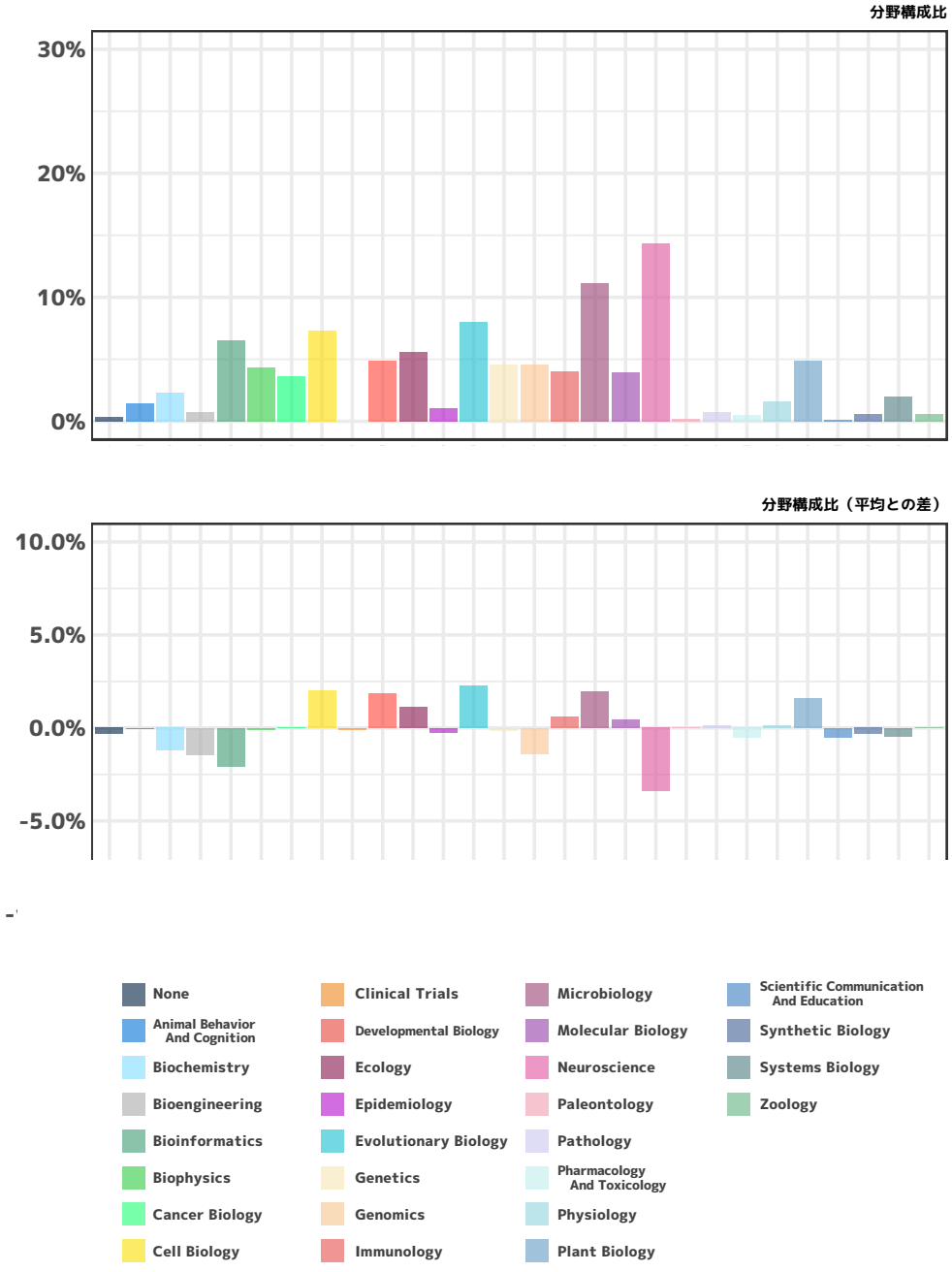


図 16　分野比率（France）

Canada

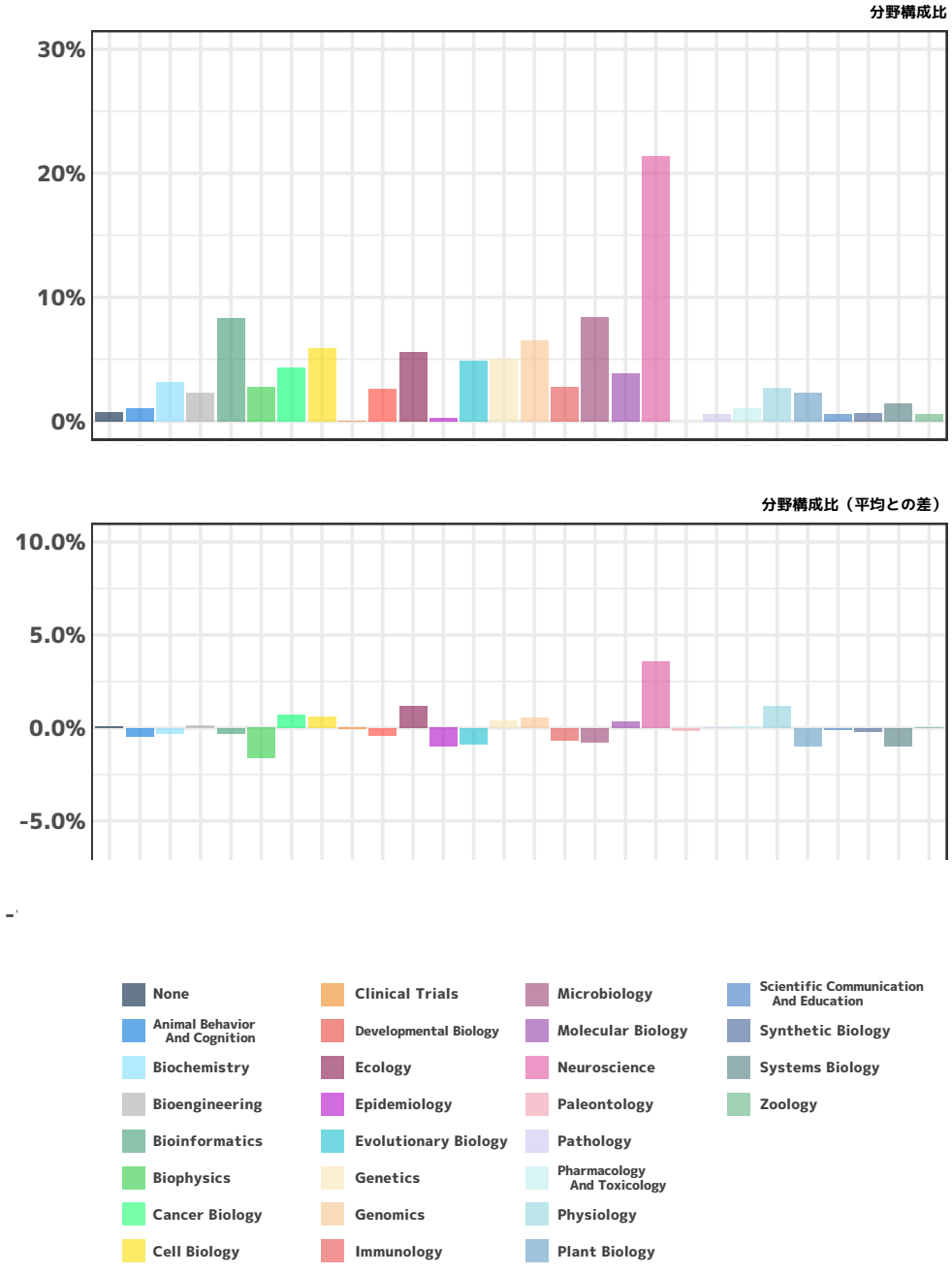


図 17　分野比率（Canada）

Australia

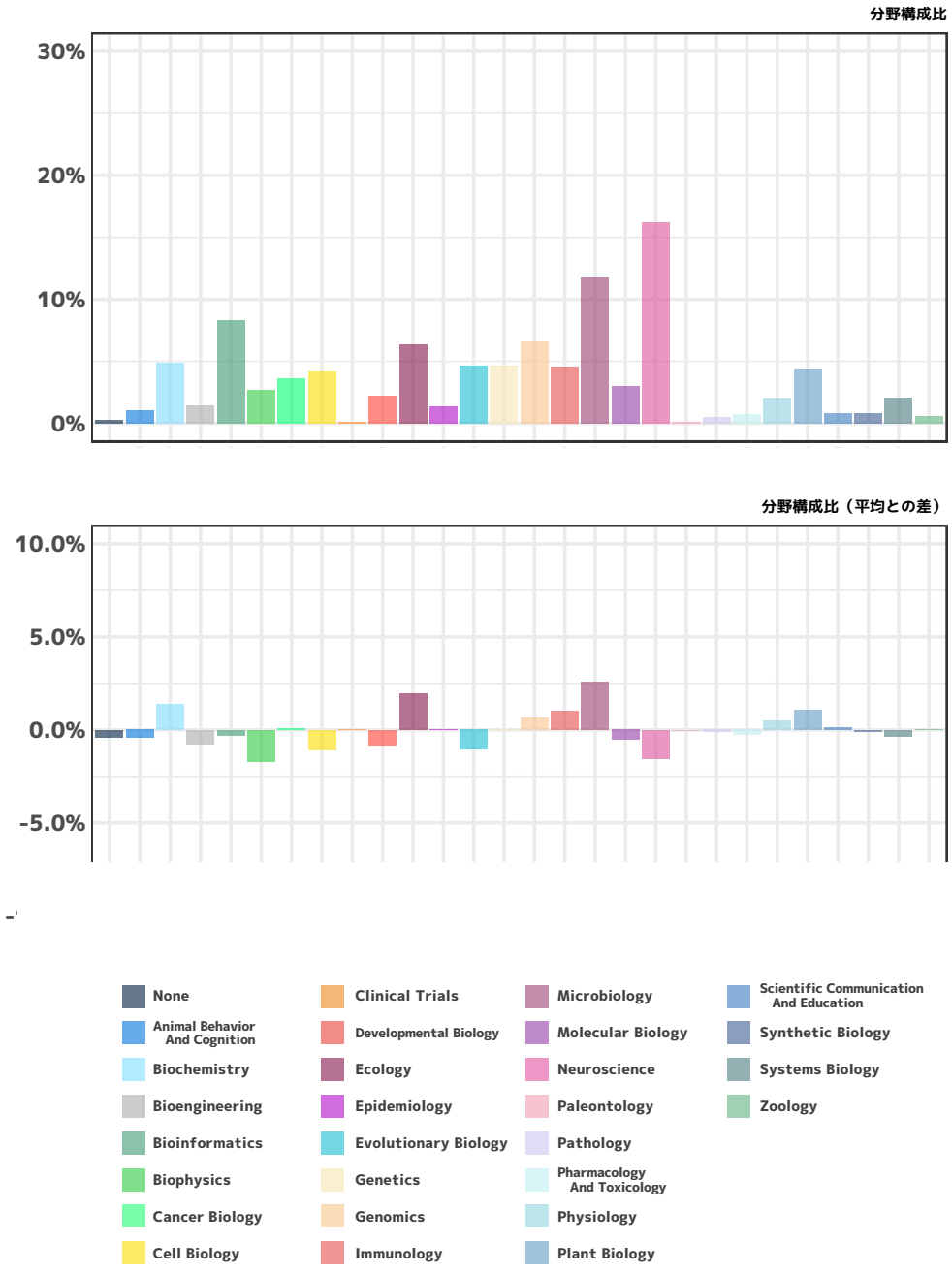


図 18　分野比率（Australia）

Japan

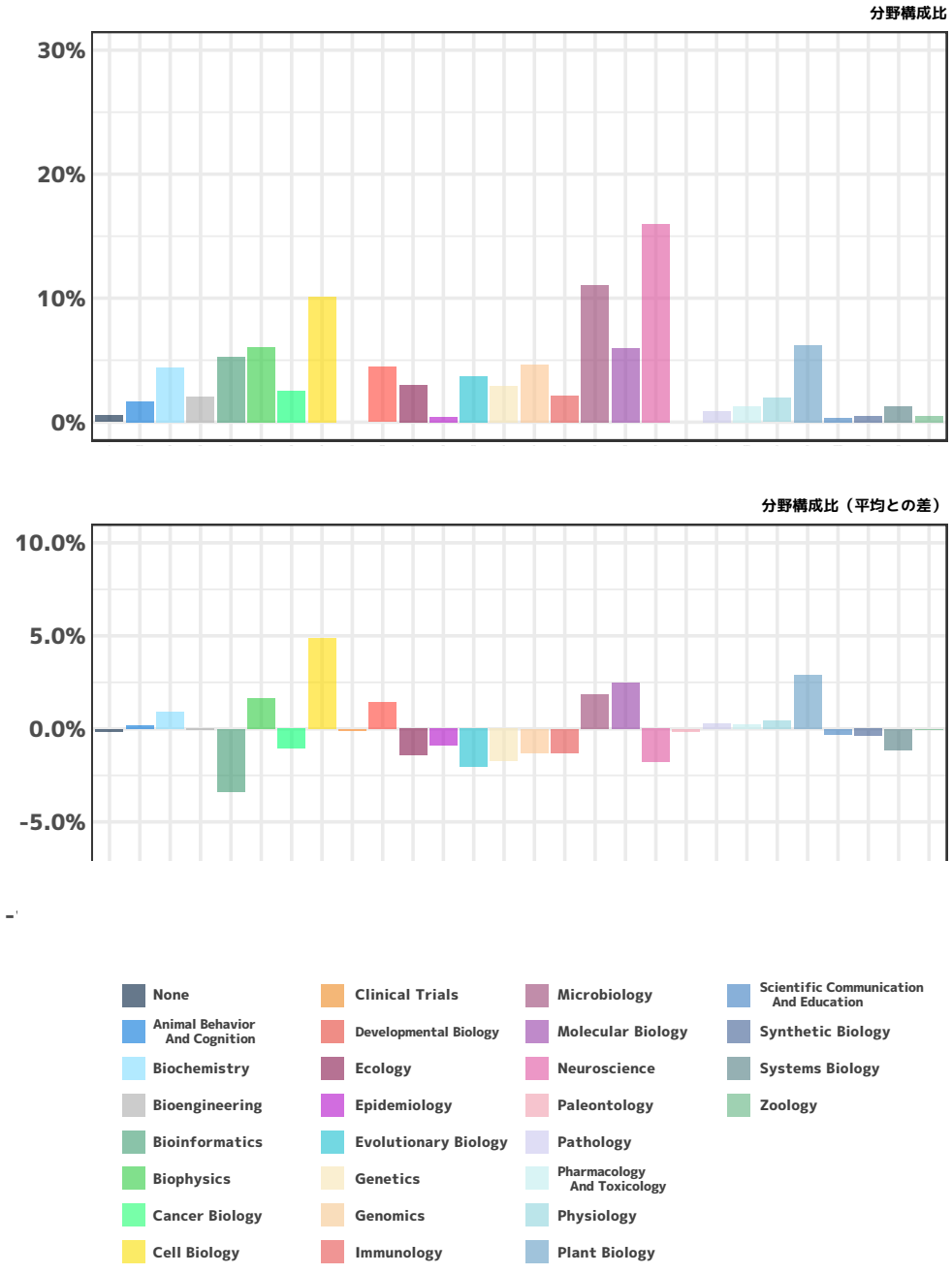


図 19　分野比率（Japan）

Switzerland

図 20　分野比率（Switzerland）

Netherlands

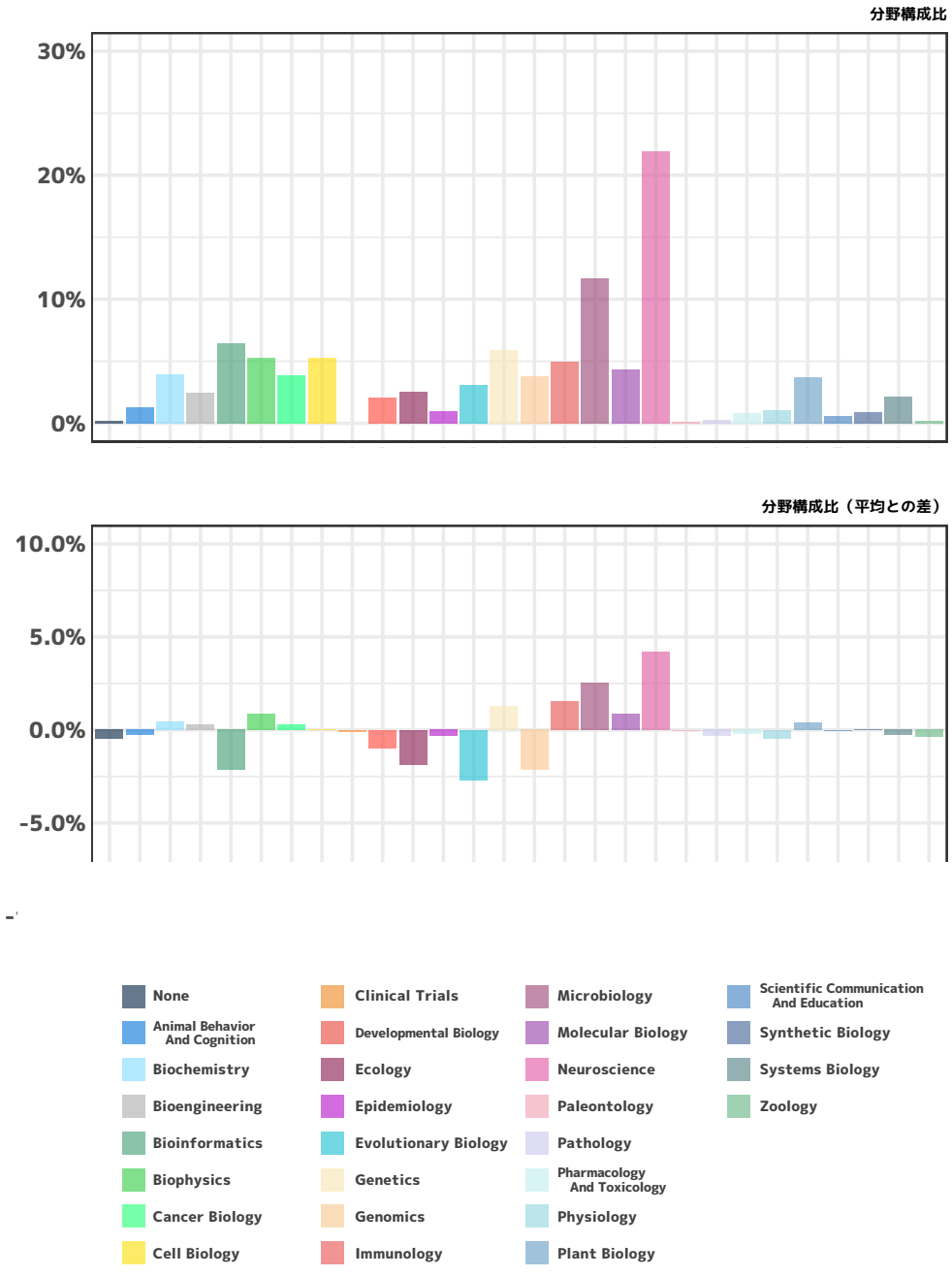


図 21　分野比率（Netherlands）

India

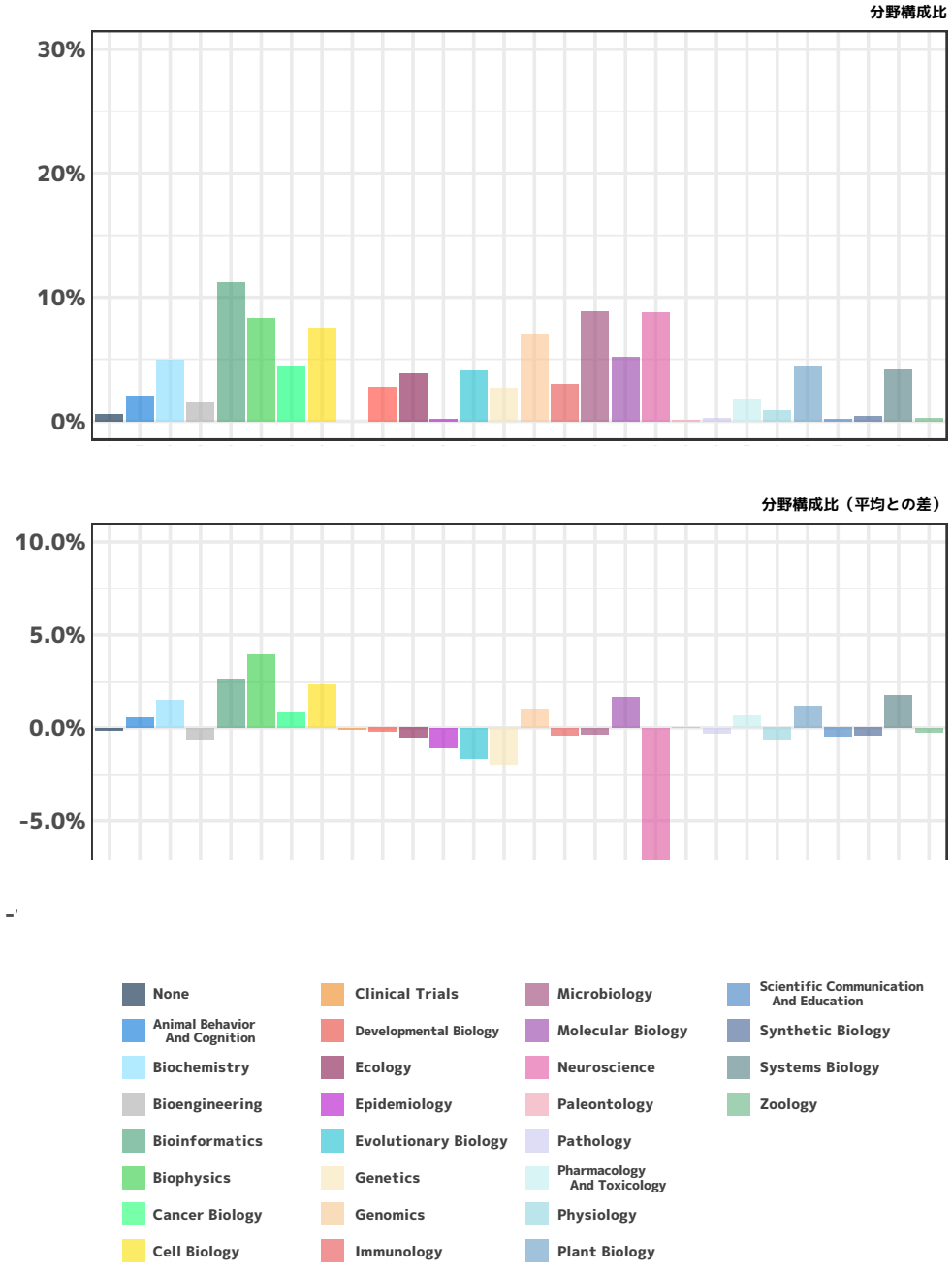


図 22　分野比率（India）

Spain

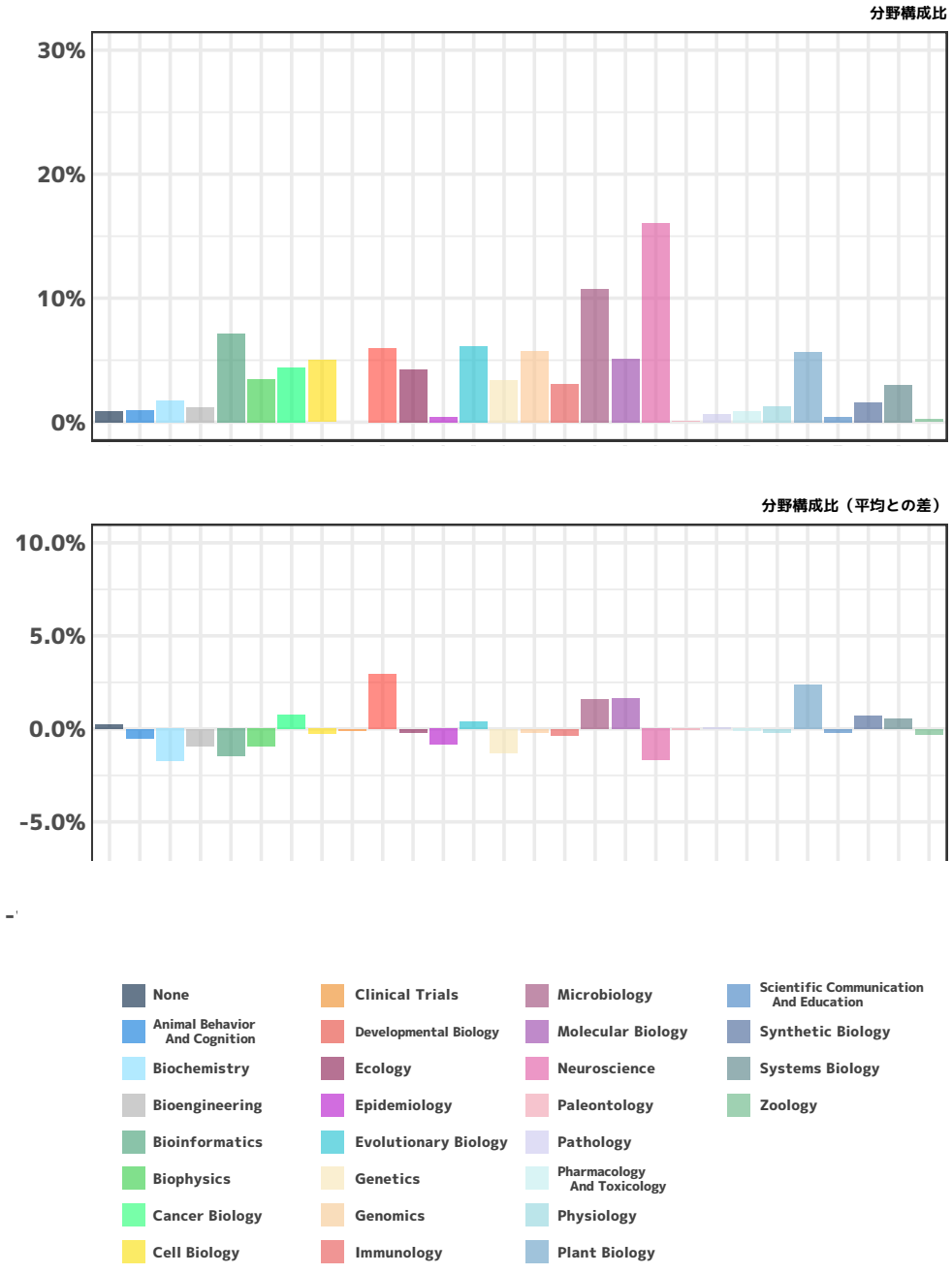


図 23　分野比率（Spain）

Sweden

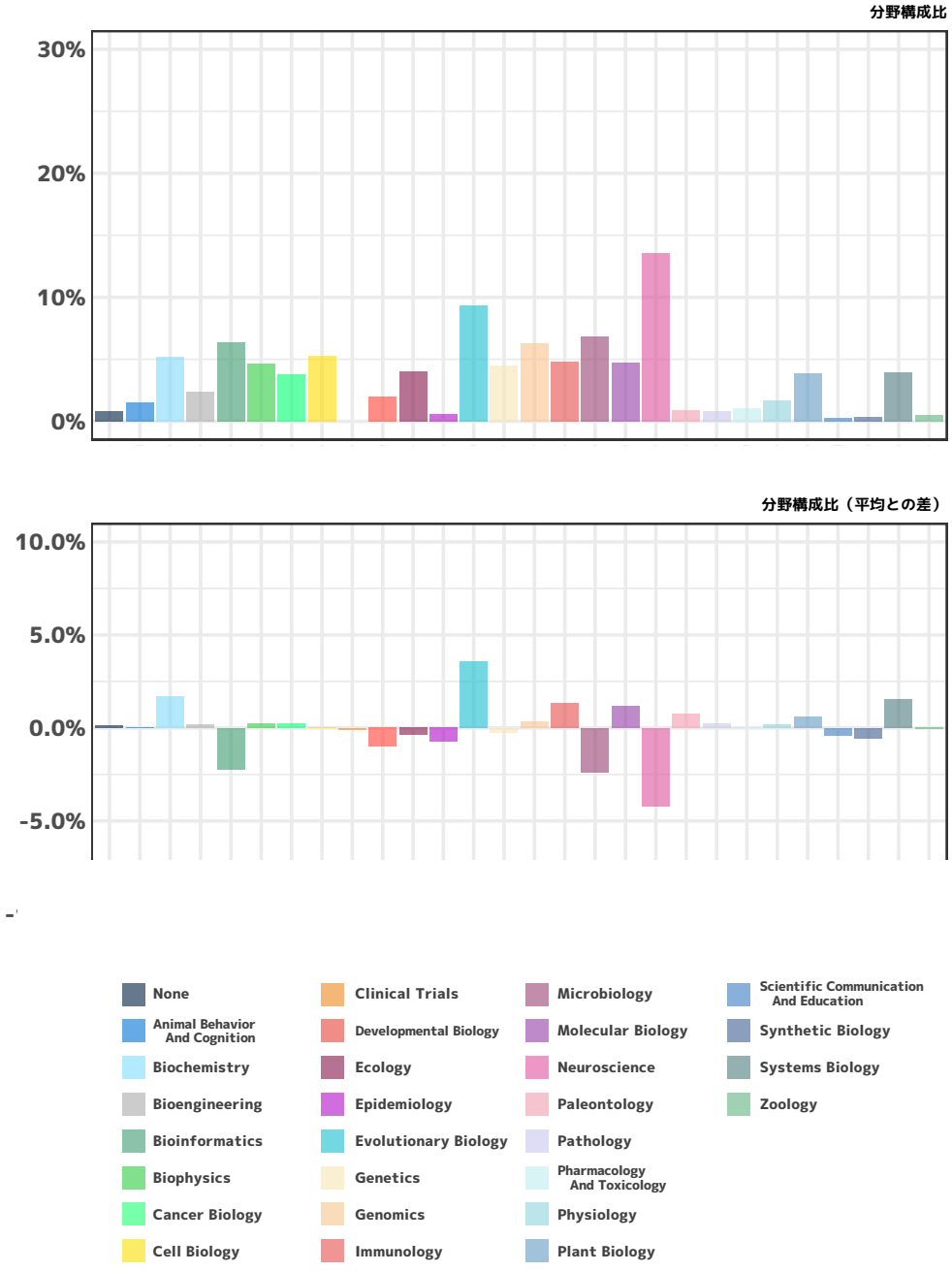


図 24　分野比率（Sweden）

Italy

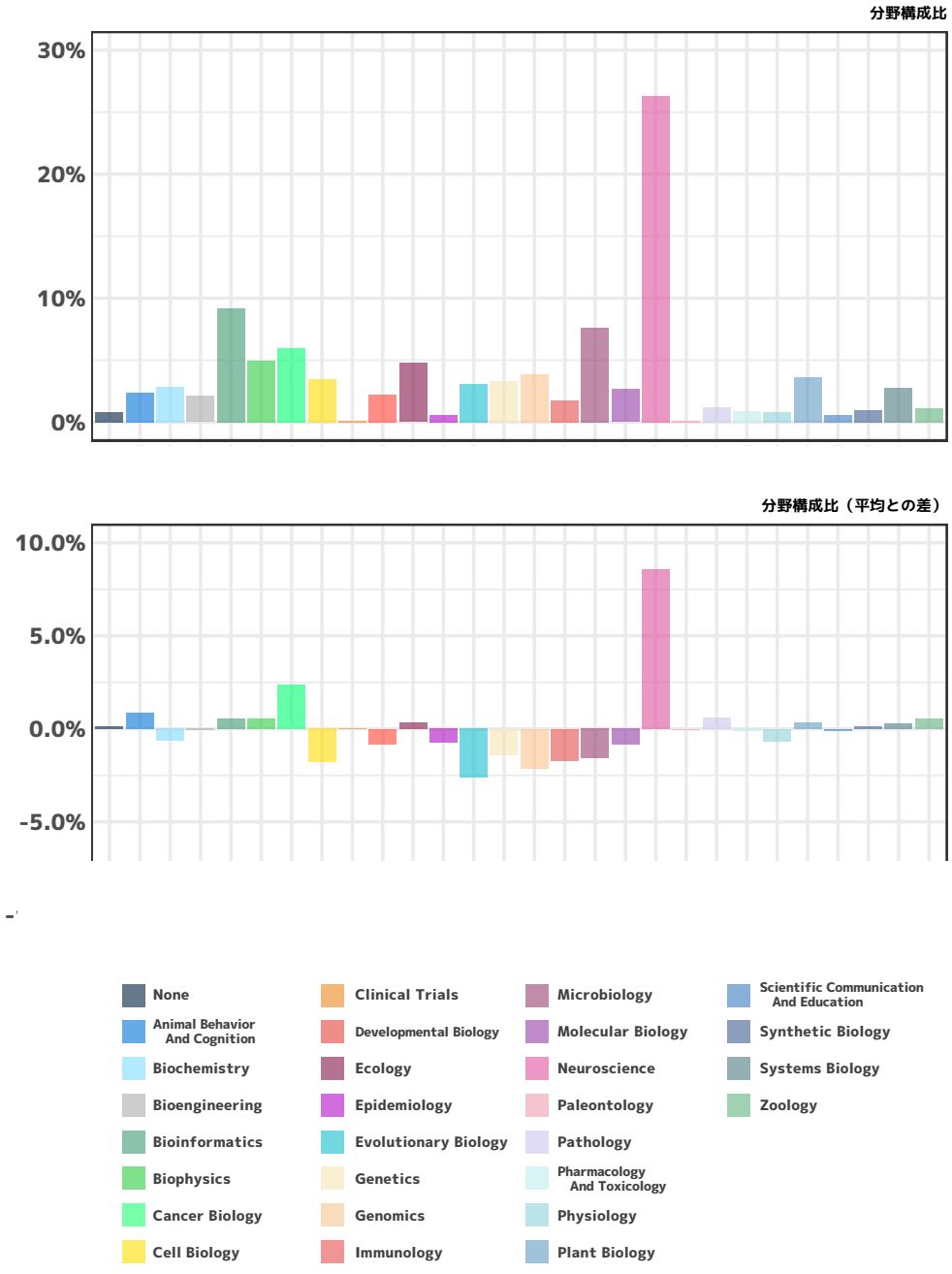


図 25　分野比率（Italy）

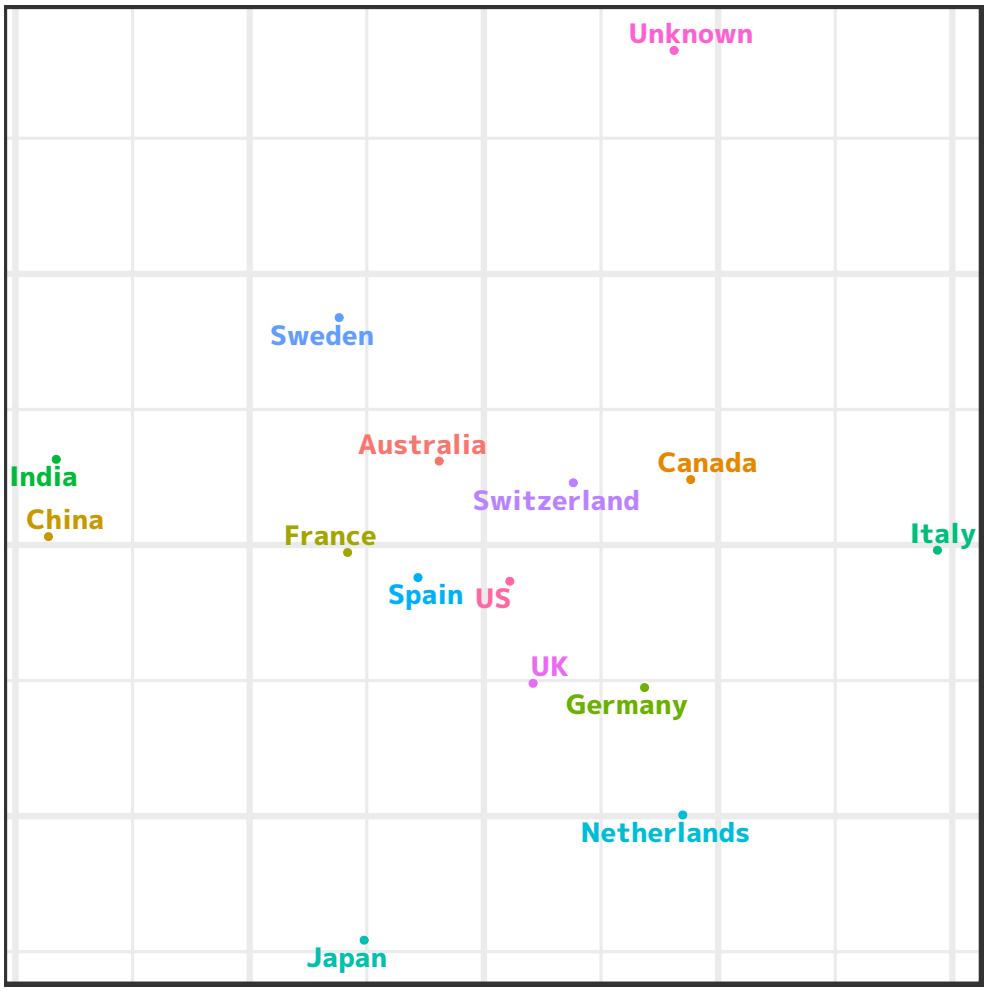


図 26　分野比率に基づく国・地域の類似性

4.8　COVID-19 の影響

既報 [4] の通り，2019 年末に端を発して 2020 年から世界的に流行した COVID-19 は学術にも大きな影響を及ぼした．この中で迅速性の観点等からプレプリントの積極的活用も観察され，bioRxiv もその一翼を担っている．

そこで，図 27 に 2020 年 1 月から 2021 年 3 月における，bioRxiv への COVID-19 関連投稿の数を示した．COVID-19 関連投稿数は既報 [4] と同様に，bioRxiv が提供するリストに掲載された記事件数である．

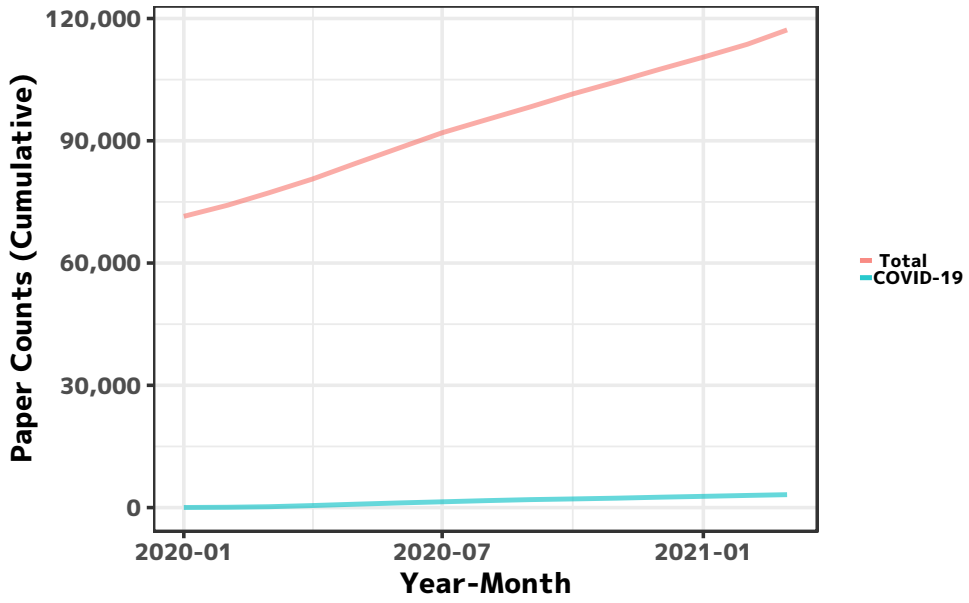


図 27　bioRxiv 全体に占める COVID-19 関連投稿の数（累積）

2020 年 3 月末までに累積で 3 千件を超える COVID-19 関連投稿がなされているが，期間中の投稿数はそれを大きく上回っており，COVID-19 関連投稿が全体に対して大きな影響を及ぼしているようには見受けられない．

参考までに同手法・同期間で医療系のプレプリントである medRxiv の COVID-19 関連投稿をまとめたものを図 28 に示す．

medRxiv は 2019 年にサービスを開始と新しく，かつ COVID-19 に直接関わる医療分野と言うこともあって COVID-19 の影響が大きいことが分かる．

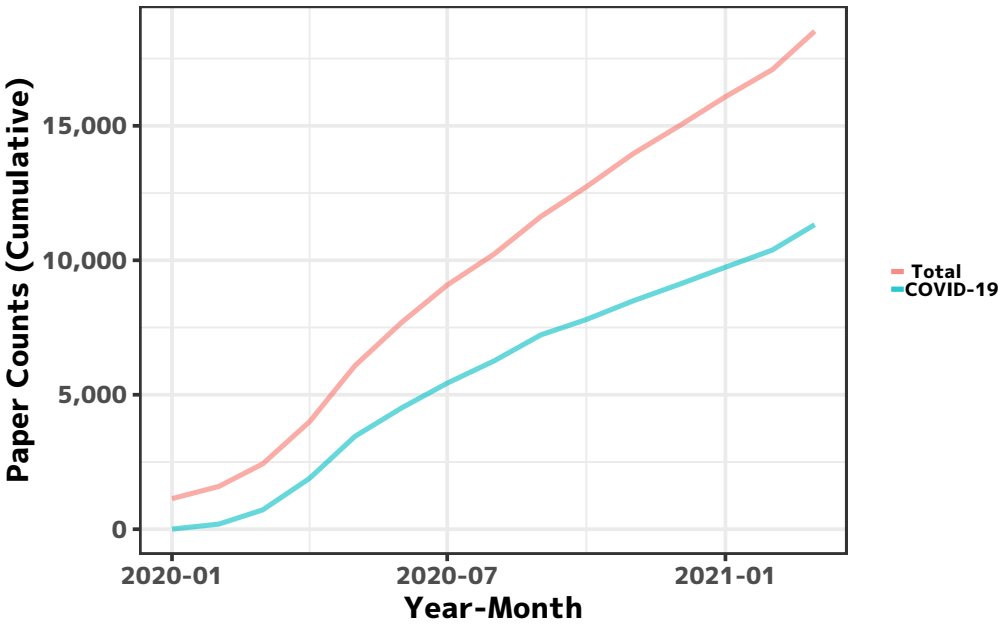


図 28　medRxiv 全体に占める COVID-19 関連投稿の数（累積）

5 考察

本章では分析の結果から得られる示唆等について他のプレプリントサーバとの比較を中心に述べる．

5.1 分野毎の差異について

今回の試行からは，bioRxiv 登録からジャーナル DOI 付与までの期間や，ジャーナル DOI 付与率などの指標について，分野間で大きな違いは見られなかった．arXiv を調査した既報 [1] ではこれらについて，分野ごとに大きな差が見られていたこととは対照的である．

arXiv は，歴史的に，高エネルギー物理学から物理学一般に広がり，さらに他の様々な分野に受け入れられる過程を経ていることに対して，bioRxiv は生物学系として比較的均質な分野を対象にしていることが背景にあると考えられる．

また，被引用数については arXiv では 1 万件近い引用を稼ぐものもあった一方，bioRxiv では最大で 1,000 件未満と少ない．この点については，研究分野そのものにおけるプレプリントの位置づけや浸透度，引用作法の違いなど，様々な要因が考えられる．

一方，COVID-19 に関するプレプリント件数の比較において, medRxiv と比較して COVID-19 に関するプレプリントの割合は小さいことが分かった．これは，bioRxiv がすでに生物学系のプレプリントサーバとして幅広く受け入れられているため，COVID-19 のような緊急性の高いトピックの影響を受けにくくなっていることを示唆する．

5.2 オープン化の進展について

まず今回の試行そのものに直接起因しない点から述べる．arXiv は物理・情報系で用いられてきたこともあり，古くから API を整備してプレプリント情報の収集・分析が容易であった．bioRxiv も 2020 年 4 月から API の公開や，データのバルクダウンロード機能が提供された[14]．ここでは JATS(Journal Article Tag Suite) と呼ばれる規格に基づいた XML の形式で全文テキストデータまで提供されている．近年オープンアクセス (OA) 論文も増加しているが，OA 論文も JATS-XML での全文テキストデータを提供するなどしており，機械可読性の高い規格を共有していることは相互運用性をはじめ様々な観点で有用といえる．この点に関して arXiv も原稿の元データを公開しているものの，歴史の長さ・コンテンツの多さや組み版ソフトの TEX を中心とする物理・情報系の論文執筆文化と相まって JATS-XML 対応は進んでいない[15]．これら API の整備や JATS-XML での全文テキストデータ提供は 2018 年 9 月に開始されたオープンアクセス出版のイニシアチブ Plan S [16] に対応するものであり，今後ますます進展していくものと考えられるが bioRxiv は比較的歴史

[14] https://connect.biorxiv.org/news/2020/04/18/tdm
[15] https://blog.arxiv.org/2019/07/18/technical-considerations-for-arxiv-compliance-with-plan-s/
[16] https://www.coalition-s.org/

が新しく，かつ一定規模の利用があることなどとも相まって比較的機動的に対応できているように観察される．

今回の試行そのものに関連する部分としては，前掲の表 2 に示したジャーナル論文化した際の収録雑誌種別があげられる．表 2 を見ると上位の多くが OA で，購読料を主体とする従来型の雑誌は極めて少ない．生物系の中でもプレプリント活用に積極的な一部の層の，さらにジャーナル論文に採択された集合のみを見ているため読み取りには注意を要するが，オープン化を牽引する研究者の一定の層を捉えている可能性もあり, 今後も注視されるべき興味深い結果となった．

5.3 政策への活用について

本調査でも示したとおり bioRxiv の利用数は順調に伸びており，このプレプリントの集合を分析することで, 生物学系の研究動向を査読付き原著論文が出版される前に把握できる可能性がある．

ただし, 多くのプレプリントを掲載しているとはいえ，生物学系の研究者全体の数や国別の分布, さらに投稿先のジャーナルを考慮すると，現在 bioRxiv の利用を行っている研究者は一部のアーリーアダプター層の可能性がある．

その上で，こうしたアーリーアダプター層は bioRxiv や OA 論文を主体に OA のエコシステム内での活動割合を伸ばしていると見ることができ，今後，研究データのオープン化などにも順次進展していく可能性が考えられる．この点については研究分野における論文を含む各種データのオープン化がどのように進展していくかのモデルケースとなる可能性があり注目を続けたい．

arXiv は主に物理・情報系の研究者が用いるプレプリントサーバである．インターネットももともとは物理系に端を発するもので情報技術に対する親和性が極めて高い．arXiv が 1990 年から運用されていることもそうした背景に起因する．研究分野全体の観点でプレプリントの利活用を見た場合，さきほどのアーリーアダプターに対応させると arXiv はイノベーター層で，重要ではあるが他の分野への横展開の可能性を伺うには難しい側面もある．

生物学でも近年情報技術の利活用は不可欠だが，情報技術そのもののを研究する分野ではない．したがって bioRxiv の利活用は情報技術そのものを専門としない多くの分野でのプレプリント活用に関する知見として有用な可能性が高い．

これら，プレプリントサーバの特色を踏まえた上で，プレプリントの分析を行い，政策づくりに役立てる必要がある．

5.4 留意点

本調査結果の読み取りに関して，留意点を以下に述べる．

まず前掲の図 3 に示したとおり，公式値と多少のズレが見られる．また，利用数は多いものの生物学系の研究者全体の数や国別等の分布を考えると現在 bioRxiv の利用を行っている研究者は一部であって，bioRxiv 上での状況をそのまま生物学系研究動向の縮図として見ることは難しい．すでに述べた OA 論文誌への高い収録率などを考えると大きなバイアスがかかっている可能性は高い．

国・地域の分析については分析方法の限界からジャーナル論文などの分析で一般的に用いられるものとは大きく異なるカウントをしている．具体的には，一般に著者全員の所属機関ベースの国・地域情報を用いるところ，1 名のみのメールアドレスに基づく国・地域情報を用いている．このため，ジャーナル論文を対象とした他の調査などとの単純比較は難しい．

今回調査した範囲では bioRxiv における COVID-19 関連投稿が占める割合は大きくないが，それでも一定のインパクトを有している．COVID-19 は危急の案件であるために，通常時以上に迅速な成果公開がなされたり，普段利用しないユーザが利用したり，という特異な行動を促した．少なくとも 2020 年に関してはこうした種別の原稿が一定数含まれることに留意が必要である．

6　まとめ

本稿では bioRxiv に登載されたプレプリントの調査を，arXiv について調査した既報 [1] に対比させて行った．結果として，bioRxiv は相対的に分野差が少なく例えばジャーナルになる割合は 4 割程度，ジャーナル誌に採録されるまでは概ね 6 ヶ月から 8 ヶ月，といった傾向が見られた．ジャーナル誌としては OA 誌に投稿されている割合が多いことが分かった．

国・地域の分析は割り付け方法が特殊なため一定の留意は必要だが，インド等を始め論文誌とは異なるような位置を占める国・地域が観察されており，国別にプレプリント対応の差が表れていることを示唆している．

bioRxiv は特に PlanS と関連してデータの収集・分析が arXiv より容易になっている．今後，科学技術・イノベーション政策の文脈においてジャーナル論文と同様に各種のプレプリントサーバの動向分析も重要度を増すことが想定される中，論文, 掲載ジャーナル, および電子ジャーナルプラットフォームの分析と同様に，分野別やプレプリントサーバごとの特性を踏まえた分析と，その分析から得られる価値と政策への展開を念頭に，引き続き調査を行う予定である．

参考文献

[1] 林和弘, 小柴等：arXiv に着目したプレプリントの分析. *NISTEP DISCUSSION PAPER* Vol.187, Aug 2020. https://doi.org/10.15108/dp187

[2] 林和弘： MedRxiv, ChemRxiv にみるプレプリントファーストへの変化の兆しと オープンサイエンス時代の研究論文. *STI Horizon 2020* **春号** Vol.6, No.1, Mar 2020. https://doi.org/10.15108/stih.00205

[3] MEXT–NISTEP プレプリント調査・検討チーム： プレプリントをめぐる近年の動向及び今後の科学技術行政への示唆. **文部科学省 科学技術・学術審議会 情報委員会 ジャーナル問題検討部会 第 *7* 回 配布資料** 資料 1-別添, Oct 2020. https://www.mext.go.jp/content/20201026-mxt_jyohoka01-000010684_2.pdf

[4] 小柴　等, 林　和弘, 伊藤裕子： COVID-19 / SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析. *NISTEP Discuttion Paper,* No.186, June 2020. http://doi.org/10.15108/dp186

DISCUSSION PAPER No.197

bioRxiv に着目したプレプリントの分析

2021 年 08月

文部科学省 科学技術・学術政策研究所
林 和弘, 小柴 等

〒100-0013 東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL: 03-3581-2391 FAX: 03-3503-3996

Analysis of preprints on bioRxiv

Aug 2021

HAYASHI Kazuhiro, KOSHIBA Hitoshi
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

https://doi.org/10.15108/dpXXX

https://www.nistep.go.jp