Classifying Sleeping Beauties and Princes Using Citation Rarity

Takahiro Miura(✉), Kimitaka Asatani, and Ichiro Sakata

Department of Technology Management for Innovation,
Graduate School of Engineering, The University of Tokyo, Bunkyo-ku, Tokyo, Japan
{miura,asatani,isakata}@ipr-ctr.t.u-tokyo.ac.jp

Abstract. The scientific community sometimes resists important scientific findings initially. This is the so-called "delayed recognition." A "sleeping beauty (SB)," a representative phenomenon of delayed recognition, is a paper reported by a Prince (PR) paper. The SB includes many key breakthrough concepts for resolving scientific problems. Although many PRs discover their SBs, it is still unknown how they do that because the citation culture differs depending on the category of the paper. This study classifies SBs and their PR pairs using citation rarity within clusters that represent a unique category of a paper. Results show that citation rarity corresponds to the types of contributions to PR papers. Rare citations explore methodological insights into PR fields. Meanwhile, common citations can lead to rediscovery of the core concepts of a sleeping beauty. Furthermore, informatics and materials sciences cover major studies that include citations for SBs, whereas biological subjects find key papers through rediscovery. Results indicate that different categories of citations yield different types of SBs.

Keywords: Bibliometrics · Cross-disciplinary · Princes · Sleeping beauties

1 Introduction

Often, some of the innovative scientific works go unnoticed for long periods. This phenomenon is known as "delayed recognition" [1–3]. New discoveries and theories are significantly important for scientific progress; however, initially, they are often restricted or neglected as the scientific community is skeptical about them [4,5]. Further, information explosion prevents important ideas from penetrating the wall of established wisdom related to a subject. Mechanisms underlying delayed recognition are always relevant to major scientific progress or groundbreaking scientific revolutions. However, how this delayed recognition occurs remains unknown.

The quantitative concept of delayed recognition, as proposed by Van Raan, can be designated simply as a sleeping beauty (SB) phenomenon [6]. Although a set of papers might go unnoticed for a long time, the same set will be suddenly noticed after a certain point a time. In addition to the original definition of SB

R. M. Benito et al. (Eds.): COMPLEX NETWORKS 2020, SCI 943, pp. 308–318, 2021.
https://doi.org/10.1007/978-3-030-65347-7_26

using depth, length, and waking up from sleep [6], several extended terms exist for the extraction of various cases of SB papers [7,8].

Initially, SB was regarded as a rare phenomenon in scientific progress, but recent research shows that it is far less exceptional than previously thought. In fact, SBs include a number of scientific finding-related information [8].

Every SB has its own PR, which wakes it up and introduces it to the wider research community by citing the SB document. The first report to cite SB is the original definition of a PR [6]. However, this definition is suitable only for cases of "coma sleep," i.e., cases wherein no attention was paid to citations [9]. The Internet makes it easy to access minor but related articles. Therefore, a co-citation criterion is appropriate for finding a PR [10].

Many studies have positioned SBs and PRs in a specific field or category [8,11]. Nevertheless, there has been no systematic approach reported till date that can find SB–PR pairs comprehensively from articles because so many patterns show how a PR discovers an SB. While examining the computer science category specifically, it has been found that SBs contribute to some methodologies. Actually, PRs have extended the model and methodology established for SBs to make them applicable in other sub-fields [11]. Comprehensive analysis of SB–PR pair findings is essential because it remains unknown whether citation distributions for different sciences are similar.

Our research specifically examines classification of the various types of scientific findings across respective scientific disciplines using SB and PR pairs in various fields. The SB and PR pairs include breakpoints of the scientific findings in the concerned field. Comparison for a case of delayed recognition reveals cross-disciplinary similarity in the structure with respect to how delayed recognition is resolved. This might be the first report related to a study analyzing the number of SB–PR papers and categorizing their types.

The driving hypothesis of this paper is that estimation of the cross-disciplinary relation between SBs and PRs is performed through citation rarity calculated from complex citation networks. For this study, we have systematically clarified the relation between SBs and PRs by categorizing them post large-scale acquisition of SB and PR pairs. As a classification technique, we have considered the inadequacy of citation of SB by PR deduced on the basis of inter-cluster distance calculated with respect to complex networks corresponding to the citations.

2 Results

2.1 Sleeping Beauties and Princes

There are various methods to identify SBs, such as an average-based approach [6,12], a quartile-based approach [13,14], and a non-parametric approach. In this research, we have used the "beauty-coefficient," which is a non-parametric method, for extracting SBs proposed by Ke [8] and, subsequently, for classifying the SB papers. This is because average-based and quartile-based approaches are strongly affected by arbitrary parameters of citation thresholds, which depend

on their categorical citation bias [15]. For specific examination of articles that have sufficient impact on the scientific community, we have extracted the top 5% citations from the Scopus comprehensive database. The number of top citation papers are 3,392,918, and the fewest citations are 67. As shown below, we calculated the beauty-coefficient score B for each paper.

$$B = \sum_{t=0}^{t_m} \frac{\frac{c_{t_m}-c_0}{t_m} \cdot t + c_0 - c_t}{\max\{1, c_t\}} \tag{1}$$

In the above equation, c_t represents the number of citations that the paper received after its publication in the tth year, and t_m represents the year in which the paper received maximum citations c_{t_m}.

The Eq. (1) penalizes early citations as the later the citations are accumulated, the higher is the value of index B. We have defined the top 1% of the B scores as SB papers, which include 33,939 papers.

For each SB paper, a candidate for the PR paper is the one with the highest number of co-citations among all the papers citing that SB. For definition of SB papers, we have used the Ke's awakening year [8], which describes the time of citation burst as follows.

$$t_a = \arg\{\max_{t \le t_m} d_t\} \tag{2}$$

$$d_t = \frac{|(c_{t_m} - c_0)t - t_m c_t + t_m c_0|}{\sqrt{(c_{t_m} - c_0)^2 + t_m^2}} \tag{3}$$

If the candidate paper was published within 5 years (i.e., around t_a, which is the awakening year of the SB papers), then it was defined as the PR paper of the SB. Thus, the number of SB–PR pairs was 14,317. Figure 1(a) presents the year-wise distribution of SB and PR. By definition, the greater the time distance between SB and PR, the larger the likely beauty coefficient. Therefore, most of SBs are papers published between 1970 and 1990. The gap year distribution reflects that (Fig. 1(b)) SBs are usually discovered after around 25 years.

2.2 Defining the SB–PR Pair Density

In this section, we have defined the SB–PR pair density with respect to its citation probability. We clustered the citation network of 67 million papers using the Leiden algorithm [16]. Citation probability is defined on the basis of the frequency of the edges between two clusters in the PR publication year. When papers in a cluster comprising a PR paper cite the particular cluster that includes the SB paper, the presence of edges between the SB and PR is not so unusual. Hence, the density in this case is high.

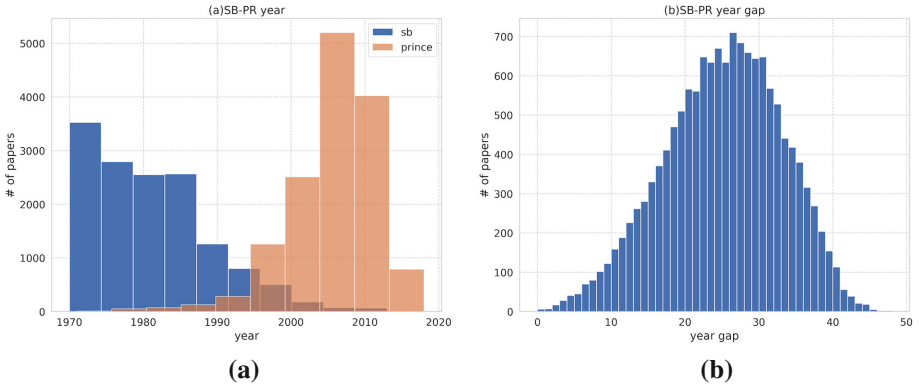

Fig. 1. (a) Annual distribution of SB and PR. (b) Gap year distribution of the SB paper and the PR paper.

We have defined the density of pairs D as follows:

$$A^{y}_{i,j} = \sum_{y} A_{y,i,j}. \tag{4}$$

$$D_{y,c_i,c_j} = \frac{A^{y}_{i,j}}{|c_i||c_j|} \quad (i \neq j) \tag{5}$$

In the above equation, $A_{y,i,j}$ indicates the number of papers in the cluster i that were published during the year y. Further, it also cites the papers in cluster j. Further, $|c_i||c_j|$ represents the possible edges between cluster i and cluster j, whereas $A^{y}_{i,j}$ showcases the actual edges between the two clusters until year y. When a PR published in the year y_p, and from the cluster c_p, cites the SB in cluster c_s, the density of this SB–PR paper is D_{y_p,c_p,c_s}. The density of the pair cannot be defined if the PR and SB are in the same cluster.

In this research, we have considered the first floor clustering of the entire citation network using the Leiden algorithm [16] as label for the papers. The purpose is to classify each paper into a unique category, as many papers exist in multiple disciplines these days.

Table 1 shows the example of each clusters. The top clusters include more than 8 million nodes, which are way too extensive to be considered under a single category. These may be covered under the basic concept of science. As we have specifically examined the cross-disciplinary SB–PR pairs in this study, we adopted the first floor clustering as a category to extract a more pointed cross section of the field. A more detailed analysis of the sub-clustering categories is necessary for future work.

2.3 Density Distribution

Among the 14,317 pairs, only 1,857 pairs are a result of cross-disciplinary findings with a citation of an SB in another cluster. Therefore, most of the SB–PR

Table 1. Cluster size and detail of the top 10 largest clusters

Cluster	Size	Label	Frequent keywords
0	8,292,009	General	Education, China
1	5,564,222	Material Science	Microstructure, Mechanical properties
2	5,427,069	Informatics	Optimization, Simulation
3	3,866,952	Life-style related diseases	Obesity, Hypertension
4	3,703,869	Cancer	Brest Cancer, Apotosis, Cancer
5	3,559,961	Biology	Taxonomy, New species
6	3,561,520	Intractable diseases	Alzheimer's disease, Schizophenia
7	2,944,362	Cell Biology	Appotosis, Asthma
8	2,947,362	Structural Chemistry	Crystal structure, synthesis

pairs are internal findings. Figure 2 presents the density distribution of cross-disciplinary SB–PR pairs. As compared to the random extraction from all cross-disciplinary citations, the distribution of SB–PR pairs is skewed to the left. This implies that SB–PR pairs include more rare collaborations than normal cross-disciplinary citations.

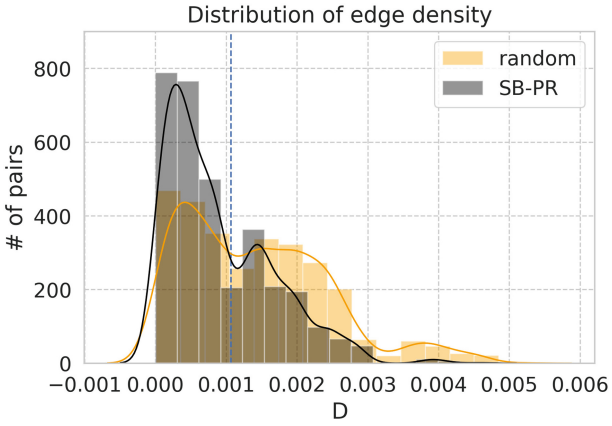

Fig. 2. Density distributions of SB and PR.

The distribution has two peaks. The first peak represents rare collaborations ($D < 1.07 \times 10^{-3}$). The most cross-disciplinary PR papers "explore" unusual categories of SB paper, thereby indicating that the PR broadens the possibilities of the field. The second peak represents common collaborations ($D \geq 1.07 \times 10^{-3}$). Even when similar papers are cited via common clusters, some PRs "rediscover" an important concept of SB papers. We have classified the bottom 66% of density under "exploring citations," which are rare collaborations that transpired until that particular year. The other 33% are "rediscovering citations," which re-evaluate the importance of common pairs of knowledge.

2.4 Rediscovering PRs and Exploring PRs

Publication of review papers frequently results in various scientific rediscoveries. Busy authors do not cite the original work; instead, they cite more recent derivative works and reviews [17]. The percentage of review papers for exploring PRs, overall PRs, and rediscovering PRs was 25%, 28%, and 35%, respectively, which increased at higher densities. Frequent citations between clusters led to the rediscovery of key findings.

Additionally, when we studied how PR papers cite SBs, we found out that discovering PRs are more likely to cite SBs in the Introduction and Results sections, whereas exploring PRs cite SBs in the Methodology section (Table 2). The introduction presents a brief description of the trajectory on which the research is based. It plays an important role in the early stage of research. Additionally, the Results section discusses core contributions toward the knowledge frontier. As a result, rediscovery of papers is presumed to extract research pairs that are linked strongly at the conceptual level. Citations in papers' Methodology section typically require an uncommon method to break the known challenges in the PR field. An SB category develops a way to solve other problems, which can be transferred to PR field problems. Moreover, among the top 100 PRs, 9 exploring PRs awaken multiple SBs, while all rediscovering PRs evoke only 1 SB. Exploring PRs have the potential to discover more than one SBs at a time.

Table 3 presents the highest and lowest examples of citation of two types of PR. Rediscovering PRs and SBs depict the field background and the comparison between the impact of the experimentally obtained results and results obtained from general studies. Exploring PRs are often used to conduct analyses that involve implementing methods that are not often used in a field. This paper has led to the popularization of this particular method of analysis in the field because this is the largest co-cited pair.

Table 2. Citation points of PR from SB for 100 articles each

Doctype	Citation point	Exploring PR	Rediscovering PR
Article, conference paper	Introduction	17	20
	Methodology	19	12
	Results, Discussion	6	19
	Others	9	3
Review		34	39
Book		14	5
Others		1	2
Total		100	100

2.5 Relation Type of SBs and Princes

Next, we identify whether the trend in SB–PR pairs varies by field. Figure 3 shows the specific rediscovering and exploring pairs that are more likely to occur

Table 3. Examples of exploring PR and rediscovering PR

D	Part	Citation sentence in PR
4.2×10^{-3} (rediscovering)	Introduction	Mitochondria are evolutionary endosymbionts derived from bacteria and contain DNA similar to bacterial DNA [19]
4.0×10^{-3} (rediscovering)	Result	This suggestion is surprising, because it is generally thought that chromatin structure does not play an important role in HSV gene transcription, largely because, unlike other viruses (e.g., SV40), newly replicated HSV genomes are not packaged into chromatin [20]
6.0×10^{-6} (exploring)	Methodology	Models with an initial percolating k-core cluster of quasi-crystalline short-range order showedshear localization at low strain rates; those without thisorder showed homogeneous deformation [21]
7.0×10^{-6} (exploring)	Methodology	LEfSe is implemented in Python and makes use of R statistical functions in the coin and MASS libraries through the rpy2 library and of the matplotlib library for graphical output [22]

between disciplines. Unlike exploring pairs, rediscovering PRs contribute largely to locally specific discipline SBs. For example, lifestyle-related diseases, cancer, cell biology, and molecular biology PRs tend to rediscover the past findings. These categories expand the specific knowledge range by leveraging references from closely related fields. In contrast, general, informatics, and materials engineering PRs are likely to use exploring citations. These clusters combine various types of knowledge through broader categories. It could be an intersection of scientific key findings.

Instead of being explored, material science is more likely to explore various types of fields, indicating that the field applies key findings obtained from other fields. As far as informatics is concerned, it applies knowledge of the environment, materials engineering, and physical astronomy. Subsequently, biological categories, such as cancer and intractable diseases, make use of the findings. We can observe the circulation of knowledge across disciplines using citation rarity. This heatmap presents a foundation or relation type application of each pair of categories.

Table 4 presents the most frequent SB–PR pairs for each finding. The disciplines that become SBs and the ones that become PRs are relative matters. Thus, the flow of knowledge is not necessarily restricted to one direction (i.e., toward the basic and applied disciplines). However, some trends exist in scientific findings among the categories. Informatics may include key PRs that explore unknown knowledge from various fields, such as physics, materials engineer-

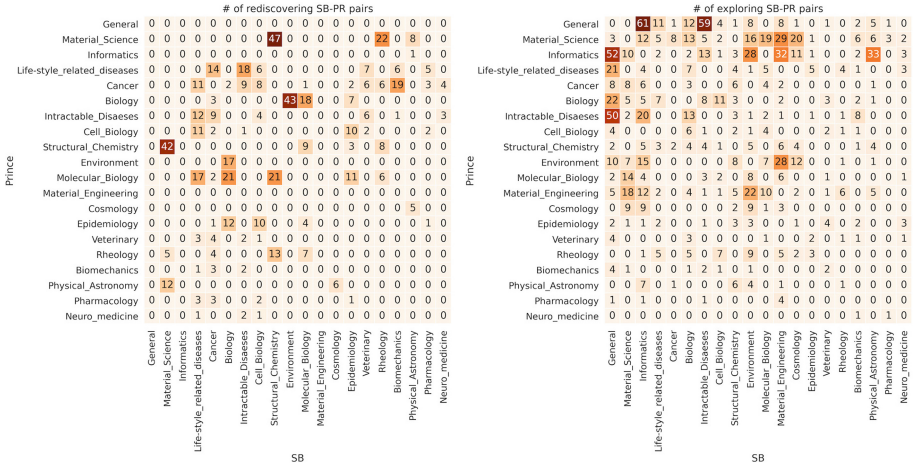

Fig. 3. Frequency of SB–PR pairs among the top 20 clusters.

ing, and environment related. Biology and chemistry, which are closely related, demonstrate rediscovery of the core concepts of the mutual findings.

Table 4. Frequent pairs of SB–PR in exploring and rediscovering collaborations

Exploring citation		Rediscovering citation	
PR	SB	PR	SB
General	Informatics	Materials Science	Rheology
General	Intractable Diseases	Materials Science	Structural Chemistry
Informatics	General	Biology	Environment
Informatics	Physics	Biology	Molecular Biology
Informatics	Materials Engineering	Structural Chemistry	Materials Science
Informatics	Environment	Molecular Biology	Biology
Intractable Diseases	General	Molecular Biology	Structural Chemistry
Materials Science	Materials Engineering	Molecular Biology	Lifestyle-related disease
Environment	Materials Engineering	Cancer	Biomechanics
Materials Engineering	Environment	Lifestyle-related disease	Intractable disease

2.6 Density vs. Citation

We hypothesize that, as an increasing number of exploration of citations occurs, the volatility of citation of PR papers increases because a rare combination unexpectedly produces revolutionary effects on research in the concerned field. However, the length of SB–PR pairs does not correlate with the citation of SBs ($R^2 = 0.00$) and PRs ($R^2 = 0.00$). We expected the citation gap, which separates

the successful papers from the failed papers, to be larger for exploring citations. However, the variance did not differ on the basis of whether the cited works were explored or rediscovered.

Furthermore, examination of key papers related to Nobel prize-winning findings selected by Mr. John Ioannidis [18] revealed that Nobel prize papers among the cross-disciplinary SB and PR papers are very few. We hypothesize that Nobel prize papers broaden the horizon of a category and they have an extremely strong impact beyond the representation of citations. Therefore, some of them may exist in SB–PR pairs. However, all SB–PR pairs include only four SBs and four PRs; cross-disciplinary pairs include only 1 SB. There was no correlation found between the impact of SB—PR papers and their density of citation. These results imply that surprising citations may not necessarily result in useful findings for the scientific community. With increasing attention being focused on the importance of cross-disciplinary research, the implications of the rarity of citations in the network are expected to be a major challenge in the future.

3 Conclusion

In this study, we have classified the types of SB–PR pairs across scientific disciplines in various fields. The relation of the pair is described on the basis of the citation rarity of the clusters that they are present in. The pairs have been broadly divided into two categories: major exploration citations and minor rediscovery citations. Rediscovering PRs contain more review articles than average. They refer to the SB in the Introduction and Results sections, which cite fundamentally important information about key findings. Meanwhile, the exploring PRs form an integral part of the Methodology section, which require an uncommon method to break the known challenges in the PR field. Furthermore, the materials science PRs, instead of being explored, are more likely to explore various types of fields, such as rheology or structural chemistry. This indicates that the field applies key findings obtained from other fields. However, biological subjects, such as cancer or cell biology, exhibit rediscovery of important papers through common clusters of SB–PR pairs.

This research contributes toward a better understanding of the delayed recognition across categories.

4 Data

We use bibliographic databases extracted from Scopus. These include 67 million papers and 1 billion citations from 27 fields covered from 1970 to 2018. The scientific fields are not fixed on the basis of time but rather expand and contract as and when they fuse and separate from other fields. Hence, we clustered the entire citation network into 1858 partitions using the Leiden algorithm [16] to identify the related category of each paper.

References

1. Garfield, E.: Premature discovery or delayed recognition - why? Essays Inf. Sci. **4**, 488–493 (1980)
2. Garfield, E.: Delayed recognition in scientific discovery: citation frequency analysis aids the search for case histories. Curr. Contents **23**, 3–9 (1989)
3. Garfield, E.: More delayed recognition. Part 2. From inhibin to scanning electron microscopy. Essays Inf. Sci. **13**, 68–74 (1990)
4. Campanario, J.M.: Rejecting and resisting Nobel class discoveries: accounts by Nobel Laureates. Scientometrics **81**(2), 549–565 (2009). https://doi.org/10.1007/s11192-008-2141-5
5. Fang, H.: An explanation of resisted discoveries based on construal-level theory. Sci. Eng. Ethics **21**(1), 41–50 (2015). https://doi.org/10.1007/s11948-013-9512-x
6. van Raan, A.F.J.: Sleeping beauties in science. Scientometrics **59**(3), 467–472 (2004). https://doi.org/10.1023/B:SCIE.0000018543.82441.f1
7. Mazloumian, A., Eom, Y.H., Helbing, D., Lozano, S., Fortunato, S.: How citation boosts promote scientific paradigm shifts and Nobel Prizes. PLOS ONE **6**(5) (2011). https://doi.org/10.1371/journal.pone.0018975
8. Ke, Q., Ferrara, E., Raduccgu, F., Flammini, A.: Defining and identifying sleeping beauties in science. Proc. Nat. Acad. Sci. U.S.A. **112**(24), 7426–7431 (2015). https://doi.org/10.1073/pnas.1424329112
9. van Raan AFJ.: Dormitory of physical and engineering sciences: sleeping beauties may be sleeping innovations. PLOS ONE **10**(10), e0139786 (2015). https://doi.org/10.1371/journal.pone.0139786
10. Du, J., Wu, Y.: A bibliometric framework for identifying "princes" who wake up the "sleeping beauty" in challenge-type scientific discoveries. J. Data Inf. Sci. **1**(1), 50–68 (2016). https://doi.org/10.20309/jdis.201605
11. Dey, R., Roy, A., Chakraborty, T., Chosh, S.: Sleeping beauties in computer science: characterization and early identification. Scientometrics **113**, 1645–1663 (2017). https://doi.org/10.1007/s11192-017-2543-3
12. Glänzel, W., Schlemmer, B., Thijs, B.: Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. Scientometrics **58**, 571–586 (2013). https://doi.org/10.1023/b:scie.0000006881.30700.ea
13. Costas, R., van Leeuwen, T.N., van Raan, A.F.J.: Is scientific literature subject to a 'Sell-By-Date'? A general methodology to analyze the 'durability' of scientific documents. J. Am. Soc. Inf. Sci. Technol. **61**, 329–339 (2010). https://doi.org/10.1002/asi.21244
14. Li, J.: Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition". Scientometrics **100**(2), 595–601 (2013). https://doi.org/10.1007/s11192-013-1217-z
15. Ioannidis, J.P.A, Baas, J., Klavans, R., Boyack, K.W.: A standardized citation metrics author database annotated for scientific field. PLOS Biol. **17**(8) (2019). https://doi.org/10.1371/journal.pbio.3000384
16. Traag, V.A., Waltman, L., van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. **9**, 5233 (2019). https://doi.org/10.1371/journal.pbio.3000384
17. Marks, M.S., Marsh, M.C., Schroer, T.A., Stevens, T.H.: An alarming trend within the biological/biomedical research literature towards the citation of review articles rather than the primary research papers. Traffic **14**(1), 1 (2013). https://doi.org/10.1111/tra.12023

18. Ioannidis, J.P.A., Cristea, I.A., Boyack, K.W.: Work honored by Nobel prizes clusters heavily in a few scientific fields. PLOS ONE **15**(7), e0234612 (2020). https://doi.org/10.1371/journal.pone.0234612
19. Oka, T., Hikoso, S., Yamaguchi, O., Taneike, M., Takeda, T., Tamai, T., Akira, S.: Mitochondrial DNA that escapes from autophagy causes inflammation and heart failure. Nature **485**(7397), 251–255 (2012). https://doi.org/10.1038/nature10992
20. Wysocka, J., Myers, M.P., Laherty, C.D., Eisenman, R.N., Herr, W.: Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3–K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. Genes Dev. **17**(7), 896–911 (2003). https://doi.org/10.1101/gad.252103
21. Schuh, C.A., Hufnagel, T.C., Ramamurty, U.: Mechanical behavior of amorphous alloys. Acta Mater. **55**(12), 4067–4109 (2007). https://doi.org/10.1016/j.actamat.2007.01.052
22. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C.: Metagenomic biomarker discovery and explanation. Genome Biol. **12**(6), 1–18 (2011). https://doi.org/10.1186/gb-2011-12-6-r60