Correspondence

A recipe for high impact

Murat Cokol*†, Raul Rodriguez-Esteban†‡ and Andrey Rzhetsky*†§

Addresses: *Department of Biomedical Informatics, and †Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA. ‡Department of Electrical Engineering, Columbia University, New York, NY 10025, USA. §Judith P. Sulzberger MD Columbia Genome Center and Department of Biological Sciences, Columbia University, New York, NY 10032, USA.

Correspondence: Andrey Rzhetsky. Email: andrey.rzhetsky@dbmi.columbia.edu

Published: 10 May 2007

Genome ***Biology*** 2007, **8:**406 (doi:10.1186/gb-2007-8-5-406)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/5/406

Abstract

Our analysis highlights common statistical features of high-impact articles; we also show how information flows among various publication types.

Every research article has at least two important ingredients: it attacks a scientific problem (topic), and invents or recycles a study technique (method). Here we quantify the relative contribution of these two elements to an article's success by sifting through myriads of time-stamped scientific texts, accumulated over decades in the permafrost of reference databases [1].

We define and analyze here three attributes associated with each scientific article: 'topic', 'method' and 'impact'. Nearly every article referenced in the PubMed database has a list of keywords reflecting its content: chosen from more than 20,000 MeSH terms and more than 150,000 chemical names [2]. We use MeSH terms and chemical names as indicators of an article's topic and method, respectively. The 'impact factor' (IF) of the journal where the article was published is provided by the Thomson ISI database [3].

Ingredients of a scholarly study

For millions of articles published in 1,757 journals we compute two parameters (separately for topic and method concepts): 'temperature' and 'novelty', as introduced in our earlier work [4], using a reference corpus of publications pre-dating each article (see Additional data file 1). When all journal-specific articles are considered together, a high temperature of a journal indicates its tendency to publish popular (hot) concepts. The novelty parameter can change between 0 and 1, and, as the name implies, reflects the proportion of new (previously unpublished) concepts in a group of texts.

We used a five-parameter linear regression model to assess contributions of topic- and method-specific estimates of temperature and novelty to a journal's IF (see Additional data file 1). We observe that high IFs correlate strongly with hotter topics and colder methods (see Figure 1a,b). Disturbingly, both method and topic novelty are unimportant for predicting IF. Despite a strong positive correlation between the popularity of article's topic and method - contributed by the bulk of the moderately influential articles (see Figure 1b, inset) - the highest-impact scientific research emerges when very popular (important) topics are tackled with unpopular methods.

Our topic and method terms have very different frequency distributions - reflecting the difference in their genesis. In the former case, it is a human expert who decides that a new concept is sufficiently frequently used to merit its addition to the controlled MeSH vocabulary. In the latter case, the list of new terms is not artificially restricted; they are allowed to be very rare (see Figure 1b). As a result, frequencies of the chemical terms follow a classical Zipf's distribution, while MeSH terms clearly deviate from this distribution due to deficiency of the rare terms (see Figure 1b).

Information flow through publication-type niches

Figure 1c,d illustrates the unique (statistically distinct) niches of distinct publication types in the space of novelty and temperature. For methods (chemicals, including drugs), information diffuses from novel-unpopular to known-popular

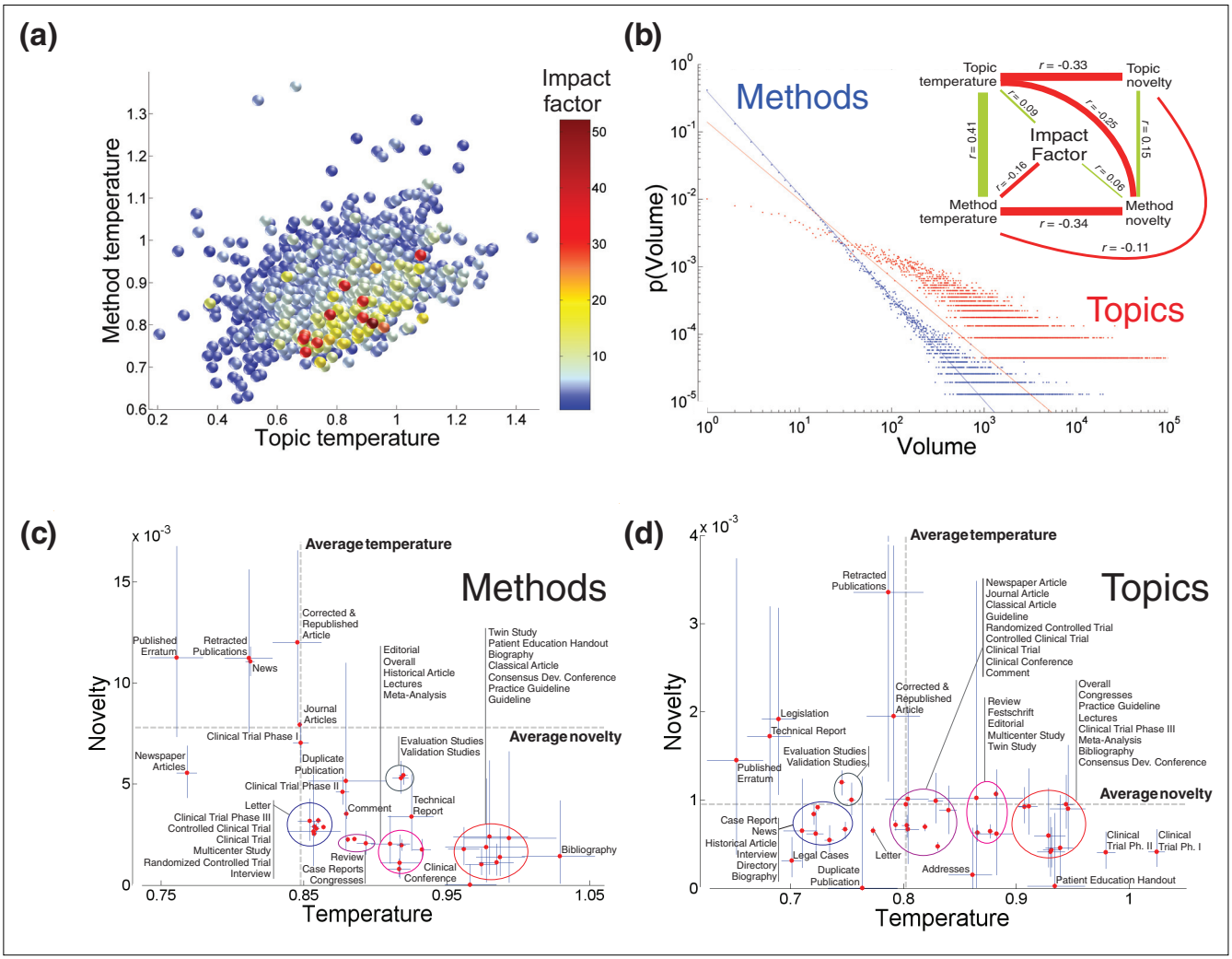

Figure 1
Contributions of topic- and method-specific estimates of temperature and novelty to a journal's impact factor. **(a)** Relationship among the method-temperature (chemical), topic-temperature (MeSH), and the impact factor of 1,757 journals. **(b)** Volume (number of mentions) distribution of topics and methods. Inset: significant ($p < 0.01$) correlations between pairs of the five parameters. Green and red lines indicate positive and negative correlations, respectively, with line width proportional to the corresponding correlation strength. **(c,d)** Estimates of temperature and novelty parameters for various publication types with 95% credible intervals. Ovals indicate closely grouped estimates; labels are listed in decreasing novelty.

publication types. 'Colder' chemicals are published first in the journal articles; some of them later make it to the warmer and less novel space of phase I clinical trials, and a subset of these drugs makes it to the significantly warmer area of phase II clinical trials (Figure 1c). Furthermore, the growth of temperature and loss of novelty progressively accelerates to reviews, lectures and biographies. Curiously, the retracted and corrected papers (Figure 1c), along with news, are champions in the novelty competition - it looks almost as if the retracted articles are too novel to be correct. For topics, we observe a similar - albeit less intuitive - picture (Figure 1d), where retracted articles again have the highest novelty. The clinical trial story shows a new twist here: most clinical trials take years; they persist long enough for their initially hot topics (at the stage of a research article and phase I clinical trial) to cool down before reaching phase II and III trials (Figure 1d) - a consequence of the time-dependence of temperature estimates that capture ephemeral fads within biological disciplines.

Our analysis highlights the importance of choice of a research topic, and of putting new work in the right context. A remarkable idea (method) presented to the world in a wrong context (topic) has little chance of being noticed. A successful idea travels through publication types much as energy flows through an ecosystem: it is typically born novel and unpopular in research articles (plants), and diffuses eventually to reviews, lectures, clinical trials, and bibliographies (top-hierarchy carnivores), where it reaches the pinnacle of popularity.

Additional data file

The method of analysis and supporting data are available with this article online in Additional data file 1.

Acknowledgements

We would like to thank Emek Demir for valuable discussions and Chani Weinreb for comments on earlier version of the manuscript. This work was supported by the National Institutes of Health (training fellowship 5-T15-LM007079 to M.C. and RO1 GM61372 to A.R.).

References

1. **Entrez PubMed** [www.ncbi.nlm.nih.gov/entrez]
2. **Medical subject headings (MESH) fact sheet** [www.nlm.nih.gov/pubs/factsheets/mesh.html]
3. **Thomson Scientific** [www.isinet.com]
4. Cokol M, Iossifov I, Weinreb C, Rzhetsky A: **Emergent behavior of growing knowledge about molecular interactions.** *Nat Biotechnol* 2005, **23:**1243-1247.