BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S249–S257

Of truth and pathways: chasing bits of information through myriads of articles

Michael Krauthammer[1], Pauline Kra[1, 2], Ivan Iossifov[1, 2], Shawn M. Gomez[2], George Hripcsak[1], Vasileios Hatzivassiloglou[4], Carol Friedman[1, 3] and Andrey Rzhetsky[1, 2]

[1]Department of Medical Informatics, Columbia University, New York, NY, 10032, USA, [2]Columbia Genome Center, Columbia University, New York, NY, 10032, USA, [3]Department of Computer Science, Queens College CUNY, Flushing, NY, 11367, USA and [4]Department of Computer Science, Columbia University, New York, NY, 10027, USA

Received on January 24, 2002; revised and accepted on April 1, 2002

ABSTRACT

Knowledge on interactions between molecules in living cells is indispensable for theoretical analysis and practical applications in modern genomics and molecular biology. Building such networks relies on the assumption that the correct molecular interactions are known or can be identified by reading a few research articles. However, this assumption does not necessarily hold, as truth is rather an emerging property based on many potentially conflicting facts. This paper explores the processes of knowledge generation and publishing in the molecular biology literature using modelling and analysis of real molecular interaction data. The data analysed in this article were automatically extracted from 50 000 research articles in molecular biology using a computer system called GeneWays containing a natural language processing module. The paper indicates that truthfulness of statements is associated in the minds of scientists with the relative importance (connectedness) of substances under study, revealing a potential selection bias in the reporting of research results. Aiming at understanding the statistical properties of the life cycle of biological facts reported in research articles, we formulate a stochastic model describing generation and propagation of knowledge about molecular interactions through scientific publications. We hope that in the future such a model can be useful for automatically producing consensus views of molecular interaction data.

Contact: ar345@columbia.edu

Keywords: statistical modelling; scientometric analysis; molecular interaction data; natural language processing

INTRODUCTION

Molecular interaction data and corresponding knowledge bases are becoming increasingly important for both academic and commercial undertakings in modern biology (Jeong *et al.*, 2001; Karp, 2000; Karp *et al.*, 1998). As these resources are used more intensively, the updating of manually curated repositories becomes an important issue. Usually, experts determine which information should be included in the repositories, and some databases, such as DIP, invite outside researchers to help curate the growing amount of data (Xenarios *et al.*, 2002). While expert consensus is certainly the de facto standard in determining true molecular interactions, it is becoming increasingly more difficult to keep up with the avalanche of information flooding research journals. Furthermore, there is some concern that biased reporting of research results in the literature may complicate the process of truth finding. Mrowka and colleagues (Mrowka *et al.*, 2001) have recently described significant discrepancies of two-hybrid protein–protein interaction datasets, which were either indirectly compiled from single research publications or directly compiled from genomewide screens. Their data shows a potential selection bias in the literature-based dataset, which ‘may have been introduced by the failure to report interactions which cannot be understood from previous publications, or by failing to perform experiments for such pairs in the first case’. Elucidating such biases, as well as other complicating factors such as contradicting research results, are the aim of this paper. Our motivation is the direct application of such insights to our system called GeneWays, which automatically collects molecular interaction data from the research literature using a natural language module called GENIES (Friedman *et al.*, 2001). Our goal is to assist experts in building a consensus representation of the extracted molecular information by automating the consensus finding process when there are biased and/or conflicting research results.

Using scientometric techniques, this paper attempts to shed light on how molecular interaction data is reported in the research literature. Unlike traditional scientometric approaches, which rely on citation data, we attempt to use individual statements on molecular interactions as they appear in the research articles. The advantage of this approach lies in the potential to track the reporting of research results over time, which enables the measurement of basic publication properties. These include so-called waiting times of research ideas, which measure the diffusion of a certain research idea from its first conception (publication) through the research literature. In this particular context, we are interested in the time spans between subsequent publications of unique research results, as well as the time spans between subsequent publications of the same research results. We find that these waiting times follow an exponential distribution. By recording negations of research results we are also able to measure the level of contradiction in research results. Through the analysis of the connectedness of a substance to neighbouring substances, we infer the importance of a statement as a function of the general interest exhibited by the research community towards a particular molecular substance. This allows for a direct measurement of selection bias that may be present in the research literature. We finally integrate these results into a model describing the generation of research ideas in the domain of molecular biology.

BACKGROUND

The immense growth of research literature in the field of molecular biology calls for methods to automatically capture and intelligently store molecular data. In recent years, many groups have worked on dedicated problems in this area, like machine-selection of articles of interests (Iliopoulos *et al.*, 2001; Shatkay *et al.*, 2000), automated extraction of information using statistical methods (Stephens *et al.*, 2001; Stapley and Benoit, 2000) or natural language processing techniques (Friedman *et al.*, 2001; Yakushiji *et al.*, 2001; Thomas *et al.*, 2000; Ng and Wong, 1999; Sekimizu *et al.*, 1998) as well as setting up specialized knowledge bases for storing molecular knowledge (Stevens *et al.*, 2000). We are working on a system called GeneWays, which combines these subtasks into an integrated system. The system targets extraction of binary statements about two interacting substances, such as proteins, RNAs, genes and small molecules, and the type of interaction between them, such as *phosphorylate*, *demethylate*, *activate* or *inhibit*. The system annotates each molecular interaction with information such as the actual journal statement, the name of the corresponding journal, the date of publication and whether the interaction is positively ('*A activates B*') or negatively ('*A does not activate B*') stated. The core of the system is a knowledge base of molecular actions (Rzhetsky *et al.*, 2000) optimized for storing information from different knowledge sources. The knowledge is provided by various system modules, which sequentially select scientific journals of interest, mark and identify substance names in the journal text (Hatzivassiloglou *et al.*, 2001; Krauthammer *et al.*, 2000) and extract interactions between these substances and other actions by means of natural language processing (NLP) (Friedman *et al.*, 2001). Finally, the stored knowledge can be queried, critiqued and visualized (Koike and Rzhetsky, 2000) by interested researchers. The system is fully functional and is collecting and processing articles from online scientific journals. We have evaluated different systems modules, such as term identification (sensitivity 78.8%, precision 71.7%), term disambiguation (85% accuracy) and relationship extraction (sensitivity 63%, precision 96%). The system and evaluations are described elsewhere in more detail (Friedman *et al.*, 2001; Hatzivassiloglou *et al.*, 2001; Krauthammer *et al.*, 2000). We are currently working on adding new functionalities to the system, such as synonym resolution and automated learning of new functional relationships as encountered in the journal articles (Hatzivassiloglou and Weng, 2002). This paper aims at a further system enhancement involving the development of a module that automatically derives consensus opinion from conflicting statements in articles.

While our work has some similarities with the concept of epidemic spread in small world networks (Watts, 1999), it can be best compared to methodologies applied in the field of citation analysis. These include modelling of observed citation frequency distributions or measuring citation diffusion rates. Attempts to model the first-citation distribution of journal articles (Burrell, 2001; Egghe, 2000), which describes the time delay between publication and the first citation, use stochastic approaches to fit a model to the observed citation distribution. Other studies are concerned with measuring the citation diffusion rate (Kortelainen, 2001), which can be used to describe the diffusion of a certain journal in an international context. Although these approaches can be compared to our methodology of modelling waiting times, there is an important qualitative difference between the two strategies. While citations stand for 'concept symbols' (Van der Veer Martens, 2001) representing past ideas or innovations, they are usually represented in a format that makes it difficult for a machine to understand the content of the 'concept symbol' (i.e., the nature of the idea or the innovation) unless the idea is explicitly stated in the title of the citation.

In contrast, our approach is based on the occurrence of research results in actual statements of the articles. For example, from the hypothetical sentence '*we thereby conclude that cbl phosphorylates abl*', the system extracts the crucial knowledge content

[action, phosphorylate, [protein,cbl], [protein,abl]]

in a frame-based knowledge representation format which is machine readable (for an explanation of the knowledge representation format see Friedman *et al.* (2001)). We are thus able to track actual research results as they are amplified throughout the research literature. Continuing the above example, a sentence such as '*we observed the phosphorylation of abl through cbl*', which appears in a journal article issued after the first publication of this molecular interaction, is crucial in determining the amplification (diffusion) of the molecular interactions. Using the same example, a sentence such as *'interestingly, cbl did not phosphorylate abl'* would help to establish a measure of contradiction involved in this molecular interaction.

The distribution of these properties (amplification and contradiction) can be attributed to well-described statistical processes. This allows for a robust statistical description of how molecular interactions are reported in the literature. From the resulting model, we draw some general conclusion about how truth evolves in the domain of molecular biology. We further show how the model can be used to infer consensus opinion automatically from contradictory research statements.

METHODS

Our methods can be subdivided into two main parts: The first part deals with the design of the 'noisy truth generator' model, which describes the evolution of a research idea from first conception and publication to amplification in the research literature. The second part consists of using data from actual research articles to justify assumptions inherent in the 'noisy truth generator' model.

The 'noisy truth generator' model

The model (Figure 1) is conceived to answer the following type of question: Given a set of statements about a particular molecular interaction, which appear in different journals articles published over the time period from first publication to the day of the analysis, what is the probability that these statements describe a true molecular interaction? That is, how well do the properties of the set of statements under study correspond to the normal life cycle of statements about true molecular interactions?

We start with the assumption that there is an imaginary device (a 'truth generator') that produces original statements about interactions between molecules (such as genes, proteins, RNA and small molecules). The generated statements can be either true or false, and they are not allowed to repeat. The statements that we have in mind are of the type *'protein X activates protein Y'* or *'small molecule Z binds protein X'*. Both molecule names and verbs (which we call 'actions') are chosen from finite but large lists.

The generated unique statements are accumulated in an imaginary vessel from which they enter various journals. Here we assume that true and false statements are entering the same journal at different (unknown) rates. Furthermore, these rates can vary for different journals.

Once a statement enters one of the journals, it is subject to amplification. We use the term amplification to indicate the repetition of already published statements in research articles. From our personal experience with the biological peer-review process, it appears that it is substantially easier to publish statements that have already appeared in publications than completely new ones. This may be because it is substantially easier for a reviewer to question the validity of new research results than already published ones.

Further, the amplified and original statements alike are subject to automatic extraction with our computer system, see Friedman *et al.* (2001). A proportion of statements is successfully extracted (this proportion is called *recall* in the natural-language processing community) and another, smaller, proportion of statements is lost. Furthermore, out of the extracted statements, only a proportion is extracted correctly (this proportion is commonly called *precision*) and the remaining, smaller, proportion is extracted with errors.

The observed data in this model are represented by a collection of extracted statements with references to corresponding sentences, articles to which these statements belong, and journals and publication dates corresponding to each article. The observed data also provide information about the number of times each statement appears in a printed form and the number of unique statements. Evaluation of the extracted information by human experts (qualified molecular biologists) can provide judgment (whether a statement is true or false) for each individual statement. Given the observed data and the model, we can estimate the model parameters via the maximization of the likelihood function (Maximum Likelihood Estimation). Then, given parameter estimates and some knowledge about the prior distribution of data, we can use the model to assign to each new unique statement (not used for training the model) a posterior probability that this statement is true.

Now we provide the analytical description of the model with mathematical specifics. We start the mathematical description by determining the probability of true and false statements entering the ith journal. The probability that N_i original statements enter the ith journal during time t

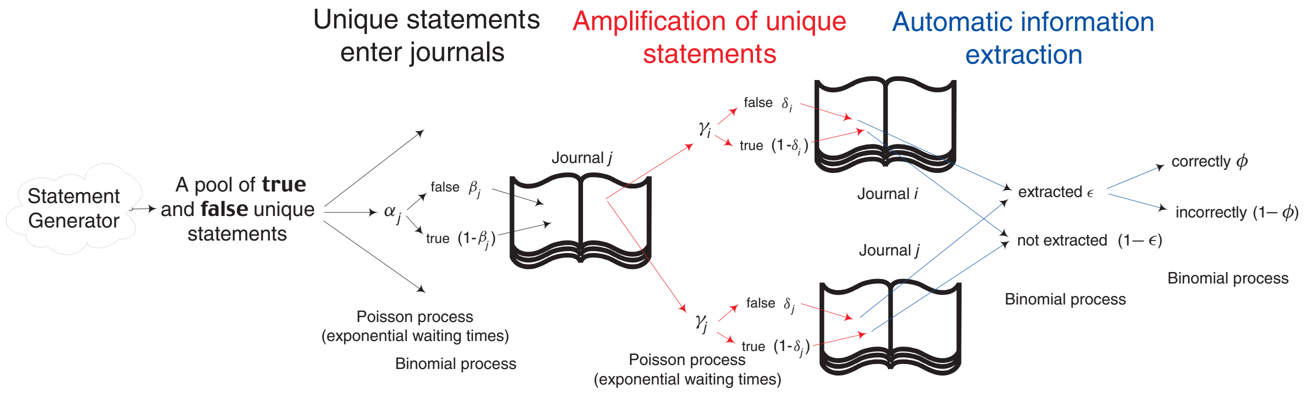

Fig. 1. The 'noisy truth generator model.'

under the model is given by

$$P(N_i|t, \alpha_i) = \frac{(\alpha_i t)^{N_i} e^{-(\alpha_i t)}}{N_i!}. \tag{1}$$

Here we assume that the statements arrive according a Poisson process with rate α_i (per unit time, so that the expected number of events during time t is $\alpha_i t$) specific to the ith journal. For this assumption to be correct, the waiting times between arrivals of individual unique statements should follow an exponential distribution.

Then, we compute the conditional probability that out of N_i individual statements $N_{i,T}$ are true and $N_{i,F}$ are false in the following way:

$$P(N_{i,T}|\beta_i, N_i) = \binom{N_i}{N_{i,T}} (1-\beta_i)^{N_{i,T}} \beta_i^{\,(N_i - N_{i,T})}. \tag{2}$$

Mathematically this is equivalent to an assumption that we have two independent Poisson processes, one, for true statements, with rate $[a_i(\beta_i - 1)t]$, and another, for false statements, with rate $[a_i \beta_i t]$.

Next, we assume that the amplification of statements also follows a Poisson process with rates that are different for true and false statements and different for distinct journals. Specifically, given an observation that a statement is amplified M_i times in the ith journal, the probability of this event, given it is true, is

$$P(M_i|T, \gamma_i, \delta_i, t) = \frac{[\gamma_i(1-\delta_i)t]^{M_i} e^{-[\gamma_i(1-\delta_i)t]}}{M_i!}. \tag{3}$$

The probability of observing M_i amplifications given the statement is false is then given by

$$P(M_i|F, \gamma_i, \delta_i, t) = \frac{[\gamma_i \delta_i t]^{M_i} e^{-[\gamma_i \delta_i t]}}{M_i!}. \tag{4}$$

The reader can see that the parameter γ_i has meaning as the average amplification rate per unit time for the ith journal, while (1-δ_i) has meaning as the proportion contributed to the total amplification for the journal by true statements.

Finally, turning to the information extraction system, the probability that the automated system extracts O_E statements from research articles while missing $[O - O_E]$ statements is given by binomial probability

$$P(O_E|\varepsilon, O) = \binom{O}{O_E} \varepsilon^{O_E} (1-\varepsilon)^{O-O_E}. \tag{5}$$

Similarly, the probability of extracting correctly O_{EC} out of O_E statements is given by

$$P(O_{EC}|\phi, O_E) = \binom{O_E}{O_{EC}} \phi^{O_{EC}} (1-\phi)^{O_E - O_{EC}}. \tag{6}$$

Now we are equipped to write an expression for the likelihood of the data for a specified set of parameter values. Given O_E extracted statements, we should assign them into groups 'extracted correctly' (O_{EC} statements) and 'extracted incorrectly' ($O_E - O_{EC}$ statements). Within the group 'extracted correctly' we should group statements into unique (the first chronological statement of each kind) and amplified statements (chronologically recent statements of the same kind) and compute for each unique statement and each (ith) journal M_i, the number of times the statements were amplified in the ith journal. Then, we compute the number of unique statements per journal, N_i. Each unique statement has to be assigned either into the class of true statements or false statements (this assignment does not need to be optimum at the first pass). The resulting likelihood is a product of the individual probabilities defined above.

The goal of this exercise is to assign extracted statements into the groups 'erroneously' and 'correctly extracted,' and for correctly extracted unique statements into the groups 'false' and 'true' that maximizes the posterior probability (given known parameter values)

$$P(\Gamma|D,\Theta) = \frac{P(D|\Gamma,\Theta)P(\Gamma)}{\sum_{\Gamma} P(D|\Gamma,\Theta)P(\Gamma)}, \qquad (7)$$

where Γ is the assignment of statements into groups of false and true, Θ is a set of parameters values, and D is the observed data.

Analysis of the research literature

We can assess whether the key assumptions in the Noisy Truth Generator model are reasonable by analysing the real pathway statements extracted from the research literature. In particular, in the model we assume that both unique and amplified statements arrive according to Poisson processes. By observing how often unique and amplified statements about molecular interactions enter the research journals we can examine the properties of the underlying stochastic process. In particular, the waiting times between consecutive arrivals of statements should follow an exponential distribution in order for the stochastic process to be Poisson.

We conducted our analysis on articles from seven peer-reviewed journals that either focus on molecular biology or cover a broad enough topic range to include articles on molecular biology. This set of journals resulted in a pool of over 50 000 full-text research articles covering a publication span from 1995 to 2001. As a first step, the GeneWays system described above processed the articles, which resulted in a dataset of over 700 000 total instances of statements about molecular interactions [†]. The sequential processing time for each article by subsequent system modules is approximately 2 minutes. The total set of articles was thus processed in 70 CPU days. We analysed the distributions of the following statistics:

- Waiting times between amplified statements: Defined as the time span between the two closest in time statements describing the same molecular interaction.
- Waiting times between unique statements: Defined as the time span between two subsequent unique statements. For the purpose of this study, a unique statement equals the first time description of a molecular interaction between two specific substances among all journals under analysis. Because the articles represented in the data set span a limited time period, a definite determination of the uniqueness of a statement is probably not possible. We therefore limited this part of the study to the time period of 1997–2001, and assumed that any statement during this period is unique unless it has been already mentioned in the time period of 1995–1996.

The waiting time for a new statement or for an amplification of that statement may be associated with the importance of the statement to the research community. One measure of importance is the connectedness of the substances involved in the statement, where connectedness is defined as the number of immediate molecular neighbours of a particular substance. A statement about a substance with many interactions may be more important to the community, and may undergo more frequent publication. We studied this hypothesis as follows:

- The connectedness of a molecular substance: Using our set of statements about molecular interactions, we calculated the connectedness of substances directly. We set a minimum threshold of ten instances mentioning a particular molecular interaction in order to count this interaction towards the connectedness of the substances involved.
- The average amplification of substances of the same connectedness: This property measures how many times substances of a particular connectedness are amplified in the research literature. More precisely, it is a measure not of the amplification of the substances per se, but rather of the statements of which these substances are a part.
- The average disagreement rate of substances of the same connectedness: This measure is based on the ratio of negatively stated molecular interactions divided by the total number of (positive and negative) statements on the same molecular interactions. This ratio is determined for each set of statements about a particular molecular interaction, which results in an average disagreement rate for the substances involved in these interactions. Because the GeneWays system does not yet extract experimental conditions for molecular interactions, disagreement between two statements does not necessarily mean that one of the statements is false. On the contrary, both statements may be true under different experimental conditions. We therefore limited this measure to so-called direct interactions between two substances, such as *phosphorylation* and *acetylation*. We assume that these kinds of interactions hold under all conditions, in contrast to indirect interactions, such as *expression*

[†] To reduce noise introduced in the GeneWays preprocessing step (term identification), we performed the data analysis on a smaller set than the initial 700 000 statements. We observed that interactions containing wrongly identified terms have a low amplification rate. We therefore limited the pool of statements by considering only statements which are repeated at least 4 times and in at least 2 journals. The final set thus consisted of approximately 100 000 statements corresponding to roughly 6500 molecular interactions. The difference between those two numbers is explained by repeated statements about the same molecular interaction.

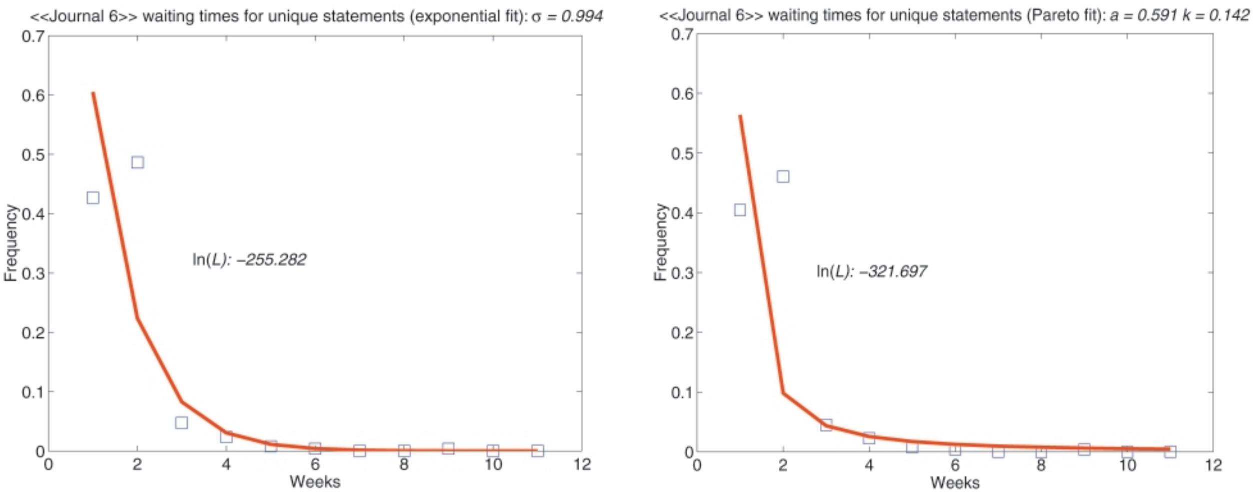

Fig. 2. Distribution of waiting times for amplified statements. Results from journal with IF 20–25, $n = 1360$ (waiting times), (a) exponential fit, (b) Pareto fit.

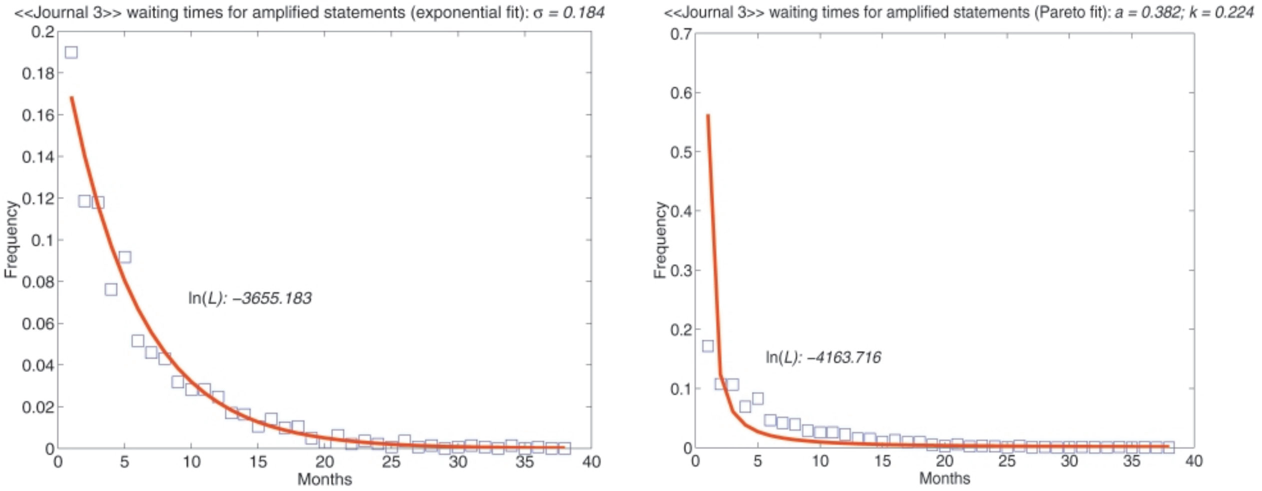

Fig. 3. Distribution of waiting times for unique statements. Results from journal with IF 5–10, $n = 254$ (waiting times), (a) exponential fit, (b) Pareto fit.

or *activation*, which are much more likely to be condition-dependent.

In order to justify our assumption that a Poisson process could describe the arrival of pathway statements we tried to fit an exponential distribution

$$P(x) = \delta e^{-\delta x} \qquad (8)$$

to the observed waiting time distributions. To estimate parameters of the exponential distribution, we used a maximum likelihood estimation procedure. The likelihood function is defined in the following way:

$$L(\delta|x_1, x_2, \ldots, x_n) = P(x_1, x_2, \ldots, x_n|\delta) = \prod_{i=1}^{n} \delta e^{-\delta x_i}, \qquad (9)$$

where each data point (x_i) corresponds to a single time interval between two closest consecutive statements published in research articles. The estimate of the parameter δ is obtained by maximizing the logarithm of likelihood function (we used a MatLab function implementing a simplex search method) and the estimate corresponds to the value of δ at the maximum of the likelihood function.

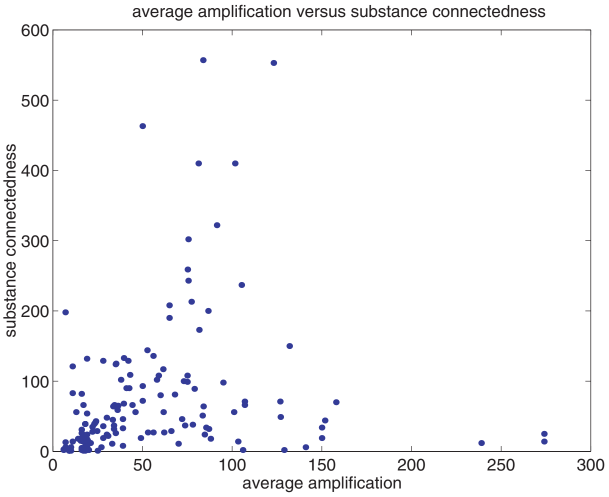

Fig. 4. Average amplification versus substance connectedness. $n = 142$ (substances).

The value of the likelihood function at the point of maximum is useful in direct comparison to the fit of rival models (with a close number of parameters) based on the same data. For example, we compared the maximum likelihood fit of a Pareto distribution to waiting times data, as described in Results.

RESULTS AND DISCUSSION

Our analysis favours the conclusion that Poisson modelling adequately portrays generation and amplification of research ideas in research literature. Figure 2 and Figure 3 show the distribution of the waiting times of unique and amplified statements as observed in the 50 000 biological articles. As expected under a Poisson model, the waiting times between events appear to follow an exponential distribution, and the maximum likelihood analysis indicates that the exponential distribution fits the data significantly better than alternatives, such as a Pareto distribution. From these data, we have calculated the Poisson parameters α and γ for unique and amplified statements, respectively (Table 1). The variances of those parameters indicate significant differences among the journals studied.

We hypothesized, tested and confirmed a few properties of the random process of knowledge generation that could be useful in future modelling. First, by limiting the analysis to statements encountered in at least 5 journals, we found that the connectedness of a substance was significantly associated ($p = 0.016$) with amplification rates of the related statements (see Figure 4) with a correlation coefficient of 0.20 (95% CI 0.04 to 0.38). We further observed that connectedness was associated ($p = 0.016$) with the level of disagreement with a correlation coefficient of -0.26 (95% CI -0.45 to -0.04), implying that statements with greater connectedness underwent less disagreement. As described earlier, if connectedness is related to a subjective measure of importance, these findings hint at some biases in the reporting of molecular interactions. We extended the latter idea and measured correlations between journal impact factor and connectedness as well as unique statements that appear in a journal of interest. The hypotheses that journals with a high impact factor discuss mostly highly connected substances (Figure 5) or introduce a higher degree of new research ideas, i.e., unique statements (Figure 6), appear to have some support in data, but more journals need to be studied for a statistically conclusive result.

Table 1. Estimated parameters for 7 journals ranked according to Impact Factor (IF). $\widehat{\alpha}$ = the estimated number of unique molecular interactions per month, $\widehat{\gamma}$ = the estimated number of amplifications of molecular interactions per month

Journal	*IF*	$\widehat{\alpha}$	$2\sqrt{\widehat{Var}(\widehat{\alpha})}$	$\widehat{\gamma}$	$2\sqrt{\widehat{Var}(\widehat{\gamma})}$
1	40 – 45	0.84	0.31	0.23	6.0×10^{-2}
2	30 – 35	6.65	0.71	0.17	5.6×10^{-3}
3	20 – 25	7.03	0.72	0.20	4.7×10^{-3}
4	10 – 15	4.75	0.60	0.18	5.3×10^{-3}
5	5 – 10	15.35	1.03	0.29	2.0×10^{-3}
6	5 – 10	0.50	0.58	0.41	7.6×10^{-2}
7	0 – 5	1.04	0.31	0.22	1.3×10^{-2}

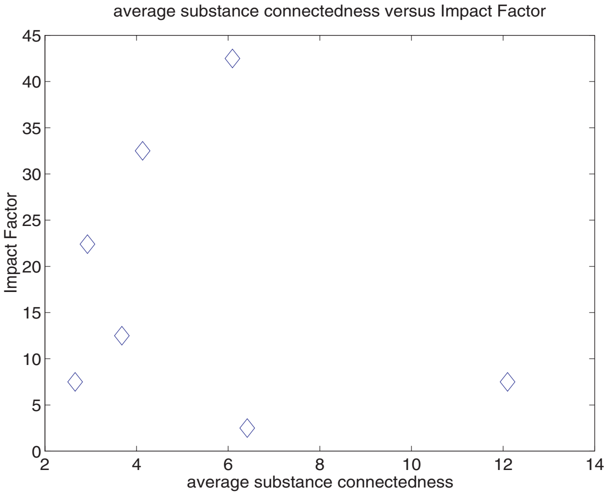

Fig. 5. Journal Impact Factor versus average substance connectedness.

Molecular interaction data are useful for characterization of protein interaction networks (Jeong *et al.*, 2001) or for inference about properties of model organisms (Karp,

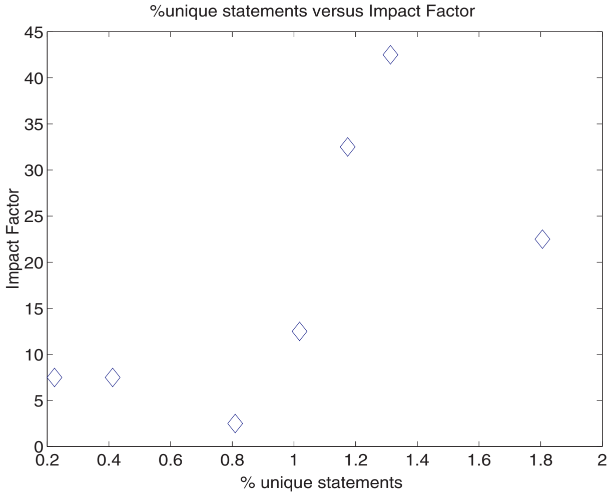

Fig. 6. Journal Impact Factor versus percentage of unique statements.

2000; Karp *et al.*, 1998) represented in functional ontologies. A common source of molecular interaction data is the research literature, and there has been a growing effort to build systems that can automatically collect such information (Friedman *et al.*, 2001; Yakushiji *et al.*, 2001; Thomas *et al.*, 2000; Ng and Wong, 1999; Sekimizu *et al.*, 1998) directly from the research articles. As research ideas and results evolve over time, this information represents a collection of potentially contradicting statements and opinions. A reasonable hope is that the 'truth' can be inferred from the pooled collection of statements on the same research topic. While domain experts are usually good at this task, without a proper computational resolution of conflicting research results, it is hardly possible to achieve a useful integration of this tremendously vast and valuable knowledge.

While, in the long term, aiming at automated resolution of conflicting research results, this paper describes and analyses the type and variation of the data typically generated by such knowledge extraction systems. These data are rather different from citation studies by tracing specific research results rather than citations, which describe much less precisely the content and results of the research under the study. We thus are able to get a unique glimpse into how research results and ideas are published and reproduced in the research literature.

As the most basic assumptions of our Noisy Truth Generator model appear reasonable from our analysis, we hope that the model will be useful for calculating a probability that a set of statements about a certain molecular interaction is true. In other words, given the scenario that we have two conflicting statements, such as '*A activates B*', which has been stated in fifty articles appearing in mostly less known journals, and a second statements, '*A does not activate B*', which is stated in ten articles appearing in high impact factor journals, the model should be able to conclusively tell which one of the two statements is more likely to be correct.

The analysis of the actual data brings about important question about distinction between the 'true statements' and statements accepted by a research community. It is quite likely that analysis of textual data alone is sufficient for inferring the latter, but not the former. However, as independent high-throughput experimental techniques deliver additional data, the inference based on integration of multiple data types should converge to truth.

As this is our first detailed analysis of literature interaction data, many questions need to be further investigated. For example, we need to address how system recall and precision influence the results of the analysis. Our system and the design of the analysis is geared towards high precision and we expect quanitative changes in parameters such as the rate of unique statements as we improve the system recall. Even with higher recall, we expect that our observations about distributions of waiting times still hold.

There is ample room for modification and adjustment of the model. For example, one natural extension to the model would be adding a probabilistic 'latent phase' between publication of an original statement and the first amplification of this statement. This would immediately allow distinguishing a true original statement with better precision from an amplification of an older (outside the scope of the corpus under analysis) original statement. However, we believe that even the simplest model outlined in this report is not devoid of utility and interest.

This work was supported by NSF ITR and NIH RO1 awards to A.R.

REFERENCES

Burrell,Q.L. (2001) Stochastic modelling of the first-citation distribution. *Scientometrics*, **52**, 3–12.

Egghe,L. (2000) A heuristic study of the first-citation distribution. *Scientometrics*, **48**, 345–359.

Friedman,C., Kra,P., Yu,H., Krauthammer,M. and Rzhetsky,A. (2001) Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**, S74–S82.

Hatzivassiloglou,V., Duboue,P.A. and Rzhetsky,A. (2001) Disambiguating proteins, genes, and rna in text: a machine learning approach. *Bioinformatics*, **17**, S97–S106.

Hatzivassiloglou,V. and Weng,W. (2002) Learning anchor verbs for biological interaction patterns from published text articles. In *Proceedings of the Workshop on Natural Language Processing in Biomedical Applications*. pp. 81–86.

Iliopoulos,I., Enright,A.J. and Ouzounis,C.A. (2001) Textquest: document clustering of medline abstracts for concept discovery in molecular biology. *Pac. Symp. Biocomput.*, 384–395.

Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Karp,P.D. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.

Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. (1998) Ecocyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **26**, 50–53.

Koike,T. and Rzhetsky,A. (2000) A graphic editor for analyzing signal-transduction pathways. *Gene*, **259**, 235–244.

Kortelainen,T.A.M. (2001) Studying the international diffusion of a national scientific journal. *Scientometrics*, **51**, 133–146.

Krauthammer,M., Rzhetsky,A., Morozov,P. and Friedman,C. (2000) Using Blast for identifying gene and protein names in journal articles. *Gene*, **259**, 245–252.

Mrowka,R., Patzak,A. and Herzel,H. (2001) Is there a bias in proteome research? *Genome Res.*, **11**, 1971–1973.

Ng,S.K. and Wong,M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform Ser Workshop Genome Inform*, Using Smart Source Parsing, **10**, pp. 104–112.

Rzhetsky,A., Koike,T., Kalachikov,S., Gomez,S.M., Krauthammer,M., Kaplan,S.H., Kra,P., Russo,J.J. and Friedman,C. (2000) A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics*, **16**, 1120–1128.

Sekimizu,T., Park,H.S. and Tsujii,J. (1998) Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome Inform. Ser. Workshop Genome Inform*, Using Smart Source Parsing, **9**, pp. 62–71.

Shatkay,H., Edwards,S., Wilbur,W.J. and Boguski,M. (2000) Genes, themes and microarrays: usinginformation retrieval for large-scale gene analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 317–328.

Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pac. Symp. Biocomput.*, 529–540.

Stephens,M., Palakal,M., Mukhopadhyay,S., Raje,R. and Mostafa,J. (2001) Detecting gene relations from medline abstracts. *Pac. Symp. Biocomput.*, 483–495.

Stevens,R., Goble,C.A. and Bechofer,S. (2000) Ontology-based knowledge representation for bioinformatics. *Briefings In BioInformatics*, **1**, 398–414.

Thomas,J., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.*, 541–52.

Van der Veer Martens,B. (2001) Do citiation systems represent theories of truth? *Information Research*, **6**, http://informationr.net/ir/6-2/paper92.html.

Watts,D.J. (1999) Small worlds: the dynamics of networks between order and randomness. *Princeton studies in complexity*. Princeton University Press, Princeton, NY.

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Yakushiji,A., Tateisi,Y., Miyao,Y. and Tsujii,J. (2001) Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.*, 408–419.