Supporting Text

1 Introduction

Every scientist knows that there is a gap (or a precipice) between a scientific observation and a published scientific assertion: Interpretation of experimental observations by an individual scientist is influenced by prior conclusions published by her peers, an effect known in economics world as "information cascade." In this work we modeled and evaluated the importance of sequences of related assertions in biological publications. We were especially interested in estimating parameters of the real process, defining the most prominent patterns of dependencies among scientific publications, and addressing the possibility of modifying the peer-review process to increase the probability that a correct assertion will be established after a few consecutive scientific publications.

We attempted mathematical modeling of the work of a biological scientific community, and focused on patterns of publications about molecular interactions, such as protein–protein binding, because these interactions have both practical importance and relative clarity of definition, and because we had available many data from our ongoing information-extraction project.

We are describing here two stochastic models (a "naïve" and a "reasoning chain" models) that can generate binary chains of variable length–resembling chains of statements about molecular interactions that are published in research articles in numerous biological periodicals.

In designing our "reasoning chain" model we strove to make it as general and practical (in the sense that all parameters *can* be estimated) as possible. The model also went through a series of alternative specifications—we started with more parameter-rich models and eliminated non-essential parameters that did not affect the model performance.

2 Outline of the modeling approach and notations

For compactness of presentation we will often denote *all* parameters of the model with symbol Θ—for example, for the reasoning chain model $\Theta = \{\alpha, \iota, \gamma, \tau, \theta, \rho, \nu, \mu, \varphi, \beta, \eta\}$—individual notations are be explained in the following section.

Our data for the two models is represented by series of chronologically arranged statements about molecular interactions. The i^{th} sequence of statements about the same interaction is denoted with c_i (c stands for *chain*). The number of missing (unpublished) statements that would be normally included into the j^{th} chain in the data, is denoted with y_j. The complete data therefore has the observed and missing components: $\mathbf{C} = \{c_1, c_2, \ldots, c_m\}$ and

$\mathbf{Y} = \{y_{\text{start}}, \{y_1, y_2, \ldots, y_m\}\}$, respectively.

We assume that the probability of data (sequences of chronologically ordered published statements about molecular interactions) given model and model parameter values, M and Θ, respectively, has two independent components, chain length and chain pattern.

$$P(chains|M,\Theta) = \prod_{all\ actions} P(chain\ length_i|M,\Theta)P(chain\ pattern_i|chain\ length_i,M,\Theta). \tag{1}$$

Equation 1 tells us that each chain of reasoning is a result of two random processes: one of them determines the length of the chain, whereas the other specifies the arrangement of zeros and ones within the chain given that length. The second random process (see figure 2 A) is responsible for generating a specific sequence of zeros and ones within a chain of a given length. The model that we propose here takes into account unequal probabilities of encountering positive and negative "truths" about molecular interactions, the possibility of having an exception to a rule, and the possibility of false-positive and false-negative errors in wet-laboratory experiments.

(In the following text for the sake of brevity we will omit the symbol for model, M, in the conditional probability equations, but, of course all parameters, Θ have meaning only in a context of a model.)

The real chains of reasoning extracted by our system look as shown in figure 6.

3 Data used in the analysis

We have to admit that we do not know the extent to which automatically extracted information from the literature accurately reflects biological knowledge. However, the results from our analysis actually provide the first look at this problem. Our "chain-reasoning" model was developed in part as an attempt to measure the accuracy to which published results reflect biological knowledge.

Before analyzing the whole database, we studied a smaller manually cleaned dataset (a few thousand interactions)—the smaller set behaved exactly in the same way (with a large variances as the whole database. It is important to note that the model was specifically designed to describe data that were automatically extracted from literature and to "clean" the data—the whole point of doing a probabilistic modeling is to factor in a possibility of an error. This allows our analysis seems to be a valid first-order approach that gives answers to previously unanswerable questions.

We did our analysis in two ways. In one version of the analysis we used only one (most frequent) statement per article—the hypothetical conflict among statements within the same article would not occur in this analysis. In the other version of analysis we used all statements available in the database. The results of the two analyses are qualitatively indistinguishable.

(It is a good place to clarify the meaning of β-parameter: it is a mean proportion of articles that repeat information about known publications (without an additional experiment) at least once. If we would take into account multiple statements in the same article, β would be larger than our estimated β.)

4 Assumptions behind the "reasoning chain" model

The "reasoning chain" model is built on eight simple and intuitive assumptions.

First, we assume that for every pair of substances, there is a general truth or rule: These substances either usually do or usually do not interact. The odds of encountering a negative rule ("A usually does not interact with B") are not necessarily the same as the odds of encountering a positive rule ("C usually does interact with D"); we denote the corresponding probabilities by $1 - \rho$ and ρ, respectively.

Second, each general rule may have an exception, with probability ϕ (e.g., proteins A and B interact in most cases, but do not interact when in tissue X).

Third, we allow experiments to produce erroneous results: They produce false-negative results with probability ν and false-positive results with probability μ.

Fourth, we assume an asymmetry in terms of ease of publication between negative and positive experimental results. Many experimentalists believe that it is more difficult to publish a negative result ("we were unable to demonstrate that A and B interact") than to publish a positive result ("we demonstrated that A and B interact"), so the model allows negative results to be discarded, without publication, with probability $1 - \eta$.

Fifth, we assume that a published statement can be based on original experiments (with probability $1 - \beta_i$) or can be a re-statement of an earlier published statement (with probability β_i). We tested two formulations of the model: the simpler version assumes that β_i is constant, while, in the more complicated version of the model, β_i is increasing as the chain grows longer: $\beta_i = 1 - i^{-\Psi}, \Psi > 0$. The more complicated formulation asserts that the chances that a scientist would experimentally re-verify an old statement drop with the growth of the available evidence. We assume that the first statement in every chain is always supported by an experiment.

Sixth, we allow an experimenters interpretation of her own data (and hence of her published result) to differ from the "unbiased" interpretation of the same data that an expert would have in the absence of prior publications. This model feature reflects our observa-

tion that, when reading about published experiments similar to their own, scientists build in their minds an equivalent of statistical prior distributions of experimental outcomes that they are using for interpreting their own experimental data. We assume that each published statement has a weight that is different for statements in reasoning chains where they are in the majority (α), are the minority (ι), and are of equal number (τ). For example, for the chain of reasoning 1, 0, 0, 1, 1, 1, every published positive statement would have weight α (because it is in the majority), whereas each negative statement would have weight ι. For the hypothetical chain 0, 1, 0, 1, the weight of each the statement would be equal (τ), because there are an equal number of zeros and ones. The weight of each published statement is non-negative and reflects the importance of published statements in influencing both a researchers choice of experiments (and thus ultimately observed results) and her interpretation of the results. We set the subjective weight of the researchers own experiment to 1.[1]

Seventh, we assume that relationship among statements related to the same molecular interaction is adequately represented with a linear structure (a chain), see figure 7. (A more complicated model that allows directed acyclic graph dependencies among published statements can be defined relatively easily but it would be very expensive computationally.)

Eighth, we assume that different chains are statistically independent.

5 Notations—in detail

The variables and parameters of the two models that we describe here are as follows.

T $\longrightarrow$ is an unknown *true rule* about interaction between a pair of molecules, $T = 1$ if the molecules usually interact under appropriate conditions and $T = 0$ otherwise. (For example "BCL-2 binds BAX" is a positive statement, while "BCL-2 does not bind BAX" is a corresponding negative statement.)

T_i' $\longrightarrow$ an instance of *true rule*—may differ from the rule (an exception).

ε_i $\longrightarrow$ a *hidden* (private for an individual experimentalist) *result* about a molecular interaction. $\varepsilon_i = 1$ if the hidden experiment suggests that the molecules do interact and $\varepsilon_i = 0$ otherwise.

O_i $\longrightarrow$ an indicator variable, that is equal to 0 if a hidden experiment with negative result ($\varepsilon_i = 0$) is discarded (is missing), and is equal to 1 if the hidden negative value ($\varepsilon_i = 0$) participates in an information cascade. According to our model, *a*ll positive hidden statements ($\varepsilon_i = 1$) are used in cascades, that is $P(O_i = 1|\varepsilon_i = 1) = 1$.

[1] We are not at all suggesting that scientists publish something that they don′t see in the experiments—on the contrary, we suggest that each reasoning chain forms a micro-paradigm that changes scientists′ choice of experiments to perform and the interpretation of their data. In the middle of a long reasoning chain scientists honestly see 1 where they would see 0 in the absence of prior publications.

H_i $\longrightarrow$ an indicator variable which is equal to 1 if researcher performs a (hidden) experiment before repeating a statement in publication, and equal to 0 if the researcher repeats the statement without additional experiments.

E_i $\longrightarrow$ a *published* statement about a molecular interaction. $E_i = 1$, if the statement is positive, and $E_i = 0$, if the statement is negative.

ρ $\longrightarrow$ $P(T = 1)$, the probability of sampling a positive rule about molecular interactions.

ν $\longrightarrow$ probability of getting a single false negative result, $P(\varepsilon_i = 0|T'_i = 1)$ in an experiment.

μ $\longrightarrow$ probability of getting a single false positive result, $P(\varepsilon_i = 1|T'_i = 0)$ in an experiment.

ψ $\longrightarrow$ probability of obtaining a false experimental result, $P(\varepsilon_i \neq T)$.

ϕ $\longrightarrow$ probability of observing an exception to a rule, $P(T \neq T'_i)$.

θ $\longrightarrow$instantaneous rate of discovery of new interactions ("innovation").

λ $\longrightarrow$ instantaneous rate of repeating known statements ("amplification").

γ $\longrightarrow$ subjective attractiveness of a popular (high copy number) statement.

η $\longrightarrow$ probability that a negative experimental observation is *published*. On average $1 - \eta$ of all negative results will not be published and would constitute *missing* data.

β $\longrightarrow$ probability of publishing a statement without doing an additional (hidden) experiment.

α $\longrightarrow$ a subjective weight of a single external statement representing the m**A**jority opinion.

ι $\longrightarrow$ a subjective weight of a single external statement representing the m**I**nority opinion.

τ $\longrightarrow$ subjective weight of a single external statement in a **T**ie situation.

$I_{0,i-1}$ $\longrightarrow$ the total number of *negative* statements in a chain of $i - 1$ statements about the same molecular interaction, $\{E_1, E_2, \ldots, E_{i-1}\}$.

$I_{1,i-1}$ $\longrightarrow$ the total number of *positive* statements in a chain of $i - 1$ statements about the same molecular interaction, $\{E_1, E_2, \ldots, E_{i-1}\}$.

$y_{\mathbf{start}}$ $\longrightarrow$ number of *missing* cascades that would start with zero.

y_i $\longrightarrow$ number of *missing* hidden negative experiments ($\varepsilon_j = 0$) in the ith cascade in the data.

6 Chain length distribution

6.1 Chain length: "Naïve" model

In our earlier paper [1] we introduced a "naïve" model of publication process.

The *naïve* model, assumes that scientists produce new assertions at an approximately steady rate; once published the first time, each new assertion has a chance of being repeated in the scientific literature (see figure 8). We refer to publication of a new assertion as "innovation," and to re-publication of a known assertion as "amplification"; we postulate that both processes are Poisson processes, with rates different for innovation and amplification, and different for amplification of *false* and *true* assertions.

Implicitly, we hope that the true assertions are both generated and amplified with higher mean rates than the false assertions. More specifically, we assume that the probability of generating l new statements over time interval t given innovation rate θ is given by

$$P(l|\theta,t) = \frac{e^{-\theta t}(\theta t)^l}{l!}, \tag{2}$$

and the probabilities of amplifying a true and a false assertions during time interval t, given amplification rates λ_{true} and λ_{false} are as follows.

$$P(k|\lambda_{true},t) = \frac{e^{-\lambda_{true}t}(\lambda_{true}t)^k}{k!}, \tag{3}$$

$$P(n|\lambda_{false},t) = \frac{e^{-\lambda_{false}t}(\lambda_{false}t)^n}{n!}. \tag{4}$$

Combining innovation and amplification steps into a single equation, we can find distributions of the number of amplified true and false setatements. Specifically, under the naïve model, the distribution of the number of times that an assertion is mentioned in the literature has the shape of a smoothed step function (equation 5).

$$P(k|\lambda_x,t) = \frac{1}{\lambda_x t}\left[1 - e^{-\lambda_x t}\sum_{j=0}^{k-1}\frac{(\lambda_x t)^j}{j!}\right]. \tag{5}$$

where x can take values *false* or *true*. The chain length distribution is mixture of distributions of false and true statements:

$$P(m|\lambda_{true},\lambda_{false},t) = \psi P(m|\lambda_{false},t) + (1-\psi)P(m|\lambda_{true},t). \tag{6}$$

where k is the length of a chain, λ is an amplification rate per time unit (subscripts $true$ and $false$ refer to true and false statements, that are assumed to have different amplification rates), and t is the time since the first statement in the series was published. Variable ψ denotes the expected proportion of false statements in the published articles, where

$$
\begin{aligned}
\psi &= P(\varepsilon_1 \neq T|\Theta) \\
&= P(T=1|\Theta)P(T'=1|T=1,\Theta)P(\varepsilon_1=0|T'=1,\Theta) \\
&+ P(T=1|\Theta)P(T'=0|T=1,\Theta)P(\varepsilon_1=0|T'=0,\Theta) \\
&+ P(T=0|\Theta)P(T'=1|T=1,\Theta)P(\varepsilon_1=1|T'=1,\Theta) \\
&+ P(T=0|\Theta)P(T'=0|T=1,\Theta)P(\varepsilon_1=1|T'=0,\Theta), \qquad (7)
\end{aligned}
$$

which leads us to

$$
\begin{aligned}
\psi &= \rho(1-\phi)\nu + \rho\phi(1-\mu) \\
&+ (1-\rho)\phi(1-\nu) + (1-\rho)(1-\phi)\mu. \qquad (8)
\end{aligned}
$$

The distribution of the combined *chain length* for true and false statements under the naïve model is as shown in Figure 9 (green line and green dots for theoretical and simulated distributions, respectively).

The *real* chain length distribution turned out to be as shown in Figure 11 in this document (approximately Zipf-Estoup or discrete Pareto distribution), which is very different from the distribution expected under the "naïve" model.

In the new *reasoning chain model*, the innovation process is not constant in time, but rather depends on the total frequency of assertions that are given more than one mention in literature; the amplification process depends on the number of published mentions of a assertion, and on a *vanity* parameter that reflects the subjective (to a researcher) attractiveness of a popular statement, γ. This part of the cascade model resembles the model of a Yule process ([2] , p. 450). This new model for chain length distribution produces exponential-like distributions of chain length if γ is small, and is similar to a Zipf-Estoup (discrete Pareto) distribution if γ is greater than 1 (see Figure 11).

The newer model ("chain of collective reasoning") for chain length distribution fits the real data much better (see the next section).

6.2 Chain length: "Reasoning chain" model

The second model is slightly more involved.

We assume a Markovian version of the previous model where the mean innovation rate (rate at which original statements are generated) can vary in time.

Postulates for our Markovian/Poisson process (which resembles Yule process, see [2], vol. 1, p. 450) are as follows. (*i*) Direct transitions from state i are possible only to state $i+1$. (*ii*) If at epoch t system is in state k, the probability of jump to state $k+1$ within short interval between t and $t+h$ equals $\lambda h k^{\lambda} p_k(t) + o(h)$, while probability of more than one jump is $o(h)$. (*iii*) Innovation process is defined as increase of probability at state 1 within

short interval between t and $t+h$ proportional to $(1-p_1(t))\theta h$. This increase in the state 1 is balances by decrease in all the remaining states proportional to $-p_k(t)\theta h$.

To make the following derivation more compact we introduce the following notation.

$$p_k(t) \stackrel{\text{def}}{=} P(k|t,\lambda,\theta,\gamma). \tag{9}$$

These assumptions are sufficient to derive the following system of differential equations

$$\begin{cases} \frac{\partial p_1(t)}{\partial t} = \theta\left[1-p_1(t)\right] - \lambda p_1(t), \\ \cdots \\ \frac{\partial p_k(t)}{\partial t} = \lambda(k-1)^\gamma p_{k-1}(t) - \left[\theta + \lambda k^\gamma\right] p_k(t), \text{ for } k>1 \\ \cdots \end{cases} \tag{10}$$

As is clear from analysis of system 10, processes of innovation and duplication are competing with each other when the chain length distribution is considered: each time a novel statement is published, the proportion of chains with length 1 (p_1) increases and therefore proportion of longer chains, $p_k(k>1)$, decreases, because the sum over all p_k is always equal to one. On the other hand, amplification process increases the proportion of chains with length more than one (because it extends chains of length 1 to longer chains).

System 10 turned out to have a nice closed-form solution:

$$p_k(t) = \frac{\lambda^{k-1}\left[(k-1)!\right]^\gamma}{\prod_{j=1}^{k}(j^\gamma\lambda+\theta)}\left[\theta + \lambda e^{-(\lambda+\theta)t}\right]. \tag{11}$$

We need time parameter in the modeling because in our database we store absolute time points (publication date of individual journals) for each statement extracted from literature—therefore we can work with real time (in weeks or days) and estimate for each journal the absolute rate of publishing totally new statements (θ) and the rate of repetition of old statements ($\lambda = \lambda_{true} + \lambda_{false}$).

Unless extremely small time intervals are considered (which is not the case in our model), it is most natural to use a continuous-time model—it would be more difficult to justify a discrete-time Markov model in this case.

The θ parameter is important in comparing multiple journals that publish novel and repeated statements at different rates. We plan to estimate such parameters for numerous journals (we have 81 journals in the current GeneWays database)—this information is invaluable for automated assignment of posterior probability of being true for all statements in the database (which is the ultimate purpose of modeling that we do).

Incidentally, we used the system of equations 10 earlier for description of a growing proteome, see [3].

7 Chain content

7.1 Chain content: "Naïve" model

Under the "naïve" model all statements within the chain are independently distributed, so that all possible patterns with the same number of zeros and ones are equally likely to occur in the published articles.

7.2 Chain content: "Reasoning chain" model

In the "Reasoning chain" model a probabilistic dependence among statements within the same chain is explicitly built into the model (see figure 1).

T is the general rule or "truth"—a statement about molecular interactions that is correct in the majority of real-life cases.

$$\begin{aligned} P(T=1|\Theta) &= \rho, & (12)\\ P(T=0|\Theta) &= 1-\rho. & (13) \end{aligned}$$

T' is a truth instance that may be different from the general rule or "truth"—for example, "BAD binds BAX" in the majority of animals, but an experimenter happened to sample an animal with a mutant form of BAD that fails to bind BAX.

$$\begin{aligned} P(T'_i=\bar{T}|\Theta) &= \varphi, & (14)\\ P(T'_i=T|\Theta) &= 1-\varphi. & (15) \end{aligned}$$

ε_i is the outcome of the i^{th} experiment about molecular interaction. With probabilities ν and μ results may be false-positive or false-negative with respect to the instance of truth, T'.

$$\begin{aligned} P(\varepsilon_i=0|T'_i=1,\Theta) &= \nu, & (16)\\ P(\varepsilon_i=1|T'_i=1,\Theta) &= 1-\nu, & (17)\\ P(\varepsilon_i=1|T'_i=0,\Theta) &= \mu, & (18)\\ P(\varepsilon_i=0|T'_i=0,\Theta) &= 1-\mu. & (19) \end{aligned}$$

So long as the instance of truth, T', can be different from the general truth, T, even a perfectly performed experiment may result in a conclusion, ε_i, that is different from the general truth T.

$$
\begin{aligned}
P(\varepsilon_i = 0|T = 1, \Theta) &= P(\varepsilon_i = 0|T_i' = 0, \Theta)P(T_i' = 0|T = 1, \Theta) + && (20)\\
&\quad P(\varepsilon_i = 0|T_i' = 1, \Theta)P(T_i' = 1|T = 1, \Theta) && (21)\\
&= (1-\mu)\varphi + \nu(1-\varphi) = 1 - \Lambda_2, && (22)\\
P(\varepsilon_i = 1|T = 1, \Theta) &= (1-\nu)(1-\varphi) + \mu\varphi = \Lambda_2, && (23)\\
P(\varepsilon_i = 1|T = 0, \Theta) &= (1-\nu)\varphi + \mu(1-\varphi) = \Lambda_1, && (24)\\
P(\varepsilon_i = 0|T = 0, \Theta) &= (1-\mu)(1-\varphi) + \nu\varphi = 1 - \Lambda_1. && (25)
\end{aligned}
$$

For the chain links following the first link, each new statement (link) can be either supported by a new experiment ($H_i = 1$) or it can be a repetition of an earlier published statement in the absence of new experimental support ($H_i = 0$).

7.3 Statements unsupported by an experiment are allowed to participate in a chain

We allow some of the statements in the reasoning chain to be unsupported by experimental data—this is because for most humans frequently repeated statements (even without additional evidence) often get etched in the memory as facts. For example, if one hears X times that mister Y is a bad person (even without evidence) one might start to believe the statement at some point. That is why we explicitly account for repeated statements about interactions, as opposed to statements describing results newly presented. These repeated statements are typically not based on general knowledge, but instead are references to specific papers and experiments, as one might find in the introduction and background section of a paper.

$$
\begin{aligned}
P(H_1 = 1|\Theta) &= 1, && (26)\\
P(H_{i>1} = 1|\Theta) &= \beta, && (27)\\
P(H_{i>1} = 0|\Theta) &= 1 - \beta. && (28)
\end{aligned}
$$

In our mathematical treatment of the sequence of statement at each point of the chain, the value of statement E_i is completely determined by the counts of conflicting statements ($I_{1,i-1}$ and $I_{0,i-1}$) preceding the i^{th} link in the chain:

$$P(E_i|E_1, \ldots, E_{i-1}, \Theta) = P(E_i|I_{1,i-1}, I_{0,i-1}, \Theta). \quad (29)$$

7.4 Interpretation of data in the chain extension

The major difference between the first and all the next links in a chain comes at the point of data interpretation. If the new link is supported by a new experiment, interpretation of this new experiment would be done in the context of the previously published data, that can bias the resulting conclusion. That is, the published statement, E_i can be different from the "unbiased" experimental result, ε_i.

$$P(E_i = x|\varepsilon_i = y, H_i = 1, I_{1,i-1} > I_{0,i-1}, \Theta) = \frac{x\alpha I_{1,i-1} + (1-x)\iota I_{0,i-1} + \delta(x,y)}{\alpha I_{1,i-1} + \iota I_{0,i-1} + 1}, \quad (30)$$

where

$$\delta(x,y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases} \quad (31)$$

$$P(E_i = x|\varepsilon_i = y, H_i = 1, I_{1,i-1} < I_{0,i-1}, \Theta) = \frac{x\iota I_{1,i-1} + (1-x)\alpha I_{0,i-1} + \delta(x,y)}{\iota I_{1,i-1} + \alpha I_{0,i-1} + 1}, \quad (32)$$

$$P(E_i = x|\varepsilon_i = y, H_i = 1, I_{1,i-1} = I_{0,i-1}, \Theta) = \frac{x\tau I_{1,i-1} + (1-x)\tau I_{0,i-1} + \delta(x,y)}{\tau(i-1) + 1}. \quad (33)$$

Clearly, the larger are the values of the momentum parameters α and ι, the smaller is correlation between the result of the new experiment (ε_i) and the resulting value of E_i.

In the case the new statement is a mere repetition of an older statement, there is no hidden experiment. If all statements in the chain agree, the repeated statements have no choice but agree with the beginning of the chain. In the case when chain contains both zeros and ones, the value of the repeated statement depends only on the counts of the conflicting statements at the beginning of the chain and values of momentum parameters:

$$P(E_i = 1|H_i = 0, I_{1,i-1} > I_{0,i-1}, \Theta) = \frac{\alpha I_{1,i-1}}{\alpha I_{1,i-1} + \iota I_{0,i-1}}, \quad (34)$$

$$P(E_i = 1|H_i = 0, I_{1,i-1} < I_{0,i-1}, \Theta) = \frac{\iota I_{1,i-1}}{\iota I_{1,i-1} + \alpha I_{0,i-1}}, \quad (35)$$

$$P(E_i = 1|H_i = 0, I_{1,i-1} = I_{0,i-1}, \Theta) = \frac{1}{2}. \quad (36)$$

Finally, if the result of experiment is negative a researcher may choose to discard it without publishing with probability $1 - \eta$.

$$P(O_i = 1|\varepsilon_i = 0, \Theta) = \eta, \quad (37)$$

$$P(O_i = 1|\varepsilon_i = 1, \Theta) = 1. \quad (38)$$

The possibility of discarding the negative results stems from observation that only 4% of published statements are negativeand it is very likely that the negative results occur in experiments much more frequently. Therefore it is very likely that a significant portion of

negative results is simply discarded. We needed to reflect this possibility in our model to test whether this feature would change our parameter estimates. This missing data feature appears to have very little influence on other parameter estimates (and is very expensive computationally), but we had no way to know this without testing.

Many negative results may not be published because the scientist feels there is no reason to publish a negative result—there, of course, can be many other reasons why a negative result is not published, including journal policy, a perceived low societal impact of negative results, and competing (for time required for producing an article) projects within the same laboratory. Importantly, the underlying reasons for "holding back" of negative results do not affect our ability to measure the actual rate of published negative results or to make estimates from our model—the only difference for the model this would make is the higher empirical estimate of the ρ parameter (the expected proportion of true positive statements).

8 Optimizing the probability of reaching the correct answer at the end of a reasoning chain

Our goal in this section to find a set of parameter values, $\hat{\Theta}$, that maximizes the probability of reaching the correct answer at the i^{th} step of a chain:

$$\hat{\Theta} = \arg\max_{\Theta}[P(E_i = T|\Theta)]. \tag{39}$$

$$\begin{aligned} P(\{E_1, \ldots, E_i\}|\Theta) &= P(\{E_1, \ldots, E_i\}|T = 0, \Theta)P(T = 0, \Theta) \\ &+ P(\{E_1, \ldots, E_i\}|T = 1, \Theta)P(T = 1, \Theta). \end{aligned} \tag{40}$$

$$P(E_i = T|\Theta) = P(E_i = 0|T = 0, \Theta)P(T = 0|\Theta) + P(E_i = 1|T = 1, \Theta)P(T = 1|\Theta). \tag{41}$$

$$P(E_i = 0|T = 0, \Theta) = \sum_{x=0}^{i-1} P(E_i = 0|I_{1,i-1} = x, T = 0, \Theta)P(I_{1,i-1} = x|T = 0, \Theta), \tag{42}$$

$$P(E_i = 1|T = 1, \Theta) = \sum_{x=0}^{i-1} P(E_i = 1|I_{1,i-1} = x, T = 1, \Theta)P(I_{1,i-1} = x|T = 1, \Theta), \tag{43}$$

$$\begin{aligned} P(I_{1,i} = x|T = y, \Theta) &= P(E_i = 1|I_{1,i-1} = x - 1, T = y, \Theta)P(I_{1,i-1} = x - 1|T = y, \Theta) + \\ &P(E_i = 0|I_{1,i-1} = x, T = y, \Theta)P(I_{1,i-1} = x|T = y, \Theta). \end{aligned} \tag{44}$$

$$P(I_{1,0} = 0|T = y, \Theta) = 1. \tag{45}$$

$$
\begin{aligned}
P(E_i = x|T = y, I_{1,i-1}, I_{0,i-1}, \Theta) = & \\
& P(E_i = x|\varepsilon_i = 0, T = y, I_{1,i-1}, I_{0,i-1}, \Theta) \\
\times & \quad P(\varepsilon_i = 0|T = y, \Theta) + \\
& P(E_i = x|\varepsilon_i = 1, T = y, I_{1,i-1}, I_{0,i-1}, \Theta) \\
\times & \quad P(\varepsilon_i = 1|T = y, \Theta).
\end{aligned} \tag{46}
$$

$$
\begin{aligned}
P(E_i = 1|T = x, I_{1,i-1}, I_{0,i-1}, \Theta) \; = & \\
& P(E_i = 1|H_i = 1, T = x, I_{1,i-1}, I_{0,i-1}, \Theta)P(H_i = 1|\Theta) + \\
& P(E_i = 1|H_i = 0, T = x, I_{1,i-1}, I_{0,i-1}, \Theta)P(H_i = 0|\Theta).
\end{aligned} \tag{47}
$$

$$
P(E_{i,i>1} = 1|T, I_{0,i-1}, I_{1,i-1}, \Theta) = \begin{cases}
(1-\beta)\frac{\alpha I_{1,i-1}+\Lambda_1}{\alpha I_{1,i-1}+\iota I_{0,i-1}+1} + \beta\frac{\alpha I_{1,i-1}}{\alpha I_{1,i-1}+\iota I_{0,i-1}} & |T = 0, I_{1,i-1} > I_{0,i-1}, \\
(1-\beta)\frac{\iota I_{1,i-1}+\Lambda_1}{\iota I_{1,i-1}+\alpha I_{0,i-1}+1} + \beta\frac{\iota I_{1,i-1}}{\iota I_{1,i-1}+\alpha I_{0,i-1}} & |T = 0, I_{1,i-1} < I_{0,i-1}, \\
(1-\beta)\frac{\tau(i-1)/2+\Lambda_1}{\tau(i-1)+1} + \beta/2 & |T = 0, I_{1,i-1} = I_{0,i-1}, \\
(1-\beta)\frac{\alpha I_{1,i-1}+\Lambda_2}{\alpha I_{1,i-1}+\iota I_{0,i-1}+1} + \beta\frac{\alpha I_{1,i-1}}{\alpha I_{1,i-1}+\iota I_{0,i-1}} & |T = 1, I_{1,i-1} > I_{0,i-1}, \\
(1-\beta)\frac{\iota I_{1,i-1}+\Lambda_2}{\iota I_{1,i-1}+\alpha I_{0,i-1}+1} + \beta\frac{\iota I_{1,i-1}}{\iota I_{1,i-1}+\alpha I_{0,i-1}} & |T = 1, I_{1,i-1} < I_{0,i-1}, \\
(1-\beta)\frac{\tau(i-1)/2+\Lambda_2}{\tau(i-1)+1} + \beta/2 & |T = 1, I_{1,i-1} = I_{0,i-1},
\end{cases} \tag{48}
$$

$$
P(E_1 = 1|T = 0, \Theta) = \Lambda_1, \tag{49}
$$

$$
P(E_1 = 1|T = 1, \Theta) = \Lambda_2, \tag{50}
$$

where

$$
\Lambda_1 = (1-\nu)\varphi + \mu(1-\varphi), \tag{51}
$$

$$
\Lambda_2 = (1-\nu)(1-\varphi) + \mu\varphi. \tag{52}
$$

The existence of a formal probabilistic model allows us to compute the probability of reaching the correct answer at an arbitrary step of publication series. Therefore, using a numerical optimization technique, we can identify the values of subjective weight parameters that optimize the probability of reaching the correct answer under a wide choice of values of other parameters; our aim here is to identify the stochastic strategy that is most successful

from the point of view of generating valid knowledge. It appears that, under all randomly selected reasonably constrained ($0 \leq \rho \leq 1$, $0 \leq \mu \leq 0.5$ and $0 \leq \nu \leq 0.5$) values of other parameters, it is advantageous to ignore the minority statements (the optimal value of ι is zero). Furthermore, the longer the series, the lower the optimum value of α (figure 12); the optimum value of α is always smaller than 1.

9 Gibbs sampling within Metropolis-Hastings update

We did analysis in the Bayesian famework; we estimated the posterior distribution of parameter values given data— from analysis of the whole dataset. Our computation is nothing but stochastic integration according to the following equation (which is a mere re-statement of the Bayes theorem).

$$P(\Theta|Data) = \frac{P(Data|\Theta)P(\Theta)}{\int_\Theta P(Data|\Theta)P(\Theta)}. \tag{53}$$

The main (oversimplified) idea of statistical integration is as follows. Imagine that you are interested in the volume of a complex shape encompassed by a multidimensional surface. All you need to estimate the volume is (1) be able to tell whether a given point is inside or outside the shape, and (2) be able to randomly pick points within a hypercube that contains your shape. Then you can compute the volume of hypercube (an easy task since you know the length of each edge of the hypercube), and the proportion of randomly distributed points that fall inside and outside your shape. The precision of the estimate will grow with the number of random points generated.

In most real-life applications a Markov chain Monte Carlo (MCMC) chain is updated one component at the time, see [4, 5, 6, 7, 8] for more detail. In our case also missing data is involved—which comes in two flavors: the number of unpublished one-link negative chains, y_{start}, and unpublished negative links $\{y_i\}_{i=1,m}$, within the corresponding observed chains, $\{c_i\}_{i=1,m}$. The missing data is *imputed* during the computation—the computation involved in data imputation is essentially the same as in the sampling parameter values from condition distributions.

The brief representation of our update schedule is as follows. Notation $A|B, C, D$ means "update value A by sampling from conditional distribution $P(A|B, C, D)$."

$$
\begin{array}{rcl}
y_{\text{start}} & | & \Theta, \mathbf{C} \\
y_i & | & \Theta,\ c_i (i = 1, \ldots, m) \\
\alpha & | & \Theta \setminus \alpha, \mathbf{C}, \mathbf{Y} \\
\iota & | & \Theta \setminus \iota, \mathbf{C}, \mathbf{Y} \\
\rho & | & \Theta \setminus \rho, \mathbf{C}, \mathbf{Y} \\
\mu & | & \Theta \setminus \mu, \mathbf{C}, \mathbf{Y} \\
\nu & | & \Theta \setminus \nu, C, \mathbf{Y} \\
\varphi & | & \Theta \setminus \varphi, \mathbf{C}, \mathbf{Y} \\
\eta & | & \Theta \setminus \eta, \mathbf{C}, \mathbf{Y} \\
\beta & | & \Theta \setminus \beta, \mathbf{C}, \mathbf{Y}
\end{array} \tag{54}
$$

10 Equations for description of missing data

The likelihood (probability of data, both observed and missing, given model and model parameters) in for our models has the following form. (We use the standard notation for likelihood $L(\Theta|Data) \stackrel{\text{def}}{=} P(Data|\Theta)$, e.g. see [9].)

$$
\begin{aligned}
P(\{c_1, \ldots, c_m\}, y_{\text{start}}, \{y_1, \ldots, y_m\}|\Theta) &= L(\Theta|\{c_1, \ldots, c_m\}, y_{\text{start}}, \{y_1, \ldots, y_m\}) \\
&= L_{\text{start}}(\Theta|\{c_1, \ldots, c_m\}, y_{\text{start}}) \prod_{i=1}^{m} L(\Theta|c_i, y_i).
\end{aligned} \tag{55}
$$

$$
L_{\text{start}}(\Theta|\{c_1, \ldots, c_m\}, y_{\text{start}}) = \binom{m + y_{\text{start}}}{y_{\text{start}}} \chi^m (1 - \chi)^{y_{\text{start}}}, \tag{56}
$$

where

$$
\begin{aligned}
\chi &= P(O_{\text{start}} = 1) \\
&= P(T = 1)P(O_i = 1|T = 1) + P(T = 0)P(O_i = 1|T = 0) \\
&= \rho\omega_1 + (1 - \rho)\omega_2,
\end{aligned} \tag{57}
$$

and

$$
\begin{aligned}
\omega_1 &= P(O_i = 1|T = 1, \Theta) \\
&= P(\varepsilon_i = 1|T = 1, \Theta)P(O_i = 1|\varepsilon_i = 1, \Theta) \\
&+ P(\varepsilon_i = 0|T = 1, \Theta)P(O_i = 1|\varepsilon_i = 0, \Theta) \\
&= (1 - \nu)(1 - \varphi) + \mu\varphi + [(1 - \mu)\varphi + \nu(1 - \varphi)]\eta,
\end{aligned} \tag{58}
$$

$$\begin{aligned}\omega_2 &= P(O_i = 1|T = 0, \Theta) \\ &= P(\varepsilon_i = 1|T = 0, \Theta)P(O_i = 1|\varepsilon_i = 1, \Theta) \\ &+ P(\varepsilon_i = 0|T = 0, \Theta)P(O_i = 1|\varepsilon_i = 0, \Theta) \\ &= (1-\nu)\varphi + \mu(1-\varphi) + [(1-\mu)(1-\varphi) + \nu\varphi]\eta.\end{aligned} \tag{59}$$

$$\begin{aligned}L(\Theta|c_i = \{E_1, \ldots, E_{n_i}\}, y_i) &= P(T=1)\binom{n_i+y_i}{y_i}\omega_1^{n_i}(1-\omega_1)^{y_i}P(c_i|T=1,\Theta) \\ &+ P(T=0)\binom{n_i+y_i}{y_i}\omega_2^{n_i}(1-\omega_2)^{y_i}P(c_i|T=0,\Theta) \\ &= \binom{n_i+y_i}{y_i}[\rho\omega_1^{n_i}(1-\omega_1)^{y_i}P(c_i|T=1,\Theta) \\ &+ (1-\rho)\omega_2^{n_i}(1-\omega_2)^{y_i}P(c_i|T=0,\Theta)].\end{aligned} \tag{60}$$

Finally,

$$P(c_i = \{E_1, \ldots, E_n\}|\Theta, T = x) = \prod_{j=1}^{n} P(E_j|I_{1,j-1}, I_{0,j-1}, T = x, \Theta). \tag{61}$$

$$\begin{aligned}&P(y_i|\Theta, c_i = \{E_1, \ldots, E_{n_i}\}) \\ &= \frac{\binom{n_i+y_i}{y_i}[\rho\omega_1^{n_i}(1-\omega_1)^{y_i}P(c_i|\Theta,T=1)+(1-\rho)\omega_2^{n_i}(1-\omega_2)^{y_i}P(c_i|\Theta,T=0)]}{\rho P(c_i|\Theta,T=1)/\omega_1+(1-\rho)P(c_i|\Theta,T=0)/\omega_2} \\ &= \mathrm{NB}(n_i+1,\omega_1)\xi_i + \mathrm{NB}(n_i+1,\omega_2)(1-\xi_i),\end{aligned} \tag{62}$$

where "NB(x, y)" stands for a "negative binomial distribution with parameters x and y", and

$$\xi_i = \frac{\rho/\omega_1}{\rho/\omega_1 + [(1-\rho)/\omega_2]\frac{P(c_i|\Theta,T=0)}{P(c_i|\Theta,T=1)}}, \tag{63}$$

$$P(y_{\text{start}}|\Theta, \{y_i\}, \{c_1, \ldots, c_m\}) = \binom{m+y_{\text{start}}}{y_{\text{start}}}\chi^{m+1}(1-\chi)^{y_{\text{start}}} = \mathrm{NB}(m+1,\chi). \tag{64}$$

Our implementation of the Metropolis-coupled Markov chain Monte Carlo closely followed [10].

11 "Multiple universes": models with multiple solutions

The goal of a typical MCMC analysis is to estimate the posterior distribution $P(\Theta|Data)$—the resulting distribution estimate provides an intuitive measure of confidence in parameter

estimates. The point estimates of the parameters can be obtained either as posterior means or maximum posterior probability values—we used the later option because our posterior distribution turned out to be bimodal. It is hard to represent the full multidimensional posterior distribution for all parameters in our model—we show pairwise joint distribution of parameters ρ, μ and ν in figure 13A (it clearly shows the bimodal nature of the estimated distribution), and correlation matrix for all parameters at figure 13B.

We divided the whole space of the permissible parameter values into "bad" and "good" neighborhoods (error rate is very high and in the bad neighborhood and low in the good neighborhood) and then computed posterior probability that the real parameter values belong to each neighborhood (we call these neighborhoods a "pessimists′ universe" and an "optimists′ universe," respectively, see figure 5).

We will try to illustrate the reason for existence of multiple model solutions (modes of posterior distribution) through a thought experiment. Imagine that you are standing on the Fifth Avenue (next to the Central Park and corner of 68^{th} East Street in Manhattan) facing East river. You are new to Manhattan and you would like to find which way is the Guggenheim museum. To be cautious, you are asking four passers by about the direction. Three are showing to your left and one to your right. You can interpret these data in two ways: you can label all statements "right" as "true" and statement "left" as "false," or other way around. These two solutions are obviously distinct because they lead to very different actions on your part (either you are walking to the south or to the north).

Similarly, optimists′ and pessimists′ universes in our analysis are drastically different—in one universe the majority of published statements are correct (and the model can be used to identify these correct statements), while in the other universe the majority of published statements are incorrect—the ramifications for practical usage of the molecular interaction data are drastically different. (The value of T for each individual chain of reasoning defines whether each published statement in the chain is true (if value of T coincides with the published interaction value) or false (if the published interaction value differs from T).

12 An attempt to classify possible data types with regard to results of data reordering

Below we will consider three types of data randomization. First, we perform what we call a *global randomization*—a data reshuffling that keeps the chain length distribution intact, but completely changes content of individual chains, preserving only the mean frequencies of zeros and ones in the real data. Second, *chain reversal*—where the natural chronological order of statements within the same chain is substituted with a reversed chronological order. Third, *chain scrambling*—where each chain retains its original length and statements that are observed in the real data set, but the chronological order of statements is substituted with a random order.

In the main text of the paper we described a few possible types of chains (generated with different combinations of values of parameters α and ι.

Chains with completely independent statements ($\alpha = \iota = 0$—*trust nobody* pattern) would be resistant to all three types of randomization (in the sense of robustness of the results of probabilistic analysis).

Super-conformism chains (large α and $\iota = 0$) would tend to be one-digit chains, such as 111111111, and therefore resistant to chain-reversal and chain scrambling. However, if both negative and positive statements are observed in the data, the *super-conformism* chains would not be resistant to the global randomization (we should expect difference in estimated parameters between the original and globally randomized data).

Anti-conformism chains (large ι and zero or small α) would tend to be resistant to chain-reversal (because the characteristic alternating pattern is preserved by the order reversal), but not to global randomization and chain scrambling.

Other types of chains should fall somewhere in between these main types: for example, the *weak skepticism* pattern (both α and ι are significantly positive, but ι is smaller than α) should combine properties of the *super-conformism* and *trust nobody* chains.

13 Data-shuffling experiments

To make the comparison with the rearranged-data experiments easier, we first provide results for the correctly arranged (chronological-order) physical-interaction chains (see figure 15). (For the purpose of an easier presentation we show only *physical* interactions here).

13.1 Global randomization

To generate a "globally" randomized dataset, we first counted the number of positive and negative statements in the real chains and computed a probability for a random statement to be positive ($p_1 = 0.9652$). Then, we populated every chain with statements by randomly and independently sampling statements based on the fixed probability of generating the positive statement (p_1); we did not alter the length of the chains (as in the real data). *The results were significantly different from the analysis of the real data* (see the figure 16)—momentum parameters ι and α are much smaller and *no longer significantly differ from* 0, and distributions for μ and ν are not any longer bimodal (only one "universe" rather than two).

Using this random model, we obtained 2,597 patterned (two-digit) cascades—twice as many as in the real data. Therefore, the real data are highly structured and dependence between statements within the same chain is highly significant. Also, the two-universe solution

for our data does not appear to be a phenomenon that is common for all possible data sets.

13.2 Local randomization: content randomization or reversal within each chain

If we reverse each chain (so that the statements go from the future to the past), we obtain figure 17.

The results are not very different (with respect to the original, unreversed data). This is a rather intuitive result if we recall that we have primarily single-state chains (either only ones or only zeros)—in the whole collection of distinct chain patterns (each pattern can occur in data multiple times) we have 323,489 uniform (single-digit) and 1,136 non-uniform (two-digit) unique chain patterns. Time reversal does not affect the uniform patterns, and information for most of the parameters is derived from all patterns combined.

Exactly the same logic applies to the shuffled-data experiments shown in the following figure 18).

14 Looking only at variable chains (chains with both zeros and ones)

To illustrate the fact that the order of statement within chains does matter, we have analyzed separately only chains that have both zeros and ones. While these experiments serve as a good control, the full model including the uniform chains is absolutely necessary to represent the data in the published literature. In general, selecting a non-random subset of data can lead to parameter estimates that are severely biased,[2] but it is very useful to see the properties of the reshuffled data in order to show that our model does not produce the same results regardless of what the data looks like. Specifically, the analysis shows that the order of reporting results in the literature (i.e. the historical patterns) can indeed make a substantial difference in the parameter estimates obtained.

Now let us consider analysis of only non-uniform (zero-one) chains. Essentially, we are taking only the "controversial" data from the literature and using the model to estimate parameters on this interesting subset. Accordingly, it is expected to see different results for the non-uniform subset of data than the complete dataset. In fact, the resulting estimates look very different: the distribution for ρ is essentially uniform. Both momentum parameters, α and ι, have much smaller means, although the α-parameter is still significantly greater than

[2]The closest analogy that we are aware of is in the inference of phylogenetic trees: if one has an alignment of homologous nucleotide sequences, in a typical reliable alignment, most of the homologous nucleotide sites in different sequences contain the same nucleotide and appear uninformative. Yet removing the constant sites can lead to inference of incorrect tree with the maximum likelihood method or distance-matrix methods, and to vast overestimation of the evolutionary distances. Very similar logic is applicable to our modeling.

0. Furthermore, the two-universe phenomenon (two modes of the posterior distribution in the analysis of the complete dataset) is no longer observed: the false-negative error estimate is reliably low, and the false-positive rate is reliably high.

The order within chains does matter when we analyze the non-uniform subset of data. We show two cases of extreme shuffling within chains. The following analysis was performed on 0/1 chains where all zeros were moved to the beginning of each chain, see Figure 20. Plots for the intact variable chains and the rearranged ones look different. We can try yet another rearrangement, putting ones at the beginning of each variable chain. This rearrangement changes the estimates yet again, see Figure 21.

The mono-digital chains (such as $1111\ldots111$) are not uninformative because they reflect the propensity of scientists to agree with the published statements; with removal of such chains the data becomes severely biased towards low values of momentum parameters.

15 Looking only at constant chains (chains with only zeros or only ones)

An additional logical experiment is to analyze only mono-digit (constant) chains separately from the variable ones.

This experiment produces also quite curious parameter estimates, see Figure 22. The two-universe effect vanished, the joint posterior distribution has only one mode. The marginal posterior distribution for the false-negative error (ν) has a mode near zero, but the corresponding 99% credible interval extends to the whole range between zero and one. Posterior distribution for the false-positive error parameter (μ) has mode near 1. Parameters α and ι have rather flat marginal posterior distributions (see Figure 22) with a constraint: the value of α is always (much) bigger than value of ι, see Figure 23,—this relationship between the values of α and ι (that the momentum of majority-statement is much stronger than the minority-statement momentum) is the logical explanation of the observed data within our model.

References

[1] Krauthammer, M., Kra, P., Iossifov, I., Gomez, S. M., Hripcsak, G., Hatzivassiloglou, V., Friedman, C. & Rzhetsky, A. (2002) *Bioinformatics* **18,** Suppl. 1, S249–S257.

[2] Feller, William (1966) *An Introduction to Probability Theory and its Applications* (Wiley, New York), 3rd. Ed.

[3] Rzhetsky, A. & Gomez, S. M. (2001) *Bioinformatics* **17**, 988–996.

[4] Metropolis, S. C., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953) *J. Chem. Physics* **21**, 1087–1092.

[5] Hastings, W. K. (1970) *Biometrika* **57**, 97–109.

[6] Gilks, W. R., Richardson, S. & Spiegelhalter, D. J., eds. (1996) *Markov Chain Monte Carlo in Practice* (Chapman & Hall/CRC, New York).

[7] Green, P. J. (1994) *Reversible Jump MCMC Computation and Bayesian Model Determination*, Technical report, University of Bristol, Bristol, U.K.

[8] Liu, J. S. (2001) *Springer Series in Statistics* (Springer, New York)

[9] Edwards, A. W. F. (1992) *Likelihood* (Johns Hopkins University Press, Baltimore), Expanded edition.

[10] Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. (2004) *Bioinformatics* **20**, 407–415.