Mining Novellas from PubMed Abstracts using a Storytelling Algorithm

Joseph Gresock[†], Deept Kumar[†], Richard F. Helm[*],
Malcolm Potts[*], and Naren Ramakrishnan[†]

[†]Department of Computer Science, Virginia Tech, VA 24061, USA
[*]Department of Biochemistry, Virginia Tech, VA 24061, USA

Abstract

Motivation: There are now a multitude of articles published in a diversity of journals providing information about genes, proteins, pathways, and entire processes. Each article investigates particular subsets of a biological process, but to gain insight into the functioning of a system as a whole, we must computationally integrate information across multiple publications. This is especially important in problems such as modeling cross-talk in signaling networks, designing drug therapies for combinatorial selectivity, and unraveling the role of gene interactions in deleterious phenotypes, where the cost of performing combinatorial screens is exorbitant.

Results: We present an automated approach to biological knowledge discovery from PubMed abstracts, suitable for unraveling combinatorial relationships. It involves the systematic application of a 'storytelling' algorithm followed by compression of the stories into 'novellas.' Given a start and end publication, typically with little or no overlap in content, storytelling identifies a chain of intermediate publications from one to the other, such that neighboring publications have significant content similarity. Stories discovered thus provide an argued approach to relate distant concepts through compositions of related concepts. The chains of links employed by stories are then mined to find frequently reused sub-stories, which can be compressed to yield novellas, or compact templates of connections. We demonstrate a successful application of storytelling and novella finding to modeling combinatorial relationships between introduction of extracellular factors and downstream cellular events.

Availability: A story visualizer, suitable for interactive exploration of stories and novellas described in this paper, is available for demo/download at https://bioinformatics.cs.vt.edu/storytelling.

1 Introduction

The use of high-throughput data screens in biology [1] is resulting in an ever-growing stream of information about genes, proteins, pathways, and even entire processes. As has been observed [13], there is a concomitant increase in the number of papers published based on these large-scale studies. There is growing acknowledgment of the need to automatically integrate information across multiple publications [7], a task that is key to gaining insight into the functioning of biological systems as a whole. In particular, such integration can help formulate biological hypotheses in areas that are still too expensive or too tentative to study by traditional experimental methods.

As a motivating example, consider elucidating an individual cell's response to an external stress. The observed response is the net result of the applied stress as well as inputs from neighboring cells. These inputs feed into signal transduction cascades originating at the plasma membrane surface and culminating typically in the nucleus, leading to a molecular response that can potentially be quantified. A single environmental stress can lead to an array of changes in transduction pathways, the net result being a distinct cell state. Many different signaling pathways can provide similar cell states, suggesting that there is convergence in signaling pathways linking cellular inputs and the resulting cellular outputs. Determining the underlying pathways and the contribution of each is presently a daunting task, yet is required if basic biomedical life science research is to provide personalized approaches to combating disease states in humans. Similar motivations can be drawn from the problems of unraveling gene interactions in complex diseases such as cancer, and designing drugs from the viewpoint of combinatorial selectivity.

To address problems such as these, we present an automated approach to biological knowledge discovery from PubMed abstracts through systematic application of a 'storytelling' algorithm followed by compression of the stories into 'novellas.' Our working assumption is that there are experiments that have been published in the literature that look at particular subsets of a biological process. Linking these papers by their common elements into a story (and investigating how stories reuse subpaths) can provide hypotheses that can be tested at the bench, potentially resulting in new insights about the combinatorial nature of the phenomenon a biologist is interested in.

Section 2 gives an example of a story we manually constructed, and how stories can suggest hypotheses for experimental investigation. Section 3 defines our underlying biological problem and its formulation as a storytelling and novella mining task. Section 4 details the various stages in our pipeline, covering algorithmic design decisions, statistical significance testing, and multiple layers of verification. Section 5 presents experimental results toward addressing the problem from Section 3 and also our interface for interactively exploring stories. Related work, and discussion are provided finally, in Sections 6 and 7.

We build upon our prior work [9] which provided a general framework for storytelling, but was not particularly geared toward finding relationships in biological abstracts. This paper generalizes the algorithm from [9] to work with bag representations of documents, solves combinatorial information integration scenarios by organizing multiple runs of storytelling, and provides novel post-processing functions, involving named entity extraction for sentence cohesion checks, frequent substructure mining in stories, and story summarization. Above all, we present the results of addressing a relevant biological problem, shedding considerable domain-specific insight.

2 Storytelling Example

As an example of the capabilities envisaged here, we explore the similarities in adaptation to metabolic arrest (quiescence) between cyanobacteria (a simple prokaryote) and complex eukaryotes such as mice. While such a correlation may appear initially to be far-fetched, metabolic arrest is a process inherent to all organisms. One would thus assume that there are process similarities across all forms of life and testing such a relationship may provide new insights into this important process. In the *manually* conceptualized example story below, we begin with a paper by the authors:

> L. Garczarek, N. Ramakrishnan, D. Kumar, R.F. Helm, and M. Potts, Global cross-over points in the genome responses of Synechocystis sp. PCC 6803, to dehydration, UV-irradiation, and other stresses, under communication to *BMC Microbiology*, 2007.

By studying its abstract, we observe the mention of the CBS (cystathionine-beta-synthase) domain. CBS domains are small intracellular modules, mostly found in two or four copies within a protein, that occur in several different proteins in all kingdoms of life. We might begin by looking for other publications that discuss CBS. Using this idea, we can reach:

> L. Schmitt and R. Tampe, Structure and mechanism of ABC transporters, *Current Opinion in Structural Biology*, Vol. 14, No. 4, pages 426–431, Aug 2004.

From this paper, we learn that CBS domains are found in glycine betaine transport proteins that mediate osmotic adjustment in cells. Following this thread and investigating CBS domains further leads to the paper:

> J.W. Scott, S.A. Hawley, K.A. Green, M. Anis, G. Stewart, G.A. Scullion, D.G. Norman, and D.G. Hardie, CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations, *Journal of Clinical Investigation*, Vol. 113, No. 2, pages 182–184, Jan 2004.

This publication reveals the nature of interactions found in molecular complexes involving CBS domains. At this point, we shift the emphasis from CBS to the function of ligands mentioned in this paper, to reach:

> C. Tang, X. Li and J. Du, Hydrogen sulfide as a new endogenous gaseous transmitter in the cardiovascular system, *Current Vascular Pharmacology*, Vol. 4, No. 1, pages 17–22, Jan 2006.

This paper indicates that in humans, hydrogen sulphide (H_2S) is produced endogenously in mammalian tissues from L-cysteine metabolism mainly by three enzymes, one of which is CBS. In addition, H_2S may not only function as a neuromodulator in the central nervous system but it also relaxes gastrointestinal smooth muscles. We now look for connections involving H_2S, leading to our intended target, a publication about mice:

> E. Blackstone, M. Morrison, and M.B. Roth, H_2S induces a suspended animation-like state in mice, *Science*, Vol. 308, No. 5721, page 518, Apr 2005.

The story thus mined, through the sequence of intermediaries, provides a continuous chain of reasoning about metabolic arrest across organisms of diverse complexity. It is important to note that the story proceeds through sulfur metabolism, a feature that we identified in a previous study as an integral part of the desiccation and recovery process of baker's yeast, *Saccharomyces cerevisiae* [14]. This application of storytelling suggests that laboratory-based efforts aimed at understanding the role of sulfur metabolism in metabolic arrest is an area worth exploring in detail.

Stories as depicted above reveal multiple forms of insights. Since intermediaries must conform to *a priori* knowledge, we can think of storytelling as a carefully argued process of removing and adding participants, not unlike a real story. Furthermore, similar to books such as Burke's *The Knowledge Web*, the network of stories underlying a domain reinforces interconnections between ideas rather than strict sequential and historical progression of discoveries. In particular, storytelling reveals if certain papers (and hence, concepts) have greater propensity for participating in some stories more than others. Such insights have great explanatory power and help formulate hypotheses for situating new data in the context of well-understood processes.

3 The Problem

Our research group is interested in the molecular mechanisms underlying organismal aging, i.e., the ability/inability of cells to have a controlled yet extended lifespan. We recently demonstrated that the supplementation of nicotinamide to the growth medium of primary fibroblasts provided a lifespan extension [12], and this lifespan extension was subsequently reported by others. In order to generate testable hypothesis to understand this lifespan extension at the molecular level, we need to elucidate relationships between cellular inputs (such as nicotinamide) and outputs (cell fate decisions such as lifespan extension).

Table 1: 18 storytelling inputs used to explore combinatorial relationships underlying ADP-ribosylation. All molecules listed are extracellular except CD38, which is bound to the outer membrane.

Molecule (Input)	Function/Comment
CD38	ADP-ribosyl cyclase 1, modulator of NAD levels
CXCL1	Growth-regulated protein alpha precursor, chemokine
IL-8	Interleukin-8, cytokine, inflammatory response
IL-1β	Interleukin-1beta, cytokine, inflammatory response
IL-6	Interleukin-6, cytokine, multiple functions
IL-13	Interleukin-13, cytokine, multiple functions
IL-24	Interleukin-13, cytokine, antiproliferative properties
MCP-1 and 2	Monocyte chemotactic proteins; mitogenic, chemotactic, and inflammatory activity
IFN-γ	Interferon-gamma; antiviral, antiproliferative, and immunoregulatory functions
Nicotinamide (NAM)	Lifespan extension in primary fibroblasts
STC-1	Stanniocalcin-1; phosphate regulator, up-regulated in nicotinamide microarray experiments
IGF-1	Insulin-like growth factor 1, downregulated in NAM supplementation studies
SFRP1	Secreted frizzled-related protein 1, function unknown, upregulated in NAM supplementation studies
Matrix metalloproteinase	Extracellular proteinases (MMP), involved in extracellular matrix degradation
MMP 3	MMP, downregulated in NAM supplementation studies, implicated in wound repair, atherosclerosis, and tumor initiation
MMP 12	Extracellular proteinase, downregulated in NAM supplementation
Serpin B-2/PAI-2	Degrades elastin, serine or cysteine proteinase inhibitor

Nicotinamide, a component of nicotinamide adenine dinucleotide (NAD+), is usually considered a cofactor in redox reactions, but is also known to be a substrate in the ‘ADP-ribosylation’ reaction. More specifically, high levels of nicotinamide inhibit the ADP-ribosylation process. ADP-ribosylation is an enzyme-mediated reaction whereby NAD+ is converted to ADP-ribose, releasing nicotinamide. The ADP-ribose group can then be attached to a protein post-translationally to form chains of ADP-ribose polymers, or to form ADP-ribose polymers free of protein. ADP-ribosylation is associated with the DNA damage response, cell death processes, as well as chromatin remodeling [4]. PARP-1, the enzyme that generates ADP-ribose chains, is one of the most abundant proteins in the nucleus of human cells and PARP-1 inhibition is presently an active line of pharmaceutical research.

Building upon our nicotinamide supplementation work, we performed a series of transcriptional profile experiments (unpublished data) to determine if there were any genes that were significantly up- or down-regulated due to the nicotinamide addition. While several genes were identified through this work, there were no obvious links in the literature, especially to ADP-ribosylation inhibition. We also evaluated primary fibroblasts for their ability to metabolically arrest for extended periods (a form of lifespan extension), and several cytokines were implicated in this response [6]. Again we sought to identify links between these extracellular molecules and ADP-ribosylation.

To computationally model this problem, we cast it as one of exploring combinatorial relationships in the literature between a set of 18 input extracellular molecules (see Table 1) and one output, i.e., (poly)ADP-ribose. How do signaling cascades and cellular responses to these molecules interact? Are there overlapping pathways or largely independent cascades of input-output relationships? For each molecule, we pick a representative document discussing it, attempt to create a story to a document discussing ADP-ribos(ylation), and mine the discovered stories to find novellas of relationships.

Formally, a document is modeled as a weighted vector over a vocabulary of T terms. Two documents $d_1 = (w_{11}, w_{12}, \cdots, w_{1T})$ and $d_2 = (w_{21}, w_{22}, \cdots, w_{2T})$ are said to be (approximate) *redescriptions* of each other (written as $d_1 \Leftrightarrow d_2$) if they share significant commonality in term usage. The quality of the redescription is typically measured by the (weighted) Jaccards coefficient $\mathcal{J}(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\|^2 + \|d_2\|^2 - d_1 \cdot d_2}$

where $d_1 \cdot d_2$ is the scalar product of two vectors and $\| \cdot \|$ is the L_2 norm. $\mathcal{J}$ reduces to the traditional Jaccards coefficient when the documents are binary vectors, i.e., when they denote sets, and is then given by $\frac{|d_1 \cap d_2|}{|d_1 \cup d_2|}$. Observe that $\mathcal{J}$ is 1 (constitutes an exact redescription) if the documents are the same, and 0 when they share no commonality in term usage. A *story* of length m is a sequence of documents $d_1, d_2, \cdots, d_m, d_{m+1}$ such that every successive pair of documents satisfy $\mathcal{J}(d_i, d_{i+1}) \geq \theta, i = 1, \cdots, m$ where θ is a specified minimum threshold on the Jaccards coefficient. Note that storytelling can progress only when redescriptions are strictly approximate, i.e., when $0 < \mathcal{J}(d_i, d_{i+1}) < 1$. Given a start document d_1 and an end document d_{m+1} the task of storytelling is to find a chain of documents that sequentially compose similarities to induce distant connections or dissimilarities. A *novella*, just as in writing, is a compact narrative that is repeatedly re-used in the mined stories and, for our purposes, is a recurring subsequence of documents from a given story collection.

4 Methods

We now outline the steps involved in mining stories and novellas to address our biological questions.

Seeding storytelling

We begin with retrieving, using Entrez ESearch, the 268,418 documents in PubMed that mention at least one of the 19 molecules = 18 input + 1 output. (This initial harvesting was done in Sep 2006, and so our study only includes publications upto that point.) Some of the molecule descriptions (e.g., 'serpinB2') were expanded to cover alternate uses and enhance coverage (e.g., to serpinB2+OR+plasminogen+activator +inhibitor-2+OR+PAI-2). Of these documents, we eliminate any that have no abstract or title (this does happen sometimes), retaining 228,466 documents. Finally we remove all review papers from consideration since they provide a 'lowest common denominator' approach to storytelling, and lead to simplistic stories involving topic dilution followed by specialization. Although PubMed's metadata tags reviews as such, there are some that slip through this filter, e.g.:

> T. Sugimura and M. Miwa, Poly(ADP-ribose): historical perspective, *Molecular and Cellular Biochemistry*, Vol. 138, No. 1-2, pages 5–12, Sep 1994.

We trapped such papers (433 of them from our base set) as those whose titles contain 'review' or whose abstracts contain the phrase 'this review' (however, other usages such as 'is reviewed' can be used in legitimate non-review documents). Together, these reduced our collection to 203,872 documents.

We then label a subset of them with the given molecules, where the respective molecule is mentioned in either the title or abstract. This was done by projecting ESearch's results for these molecules onto our seed set, and limiting each molecule to have a maximum of (top-ranked) 500 representative documents, for the input molecules, and a maximum of 20 documents for the output molecule, i.e., poly(ADP)-ribose. In total, 4737 documents were labeled with the input molecules, taking care to ensure that no document was labeled with more than molecule, breaking ties arbitrarily. A different 20 documents were labeled with the output molecule. The lack of overlap among these labeled documents ensures that we do not *a priori* plant overlapping or merging story patterns in our results.

Document modeling for similarity search

As mentioned before, we use a bag-of-words (vector) representation for each document (abstract), selecting 96,218 terms after Porter stemming, and removing stop words, numerals (e.g., measurements), and

DNA sequences (e.g., 'AAAGGT'). Next, we experimented with many IR term weighting strategies, designed to capture variations in document frequency, collection frequency, and document length. Recall that our goal is not to achieve effectiveness of document retrieval given term inputs, but rather to improve the quality of similarity search results, since they will be composed into stories. We sought to mimic PubMed's ELink 'related articles' function, which uses a combination of TFIDF measures and MeSH taxonomic identifiers to model documents. However, we desired to reserve MeSH to validate the stories. Hence we tuned our document modeling to weight just TFIDF information and replicate, as much as possible, the ELink results. Another reason to craft our own apparatus is that ELink searches over *all* indexed documents, whereas we desire to restrict our search to only documents in our seed collection (although these are a subset of the entire collection, ELink truncates the results list much before our documents could appear).

Based on statistical testing (details below), we adopt the weighting scheme for the ith document and j term as:

$$w_{ij} = \frac{(\log(\text{tf}_{ij}) + 1.0) \times \text{idf}_j}{\sum_{k=1}^{n_i}((\log(\text{tf}_{ik}) + 1.0) \times \text{idf}_k)^2}$$

where n_i is the number of terms occurring in document i, tf_{ik} is the term frequency of the kth term in document i, and idf_k is the inverse document frequency of the kth term across the whole collection. With this weighting scheme, we employ the Jaccards similarity measure as stated before to produce ordered lists of top similar documents.

This measure was validated by producing an ordered list of 75 top similar documents for each of 1,000 randomly selected documents from our seed set and comparing it, using the Kendall's τ measure, to the (projected) ranking from PubMed. A τ value of 1 indicates perfect concordance between rankings, whereas a value of -1 indicates perfect discordance (i.e., reversal). A p value can then be derived by first mapping each of the 1000 τ values to a z-score using

$$z(\tau) = \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}}$$

where N is the number of elements ranked (here, 75). This z-score is normally distributed as $\mathcal{N}(0, 1)$. Hence we can produce an aggregated z-score for all the rankings as $z = \frac{1}{\sqrt{1000}} \sum_{i=1}^{1000} z_i$, which is also normally distributed as $\mathcal{N}(0, 1)$. Converting this to a p-value results in a magnitude less than 10^{-70}, which gives us high confidence in the redescriptions derived from our similarity measure.

Storytelling

As described in [9], the similarity search algorithm is embedded within an A* search to proceed from a given start document toward the end document, such that successive documents satisfy the Jaccards threshold. For this stage, the quality metric for stories is taken to be the story length although other measures (see last section) might be more appropriate in other situations. At each step, thus, among all documents that meet the Jaccards criterion, the A* search heuristically assesses the number of steps remaining to reach the destination, and chooses the redescription that is predicted to yield the least total story length. Ties in this evaluation are broken by choosing the redescription that induces most similarity with the end document. Such a search can also be constrained in domain-specific ways (e.g., to follow chronological order of publications), but we did not explore these in this study.

As Fig. 4 (left) shows, there is an inherent tradeoff between the minimum Jaccard's threshold and the length of stories. Lower thresholds generate shorter stories but higher thresholds yield higher average similarity and, hence, greater story coherence. Observe that the average story length monotonically increases

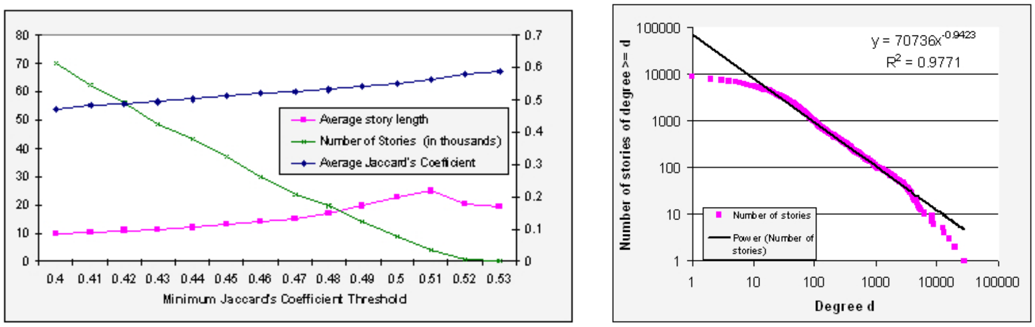


Figure 1: (left) The effect of Jaccards threshold on the number of stories mined, average story length, and average (observed) Jaccards coefficient. (right) The cumulative degree distribution of the links used for storytelling follows a power-law, and hence so does the underlying degree distribution.

upto a point (as expected), then drops because fewer stories are found and those that are pruned out are ones that would have involved longer chains. Also, as Fig. 4 (left) shows, the number of mined stories decreases monotonically with Jaccards threshold, as expected. In this study, we picked all thresholds from 0.4 to 0.53 (in steps of 0.01) and resolve to eliminate duplicate stories and story fragments subsequently. All redescriptions utilized in the mined stories are whetted to have a p-value significance of 10^{-5} or better. This is achieved by simulation, wherein we generate 50,000 randomized versions of each redescription, involving the same number of terms as either side and obeying the same underlying term-document distribution (but making random selections of terms), and assess the likelihood of obtaining the Jaccards coefficient by chance. To guard against multiple hypothesis testing, we adopt a q-value threshold of 10^{-7} (details omitted for space considerations).

Preliminary results

We mined a total of 313,591 stories through 16,221 documents. From these, for ease of interpretation, we focused on only those stories with lengths less than or equal to 10, yielding 85,967 stories through 8,761 documents. The degree distribution of the resulting graph of document interconnections, minus the source and destination documents, follows a power law (see Fig. 4 (right)), with a few documents serving as 'hubs' and most documents participating in at most a few links (similarities). The most popular hub was PubMed ID 8064725—'Altered poly(ADP-ribose) metabolism in family members of patients with systemic lupus erythematosus'—with a degree of 37,870 and lying typically two or three redescriptions before (some) end document. The second most popular hub was PubMed ID 2684169—'Two types of antibodies inhibiting interleukin-2 production by normal lymphocytes in patients with systemic lupus erythematosus'— with degree of 26,139 and is linked just after ID 8064725, toward the destination documents. The mention of lupus erythematosus in both these hubs, a chronic autoimmune disease involving inflammation and tissue damage, reinforces the underlying context of our storytelling problem.

Pruning stories using the OGM measure

We then use the optimistic genealogy measure (OGM) [3] over the MeSH taxonomy to posit a further degree of coherence among the mined stories. For every (input, output) molecule pair, we compute the OGM for every link in every story involving these molecules, and use this distribution to characterize significance cutoffs. The OGM score distributions were tested for normality and significance thresholds were picked that would retain the top 1% of stories for most molecules (e.g., CD38; corresponds to an OGM cutoff of 0.55) and the top 5% for others (e.g., CXCL1, which have fewer numbers of stories in general, with an OGM cutoff of 0.5). OGM is an intuitive similarity measure that exploits hierarchical structure to posit relationships between documents. For two documents indexed in MeSH, OGM involves a weighted average of similarities assessed at every level of taxonomic classification. Although MeSH is not a tree, we effectively model it as a tree for OGM purposes, breaking ties between parents arbitrarily and also pruning headers deemed irrelevant to our problem domain (e.g., MeSH categories such as Anthropology, Education, Sociology and Social Phenomena [I] and Humanities [K]).

Frequent episode mining

Next, we analyze the ordered sequences of documents in story chains to find commonly recurring subsequences, or frequent episodes. Here we define a frequent episode as a document d_i, followed by a document d_j (not necessarily consecutively) in at least s stories (where s is the support, which we set to be a minimum of 2). We also stipulated that there can be at most three documents in between d_i and d_j (since we have run storytelling with multiple thresholds, it is realistic that there would be subpaths of different lengths running between the same pair of documents). We employ the hidden Markov model formulation of [11] to find 115,891 frequent episodes, which are assessed to be significant at a p value of 10^{-3}. The average number of documents represented by a frequent episode is 2.127, and the average number of frequent episodes in a story is 3.9444. An example episode is:

$$\text{PubMed ID 16430457} \rightarrow \cdots \rightarrow \text{PubMed ID 1386861}$$

and makes a connection from a document discussing soluble CD23 in B-cell chronic lymphocytic leukemia to a document discussing regulation of soluble CD23 by the granulocyte-macrophage colony-stimulating factor (more on this later).

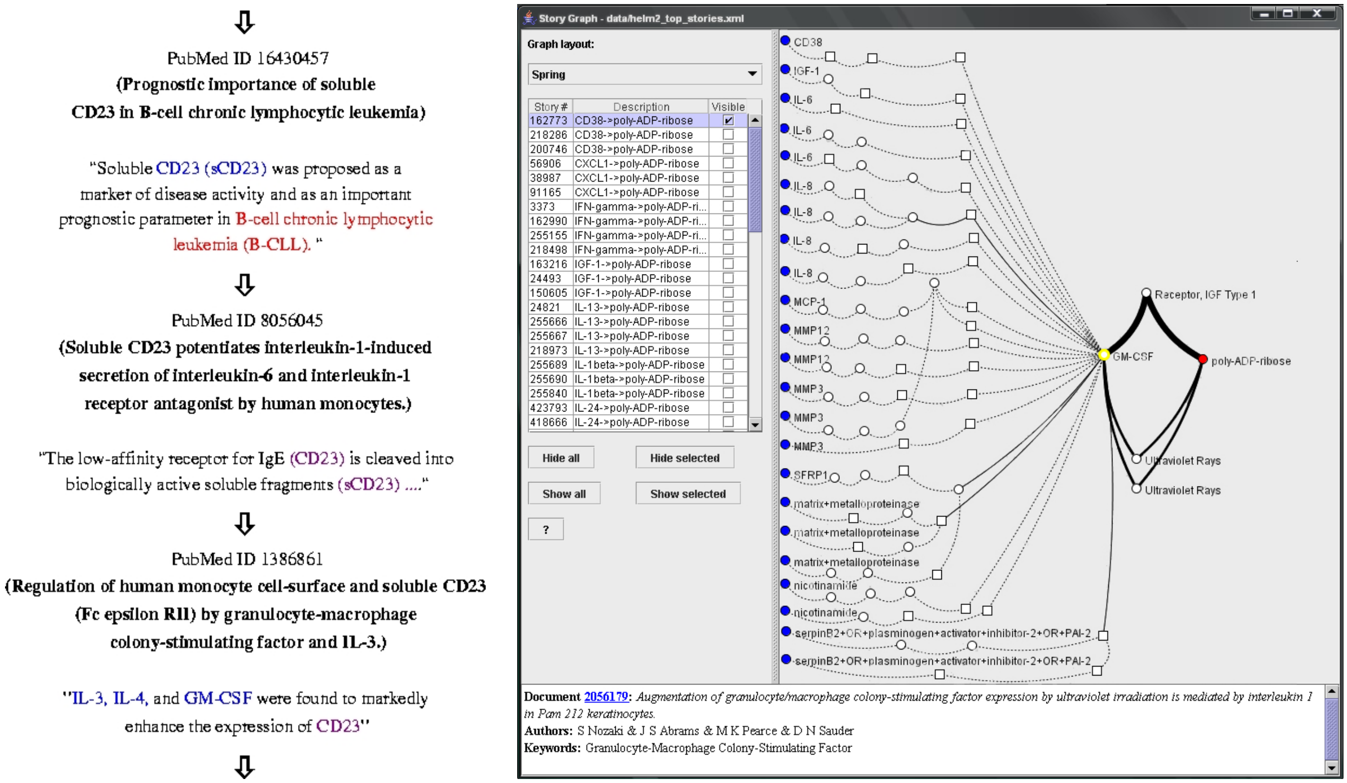
Story compression into novellas

Frequent episodes not only shed insight into commonly reused substories but also suggest an approach to *story compression*, to yield novellas. First, we relabel the document subsequences comprising episodes with a single episode label, iterating through the episodes in order of decreasing support s; all but 13 stories are relabeled at the end of this step. Stories that have the same end-to-end signature (stated in terms of both document IDs and episode labels) are then clustered together and replaced with a single story from the cluster, the one with the highest average Jaccards coefficient. Finally, stories that don't have the same signature but do have close enough signatures (i.e., differ in at most one document ID or episode symbol) are clustered together as well and the story with the lower average Jaccards coefficient is removed.

Sentence cohesion check and story summarization

Our last step is to be able to summarize a story by picking key sentences from the abstracts (one from each), and tiling them from start to end documents. We use the LingPipe named entity recognizer to find biological entities in all titles and abstracts (from the original seed set), and create document-sentence and sentence-named entity indices. We then attempt to find paths mirroring the storyline but spanning nodes representing the sentence IDs, and where the Jaccards coefficient is defined by overlap in terms of named entities. All stories that do not have such a mirroring path involving named entities are eliminated (observe that this can happen because the term-document modeling was done prior to named entity extraction). Each of the remaining stories is then summarized by the sentence path going through named entities with the highest (set) Jaccards coefficient between neighboring sets of named entities. An example is provided in Fig. 2 (left).

At the end of this pipeline, we obtain 277 stories over 944 documents and 604 unique episodes.



⇩

PubMed ID 16430457

(Prognostic importance of soluble CD23 in B-cell chronic lymphocytic leukemia)

"Soluble CD23 (sCD23) was proposed as a marker of disease activity and as an important prognostic parameter in B-cell chronic lymphocytic leukemia (B-CLL). "

⇩

PubMed ID 8056045

(Soluble CD23 potentiates interleukin-1-induced secretion of interleukin-6 and interleukin-1 receptor antagonist by human monocytes.)

"The low-affinity receptor for IgE (CD23) is cleaved into biologically active soluble fragments (sCD23) ...."

⇩

PubMed ID 1386861

(Regulation of human monocyte cell-surface and soluble CD23 (Fc epsilon RII) by granulocyte-macrophage colony-stimulating factor and IL-3.)

"IL-3, IL-4, and GM-CSF were found to markedly enhance the expression of CD23"

⇩

Story Graph - data/helm2_top_stories.xml

Graph layout:

Spring

Story #	Description	Visible
162773	CD38->poly-ADP-ribose	☑
218286	CD38->poly-ADP-ribose	☐
200746	CD38->poly-ADP-ribose	☐
56906	CXCL1->poly-ADP-ribose	☐
38987	CXCL1->poly-ADP-ribose	☐
91165	CXCL1->poly-ADP-ribose	☐
3373	IFN-gamma->poly-ADP-ri...	☐
162990	IFN-gamma->poly-ADP-ri...	☐
255155	IFN-gamma->poly-ADP-ri...	☐
218498	IFN-gamma->poly-ADP-ri...	☐
163216	IGF-1->poly-ADP-ribose	☐
24493	IGF-1->poly-ADP-ribose	☐
150605	IGF-1->poly-ADP-ribose	☐
24821	IL-13->poly-ADP-ribose	☐
255666	IL-13->poly-ADP-ribose	☐
255667	IL-13->poly-ADP-ribose	☐
218973	IL-13->poly-ADP-ribose	☐
255689	IL-1beta->poly-ADP-ribose	☐
255690	IL-1beta->poly-ADP-ribose	☐
255840	IL-1beta->poly-ADP-ribose	☐
423793	IL-24->poly-ADP-ribose	☐
418666	IL-24->poly-ADP-ribose	☐

Document 2056179: *Augmentation of granulocyte/macrophage colony-stimulating factor expression by ultraviolet irradiation is mediated by interleukin 1 in Pam 212 keratinocytes.*
Authors: S Nozaki & J S Abrams & M K Pearce & D N Sauder
Keywords: Granulocyte-Macrophage Colony-Stimulating Factor

Figure 2: (left) A mined story (story ID 162773) summarized by key sentences, following sentence cohesion check. Only a fragment is shown. Overlapping sentence fragments are highlighted. (right) Storygrapher interface, available for demo/download at https://bioinformatics.cs.vt.edu/storytelling.

5 Discussion of results

Fig. 2 (right) describes our Storygrapher interface for interactively exploring the mined stories and novellas. Users can selectively display or hide stories, open up episode symbols in novellas, and investigate storylines by automatically generated links to PubMed (bottom portion of Fig. 2 (right)). For each pair of input-output molecules, we focused our exploration on the top 20% of the stories, as judged by the average Jaccards coefficient, and retaining at least three stories for each pair. We will discuss a story in detail before presenting a novella.

Story 162773, depicted in part in Fig. 2 (left), links the membrane bound CD38 to poly(ADP-ribose) via a combination of the leukemia and keratinocyte (skin cells) literature. Chronic lymphocytic leukemia (CLL) causes an increase in the number of B lymphocytes in the bone marrow. These rapidly proliferating and hence cancerous cells spread from the bone marrow to lymph nodes as well as other organs, resulting in the CLL disease state. Clinical researchers have been interested for many years in developing biomarkers for the disease, and one of these early biomarkers was CD38. The first article in the story provides an evaluation of CD38 as a marker for CLL. Another biomarker of interest is the soluble form of the protein CD23 (sCD23), and the next set of articles, summarized into an episode, provides a connection between sCD23 and the inflammatory cytokine interleukin-1 (IL-1). This episode is linked to additional episodes that provide a connection between monocytes (macrophage precursor cells responsible for the immune response), CD23, and tumor necrosis factor-alpha (TNF-α). Interestingly, episode e83501 provides a direct link between TNF-α levels and GM-CSF production in monocytes. This episode then jumps to the relationship of GM-CSF to the response of keratinocytes (the major component of epidermis or skin) to UV-light. This reference is related to the IGF-1 receptor as the next paper provided evidence that keratinocyte survival under UV stress is related to activation of the IGF-1 receptor. The link between the IGF-1 receptor and poly(ADP-ribose) is due to the fact that high levels of poly(ADP-ribose) have been associated with keratinocyte apoptosis brought about by UV and oxidant stress.

We then proceeded to see how (and where) other stories, from different input molecules, might overlap with Story 162773; Fig. 2 (right) describes the results of our exploration using Storygrapher, with the darkened lines indicating a novella, involving the granulocyte-macrophage colony stimulating factor (GM-CSF). The GM-CSF document (PubMed ID 2056179) has a frequently re-used redescription with PubMed ID 9935186, which is a paper that discusses activation of the insulin-like growth factor-1 receptor. Note that we had previously identified the insulin-like growth factor-1 gene (IGF-1) as being downregulated in NAM supplementation studies (see Table 1). Document 2056179 appears in 24 different stories, all near the end of the 'storyline.' GM-CSF (also referred to as CSF2) is an extracellular protein that controls the production, differentiation, and function of hematopoietic precursor cells such as granulocytes, macrophages, eosinophils, and erythrocytes. Thus the cytokines implicated in extension of cellular lifespan and metabolic arrest, based upon storytelling analysis, have a strong relationship to the activities of GM-CSF. This is an area of research that has been addressed to some degree already as a PubMed search utilizing the terms 'granulocyte AND PARP' provides 302 papers! We previously evaluated GM-CSF levels in three different cells lines (primary fibroblasts, adenovirally-infected kidney cells, and a brain glioblastoma cell line) during the course of metabolic arrest and recovery experiments [6], but found no significant changes in levels of this cytokine. Further co-culture, media supplementation, and/or genetic work is required to evaluate the connection between fundamental metabolic processes in primary fibroblasts and cancer cell lines and the GM-CSF/ADP-ribosylation storyline, as recently suggested [4].

There are several potential limitations to our current methodology of storytelling and novella finding, which we summarize below. First, we have mined stories and novellas without modeling cell types, and it is conceivable that stronger relationships can be obtained, first via analysis of the behaviors of specific cell types, followed by the overlay of responses with different cell lines. Second, although our stories bridge extracellular factors and an intracellular molecule, there is no requirement for the mined stories to mirror a signal transduction cascade (of course, in the interest of serendipitous discovery, we might not want to enforce such constraints). Third, many research areas and subtopics will be dominated by particular types of experimentation and cell lines; while providing enrichment in certain pathways, such papers might hide potentially promising leads. The numerous studies relating to UV stress and ADP-ribosylation is a case in point. The first major body of work in the ADP-ribose research community was evaluation of the DNA damage response, with UV-light used as the effector of DNA damage. These investigations are indeed enriched in our stories (see Fig. 2 (right)). Subsequent studies revealed that ADP-ribosylation is also involved in the death response as well as chromatin remodeling. However, these studies are not as numerous as the ones evaluating DNA damage. Finally, issues pertaining to the quality of the published work, author bias (in abstracts), and the messiness underlying information integration (e.g., lack of consistent nomenclature, synonymy/polysemy, false and missing data) have to be addressed. Nevertheless, the storytelling and novella mining approach provides a systematic procedure by which the scientist can interact directly with the published literature, stepping back from the specific directions taken by the research community to uncover trends and processes that were previously not considered.

6 Related Research

Storytelling and novella mining generalize related work in many areas. In **information visualization** (e.g., see [8]) storytelling has been viewed, not as a data mining tool, but as an information organization tool based on narrative structures from real life. The emphasis of software developed here is to provide templates and diagrammatic semantics that make the *manual* process of constructing stories easier, whereas we focus on *automated* approaches to mine stories. In **topic tracking** [10], the goal is to post-process search results into storylines by analyzing bipartite graphs of document-term relationships. Here a story is a thread of related documents with temporal as well as semantic coherence. These works are focused on *unsupervised* discovery of all threads whereas we focus on *directed* (but not necessarily temporal) stories between given start and end points. Our formulation of storytelling is closest in spirit to **literature-based discovery systems** such as Swanson's Arrowsmith [17]. In 1985, Don Swanson, quite by accident, connected two different pieces of information across medical literature that led him to the hypothesis that magnesium deficiency may play a role in certain types of migraine, a result since subsequently proven. The Arrowsmith project aims to automate this process by looking for relationships among articles in biomedical literature (these works have been recently enhanced with profile-based representations of topics and topic co-occurrence modeling [16]). In this paper, we systematically explore connections between publications first using both data mining and linguistic methods, and do not engage in data reduction until the last stages of analysis. As demonstrated in the results, this yields fine-grain representations of molecular relationships culled from individual sentences.

Storytelling also tempts comparison to **'meandering search engines'** such as omnipelagos.com. Omnipelagos indexes Wikipedia documents and finds chains of hyperlinks from one document to another. This is similar to our idea in that the result is a path, rather than a single document. Further, since the paths follow manually curated links, the reconstructed stories are quite meaningful (e.g., try to compute a chain between 'magnesium' and 'migraine'). However, the links in storytelling are mined automatically; omnipelagos.com views this as merely a problem of finding a path through manually created links. Finally, the sentence summaries revealed by our pipeline are akin to those exposed by **probabilistic hidden story models** [2]. These notions can be productively combined with our pipeline, e.g., to select one or more sentences from each abstract, to compose into a story.

The Storygrapher interface exposes functionalities similar to **text tiling** [5] and builds upon capabilities provided **similarity browsing interfaces** [15]. Whereas text tiling summarizes documents using paragraph segmentation, we tile a story using automatically picked sentences from each abstract. Similarity browsing encourages exploration of related documents akin to manual relevance feedback, whereas we present automatically computed chains for exploration. In future, we seek to support complete information integration schemas using the Storygrapher, as ways to organize and visualize results from multiple runs of storytelling. This will be especially pertinent in combinatorial studies such as studied here.

7 Discussion

Storytelling and novella mining help biologists rapidly explore links in the published literature, yielding greater understanding of the relationships between biological entities, and driving the development of new biological hypotheses that can be experimentally tested. Our work also promotes a 'compositional data mining' approach to prototype complex mining algorithms (storytelling) from simpler algorithms (i.e., similarity search). In future, we plan to incorporate more expressive ways of linking documents, e.g., using relationships derived by language modeling. We also plan to investigate pushing constraints into the storytelling algorithm, e.g., syntactic constraints on story paths, such as must-have and must-not-have requirements. For instance, the biologist might desire a story that does not involve the concepts of UV stress, or one that deliberatively moves from upstream events (stimuli, signals) to downstream concepts (transcription factors, feedback molecules). We also plan to explore the implementation of structured stories where the biologist specifies the slots and roles but not the participants. Finally, new algorithms for compressing multiple stories in ways that expose analogical lines of reasoning, and new approaches to evaluate stories must be investigated. For instance, stories could be judged based on the number of intermediate participants brought into the story, and by their conformance to prior background knowledge.

References

[1] A.E. Carpenter and D.M. Sabatini. Systematic Genome-Wide Screens of Gene Function. *Nature Reviews Genetics*, Vol. 5(1):pages 11–22, Jan 2004.

[2] P. Fung and G. Ngai. One story, One Flow: Hidden Markov Story Models for Multilingual Multi-document Summarization. *ACM Transactions on Speech and Language Processing*, Vol. 3(2):pages 1–16, July 2006.

[3] P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting Hierarchical Domain Structure to Compute Similarity. *ACM Transactions on Information Systems*, Vol. 21(1):pages 64–93, Jan 2003.

[4] P.O. Hassa, S.S. Haenni, M. Elser, and M.O. Hottiger. Nuclear ADP-ribosylation Reactions in Mammalian Cells: Where are we today and where are we going? *Microbiol. Mol. Biol. Rev.*, Vol. 70:pages 789–829, 2006.

[5] M.A. Hearst. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, Vol. 23(1):pages 33–64, 1997.

[6] G.D. Jack, M.C. Cabrera, M.L. Manning, S.M. Slaughter, M. Potts, and R.F. Helm. Activated Stress Response Pathways within Multicellular Aggregates utilize an Autocrine Component. *Cell. Signal.*, doi:10.04017/j.cellsig.2006.10.005, 2007.

[7] P. Kersey and R. Apweiler. Linking Publication, Gene, and Protein Data. *Nature Cell Biology*, Vol. 8:pages 1183–1189, 2006.

[8] A. Kuchinsky, K. Graham, D. Moh, A. Adler, K. Babaria, and M.L. Creech. Biological Storytelling: a Software Tool for Biological Information Organization based upon Narrative Structure. *ACM SIGGROUP Bulletin*, Vol. 23(2):pages 4–5, Aug 2002.

[9] D. Kumar, N. Ramakrishnan, R.F. Helm, and M. Potts. Algorithms for Storytelling. In *Proc. KDD'06*, pages 604–610, Aug 2006.

[10] R. Kumar, U. Mahadevan, and D. Sivakumar. A Graph-Theoretic Approach to Extract Storylines from Search Results. In *Proc. KDD'04*, pages 216–225, 2004.

[11] S. Laxman, K.P. Unnikrishnan, and P.S. Sastry. Discovering Frequent Episodes and Learning Hidden Markov Models: A Formal Connection. *IEEE TKDE*, Vol. 17(11):pages 1505–1517, Nov 2005.

[12] C.S. Lim, M. Potts, and R.F. Helm. Nicotinamide extends the Replicative Life Span of Primary Human Cells. *Mech. Aging Dev.*, Vol. 127:pages 511–514, 2006.

[13] H. Shatkay and R. Feldman. Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology*, Vol. 10(6):pages 821–855, 2003.

[14] J. Singh, D. Kumar, N. Ramakrishnan, V. Singhal, J. Jervis, A. Desantis, J. Garst, S. Slaughter, M. Potts, and R.F. Helm. Transcriptional Response of *Saccharomyces cerevisiae* to Desiccation and Rehydration. *Applied and Environmental Microbiology*, Vol. 71(12):pages 8752–8763, Dec 2005.

[15] M.D. Smucker and J. Allan. Find-similar: Similarity Browsing as a Search Tool. In *Proc. SIGIR'06*, pages 461–468, Aug 2006.

[16] P. Srinivasan and B. Libbus. Mining MEDLINE for Implicit Links between Dietary Substances and Diseases. In *Proc. ISMB/ECCB*, pages 290–296, 2004.

[17] D.R. Swanson and N.R. Smalheiser. An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. *Artificial Intelligence*, Vol. 91(2):pages 183–203, 1997.