AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

Generating Interpretable Hypotheses Based on Syllogistic Patterns

Takuya Hagimura and **Kazuhiro Seki** and **Kuniaki Uehara**
Graduate School of System Informatics, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
seki@cs.kobe-u.ac.jp

Abstract

The ever-growing literature in biomedicine makes it virtually impossible for individuals to grasp all the information relevant to their interests. Since even experts' knowledge is likely to be incomplete, important associations among key biomedical concepts may remain unnoticed in the flood of information. Discovering those implicit, hidden knowledge is called *hypothesis discovery*. This paper reports our preliminary work on hypothesis discovery, which takes advantage of a syllogistic chain of relations extracted from existing knowledge (i.e., published literature). We consider such chains of relations as implicit patterns or rules to generate potential hypotheses. The generated hypotheses are then ranked according to their plausibility judged from the reliability of the rule which generated the hypothesis and the analogical resemblance between new and existing knowledge. We discuss the validity of the proposed approach on the entire Medline database.

Introduction

The amount of scientific knowledge is rapidly growing beyond the pace one could digest. For example, Medline,[1] the most comprehensive bibliographic database in life science, currently contains over 19 million references to journal articles and 2,000–4,000 completed references are added each day. Given the substantial volume of the publications, it is virtually impossible for any individuals to deal with the information without the aid of intelligent information technologies, such as text data mining (TDM) (Ananiadou, Kell, and Tsujii 2006; Kostoff et al. 2009).

TDM aims to discover heretofore unknown knowledge through an automatic analysis on textual data. A pioneering work in TDM, also known as literature-based discovery or hypothesis discovery, was conducted by Swanson in the 1980's. He argued that there were two premises logically connected but the connection had been unnoticed due to overwhelming publications and/or over-specialization. To demonstrate the validity of the idea, he manually analyzed a number of articles and identified logical connections implying a hypothesis that fish oil was effective for clinical treatment of Raynaud's disease (Swanson 1986). The hypothesis was later supported by experimental evidence (DiGiacomo, Kremer, and Shah 1989).

This study is motivated by the series of Swanson's work (Swanson 1987; 1988; 1990; Swanson and Smalheiser 1997; Swanson, Smalheiser, and Torvik 2006) and attempts to advance the research in hypothesis discovery. Specifically, we aim to address two problems that the existing work has generally suffered from. One is the unknown nature of a generated hypothesis. Most approaches only identify two potentially associated concepts but the meaning of the association is unknown, calling for expertise to interpret the hypothesis. This vagueness significantly limits the utility of hypothesis discovery. To cope with the problem, we derive hypothesis generation rules from numerous known facts extracted from the biomedical literature. Each rule explicitly states the meaning of an association as a predicate and is able to produce a hypothesis in the form of "N_1 V N_2", where N and V denote a noun and a verb phrase, respectively.

The second problem is the large number of generated hypotheses. Typically, most of the hypotheses are spurious and only a small fragment is worth further investigation. Because the latter is far outnumbered by the former and thus is difficult to find, it is crucial to prioritize the hypotheses according to their plausibility. To this end, we define a plausibility measure based on the reliability of the hypothesis generation rules and the semantic similarities between concepts associated with a discovered hypothesis and those with the hypothesis generation rule.

Through an experiment on the Medline database, we discuss the validity of the hypothesis generation rules and the proposed plausibility.

Related Work

Swanson has argued the potential use of a literature to discover new knowledge that has implicitly existed but been overlooked for years. His discovery framework is based on a syllogism; That is, two premises, "A causes B" and "B causes C," suggest a potential association, "A causes C," where A and C do not have a known, explicit relationship. Such an association can be seen as a hypothesis testable for verification to produce new knowledge, such as the aforementioned association between Raynaud's disease and fish oil. For this particular example, Swanson manu-

[1] http://www.ncbi.nlm.nih.gov/entrez

ally inspected two groups of articles, one concerning Raynaud's disease and the other concerning fish oil, and identified premises that "Raynaud's disease is characterized by high platelet affregability, high blood viscosity, and vasoconstriction" and that "dietary fish oil reduces blood lipids, platelet affregability, blood viscosity, and vascular reactivity," which together suggest a potential benefit of fish oil for Raynaud's patients. Based on the groundwork, Swanson himself and other researchers developed computer programs to aid hypothesis discovery. The following introduces some of the representative studies.

Weeber et al. (2001) implemented a system, called DAD-system, taking advantage of a natural language processing tool. The key feature of their system is the incorporation of the Unified Medical Language System (UMLS) Metathesaurus[2] for knowledge representation and pruning. While the previous work focused on words or phrases appearing in Medline records for reasoning, DAD-system maps them to a set of concepts defined in the UMLS Metathesaurus using MetaMap (Aronson 2001). An advantage of using MetaMap is that it can automatically collapse different wordforms (e.g., inflections) and synonyms to a single Metathesaurus concept. In addition, using *semantic types* (e.g., "Body location or region") under which each concept is categorized, irrelevant concepts can be excluded from further exploration if particular semantic types of interest are given. This filtering step can drastically reduce the number of potential associations, enabling more focused knowledge discovery. Pratt and Yetisgen-Yildiz (2003)'s system, LitLinker, is similar to Weeber's, also using the UMLS Metathesaurus but adopted a technique from association rule mining (Agrawal et al. 1996) to find two associated concepts.

Srinivasan (2004) developed another system, called Manjal, for hypothesis discovery. A primary difference of Manjal from the previous works is that it solely relies on MeSH[3] terms assigned to Medline records, disregarding all textual information. MeSH is a controlled vocabulary consisting of sets of terms (MeSH terms) and was developed for manually indexing articles in life science by the National Library of Medicine (NLM). Manjal conducts a Medline search for a given concept and extracts MeSH terms from the retrieved articles. Then, according to predefined mapping, each of the extracted MeSH terms is associated with its corresponding UMLS semantic types. Similar to DAD-system, the subsequent processes can be restricted only to the concepts under particular semantic types of interest, so as to narrow down the potential pathways. In addition, Manjal uses the semantic types for grouping resultant concepts in order to help users browse system output.

More recently, Liu et al. (2011) proposed an approach to hypothesis discovery based on hypergraphs. In the approach, concepts and their direct associations (co-occurrences) are represented by nodes and edges, respectively, and the strength of direct/indirect associations between two concepts were defined using their commute time (the time taken for a random walk to make a round trip between two nodes) or inner product. They evaluated the approach on synthetic small data, along with shopping basket and clinical note data, and showed that Swanson's hypothesis regarding fish oil can be replicated.

Despite the prolonged efforts, however, the research in hypothesis discovery is still at an early stage of development, leaving much room to improve. Most of the previous works only suggest two concepts indirectly associated without indicating their nature of the association. Also, their evaluation was typically limited only to a small number of known hypotheses reported in Swanson's work. In contrast, this study generates hypotheses with explicit meaning and quantitatively evaluates them against a number of known associations automatically extracted from Medline.

Proposed Approach

The focus of this work is twofold. One is to develop a framework to derive hypothesis generation rules from known facts or relations represented in predicate argument structure extracted from a corpus of texts. The other is to define a plausibility measure to rank hypotheses generated from the rules.

Deriving Hypothesis Generation Rules

Much previous work generates hypotheses simply based on co-occurrences of two terms or concepts. Although such approaches may produce valid hypotheses, they also produces even more spurious ones, making it more difficult to spot truly important hypotheses. This study takes into account the meaning of the relation between two concepts instead of their simple co-occurrence and only produces more reasonable hypotheses in consideration of the existing knowledge. In this study, each known fact or relation extracted from existing knowledge is expressed as a predicate-argument structure "N_1 V N_2", where V is a predicate and N_1 and N_2 are subjective and objective arguments, respectively. Based on the same arguments, these relations are merged to identify a *chain of relations* described shortly to derive a hypothesis generation rule.

Knowledge Extraction. To extract known relations from the literature, this study relies on publicly available NLP tools, specifically, a syntactic parser and a named entity (NE) recognizer. The former identifies predicate-argument structure and the latter identifies biomedical entities, including proteins and RNA, so as to generate biomedically meaningful hypotheses. Hereafter, the set of extracted relations is refereed to as the *knowledge base*.

Rule Induction and Hypothesis Generation. From the knowledge base, a hypothesis generation rule is derived as a sequence of predicates. The basic idea is to identify a syllogistic pattern composed of three relations corresponding to two premises and one conclusion in Swanson's syllogism (see the *Related Work* section) by merging the same arguments. For example, suppose that two relations were extracted from the literature: "N_1 inhibits N_2" and "N_2 directs N_3". The objective and subjective arguments of the former

[2] UMLS is an NLM's project to develop and distribute multipurpose, electronic knowledge sources and its associated lexical programs.

[3] http://www.nlm.nih.gov/mesh/

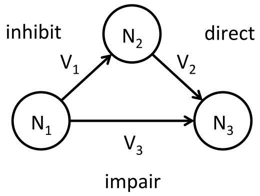

Figure 1: A chain of relations leading to a hypothesis generation rule.

and latter, respectively, are the same (i.e., N_2) and thus form a chain of two relations by merging them: "N_1-inhibit-N_2-direct-N_3". Here, the previous work in hypothesis discovery may suggest that N_1 and N_3 have *some* implicit association without specifying the meaning of the association. Instead, we take a further step to search for another relation involving N_1 and N_3, such as "N_1 impairs N_3".[4] This time, the subjective and objective arguments are the same as those of the first two relations, respectively. Further merging the arguments produces a triangular chain of relations as shown in Figure 1.

These known relations collectively suggest that a generalized rule below may apply:

> **Rule:** If "x inhibit y" and "y direct z", then "x impair z",

where x, y, and z can be any noun phrases. Note that the rule only indicates a possible association that may not be valid. However, because the possible association follows a more reasonable logic than mere co-occurrences of concepts, fewer spurious hypotheses would be expected than the previous work.

These rules can be easily identified by first finding two predicate-argument structures that share the same argument as the object and subject (i.e., N_2), and then finding another predicate-argument structure having the other two arguments (N_1 and N_3) as its subject and object. Once such rules are exhaustively identified in the knowledge base, they can be applied back to the knowledge base to generate hypotheses in which the exact meaning of the association between two concepts is explicitly stated as a predicate (i.e., "impair" in the above example).

Ranking Generated Hypotheses

The number of hypotheses to be generated from the rules will be much smaller than co-occurrence-based approaches. Still, there will be many hypotheses that hinder manual investigation. Therefore, it is crucial to sort the generated hypotheses in the order of their plausibility. To this end, we define a plausibility score and compute it for each generated hypothesis.

There are two types of information that can be used to determine the plausibility $P_{r,h}$ of each hypothesis h generated by a rule r. One is associated with r itself and the other is associated with r and h. For the former, we consider the reliability of r (denoted as P_r), and for the latter, consider the applicability of r to h (denoted as P_h). For this work, we simply define $P_{r,h}$ as the weighted average of P_r and P_h as follows:

$$P_{r,h} = t \times P_r + (1-t) \times P_h, \tag{1}$$

where t is a parameter to control the relative importance of P_r with respect to P_h. The following discusses the definitions of P_r and P_h in turn.

Reliability of hypothesis generation rules. All hypothesis generation rules are not equally reliable; Some may well represent logical connections among biomedical concepts and may lead to true hypotheses, and the others may have been derived by pure coincidence and may result in false ones. In order to distinguish more plausible hypotheses from the rest, it is important to consider the reliability of hypothesis generation rules.

To quantify the reliability of rule r, this study borrows the concepts developed in association rule mining (Agrawal et al. 1996), specifically *support* and *confidence*. Briefly, support is an empirical probability that an event occurs, and confidence is an empirical conditional probability that an event occurs when another event occurs. Rules with both high support and confidence are generally more reliable in association rule mining.

To define the reliability of rule r, we focus on the fact that a pair of premises often lead to different conclusions. That is, there are often many conclusions, N_1 V_3 N_3, with the same N_1 and N_3 but with different V_3 (see Figure 1). Intuitively, if a particular verb appears as V_3 more often than others, the reliability of the rule involving the verb should be higher. Following this intuition, we define the reliability of r as the confidence as follows:

$$P_r = conf_r = \frac{n(p_1, p_2, c)}{n(p_1, p_2)}, \tag{2}$$

where p_1, p_2, and c denote two premises and a conclusion, respectively, and $n(\cdot)$ is the number of chains of relations containing the arguments.

It is also important to consider the support $supp_r$ of rule r defined as:

$$supp_r = \frac{n(p_1, p_2)}{n_{all}}, \tag{3}$$

where n_{all} is the total number of chains of relations identified in the knowledge base. Low $supp_r$ indicates that r is endorsed by only a small number of cases. Using $supp_r$ as a threshold could filter out unreasonable rules that may have been derived by chance.

Applicability of rules to hypotheses. A hypothesis is thought to be more plausible if the rule which generated the hypothesis is more appropriate to the context (two premises) to be applied. We define this applicability of a rule as analogical resemblance between the chain of relations from which the rule was derived and that associated with the generated hypothesis. Figure 2 illustrates the idea.

[4] In fact, three relations, "actinomycin D inhibits mRNA", "mRNA directs protein synthesis", and "actinomycin D impairs protein synthesis", were extracted from Medline.

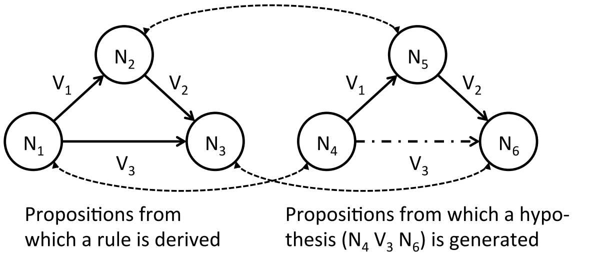
N$_1$
N$_2$
N$_3$
V$_1$
V$_2$
V$_3$
N$_4$
N$_5$
N$_6$
V$_1$
V$_2$
V$_3$
Propositions from which a rule is derived
Propositions from which a hypothesis (N$_4$ V$_3$ N$_6$) is generated

Figure 2: Analogical resemblance between two chains of relations. A dashed line connects two concepts that play the same role in the syllogistic rule represented by V$_1$, V$_2$, and V$_3$.

In Figure 2, the left triangle is the chain of relations from which a rule is derived and the right triangle is the two premises "N$_4$ V$_1$ N$_5$" and "N$_5$ V$_2$ N$_6$" from which a possible hypothesis "N$_4$ V$_3$ N$_6$" is inferred. A dashed line connects two concepts that play the same role in the syllogistic rule represented by a sequence of predicates V$_1$, V$_2$, and V$_3$. If the connected concepts are semantically more similar to each other, the rule is likely to be more applicable to the right triangle with concepts N$_4$, N$_5$, and N$_6$. Thus, we define the applicability of rule r based on the semantic similarity between concepts.

There is much work in estimating the semantic similarity of two concepts. This study adopts a corpus-based approach, specifically, *Normalized Google Distance* (NGD) (Cilibrasi and Vitanyi 2007). NGD is an approximation of Normalized Information Distance (NID) and replaces the Kolmogorov complexity in the formulation with the number of Google hits as defined in

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}, \quad (4)$$

where $f(x)$ is the number of hits by Google search with query x and N is the number of web pages indexed by Google. NGD(x,y) ranges from 0 to ∞ and NGD(x,y)=0 means that they are identical.

Instead of Google, however, we use Medline, which would better reflect the domain knowledge and also ensures that $f(x)$ for any concept x will exist (non-zero) since the concepts in our propositions are extracted from Medline. Although the formulation is exactly the same, we call the distance used with Medline *Normalized Medline Distance* (NMD) to avoid unnecessary confusion. It should be mentioned that Lu and Wilbur (2009) also used Medline for computing NGD.

As with NGD, NMD is a distance measure, taking a value between 0 and ∞. We define a similarity measure based on NMD as follows:

$$\mathrm{Sim}(x,y) = 1 - \frac{\mathrm{NMD}(x,y)}{M}, \quad (5)$$

where M is the maximum NMD value between two concepts from all the propositions extracted from Medline.

We combine three similarity values computed for three pairs of concepts (see Figure 2) by the harmonic mean to define the applicability and take it as P_h.

$$P_h = \frac{3}{\frac{1}{\mathrm{Sim}(N_1,N_4)} + \frac{1}{\mathrm{Sim}(N_2,N5)} + \frac{1}{\mathrm{Sim}(N_3,N_6)}} \quad (6)$$

The choice of the harmonic mean is experimental considering its characteristic that all values (similarity) need to be high to yield high applicability. Incidentally, for the same rule that has been derived from different propositions, the maximum value of Equation (6) is used as P_h.

Evaluation

Experimental Design

Evaluating generated hypotheses is difficult because hypotheses are by definition not known to be true or false. Investigating the validity of each hypothesis is cumbersome and often practically impossible without actual laboratory experiments. This also hinders a large scale quantitative evaluation of a given approach.

This study took a certain period of knowledge (literature) and used it as seed knowledge to induce hypothesis generation rules and used it again to generate hypotheses by applying back the rules. To validate the hypotheses, a more recent period of literature was used as test data. From the test data, all the relations were extracted in a predicate-argument structure. Then, the relations residing only in the later period were considered as true hypotheses. If the generated hypotheses were found in the true hypotheses, they were thought to be validated.

As an evaluation metric, we used *area under the ROC curve* (AUC). AUC is often used to evaluate and compare the performance of classifiers and takes a value between 0 and 1, with 1 being the perfect classification and 0.5 being random guess. In terms of classification, positive examples are the true hypotheses from the newer literature.

Experimental Procedure and Settings

The experiment was carried out in the following procedure and settings. First, existing knowledge was extracted as a predicate-argument structure from the abstracts of the Medline records published between 1949 and 2009 (58 years) using the Enju syntactic parser (Miyao and Tsujii 2008). In our knowledge representation, nouns and verbs were converted to their lemmas such that the same word with different tense or number can be matched. For passive voice, relations were extracted as in "N$_1$ be-V N$_2$". Also, relations containing general verbs were disregarded. We considered "be", "play", "have", and "take" to be too general to yield meaningful, specific hypotheses. We also discarded the relations containing no biomedical named entity so as to produce biomedical hypotheses. For this purpose, we used the GENIA Tagger (Tsuruoka and Tsujii 2005) to identify biomedical entities, including proteins, DNA, and RNA. This procedure resulted in 2,999,350 relations. Among them, 2,421,832 relations extracted from the publications until 2006 were used to derive hypothesis generation rules as described next.

Table 1: Performance of the proposed approach in AUC for different values of minimum support (MS) and parameter t.

MS	t	AUC
0.0004	0 (P_h only)	0.851
	1 (P_r only)	0.452
	0.5 ($P_{r,h}$)	0.732
0.0009	0 (P_h only)	0.827
	1 (P_r only)	0.492
	0.5 ($P_{r,h}$)	0.760
0.0019	0 (P_h only)	0.738
	1 (P_r only)	0.623
	0.5 ($P_{r,h}$)	0.763

Biomedical entities such as genes and proteins typically have many different names including aliases and acronyms. Since our approach relies on surface clues, it is important to normalize them before deriving rules so as to treat the same concept consistently. We used the Entrez Gene database[5] to compile a gene name dictionary, which includes, for example, the following mappings:

- MMP3 → matrix metalloproteinase 3,
- crr1 → cytokinin response regulator 1,
- pip1b → plasma membrane intrinsic protein 1.

Such mapping rules were applied to the extracted knowledge base and 559,849 entities were normalized. Then, 10,446 hypothesis generation rules were identified. (As compared with the case where normalization was not done, 172 more rules were found.)

The rules were then applied back to the knowledge base from which they were derived to exhaustively find new hypotheses. In generating hypotheses, the support of each hypothesis was computed and those below a predefined threshold (minimum support) were discarded. The minimum support values were experimentally set to 0.0019, 0.0009, and 0.0004 corresponding to the raw counts of 20, 10, and 5, respectively. The generated hypotheses were then ranked according to the plausibility score $P_{r,h}$ as defined in Equation (1). Parameter t was set to 0.5, giving equal importance to P_r and P_h.

Results

Table 1 summarizes the results for different values of minimum support, where the results for parameters t=1 and t=0 are also presented; They correspond to cases where only P_r and P_h are used, respectively.

Table 1 shows that, with t=1, namely, only P_r was used for ranking hypotheses, smaller minimum support (MS) resulted in poorer performance. Especially, MS=0.0004 and 0.0009 yielded the results poorer than random guess. This result suggests that for P_r to be reliable and useful, there must be certain amount of cases (chains of relations) sharing the same premises.

When t was set to 0 (i.e., only P_h was used for ranking), the AUC are generally greater than the other cases, indicating the usefulness of analogical resemblance in measuring the plausibility of hypotheses. Contrary to the previous results with t=1, however, AUC decreased with larger MS values. The change of AUC could contribute to the interaction between P_h and MS or to MS alone. We incline to think the latter since P_h is basically independent of MS in contrast to P_r.

Lastly, when t=0.5, meaning that both P_r and P_h were equally taken, AUC increased with larger MS. With MS=0.0019 (the largest in our experiments), using both P_r and P_h performed better than using either plausibility measure, although the AUC is not the highest across all the settings.

Overall, the result is somewhat confusing in that it is not clear if a larger MS limit or combining P_r and P_h was effective since using only P_h with a lower MS limit yielded the best AUC. To better understand the results, we looked at the validated hypotheses (i.e., true hypotheses). Table 2 shows the five true hypotheses from the top of the rankings by each parameter setting, where predicates are shown in boldface.

Although these hypotheses are all validated ones (meaning that they were also found as true hypotheses in the test data), some hypotheses with low MS (0.0004) and t=0 tend to contain somewhat general concepts, such as "ligand", "cytokine", and "immune response", which are unlikely to be very informative. On the other hand, larger MS with t=0 lead to hypotheses involving more specific entities, such as "matrix metalloproteinase 3", "COX 2 expression", and "human neutrophil elastase". A possible interpretation of the greater AUC with lower MS and t=0 (P_h only) is that low MS resulted in more true hypotheses but with broader concepts. More analysis is needed to understand the role of MS, P_r, and P_h in hypothesis generation.

Conclusion

This paper focused on syllogistic patterns of relations and explored a new approach to literature-based discovery. The key intuition is that a generalized rule can be abstracted from such syllogistic patterns of existing knowledge. The rule can then be applied back to the existing knowledge to generate hypotheses. To validate the idea, we exhaustively identified such hypothesis generation rules in over 50 years' worth of Medline records and applied the rules to generate hypotheses. Furthermore, we developed a ranking criteria to prioritize the resulting hypotheses and quantitatively evaluated how well true hypotheses could be ranked to the top. Through the experiment, it was found that the plausibility measures, especially P_h quantifying the applicability of the rule, appeared to be useful to make true hypotheses more visible. Also, limiting rule generation by minimum support was found necessary to make P_r reliable.

Although the results seem promising, the present work is largely preliminary and much work needs to be done. For example, the test data were automatically extracted relations from Medline and their quality is unknown, which may pose a question of how precise the reported performance is. Also, hypothesis ranking could be improved by employing a supervised learning model and by incorporating other types of information to judge the plausibility of hypothesis. We will continue this work to resolve these issues.

[5] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

Table 2: Top five hypotheses in the rankings by different parameter settings. Predicates are shown in boldface.

MS	t	Generated hypotheses
0.0004	0	ligand **bind** immunoglobulin G mutant protein **bind** ribsomal rna b cell **be stimulated by** cytokine neutrophil **be coincubated with** eosinophil cytokine **be regulated in** immune response
	1	hybrid receptor **bind** insulin like growth factor 1 substance **recruit** eosinophil mouse **be depleted of** macrophage interleukin **activate** p38 mapk interleukin **activate** t cell
	0.5	ligand **bind** immunoglobulin G mutant protein **bind** ribsomal rna b cell **be stimulated by** cytokine cytokine **be regulated in** immune response neutrophil **be coincubated with** eosinophi
0.0009	0	tumor necrosis factor alpha **be increased in** diabetes latent membrane protein1 **stimulate** STAT3 thrombin **stimulate** matrix metalloproteinase 3 ethanol **stimulate** COX 2 expression interleukin **be elevated in** all patient
	1	hybrid receptor **bind** insulin like growth factor 1 substance **recruit** eosinophil mouse **be depleted of** macrophage interleukin **activate** t cell cell **be preincubated with** tumor necrosis factor alpha
	0.5	tumor necrosis factor alpha **be increased in** diabetes thrombin **stimulate** matrix metalloproteinase 3 interleukin **be elevated:elevate in** all patient mouse **be depleted of** macrophage interleukin **activate** t cell
0.0019	0	latent membrane protein1 **stimulate** STAT3 thrombin **stimulate** matrix metalloproteinase 3 ethanol **stimulate** COX 2 expression human neutrophil elastase **activate** pro matrix metalloproteinase 9 C. pneumoniae **activate** macrophage
	1	mouse **be depleted of** macrophage cell **express** cytokine and chemokine Th2 cell **inhibit** Th1 cell lps **activate** nf kappab and mapk pathway C. pneumoniae **activate** macrophage
	0.5	thrombin **stimulate** matrix metalloproteinase 3 mouse **be depleted of** macrophage lps **activate** nf kappab and mapk pathway C. pneumoniae **activate** macrophage macrophage **activate** nk cell

Acknowledgments

This project is partially supported by the Kayamori Foundation grant #K23-XVI-363.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A.; et al. 1996. Fast discovery of association rules. *Advances in knowledge discovery and data mining* 12:307–328.

Ananiadou, S.; Kell, D. B.; and Tsujii, J. 2006. Text mining and its potential applications in systems biology. *Trends in Biotechnology* 24(12):571–579.

Aronson, A. R. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of AMIA Symposium 2001*, 17–21.

Cilibrasi, R. L., and Vitanyi, P. M. B. 2007. The Google similarity distance. *IEEE Trans. on Knowl. and Data Eng.* 19:370–383.

DiGiacomo, R. A.; Kremer, J.; and Shah, D. 1989. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *The American Journal of Medicine* 86(2):158–164.

Kostoff, R. N.; Block, J. A.; Solka, J. L.; Briggs, M. B.; Rushenberg, R. L.; Stump, J. A.; Johnson, D.; Lyons, T. J.; and Wyatt, J. R. 2009. Literature-related discovery. *Annual Review of Information Science and Technology* 43(1):1–71.

Liu, H.; Le Pendu, P.; Jin, R.; and Dou, D. 2011. A hypergraph-based method for discovering semantically associated itemsets. In *Proceedings of the 11th IEEE ICDM*, 398–406.

Lu, Z., and Wilbur, W. J. 2009. Improving accuracy for identifying related PubMed queries by an integrated approach. *Journal of Biomedical Informatics* 42(5):831–838.

Miyao, Y., and Tsujii, J. 2008. Feature forest models for probabilistic hpsg parsing. *Computational Linguistics* 34(1):35–80.

Pratt, W., and Yetisgen-Yildiz, M. 2003. Litlinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd international conference on Knowledge capture*, 105–112. ACM.

Srinivasan, P. 2004. Text mining: generating hypotheses from Medline. *JASIST* 55(5):396–413.

Swanson, D. R., and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* 91(2):183–203.

Swanson, D. R.; Smalheiser, N. R.; and Torvik, V. I. 2006. Ranking indirect connections in literature-based discovery: the role of medical subject headings. *JASIST* 57(11):1427–1439.

Swanson, D. R. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30(1):7–18.

Swanson, D. R. 1987. Two medical literatures that are logically but not bibliographically connected. *JASIS* 38(4):228–233.

Swanson, D. R. 1988. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine* 31(4):526–557.

Swanson, D. R. 1990. Somatomedin C and arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine* 33(2):157–179.

Tsuruoka, Y., and Tsujii, J. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of HLT/EMNLP*, 467–474.

Weeber, M.; Klein, H.; de Jong-van den Berg, L. T. W.; and Vos, R. 2001. Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *JASIST* 52(7):548–557.