Discovering Protein Similarity using Natural Language Processing

Indra Neil Sarkar, M.Phil.[1] and Thomas C. Rindflesch, Ph.D.[2]

1. Department of Medical Informatics, Columbia University College of Physicians and Surgeons, New York, NY, USA

2. Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Extracting protein interaction relationships from textual repositories, such as MEDLINE, may prove useful in generating novel biological hypotheses. Using abstracts relevant to two known functionally related proteins, we modified an existing natural language processing tool to extract protein interaction terms. We were able to obtain functional information about two proteins, Amyloid Precursor Protein and Prion Protein, that have been implicated in the etiology of Alzheimer's Disease and Creutzfeldt-Jakob Disease, respectively.

INTRODUCTION

Comparing attributes of proteins serves as a valuable underpinning for research in molecular biology and may provide insight into further studies, such as the etiology of disease. Protein similarity is conventionally performed using primary or secondary sequences and structural motifs. Considerable effort has also been devoted to discovering protein secondary and tertiary structure, based on just sequential information. However, protein function does not necessarily correlate simply with structure, and purely structural comparisons often lead to confounding results.[1]

Biological processes are complex systems of interactions between the functions of various proteins and protein complexes. These networks of information, which may be encoded in genes, have been an active area of recent research. In hopes of attempting to determine how two particular proteins function, biomedical researchers presently are required to perform extensive literature reviews.

Scientists often use textual databases, such as MEDLINE, to ascertain further information about specific biological entities, such as proteins. MEDLINE is a comprehensive textual database containing over 10 million citations dating back to 1966. Through searching MEDLINE, scientists are able to infer relationships between proteins that may otherwise not be categorized as being similar (e.g., sequence similarity, etc.). This process is outlined in Panel A of Figure 1.

For example, it is known that both Amyloid Precursor Protein (APP) and Prion Protein (PrP) are both involved in similar diseases, namely Alzheimer's Disease (AD) and Creutzfeldt-Jakob Disease (CJD). Both of these proteins are known to be influenced by copper and zinc ions. They are also known to have similar pathological mechanisms. Furthermore, specific neuritic plaques are characteristic in both AD and CJD, as a result of APP and PrP, respectively.[2]

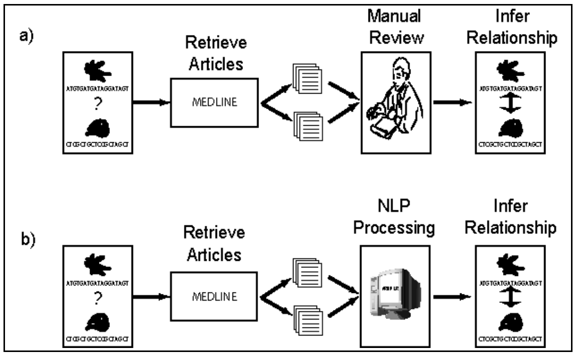

Figure 1. Panel A shows how traditional inferences about sequentially and structurally different proteins is done. Panel B shows how by using NLP processing, manual review can be reduced.

A natural language tool may be used to complement this process (Panel B). Considerable effort is being devoted towards applying NLP techniques towards extracting molecular biology information from the research literature.[3] This includes finding protein interactions[4], inhibition relations[5], enzymatic and metabolic pathways and protein structure[5], and sequence homology[7]. An overview of information extraction from biomedical texts can be found in Andrade and Bork[8]. NLP techniques currently being employed range from highly statistical to highly symbolic. Underlying grammar formalisms include semantic grammars, HPSG, and categorical grammars, as well as template-filling methodologies.[9,10,11,12,13]

In this paper we discuss a pilot project that investigates the possibility of using NLP to focus on the functional attributes of two proteins, which are not structurally similar, but are related through causing similar types of neuronal disorders: APP and PrP. The NLP method we use relies on an underspecified syntactic analysis and then draws on semantic knowledge and general rules for argument identification in English.

METHODS

A previously existing NLP application[14,15] was adapted and generalized in the development of a program (called arbiter_pi) to recover protein functional relationships from a selected set of biomedical titles and abstracts. We focused on three specific types of relationships: INDUCE, INHIBIT, and REGULATE. So, for example, we attempt to extract from the sentence below the fact that 'abeta amyloidosis' induces 'tau accumulation in APP(Sw) mice'.

> **Abeta amyloidosis *induces*** the initial stage of **tau accumulation** in APP(Sw) mice.

Dataset. A domain expert selected 45 MEDLINE citations concerned with either PrP or APP. Of the 386 sentences in these abstracts, 70 were selected for testing, and the remaining 316 were used for development of arbiter_pi. From the 70 sentences that had been set aside for testing, 40 protein interaction predications were marked by hand, and these sentences were used as a gold standard for a preliminary, informal evaluation.

Interaction Determination. The original program had been concerned with a single relationship, binding. Arbiter_pi was modified to apply to addition protein interactions and must be able to recognize assertions about proteins interacting with processes, in addition to other substances, as well as processes interacting with processes. Part of the effort in developing arbiter_pi was directed at recognizing verbs and nominalizations that cue such relationships.

Such syntactic predicates were determined from the development dataset based on the semantic characteristics of the target predications within the individual sentences. In reviewing the text, each relevant verb or nominalization was manually classified as indicating either INDUCE, INHIBIT, and REGULATE. Some examples of such verbs, based on findings from the training set, are:

INDUCES –	induce, activate, stimulate, cause, increase
INHIBIT –	inhibit, attenuate, block, damage, disrupt, impair
REGULATE –	regulate, participate, modulate, mediate

Further work was devoted to allowing arbiter_pi to characterize the arguments that participate in protein interaction predications. Arbiter_pi relies on input noun phrases being mapped[16] to concepts in the Unified Medical Language System (UMLS) Metathesaurus[17]. Concepts in the Metathesaurus are assigned one or more semantic types, which provide allowable semantic categories for the arguments of protein interaction predications. Some examples of the semantic types Arbiter_pi calls on are:

Amino Acid, Peptide, or Protein
Biologically Active Substance
Biologic Function
Cell Function
Cell or Molecular Dysfunction
Molecular Function
Organic Chemical
Organism Function

Other noun phrase heads that cue arguments for protein interaction predications were added, based on scrutiny of the training set. Arbiter_pi also has access to a module that identifies potential protein names by referring to their morphological shape[18]. We intend to pursue more effective recognition for protein name identification[19].

Of the 316 sentences in the training and development set, 124 protein interaction relationships were identified, with distribution as follows:

75	**INDUCE**
32	**INHIBIT**
12	**BIND**
5	**REGULATE**

Some examples of the predications identified by Arbiter_pi in the training set are shown below.

Original Sentence
CT 105 rendered SK-N-SH cells and rat primary cortical neurons more vulnerable to glutamate-induced excitotoxicity.

Arbiter_pi Processed Output
rat primary cortical neuron-INHIBITED_BY-glutamate - induced excitotoxicity

ct 105 rendered sk - n - sh cell-INHIBITED_BY-glutamate - induced excitotoxicity

glutamate-INDUCES-excitotoxicity

Original Sentence
Collectively, these results suggest that PrP(C) can participate in signal transduction in human T lymphocytes.

Arbiter_pi Processed Output
prp (c)-REGULATES-signal transduction, human t lymphocytes

Focusing on the goal of looking for a relationship between APP and PrP, we examined retrieved interaction predications looking for common function for both of these proteins.

RESULTS

Several characteristics of an emerging functional profile of APP and PrP were discernible from the arbiter_pi output of even this small pilot project. The pattern that was discernible was that APP and PrP do in fact share a number of similar properties, and this in fact has been cited in the literature as a justification for using animal (mouse) models for studying Alzheimer-like disorders.[2] Some of the relations were expected, such as the inducement of neuronal cells, as shown in the following two sentences and retrieved predications:

Original Sentence
Furthermore, treatment of cultures with 4-methylumbelliferyl-beta-D-xyloside, a competitive inhibitor of proteoglycan glycanation, inhibited APP-induced neurite outgrowth but did not inhibit laminin-induced neurite outgrowth.

Arbiter_pi Processed Output
precursor amyloid protein-INDUCES-neurite outgrowth

Original Sentence
PrP106-126 a peptide fragment of the prion protein induces proliferation of astrocytes.

Arbiter_pi Processed Output
peptide fragment, prion protein-INDUCES-proliferation, astrocyte

Based on a manual review, we found some rather subtle findings of similarity. The following two sentences and extracted relationships indicate that both APP and PrP are involved in inducing DNA synthesis. In this case, the inferential evidence is based on the fact that DNA synthesis occurs right after or during the late-G1 phase of the cell cycle.

Original Sentence
Then, we examined the effect of the amino-terminal fragment of sAPP and the epitope peptide of 22C11 antibody, and found that both of them also promoted DNA synthesis, suggesting that the amino-terminal region of sAPP is responsible for the biological activity.

Arbiter_pi Processed Output
fragment of sAPP-INDUCES-dna synthesis

Original Sentence
PrP106-126 induces increased progression through the cell cycle to late G1 and enhances the level of both p53 and phosphorylated ERKs in astrocytes.

Arbiter_pi Processed Output
prp106 - 126-INDUCES-cell cycle, late g1

DISCUSSION

It is important to note, that in the present study, a great deal of manual user intervention was required in order to develop the rule sets. However, when evaluating the automated results on 70 sentences containing 40 marked predications in the gold standard, arbiter_pi identified 27 protein interaction relationships. Of these, 18 were correctly identified. Recall was thus 45% and Precision was 67%. Errors were noted in several categories.

While the results of this small sample are not definitive, they provide valuable guidance regarding the potential for using this automated approach. As specific issues with natural language are addressed, it is conceivable that an NLP system could be created that would incorporate the domain knowledge required for building such a tool.

Linguistic phenomena that represent a particular challenge to NLP techniques are coordination and anaphora. The following example shows a false negative due to an error in processing coordination. The program failed to note that *APP* and *APLP2* are coordinate in this sentence and that both thus modify *expression*. *APP expression* was then not interpreted as an argument of *modulates*.

Original Sentence
These findings indicate ***APP and APLP2 expression specifically modulates copper homeostasis*** in the liver and cerebral cortex, the latter being a region of the brain particularly involved in AD.

Arbiter_pi Processed Output
aplp2 expression-REGULATES-copper homeostasis
-FN→APP expression-REGULATES-copper homeostasis

The following false negative is due to the fact that we have not yet addressed this form of anaphora. *This activity* in the sentence refers to the predication *APP reduces copper*.

Original Sentence
The activity of the copper binding domain (CuBD) is unknown, however, ***APP reduces copper*** (II) to copper (I) and ***this activity*** could promote copper-mediated neurotoxicity.

Arbiter_pi Processed Output
-FN→APP reduces copper-INDUCES-copper-mediated neurotoxicity

In several instances, although the program did not exactly match the predication marked in the gold standard, it was close. In the following example, the prepostional phrases following *CT10,* namely *for 24 h* and *at a 10 microM concentration* were wrongly taken to modify *CT104* rather than *pretreatment*. Due to the complexity of such situations, it would be difficult to get this type of modification absolutely correct. This may be addressed by loosening the restrictions with how matches are compared to the gold standard, as the overall meaning is still correct (without the modifications).

Original Sentence
We report here that the pretreatment with **CT 105** *for 24 h at a 10 microM concentration* increases intracellular calcium concentration by about twofold in SK-N-SH and PC 12 cells, but not in U251 cells, originated from human glioblastoma.

Arbiter_pi Processed Output
-FN→ct105-INDUCES--intracellular calcium concentration

-FP→ct 105, 24 h, 10 microm concentration-INDUCES-intracellular calcium concentration

CONCLUSION

In many ways, discovering functional similarity from textual information is parallel to what many researchers do at present in order to generate new hypotheses. Often before protein similarity tests can be generated, a scientist must examine the literature. Assistance from natural language processing has the potential to not only increase the number of articles that can be automatically reviewed (e.g., all of MEDLINE) but also the extract potential functional properties about certain proteins that had not previously been noticed.

It is important to state that linguistic approaches will never eliminate the need for experimental validation. Linguistic tools will also never replace biomedical researchers. However, using NLP tools to help generate testable hypothesis may very well prove to be a boon for biomedical scientists. The number of hypotheses that could conceivably be generated could provide researchers with numerous scientifically valid hypotheses that may very well be valid. The results from arbiter_pi could profitably cooperate with research aimed at developing tools to support the use of automatically-generated data to stimulate scientific discovery[20,21,22].

Based on our preliminary results, we were able to hint at similar function for two proteins that do not necessarily share sequence or structural similarity. Future work will focus on continued development of the linguistic capabilities of arbiter_pi aimed at increasing both recall and precision. In addition, we intend to investigate the use of statistical clustering and graphing methods for more effectively managing arbiter_pi output and displaying it insightfully to the molecular biology scientist[23,24].

Acknowledgments

The authors would like to thank Dr. Marin Guentchev, Dr. Marc Weeber, and Dr. Alan Aronson for there insight and assistance for the work presented here. Special thanks goes to Dr. Carol Friedman for her willingness to provide us with valuable comments towards the writing of this manuscript. Special thanks also goes to conversations at Columbia University's Laboratory for Biomedical Language (LABL). INS is supported by a Medical Informatics NLM Training Grant LM-07079-09. This work was done as part of a rotation at the Lister Hill National Center for Biomedical Communications.

References

1. Baker D. and Sali, A. Protein Structure and Structural Genomics. Science, 2001, 93-96.
2. DeArmond S.J. Alzheimer's disease and Creutzfeldt-Jakob disease: overlap of

pathogenic mechanisms. Curr Opin Neurol. (6):872-81. 1993.

3. Hahn U, Romacher M, Schulz S. Creating knowledge repositories from biomedical reports the MEDSYNDIKATE text mining system. Pac. Symp. Biocomput., 2002, 338-49.
4. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;1(1):1-9.
5. Pustejovsky J, Castaño J., Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: Extracting inhibit relations. Pac. Symp. Biocomput., 2002, 362-73.
6. Humphreys K, Demetriou G, Gaizauskas R. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. Pac. Symp. Biocomput., 2000:502-513.
7. Chang JT, Raychaudhuri S, Altman RB. Including biological literature improves homology search. Pac. Symp. Biocomput., 2001:374-383.
8. Andrake M.A. and Bork P. Automated Extraction of Information in Molecular Biology. FEBS Letters, 2000:12-17.
9. Blaschke C., Andrade M. A., Ouzounis C., and Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. Intelligent Systems for Molecular Biology, 1999:60-7.
10. Leroy G, Chen H, Filling preposition-based templates to capture information from medical abstracts. Pac. Symp. Biocomput., 2002:350-361.
11. Park JC, Kim HS, Kim JJ. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. Pac. Symp. Biocomput., 2001:396-407.
12. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M. Automatic extraction of protein interactions from scientific abstracts. Pac Symp. on Biocomp., 2000: 538-49.
13. Yakushiji A, Tateisi Y, Miyao Y, Tsujii J. Event extraction from biomedical papers using a full Parser. Pac. Symp. Biocomput., 2001:408-419.
14. Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relationships from biomedical text. Appl. Nat. Lang. Process., 2000:188-95.
15. Rindflesch TC, Hunter L, Aronson AR. Mining molecular binding terminology from biomedical text. Proc. AMIA Symp., 1999:127-31.
16. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc. AMIA Symp., 2001:17-21.
17. Humphreys B. L., Lindberg D. A. B., Schoolman H. M., and Barnett G. O. (1998) The Unified Medical language System: An informatics research collaboration. JAMIA 1998;5(1):1-13.
18. Fukuda K, Tsunoda T, Tamura A, Takagi T. Toward information extraction: Identifying protein names from biological papers. Pac. Symp. Biocomput., 1998, 707-18.
19. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics, in press.
20. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. 1997;91:183-202.
21. Weeber M, Klien H, Aronson AR, et al. Text- based discovery in biomedicine: the architecture of the DAD-system. Proc. AMIA Symp., 2000:903-7.
22. Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Medinfo., 2001:1344-8.
23. Stapley BJ, Benoit G. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp. on Biocomp., 2000:526-37.
24. Stephens M, Palakal M, Mukhopadhyay S, Raje R, and Mostafa J. Detecting Gene Relations from MEDLINE Abstracts. Pac. Symp. Biocomput., 2001:483-96.