Uncovering the Effects of Genes, Proteins, and Medications on Functions of Wound Healing: A Dependency Rule-Based Text Mining Approach Leveraging GPT-4 based Evaluation

Jayati H. Jui
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
jaj146@pitt.edu

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
milos@pitt.edu

***Abstract*—Wound healing is a complex biological process characterized by intricate cellular and molecular interactions. Understanding the underlying mechanisms and the effects of different biological entities, such as genes, proteins, and medications, on the cellular and biological functions of wound healing is of paramount importance for the development of effective therapeutic interventions. In this paper, we present a text-mining approach aimed to explore and unravel the complex regulatory relationships of genes, proteins, and medications with the biological mechanisms of wound healing. Our approach relies on a set of predefined dependency rules to capture the relationships between biological entities and their target functions from text. By leveraging advanced AI technology like Generative Pre-trained Transformer 4 (GPT-4), also known as ChatGPT, we evaluate the accuracy and quality of the extracted relations. We present a detailed discussion of the encouraging preliminary results that validate the efficacy of our model in identifying potential therapeutic targets in the complex biological system.**

***Index Terms*—Relation Extraction, GPT-4, Wound Healing, Biological Function, Medline**

I. Introduction

Wound healing is a multifaceted biological phenomenon, characterized by a series of intertwined cellular and molecular interactions. This process encompasses a variety of sequential overlapping stages including inflammation, proliferation, and remodeling, each defining a unique set of cellular events regulated by the synergistic influence of numerous genes, proteins, and potential drug targets [1] . Understanding the roles and interactions of these key targets is crucial to effectively control and improve wound healing outcomes. In this work, we focus our attention on biomedical text mining methods that can help us extract relationships between potential regulators and different processes and cellular functions involved in wound healing based on the published scientific literature.

Supported by the Defense Advanced Research Projects Agency (DARPA) through Cooperative Agreement D20AC00002 awarded by the U.S. Department of the Interior, Interior Business Center. The content of the article does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Advanced natural language processing and machine learning methods allow us to effectively navigate extensive text collections, extract target entities and unravel complex relations. In the last decade, numerous methods have been developed for mining complex biological relations among genes, proteins, miRNAs, and drugs, including many rule-based approaches [2]–[5]. Gupta et al. proposed a dependency rule-based technique to draw connections between diseases and mRNA expressions [3]. Ravikumar et al. developed a method for extracting Biological Expression Language statements from evidence sentences gathered from biological literature [5]. Among text mining approaches focused on a specific topic, Eftimov et al. introduced a rules-based named entity recognition system tailored to extract evidence-based dietary recommendations [6]. In this study, we have developed a rule-based relation extraction pipeline that aims to identify (positive and negative) regulatory relations among genes, proteins, pharmaceutical compounds, and different biological processes and cell functions important for wound healing. To limit false positive extractions, we verified the extracted relations using GPT-4 queries. The performance of large-language models (LLMs) like GPT-4 on a wide variety of tasks and queries offers great opportunities to use it as a reliable source of knowledge, including annotation and verification. We have tested our framework on a large collection of Pubmed abstracts.

II. Materials and Methods

A. Corpus

The corpus used in this study is composed of scientific abstracts sourced from Medline, a highly recognized Pubmed-based collection of biological and biomedical literature indexed with Medical Subject Headings (MeSH) terms. We extracted the titles and abstracts from all articles published in the last ten years that were indexed with the Mesh term “wound healing”. The corpus is a collection of $\sim$108K abstracts including a wide range of articles, from basic studies to clinical trials, spanning molecular biology, genomics, proteomics, and

pharmacology relevant to wound healing. The corpus used in this study is available on GitHub[1].

B. System Architecture

Our biomedical relation extraction pipeline is depicted in Fig. 1. It is composed of four main components as described below:

Text Preprocessing: The text preprocessing component is composed of a sentence tokenizer, a word tokenizer, and a parser. Sentences were filtered using a predefined set of trigger words that describe relations, and these filtered sentences were then used for further analysis.

Entity Recognition and UMLS Linking: We performed Named Entity Recognition (NER) using a Python package called **scispaCy**, specifically designed for the extraction of biological and biomedical text. These named entities were linked to UMLS concepts via scispaCy's "UMLS Entity Linker" module. In addition, we applied a filter on these entities based on specific UMLS semantic types, so as to only retain entities that symbolize genes, proteins, drugs, therapies, and cellular, molecular, and biological functions. The set of relation-defining trigger words and UMLS semantic types can be found in Supplementary Document 1[2].

Entity Phrase Augmentation and Merging: Biological functions are often written as multi-word concepts in free text. Named Entity Recognition (NER) frequently encounters difficulties when processing multi-word concepts due to their complex syntactic and semantic structures. Often, these structures complicate the NER's ability to identify an entire entity as a single unit. For instance, the phrase "fibronectin fibril formation" is split into two separate entities, "fibronectin" and "fibril formation", by scispaCy's NER module. The semantic meaning of a multiword concept usually extends beyond the mere aggregation of its parts. Additionally, these phrases can be fragmented by intervening words which can result in the NER incorrectly identifying them as separate entities. For example, "proliferation of fibroblasts" carries a different meaning compared to the separate concepts of "proliferation" and "fibroblasts", as identified by scispaCy's NER. To address such challenges, we developed a custom Entity Phrase Augmentation module that resolves three types of broken entity phrase identification by scispaCy's NER.

- Instances where two entities are adjacent and the first entity is a compound of the second entity in the dependency tree. Examples: 3T3 fibroblast wound healing, bladder cancer cell proliferation, etc.
- Instances where two entities are separated by a prepositional token. Examples: formation of actin stress fibers, cell apoptosis of osteoblasts, etc.
- Instances where an entity is followed by a noun and the entity's dependency on the noun is compound. Examples: IL-6 inhibition, TGF-beta1 gene expression.

Finally, the tokens of the entity phrases are merged together so that each entity phrase is considered as a single token.

[1]https://github.com/juijayati/Rule-based-RE-with-GPT-eval

[2]https://tinyurl.com/yeykc9sy

Relation Extraction and Resolution: Using ScispaCy's Dependency Matcher, we employed a rule-based relation extraction method that relies on the inherent dependency structures found in natural language. We have considered relationships of two types: positive and negative, based on the effects of the targets (genes, proteins, and medications) on specific functions (e.g. biological or cellular functions). The identification of positive and negative relations was determined by a predefined set of trigger words.

- Positive relation trigger words: promote, increase, elevate, up-regulate, boost, etc.
- Negative relation trigger words: reduce, suppress, inhibit, down-regulate, hinder, etc.

We utilized the Dependency Matcher to extract relations from two types of sentences:

- Type 1 Sentence: These sentences explicitly state the relationship between a target entity and a target function entity using a single verb as a trigger. For example, "Prostaglandin E2 inhibits collagen synthesis in dermal fibroblasts".
- Type 2 Sentence: These sentences explicitly state the relationship between a target entity and a target function entity using a verb followed by a noun as triggers. For example, "Polaprezinc induced upregulation of osteogenesis-related genes."

The dependency parse of Type 1 and Type 2 sentences are shown in Fig. 2.

The entity phrases often contain trigger nouns that can influence the directions of the extracted relations. For instance, in the sentence "Global IL-6 inhibition in the early phase after fracture reduced systemic inflammation", the Dependency Matcher initially extracts a negative relation between "IL-6 inhibition" and "systemic inflammation." However, a positive effect of the gene Interleukin-6 (IL-6) on the biological function called "systemic inflammation" can be inferred from the sentence. The relation extraction module resolves such relations through inference by taking into account the trigger nouns in the entity phrases and the relation trigger verbs.

III. Results

A. Evaluation

To the best of our knowledge, there is currently no publicly accessible dataset that provides a comprehensive summary of the positive and negative impacts of genes, proteins, and medications on biological functions related to wound healing. Hence for the assessment of our model, we utilized GPT-4, a groundbreaking LLM developed by OpenAI commonly referred to as ChatGPT. There are many benefits to evaluating models using ChatGPT queries. GPT-4 offers powerful natural language understanding and generation capabilities to generate and verify statements that are both unique and rich in context. By combining the GPT-4 model with constructive queries, we can generate useful synthetic data for evaluating the model's performance.

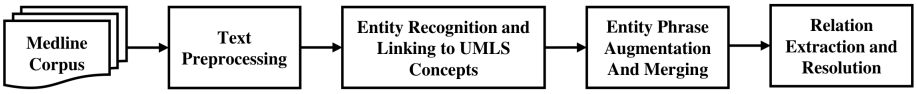
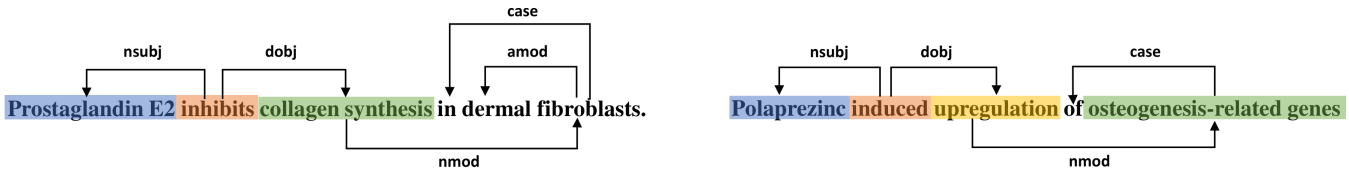


Fig. 1. Relation Extraction Pipeline



(a) Type 1 Sentence

(b) Type 2 Sentence

Fig. 2. Dependency parsing of type 1 and type 2 sentences

We employed both non-contextual and contextual queries to evaluate the accuracy of the extracted relations. Non-contextual queries involved asking ChatGPT to confirm whether a relation was true, false, or unknown whereas contextual queries involved providing ChatGPT with the supporting evidence sentence (context) from which the relation was extracted using our rule-based model.

B. Results and Discussion

We focused our analysis on six major cellular and biological processes involved in wound healing that we selected randomly. The model uncovered 836 relations between genes, proteins, and drugs that modulate these six processes. A comprehensive summary of the model's performance based on the GPT-4 evaluation can be found in Table I. Here, we can see that the inclusion of contextual queries in our evaluation demonstrated a notable enhancement in the model's performance. As shown in Table I, the utilization of contextual queries led to an increase in the proportion of verified relations categorized as "True" and a decrease in relations labeled as "Unknown." Contextual queries provided ChatGPT with the necessary evidence to generate more confident and definitive answers when verifying relations. Consequently, the percentage of relations categorized as "Unknown" decreased significantly with the incorporation of contextual queries. The model's performance with non-contextual queries was particularly strong when evaluating more specific wound healing functions such as angiogenesis, fibroblast proliferation, collagen synthesis, and epithelial-mesenchymal transition. However, non-contextual queries encountered challenges when assessing relations related to the broader term "apoptosis". This difficulty arises from the fact that apoptosis can be influenced by various factors that differ among different cell types. Consequently, verifying relations for such generic becomes more challenging due to the complexity and variability of its underlying mechanisms across different cells. A complete list of extracted relations is available in the project repository[1].

Table II provides a comprehensive summary of the aggregated results for all gene/protein and medication targets. The analysis of the results reveals that 83% of the gene and protein relations and 90% of the medication relations were confirmed as "True" through ChatGPT queries when contextual information was provided. Additionally, the inclusion of context in the queries resulted in a notable decrease in the percentage of "Unknown" relations. The agreement between the two evaluation approaches is depicted in a confusion matrix, as shown in Table III. Based on the results, it can be observed that in this case, 60% of the relations were classified as "True" by both the non-contextual and contextual evaluations. Additionally, 3% of the relations were classified as "False" and less than 1% of the relations were classified as "Unknown" by both evaluations. However, approximately 39.2% of the relations were labeled differently by the two evaluation approaches. To evaluate the relations that were labeled differently by the two evaluation methods, we conducted a manual annotation of 50 randomly selected samples. The evaluation accuracy, considering the manual annotations, was 78% when using contextual information, indicating a high level of agreement between the manual labeling and the contextual evaluation. However, the accuracy dropped significantly to 6% when evaluating the relations without context. This stark contrast in accuracy underscores the difficulty in accurately predicting relationships without the necessary contextual information. The list of manually annotated relations can be found in Supplementary Document 2[3].

Although our method shows promising initial results, our model evaluation draws upon GPT-4's capabilities, alongside its limitations. Evaluating the accuracy of our model or GPT-4 against a meticulously curated dataset of human annotations would be needed to further validate it. Another major limitation of GPT-4 based evaluation is the usage cap imposed on GPT-4, primarily due to resource constraints and computa-

[3]https://tinyurl.com/58p89xvt

TABLE I
MODEL EVALUATION RESULTS USING NON-CONTEXTUAL AND CONTEXTUAL QUERIES

Functions	Relations		GPT-4 Confirmations without Context			GPT-4 Confirmations with Context		
	Target	n	True	False	Unknown	True	False	Unknown
Angiogenesis	Gene/ Protein	193	144 (75%)	13 (7%)	36 (19%)	169 (87%)	19 (10%)	5 (2%)
	Medication	173	119 (69%)	17 (10%)	37 (21%)	159 (92%)	11 (6%)	3 (2%)
Apoptosis	Gene/ Protein	156	91 (58%)	23 (15%)	42 (27%)	112 (72%)	43 (27%)	1 (1%)
	Medication	145	105 (72%)	19 (13%)	21 (14%)	127 (87%)	16 (11%)	2 (1%)
Bone Formation	Gene/ Protein	40	19 (47%)	7 (17%)	14 (35%)	37 (92%)	2 (05%)	1 (2%)
	Medication	38	25 (66%)	2 (5%)	11 (29%)	33 (87%)	2 (5%)	3 (8%)
Fibroblast Proliferation	Gene/ Protein	35	19 (54%)	1 (3%)	15 (43%)	34 (97%)	1 (3%)	0
	Medication	37	26 (70%)	1 (3%)	10 (27%)	34 (92%)	2 (5%)	1 (3%)
Collagen Synthesis	Gene/ Protein	17	11 (65%)	3 (18%)	3 (18%)	15 (88%)	1 (6%)	1 (6%)
	Medication	35	22 (63%)	7 (2%)	6 (17%)	32 (91%)	2 (6%)	1 (3%)
Epithelial-Mesenchymal Transition	Gene/ Protein	37	23 (62 %)	5 (13%)	9 (24%)	32 (86%)	4 (12%)	1 (3%)
	Medication	11	8 (73%)	1 (9%)	2 (18%)	11 (100%)	0	0

TABLE II
SUMMARIZED EVALUATION RESULTS OF GENE/ PROTEIN AND MEDICATION TARGETS USING NON-CONTEXTUAL AND CONTEXTUAL QUERIES

Relations		GPT-4 Confirmations without Context			GPT-4 Confirmations with Context		
Target	n	True	False	Unknown	True	False	Unknown
Gene/ Protein	478	307 (64%)	52 (11%)	119 (25%)	399 (83%)	70 (15%)	9 (2%)
Medication	439	305 (69%)	47 (11%)	87 (20%)	396 (90%)	33 (7%)	10 (2%)

TABLE III
CONFUSION MATRIX OF THE AGREEMENT BETWEEN GPT-4 CONFIRMATIONS WITH AND WITHOUT CONTEXT

		Without Context		
		True	False	Unknown
With Context	True	0.60	0.07	0.18
	False	0.05	0.03	0.04
	Unknown	0.02	0.002	0.003

tional limitations. The usage cap limits the length of queries and the number of interactions that can be processed in a given time period. This limitation can hinder evaluation, particularly when dealing with large relation sets that require extensive queries. To allow for a more comprehensive evaluation, less restrictive access to the GPT-4 API is required.

IV. CONCLUSIONS

In conclusion, this study showcased a novel text mining model geared towards recognizing and interpreting gene, protein, and drug targets that have an impact on the cellular, molecular, and biological processes inherent in wound healing. We introduced a dependency rule-based relation extraction model designed to identify complex functional entities and decipher their relationship with prospective biological targets. Utilizing the power of GPT-4 and employing both contextual and non-contextual queries, we could assess the quality of the extracted relations and validate their accuracy. Our findings highlight our model's capacity to elucidate potential therapeutic approaches for complex biological systems. Going forward, our emphasis will be on incorporating the multi-word concept embedding of complex functional entities to improve relation extraction. While in this study we utilized GPT-4 for the evaluation and validation of the mined relations, we envision the integration of various GPT-based functionalities with other tools to enhance their performance in subsequent endeavors.

REFERENCES

[1] R. J. Crum, S. A. Johnson, P. Jiang, J. H. Jui, R. Zamora, D. Cortes, M. Kulkarni, A. Prabahar, J. Bolin, E. Gann *et al.*, "Transcriptomic, proteomic, and morphologic characterization of healing in volumetric muscle loss," *Tissue Engineering Part A*, vol. 28, no. 23-24, pp. 941–957, 2022.

[2] À. Bravo, J. Piñero, N. Queralt, M. Rautschka, and L. I. Furlong, "Befree: a text mining system to extract relations between genes, diseases and drugs for translational research," *SMBM 2014*, vol. 79, 2014.

[3] S. Gupta, H. Dingerdissen, K. E. Ross, Y. Hu, C. H. Wu, R. Mazumder, and K. Vijay-Shanker, "Dexter: disease-expression relation extraction from text," *Database*, vol. 2018, 2018.

[4] Y. Hou, Y. Xia, L. Wu, S. Xie, Y. Fan, J. Zhu, T. Qin, and T.-Y. Liu, "Discovering drug–target interaction knowledge from biomedical literature," *Bioinformatics*, vol. 38, no. 22, pp. 5100–5107, 2022.

[5] K. Ravikumar, M. Rastegar-Mojarad, and H. Liu, "Belminer: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences," *Database*, vol. 2017, 2017.

[6] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS one*, vol. 12, no. 6, p. e0179488, 2017.