Technische Universität München

Ludwig-Maximillians-Univesität München

Faculty of Informatics

Bachelor's Thesis in Bioinformatics

PubSeq: Amino Acid-based Search Engine for MEDLINE Abstracts

PubSeq: Aminosäuresequenz basierte Suchmachine für MEDLINE Abstrakten

---

Author:
Pandu Raharja

Supervisor:
Prof. Dr. Burkhard Rost
Advisors:
Dr. Guy Yachdav
Juan Miguel Cajuela

Technische Universität München
Faculty of Informatics

July 2015

Declaration of Authorship

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Signed: _____

Date: _____

"People usually think that progress consists in the increase of knowledge, in the improvement of life, but that isn't so. Progress consists only in the greater clarification of answers to the basic questions of life. The truth is always accessible to a man. It can't be otherwise, because a man's soul is a divine spark, the truth itself. It's only a matter of removing from this divine spark (the truth) everything that obscures it. Progress consists, not in the increase of truth, but in freeing it from its wrappings. The truth is obtained like gold, not by letting it grow bigger, but by washing off from it everything that isn't gold."

L. N. Tolstoy

Abstract

Background

In genetic research, it is imperative for biomedical researcher to stay updated on the current state of identified proteins. It was hard – and is getting harder, especially after widespread use of Next-Generation Sequencing (NGS) – for researcher to keep updated on the research into protein he/she is currently investigating. This is furthermore exacerbated by the fact that existing search engines only allow querying abstracts using protein names.

Methods

In this project, I present the first search engine that allows user to find all publications mentioning proteins that are similar or identical to the one he/she's interested in. To achieve this, I created a Solr Index that lists down all gene names that were mentioned in each of MEDLINE abstracts and titles. I then populated the index by scanning the whole MEDLINE corpus, tagging protein names found in title and abstract, normalizing those names into UniProt IDs and pushing the ID mentions onto Solr index. Given user's sequence query, the program runs a BLAST on the sequence and normalizes blast results to UniProt IDs. The program then retrieves articles mentioning this ID and return these to user. For the good usability I offer the whole service in a web interface available in following address.

Zusammenfassung

Hintergrund

In genetischer Forschung ist es erzwingend, dass der/die biomedische ForscherIn mit der aktuellen Landschaft von identifizierten Protein sich ständig informiert. Es war schwierig – und wird immer schwieriger sein, vor allem nach dem verbreiteten Ansatz von Next Generation Sequencing (NGS) Technoligien, um der/die Forscherin mit dem Protein von der Interesse in aktuellem Zustand zu halten. Die Tatsache, dass die aktuelle Suchmaschine von den Artikeln nur Namenbasierte Suche unterstützt, hilft leider nicht weiter.

Methoden

In diesem Projekt stellen wir eine Suchmaschine vor, die erlaubt den Benützer, basiert auf Aminosäuresequenz nach den Artikeln suchen, die das Protein oder die Ähnliche erwähnen. Um dies zu erreichen hatten wir einen Solr Index erstellt, der alle erwähnte Proteine innerhalb jedes MEDLINE Artikels auflistet. Wir füllen sich diesen Index in dem wir den gesamten MEDLINE Corpus durchscannen und alle Proteinname mithilfe eines NLP-Programs detektieren. Wir wurden dann diese Namen in UniProt IDs normalisieren. Diese normalisierte Namen wurden schlielich in unserem Solr Index hinzufügen. Um die Benutzbarkeit dieser Dienstleistung zu maximieren hatten wir auch eine Webschnittstelle entwurfen, die in folgender Addresse verfügbar ist.

Acknowledgements

First and mostly, I would like to thank Prof. Burkhard Rost for the holistic supports provided, be it through the lab infrastructures or himself personally. I would also to thank two of my advisors, Dr. Guy Yachdav and Juan Miguel Cajuela, who have in spite of their busy schedules and great distances (and time differences) patiently advised me through the project. Knowing that both are about to finish their PhD programs, I wish them all the best of luck in their future endeavors.

Also, I would like to thank Tatyana Goldberg for administrative support during my stay at the lab. Also my gratitude for Tim Karl, our awesome system administrator, who has helped us tremendously in incorporating each of the cogs in our pipeline into one coherent system. While not involved in our project personally, I would like to thank Prof. Lars Juhl Jensen of University of Copenhagen for giving us access to his tagger program.

Finally I would also thank all Rostlab members and its extensions, without whom this work would all but possible.

Contents

List of Figures

List of Tables

Abbreviations

BLAST	**B**asic **A**lignment **S**earch **T**ools
MEDLINE	**Med**ical **Li**terature Analysis Retrieval Onli**ne**

For Dad, my unsung hero.

Chapter 1

Introduction

1.1 An Easier Biomedical Research

I'll try to present the main idea of this project in following story. Imagine you in the position as a biomedical researcher, are currently investigating some unknown enzymes that somehow were over-expressed in a patient with medical conditions. Upon some more investigating, you managed to get the sequence of several proteins. Without prior knowledge of the proteins, you would naturally BLAST the sequences and wait a little while while the BLAST is searching the sequence against your local database or some online service. Upon the results were coming, you would naturally want to check the resulting proteins one by one, at least the best matching ones. For each protein, you would want to search for articles that have dealt with this protein before.

Imagine that, instead of having to go through blasting the sequence manually and searching for articles one by one, you could just put in a sequence in a website, wait for a while and get the site returns a list of articles that mention the proteins with similar or exact sequence to the one you have. Not only you would save time and resource during the parts that were handled by website itself, you as a researcher could focus more on the substantial part of the research – that is, finding as much essential information about the unknown protein in as little overhead as possible. Therefore, I created a web service that realizes this. In the service, user would only have to put in the sequence of unknown protein, press the search button and receive at the end a list of articles that mention proteins with identical or similar sequence to queried proteins.

With this small contribution, I hope not only to bridge the gap between sequence and knowledge discovery in biomedical research, but also give researcher more flexibility and insights in their literature research. With also ongoing feature extensions and updates, I would also hope that the service would serve more researchers with more conveniences both in medium and long run.

1.2 BLAST

As the title of this thesis already conveyed, the main idea of this project is to bridge the accessibility and knowledge gap between sequence and the main source of knowledge and reference of previous discoveries – a vast corpora of publications in natural sciences – through a modern search engine. Given a sequence of amino acids, it would be impossible for a human to directly identify directly the protein, let alone the characteristics and the functions and the characteristics of the protein.

Several attempts on bridging one component of the gap was done in eighties and earlier nineties. In 1985, Lipman and Pearson published the first paper mentioning the DNA and protein sequence alignment program FASTA [1]

Further down the road, in 1990, Altschul et al. published the Basic Alignment Research Tool [2]

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

[1] David J Lipman and William R Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985.

[2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215 (3):403–410, 1990.