PubSeq: Amino Acid-based Search Engine for MEDLINE Abstracts

Pandu Raharja pandu.raharja@tum.de

July 20, 2015

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, we 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, we create a database that lists down all gene names that were mentioned in each of MEDLINE abstracts and titles. We populate the database by scanning the whole MEDLINE, tag protein names found in title and abstract, normalize those names into UniProt IDs and index the ID mentions within our database. Given user's sequence query, the program runs a blast on the sequence and normalizes blast results to UniProt IDs. We then retrieve articles mentioning this ID and return these to user.

Abstract

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,.

In this project, we 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, we create a database that lists down all gene names that were mentioned in each of MEDLINE abstracts and titles. We populate the database by scanning the whole MEDLINE, tag protein names found in title and abstract, normalize those names into UniProt IDs and index the ID mentions within our database. Given user's sequence query, the program runs a blast on the sequence and normalizes blast results to UniProt IDs. We then retrieve articles mentioning this ID and return these to user.