Mutation mining: automated extraction of mutation data from scientific publications.

Krzysztof Rataj[a], *Jagna Witek*[a], Tomasz Kosciolek[a], Stefan Mordalski[a], Sabina Smusz[a], Andrzej J. Bojarski[a]

[a] *Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland*
e-mail: krzysztof.rataj@uj.edu.pl

Introduction

Extraction of mutagenesis data is a slow and laborious process, thus manual revision of publications may be not only inconvenient, but also lead to mistakes and oversights. However, automated selection needs careful revision, since mutations are recorded in diversified manner, and often include data about multiple targets.
The existing applications require manually created mutation database,[1] and do not provide thorough publication investigation. Creation of an accessible and reliable mutation mining software, would become an asset in gathering data and database construction. In this research we present a homemade tool handling the extraction of mutagenetic data from text – MutMiner.

About

MutMiner is a metaserver designed for acquisition of mutagenesis data from user-supplied PDF files. It utilizes UniProt[2] and HyperCLDB[3] data to cross-check extracted mutations with coresponding protein sequence. Advanced search engine enables to yield the data about multiple targets, being convenient tools for processing reviews on mutagenesis.

The algorithm

The Mutation Miner decrypts submitted PDF file to plain text, retains protein sequence and processes it by means of regular expressions to single out point mutations. In parallel it performs text scanning to recover occuring organisms or proteins, and consults the findings with UniProt[2] and HyperCLDB[3] databases. The results are then processed through scoring matrix, sorted by their reliability and displayed in a form of protein list with matching mutations.

In order to customize the search and reduce the computational time, input of protein's UniProt ID, Accession Number or Gene Name is strongly recommended. Merely Organism Name is sufficient to perform viable analysis, however in such case search would be narrowed to proteins only. Time consumption may be significantly decreased when scanning is reduced to definite mutation, but obtained data need to be validated manually.

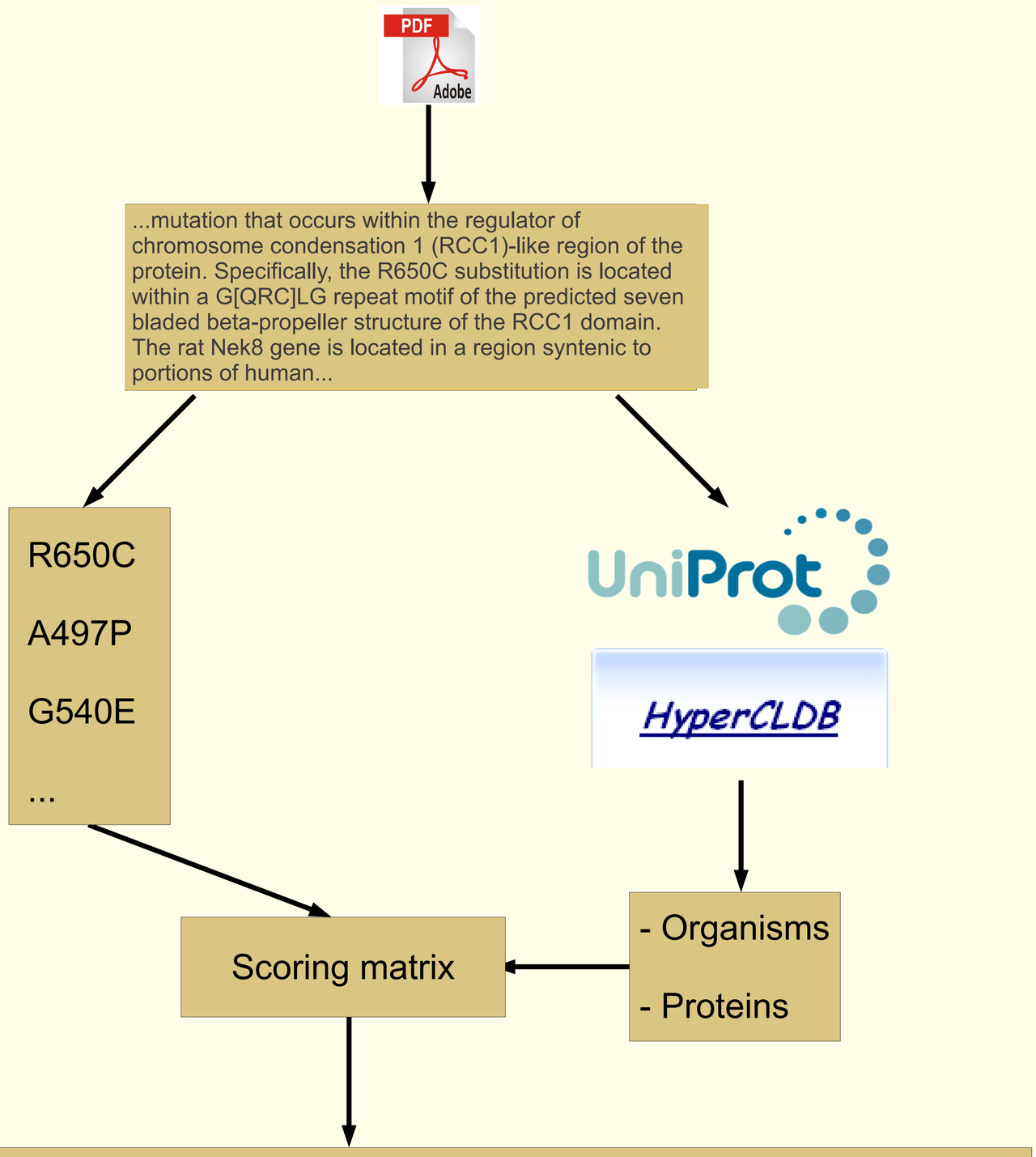
Obtained data is stored in an internal database, so submission of a publication that was already processed, results in an instant data retrieval. Its mining in search of specific mutation is possible, although the database shall grow with the metaserver usage.

Accessing MutMiner

The Mutation Miner is available alongside MutCheck tool at:

anatema.if-pan.krakow.pl.

The software is still in the testing phase, and there are a few options and methods that need to be implemented and optimized. Users are welcome to submit searches, since it would help to expand the mutations database, and optimize the software.



Scheme 1. MutMiner workflow.

Literature

[1] Horn F, Lau AL, Cohen FE.: Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics. 2004 Mar 1;20(4):557-68. 2004 Jan 22.

[2] Magrane M. and the UniProt consortium: **UniProt Knowledgebase: a hub of integrated protein data;** Database, 2011: bar009 (2011).

[3] P. Romano, A. Manniello, O. Aresu, M. Armento, M. Cesaro, B. Parodi: Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. Nucleic Acids Research 2009 37(Database issue):D925-D932. DOI: doi:10.1093/nar/gkn730; PMID: 18927105

Acknowledgments

This study is supported by project UDA-POIG.01.03.01-12-100/08-00 co-financed by European Union from the European Fund of Regional Development (EFRD); http://www.prokog.pl

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY
EUROPEAN UNION
EUROPEAN FUND OF
REGIONAL DEVELOPMENT