excelra

Building a text-mining-based post-translational modification (PTM) database for SUMOylation

CASE STUDY

Purpose

To develop an integrated database for SUMOylation-a post-translational event of interest-from publicly available databases and literature.

Client

Industry
Biotech

Location
US

Sector
Agnostic

Specification

The client wanted to integrate all publicly available data with internal information pertaining to the post-translational modification of their interest (SUMOylation) into a structured format. The data for SUMOylation is very scattered and must be collated and standardized prior to any landscape analysis or target identification. We were engaged to collate information, standardize it, and develop the lexicon.

Input

The SUMOylation lexicon was developed using Excelra's text-mining algorithm.

Workflow

The curation exercise included the identification of databases related to SUMOylation, the recording of common variables and the development of a structured file with the integrated data (fig. 1).

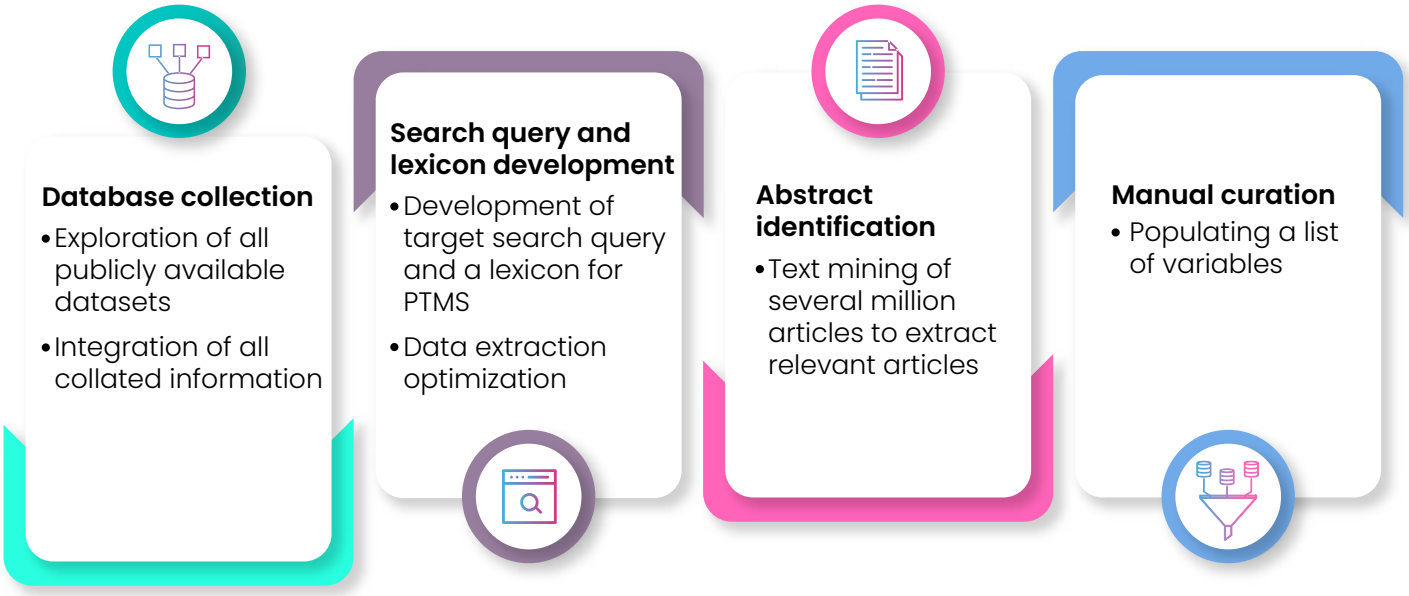

Figure 1: Stages of data collation and variable generation workflow

Excelra's curation environment ensures seamless, quality-assured curation of biomedical data. For the pilot stage of this project, 100 free full-text articles spanning around 22 SUMOylation-mediating enzymes were identified from PubMed central. Article identification was completed using Excelra's literature-mining algorithm and SUMOylation lexicon. The curation exercise was then conducted to shortlist the variables, develop the structured file with variables data, and standardize the SUMOylation and SUMO enzyme lexicon. Finally, the manually curated data and data from database were integrated based on the uniprot ID and the position of SUMOylated amino acid (fig. 2)

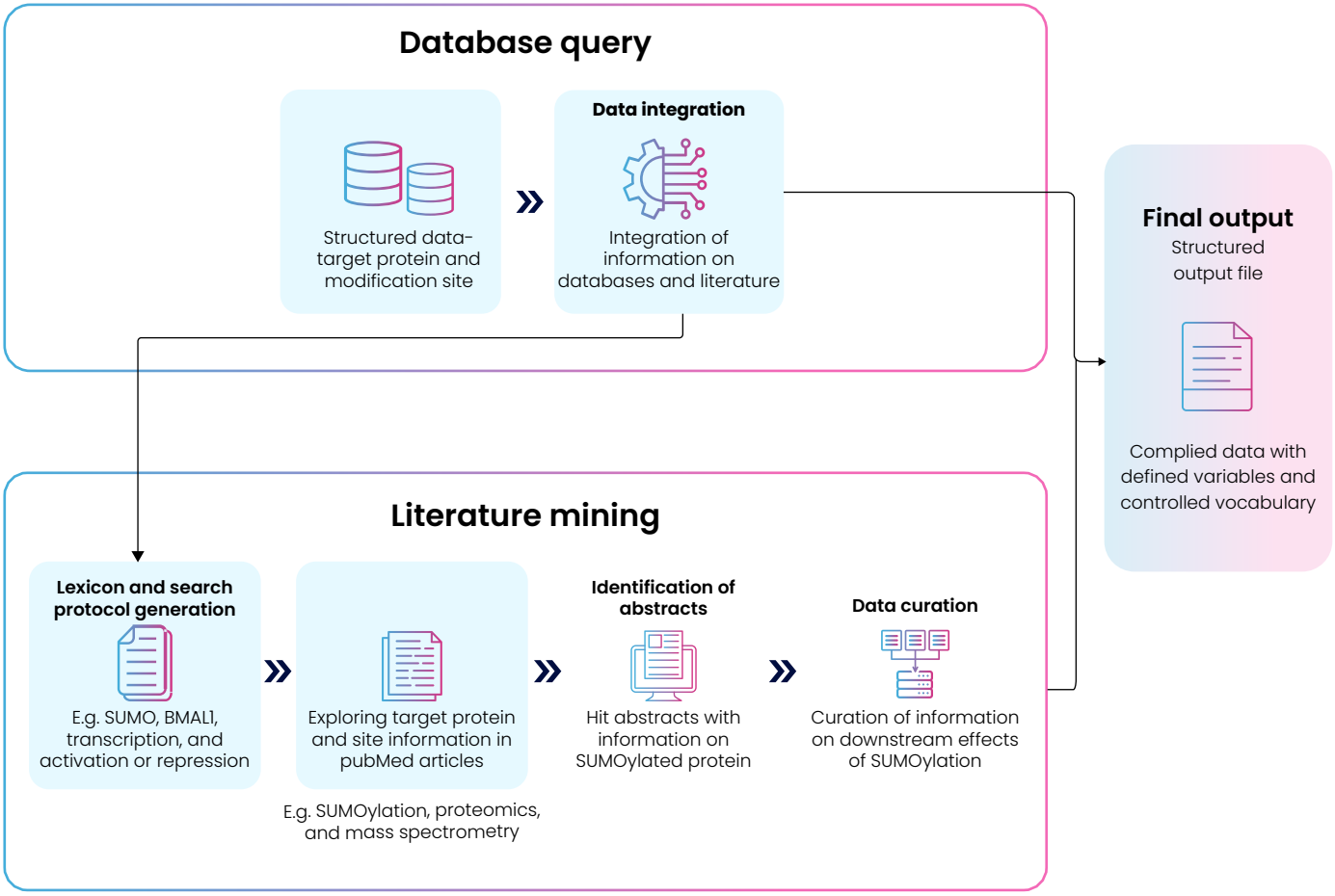

Figure 2: Workflow for data compilation and development of structured output file

Our contribution

Biocuration is imperative to creating a well-annotated and validated database. By leveraging our substantial in-house curation expertise, we were able to identify and collate SUMOylation data from 27 PTM databases and compile into a structured format.

The integrated database was also enriched with data points curated from text mining. This ensured that structured, up-to-date content was provided to the client for eventual deployment in their work environment.

Our service portfolio

Data

Data curation
Filter out the noise, focus your attention

Clinical data
Analysis-ready data for informed clinical decision-making

Semantic data
Refine your decisions, find your value

Insights

Bioinformatics
Illuminate the path to faster discoveries

Data science
Unlock the power of data

Visualization
Pictures paint a thousand words

R&D technology

Product design and development
Unlock your potential with data-driven design and development

Cloud enablement
Optimize your output on the cloud

Data engineering
Mitigate risks, protect your data, and rationalize your portfolio and processes.

Where data means more

excelra

BOSTON | UTRECHT | HYDERABAD

Connect with our experts: marketing@excelra.com

www.excelra.com