Novel Similarity Graphs

Using Neo4j, ChemAxon and Tom Sawyer Perspectives

ChemAxon UGM 2019

SANOFI

Dan Dragos Stefanescu
Scientific Computing

Business Needs

- **Efficient exploration of chemical space around biologically active chemical matter**
  - Integration of diverse information linked to compounds
    - Activity, related drugs, commercially available compounds…
  - Efficient navigation (traversing) and visualization
    - Exploitation of neighborhood relationships

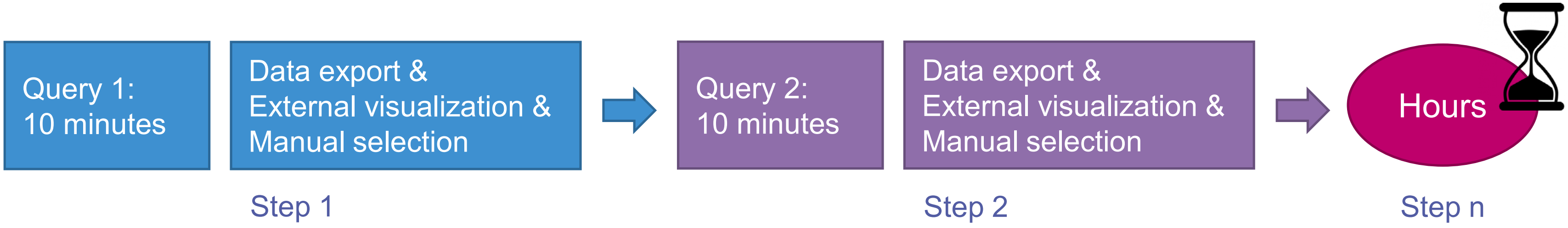
- **Highly interactive and visual data traversing of the chemical space**
  - Excellent **performance** to retrieve data from large data sets
  - High-end **visualization capabilities** to depict complex relationships
  - Benefits include new insights that might have otherwise been overlooked and increased creativity

Business Needs

- **Researchers need highly interactive (responsive) and user-friendly tools to answer questions like:**
  - What are the nearest neighbors to a given compound A that contain scaffold A and show a high permeability?
  - Which compounds show activities on targets A and B and have a reasonable ADME profile?
  - Is there a commercially available compound similar to compound A that comes with pharmacological data that might be used as a tool compound?

Previous Situation Had Technology Gaps

- Data was only stored in relational databases
- A single Nearest Neighbor Search may have taken minutes
- A compound collection walk-through required a series of successive searches that may have taken hours



Steps in Building the Similarity Graph Tool

- **Calculation of FCFP4 fingerprint (Tanimoto) similarities**
  - With 10 Nearest Neighbors, Canonical SMILES, INCHI keys, and structure pictures for all Sanofi screening collection compounds
- **Using the new ChemAxon4Neo4j plugin for substructure and similarity searches**
  - Avoid redundant storage of structures in Oracle (cartridge)
- **Compound annotations**
  - Physical Chemistry data (logD HPLC Mean, SOLUBILITY Mean) and also calculated properties
  - eADME data (PT Max Mean, METABOLISM Human Mean, METABOLISM Rat Mean)
  - Related Sanofi project names

Steps in Building the Similarity Graph Tool

- **Loading the data into the Neo4j graph database**
- **Using Tom Sawyer Perspectives by Tom Sawyer Software to build the web application**
  - Selected due to its advanced data integration and graph visualization capabilities
- **Integration of ChemAxon MARVIN JS sketcher for drawing structures for substructure search**

Features of the Similarity Graph Tool

- Retrieve Nearest Neighbors of a molecule
- Highlight highest, second highest…chemical similarity edge of a molecule node for interactive graph traversal
- Allow scientist to track the path and order of visited compounds
- Export selected compound IDs for further analysis in other tools
  - For example, Certara D360
- Allow filtering on edge and node properties
- Apply color coding (rules) to molecule nodes

Features of the Similarity Graph Tool

- **Find shortest path(s) between two molecules respecting the biological context**
  - Consider visible nodes of the currently displayed graph or all database nodes
- **Enrich nodes with data from CSV files**
  - For example, link by compound ID
- **Display scaffolds**
- **Show compounds with similar SAR**
  - Same biological function, but low chemical similarity
- **Integration of CHEMBL data**
  - 1.8 million compounds

Acknowledgements

- **ChemAxon**
  - Annamaria Kovacs
  - Andras Volford
  - Balazs Zaicsek
  - Tamas Varga
  - Janos Fejervari
- **Tom Sawyer Software**
  - Brendan Madden
  - Rudolfs Opmanis
  - Margers Kietis
  - Deborah Baron
  - Madisen Joseph
- **Neo4j**
  - Sven Janko
  - Bruno Ungermann
- **Sanofi**
  - Christian Buning
  - Christine Rudolph
  - Sven Ruf
  - Hans Matter
  - Peter Monecke
  - Jürgen Kammerer
  - Norbert Krass
  - Gerhard Hessler