Vojtěch Šípek
2014

Vizuální dotazování do chemických databází pomocí SMARTS vzorů

Abstrakt

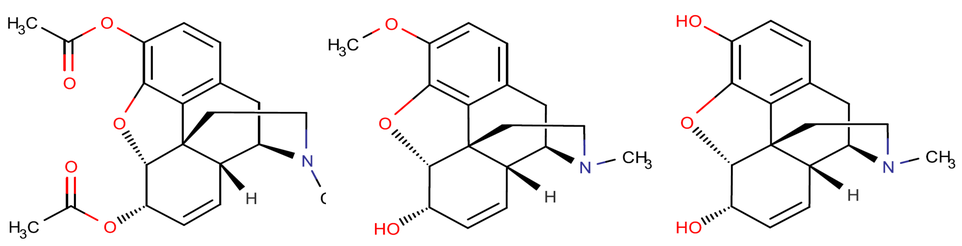
Cílem této práce je vytvoření frameworku pro vizuální dotazování do chemických databází, který bude implementován jako webová aplikace. Pomocí grafického editoru v klientské části aplikace uživatel vytváří dotazy, které jsou převedeny do chemického dotazovacího jazyka SMARTS. Tento dotaz je následně zpracován na aplikačním serveru, který je napojen na chemickou databázi. Součástí frameworku je i sada nástrojů na vytváření databáze a indexu, který je nad ní postavený.

Klientská část

- Grafický editor pro tvorbu SMARTS dotazů
  - Implementace v JavaScriptu
  - Kreslení na HTML5 canvas za podpory knihovny *Easel.js*
  - Serializace grafického vstupu do jazyka SMARTS
  - Odeslání dotazu na server
- Zobrazení výsledků dotazování
  - Podpora stránkování

Vyhledávání podstruktur

Jednou ze stěžejních oblastí chemoinformatiky je vyhledávání podstruktur, které je hojně využíváno především ve farmaceutickém průmyslu. Proces vývoje nových léků čelí stále narůstajícímu počtu dostupných sloučenin, které jsou uchovávány ve velkých databázích. Vyhledávání podstruktur v těchto databázích může uživateli pomoci získat množinu sloučenin s určitými chemickými vlastnostmi. Pokud umíme určit vlastnosti nějaké molekuly na základě její chemické struktury, pak se očekává, že molekuly s podobnou strukturou budou mít podobné vlastnosti.



Heroin Kodein Morfin

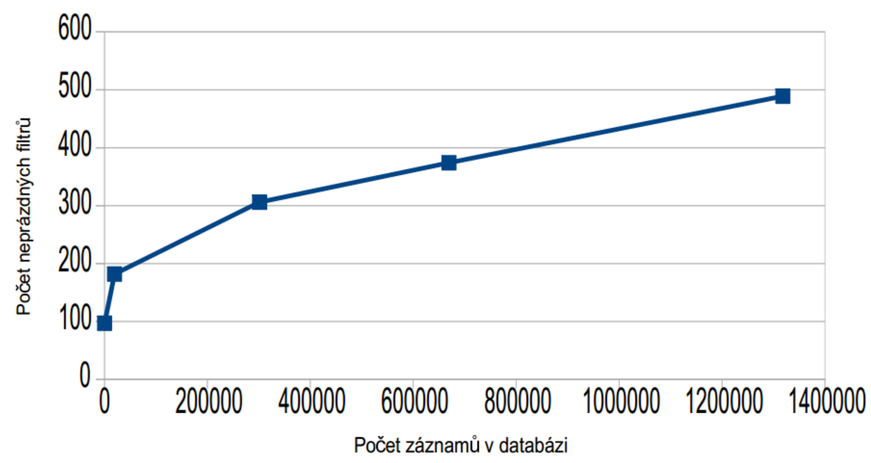
Příklad můžeme vidět na přiloženém obrázku, kde vidíme molekuly tří opiátů - heroinu, kodeinu a morfinu. Vidíme, že jejich struktura má mnoho společných rysů. Navíc víme, že všechny tři patří do skupiny opiátů, o kterých se ví, že mají podobné účinky na lidské tělo.

Jazyky SMILES a SMARTS

- SMILES
  - Simplified molecular-input line-entry systém
  - Jazyk pro 1D popis chemických struktur
  - Formát podobný „lidskému“ zápisu molekul
- SMARTS
  - SMiles ARbitrary Target Specification
  - Chemický dotazovací jazyk
  - Podobnost regulárním výrazům
  - Každý validní SMILES řetězec je zároveň validním SMARTS řetězcem

Databáze a index

- Databáze je ve formě souboru, kde každý řádek obsahuje jeden záznam
  - Záznam = SMILES popis + URL s jeho popisem
- Index nad databází obsahuje filtry pro prefiltraci databáze na základě konkrétního dotazu
  - Každý soubor indexu obsahuje jeden filtr
  - Každý filtr obsahuje informace o nějaké vlastnosti molekul (výskyt konkrétního prvku, výskyt nějakého typu vazby atd…)
  - Filtr s vlastností X je ve formě bitového řetězce, kde n-tý bit indikuje, zda n-tý záznam v databázi má vlastnost X
- Pro úsporu paměti ukládáme pouze neprázdné filtry (filtry obsahující alespoň jeden bit nastavený na 1).
  - Velikost indexu je nutné redukovat kvůli tomu, že celý index načítáme do paměti
  - Reálné databáze budou mít pouze zlomek neprázdných filtrů – celkový počet vytvářených filtrů je 27 577.
  - Následující graf ukazuje měření na reálné databázi ChEMBL



Serverová část

- Zpracování dotazu z klienstké strany
- Prefiltrace databáze na základě dotazu
- Samotné porovnávání SMARTS vzorů se SMILES řetězci z databáze pomocí knihovny *Chemistry Developtment Kit*