ChemAxon’s novel Biomolecule Toolkit is bringing the HELM standard to life. The ChEMBL v21 DB contains ~20000 peptides published in both MOL and HELM format, which we used as a benchmark for measuring the ability of our tool to start from a set of chemical representations, convert it to HELM, and index it in a database. Finally we performed queries on the indexed content, which were not possible previously using the default ChEMBL web interface.

Along the way we encountered a few challenges and data issues, of which a select few are detailed below. Our tool provides a solution for these and may help to improve ChEMBL and other datasets.

**Import / standardize**

![Diagram of import and standardization process]

Challenges
- Ensure overall molecule and functional groups are consistent with monomer library

Solution
- Standardization of input MOL files (un-group S-groups, strip salt/solvent, remove explicit hydrogens, neutralize structure, aromatize, standardize functional groups)

**Search**

![Diagram of search process]

Challenges
- Ability to query indexed and stored content for (sub-)sequences, metadata and/or chemical substructures on the example of Oxytocin (CHEMBL395423)

Solution
- Use canonicalization function of the Biomolecule Toolkit

**Summary**

The chemical representation of 19773 biotherapeutics published in ChEMBL v21 was converted into HELM format using the accompanying HELM monomer dictionary (2802 unique monomers) and ChemAxon’s Biomolecule Toolkit. Prior to conversion, KNIME® with ChemAxon’s cheminformatic nodes was used to standardize (remove salts/solvent, neutralize, aromatize, standardize functional groups) the input structures. Prior to direct comparison canonical HELM notations of the ones published in ChEMBL were generated, affecting ~64,000 molecules. We identified 3 wrong conversions on ChEMBL side. ChemAxon IDs 439175 (wrong sequence), and 56684/2233118 (GU residue instead of gamma-Glu in sequence). For later querying, molecules were indexed and stored in a DB, filtering out 284 duplicate structures.

We believe that curating the HELM monomer dictionary would further improve the content. Standard protection groups (e.g., Fmoc) are found on >169 monomers, which adds redundancy by combinatorial explosion. As of yet, there seems to be no common view as to what precisely defines a HELM monomer. In the meantime, however, our toolkit will work with this or any other custom dictionary.

---

1. [Website: www.chemaxon.com]
2. [Website: www.chemaxon.com]
3. [Website: www.chemaxon.com/products/biomolecule-toolkit]