Journal of Cheminformatics

POSTER PRESENTATION Open Access

PubChem: atom environments for molecule standardization

Volker Hähnke*, Evan E Bolton, Stephen H Bryant

From 8th German Conference on Chemoinformatics: 26 CIC-Workshop
Goslar, Germany. 11-13 November 2012

PubChem is an open repository for molecular structures, their properties and biological activities [1]. The number of deposited structures has been steadily increasing since its creation in 2004. Today, it contains more than 92 million substances (PubChem Substance) with 32 million unique small molecules (PubChem Compound). Consequently, visual inspection of every structure and correction of errors by hand to detect structure equivalencies and to ensure data quality are not feasible. Efficient and reliable automated methods for standardization are necessary during the registration process to compensate for alternating representations of as well as errors and artifacts in (sub)structure representations caused by diverging business rules, personal preferences, data format conversion, disagreements between aromaticity definitions and automated library generation. At PubChem, we are developing a new standardization approach that is based on rules for atom environment transformation. Those rules are obtained from a statistical analysis of atom environment transformations observed with a reference workflow combining chemical reasonability checks, valence filters, canonical tautomer determination and aromaticity normalization. Additional atom environment mappings are provided by hand curation. In the first application of our technique to PubChem we concentrate on purely organic compounds. Those represent 97% of the deposited structures and account for the majority of atom environments as well. Here, we present the first results obtained with our approach, highlighting the methodology, challenges, benefits and future possibilities.

Published: 22 March 2013

* Correspondence: volker.hahnke@nih.gov
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, 20894, USA

Reference

1. Bolton E, Wang Y, Thiessen PA, Bryant SH: **PubChem: Integrated Platform of Small Molecules and Biological Activities.** In *Annual Reports in Computational Chemistry, Volume 4, Chapter 12.* Oxford:Elsevier;Wheeler RA, Spellmeyer DC 2008:217-241.

doi:10.1186/1758-2946-5-S1-P38
Cite this article as: Hähnke *et al.*: PubChem: atom environments for molecule standardization. *Journal of Cheminformatics* 2013 **5**(Suppl 1):P38.