UniProt, ChEBI and IDSM Sachem: exploring biologically relevant ligands

Sebastien Gehant[1], Elisabeth Coudert[1], Anne Morgat[1], Jerven Bolleman[1], Nicole Redaschi[1], Alan Bridge[1,*] and UniProt Consortium[1,2,3,4]

[1]*Swiss-Prot group, SIB Swiss Institute of Bioinformatics, CMU, 1 Michel Servet, 1211 Geneva 4, Switzerland*

[2]*European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK*

[3]*Protein Information Resource (PIR), Georgetown University Medical Center, 3300 Whitehaven Street, NW, Suite 1200, Washington, DC 20007, USA*

[4]*Protein Information Resource (PIR), University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA*

Abstract

UniProt has recently standardized the annotation of biologically relevant small molecule ligands – essential for protein structure and function – using the chemical ontology ChEBI, allowing UniProt users to retrieve small molecule ligands that match a given chemical structure, or that are members of a given chemical class (as defined by the ChEBI ontology), via the UniProt website and APIs (www.uniprot.org). In this work we use SPARQL to further extend the chemical structure search capabilities of UniProt by federation of UniProt to the IDSM Sachem SPARQL endpoint, which supports chemical similarity and substructure searches. Protein structures – experimentally determined and predicted using AI methods – lack biologically relevant ligands. This work provides a simple means to identify them for the purposes of protein structure annotation and modeling.

Keywords

UniProt, ligands, Sachem, similarity search, federation, SPARQL

1. Introduction

Experimentally determined protein structures found in PDB[1] often contain ligands that were artificially modified to enable crystallization. Accurate protein structure determination and modeling therefore requires that these artificial ligands are replaced by their biologically relevant (cognate) equivalents.

UniProt [2] announced that in release 2022_03 it significantly improved the annotation of cognate ligands [3], replacing thousands of existing textual descriptions of ligands with their equivalents from the ChEBI ontology [4]. These annotations are now available as part of the core UniProt dataset on its public SPARQL endpoint at https://sparql.uniprot.org/. Sachem is

SWAT4HCLS 2023

[*] Corresponding author.

sebastien.gehant@sib.swiss (S. Gehant); elisabeth.coudert@sib.swiss (E. Coudert); anne.morgat@sib.swiss (A. Morgat); jerven.bolleman@sib.swiss (J. Bolleman); nicole.redaschi@sib.swiss (N. Redaschi); alan.bridge@sib.swiss (A. Bridge)

0000-0003-1608-9954 (S. Gehant); 0000-0001-8314-404X (E. Coudert); 0000-0002-1216-2969 (A. Morgat); 0000-0002-7449-1266 (J. Bolleman); 0000-0001-8890-2268 (N. Redaschi); 0000-0003-2148-9135 (A. Bridge)

a set of tools that support chemical substructure and similarity searches on a large corpora of known chemical compounds imported from PubChem, ChEMBL and ChEBI [5]. Sachem exposes these capabilities as a W3C standard SPARQL 1.1 endpoint[6].

We show with an example how the annotations in UniProtKB and Sachem data can be combined via federated SPARQL queries in order to find potential cognate ligands for AMP-PCP (PDB ligand code ACP), an ATP analog commonly used in crystallography. The ChEBI identifiers of all chemicals similar to AMP-PCP above a given similarity score (e.g. 0.8) are obtained by querying Sachem with the SMILES representation of the artificial ligand (SMILES, or Simplified Molecular-Input Line-Entry System, is a line notation for describing the structure of chemical species). The list of ChEBI identifiers, bound to the variable *?ligand*, is then reduced through the additional constraint that the variable must also correspond to a cognate ligand annotated in UniProtKB.

2. Query

```
PREFIX up:     <http://purl.uniprot.org/core/>
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
PREFIX rdfs:   <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT
  ?ligandSimilarityScore
  ?ligand
WHERE {
  SERVICE <https://idsm.elixir-czech.cz/sparql/endpoint/chebi>
      {
    [ sachem:compound ?ligand;
      sachem:score ?ligandSimilarityScore ]
      sachem:similaritySearch [
        sachem:query "c1nc(c2c(n1)n(cn2)[C@H]3[C@@H]([C@@H]([
            C@H](O3)CO[P@@](=O)(O)O[P@](=O)(CP(=O)(O)O)O)O)O)N";
            #Isomeric SMILES of AMP-PCP
        sachem:cutoff "8e-1"^^xsd:double ;
        sachem:similarityRadius 1 ]
  }

  ?uniprot up:annotation ?annotation .
  ?annotation a up:Binding_Site_Annotation ;
      up:ligand/rdfs:subClassOf ?ligand .
}
ORDER BY DESC(?ligandSimilarityScore)
```

Find cognate ligands similar to the artificial ligand AMP-PCP (PDB ACP) using Sachem and UniProtKB.

3. Funding & acknowledgements

This work was supported by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), National Institute of Mental Health (NIMH), and National Cancer Institute (NCI) of the National Institutes of Health (NIH) under grant U24HG007822. Additional support for the EMBL-EBI's involvement in UniProt comes from European Molecular Biology Laboratory (EMBL) core funds, the Alzheimer's Research UK (ARUK) grant ARUK-NAS2017A-1, the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/T010541/1] and Open Targets. UniProt activities at the SIB are additionally supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI. PIR's UniProt activities are also supported by the NIH grants R01GM080646, G08LM010720, and P20GM103446, and the National Science Foundation (NSF) grant DBI-1062520.

We used Sachem IDSM, kindly provided by the Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences.

References

[1] S. K. Burley, H. M. Berman, J. M. Duarte, Z. Feng, J. W. Flatt, B. P. Hudson, R. Lowe, E. Peisach, D. W. Piehl, Y. Rose, A. Sali, M. Sekharan, C. Shao, B. Vallat, M. Voigt, J. D. Westbrook, J. Y. Young, C. Zardecki, Protein data bank: A comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students, Biomolecules 12 (2022).

[2] UniProt Consortium, UniProt: the universal protein knowledgebase in 2023, Nucleic Acids Res. (2022).

[3] E. Coudert, S. Gehant, E. de Castro, M. Pozzato, D. Baratin, T. B. Neto, C. J. A. Sigrist, N. Redaschi, A. Bridge, The UniProt Consortium, Annotation of biologically relevant ligands in UniProtKB using ChEBI, 2022.

[4] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck, ChEBI in 2016: Improved services and an expanding collection of metabolites, 2016.

[5] M. Kratochvíl, J. Vondrášek, J. Galgonek, Sachem: a chemical cartridge for high-performance substructure search, J. Cheminform. 10 (2018) 27.

[6] M. Kratochvíl, J. Vondrášek, J. Galgonek, Interoperable chemical structure search service, J. Cheminform. 11 (2019) 45.