Bioinformatics, 37(3), 2021, 424–425
doi: 10.1093/bioinformatics/btaa693
Advance Access Publication Date: 31 July 2020
Applications Note

Structural bioinformatics

GeoMine: interactive pattern mining of protein–ligand interfaces in the Protein Data Bank

Konrad Diedrich , Joel Graef , Katrin Schöning-Stierand and Matthias Rarey *

Universität Hamburg, ZBH – Center for Bioinformatics, 20146 Hamburg, Germany

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on May 29, 2020; revised on July 12, 2020; editorial decision on July 22, 2020; accepted on July 24, 2020

Abstract

Summary: The searching of user-defined 3D queries in molecular interfaces is a computationally challenging problem that is not satisfactorily solved so far. Most of the few existing tools focused on that purpose are desktop based and not openly available. Besides that, they show a lack of query versatility, search efficiency and user-friendliness. We address this issue with GeoMine, a publicly available web application that provides textual, numerical and geometrical search functionality for protein–ligand binding sites derived from structural data contained in the Protein Data Bank (PDB). The query generation is supported by a 3D representation of a start structure that provides interactively selectable elements like atoms, bonds and interactions. GeoMine gives full control over geometric variability in the query while performing a deterministic, precise search. Reasonably selective queries are processed on the entire set of protein–ligand complexes in the PDB within a few minutes. GeoMine offers an interactive and iterative search process of successive result analyses and query adaptations. From the numerous potential applications, we picked two from the field of side-effect analyze showcasing the usefulness of GeoMine.

Availability and implementation: GeoMine is part of the ProteinsPlus web application suite and freely available at https://proteins.plus.

Contact: rarey@zbh.uni-hamburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The understanding, manipulation and modulation of protein function require substantial structural knowledge of the protein binding sites. One of the main sources for structural data is the Protein Data Bank (PDB) (Berman, 2000). Despite its substantial growth, improvement of data quality and potential as a knowledge source, there is only a small number of tools featuring 3D geometric searching for protein–ligand interfaces based on user-defined queries (Angles *et al.*, 2020; Hendlich *et al.*, 2003; Korb *et al.*, 2016; Mobilio *et al.*, 2010; Weisel *et al.*, 2012). The first has been Relibase which was suspended in 2018. To our knowledge, Relibase was the only tool supporting atomic precision for both protein and ligand parts within the query. Besides GSP4PDB, the tools are desktop applications and not freely available. Reasonably short runtimes can be observed in the case of Prolix and CrossMiner due to their use of fingerprint techniques. The query versatility of all tools is however limited. In CrossMiner, a query consists of pharmacophore spheres which represent predefined features. Prolix and PRDB do not support an atom-level precision for protein parts of the query. GSP4PDB lacks atomic query precision. In Prolix and GSP4PDB, the user can design a query using a 2D sketcher. PRDB requires a query in SQL format. Only CrossMiner provides the possibility to construct queries in a 3D representation of a protein–ligand complex. Regarding the lack of existing solutions, we developed GeoMine, which is based on an enhanced version of PELIKAN (Inhester *et al.*, 2017).

GeoMine is publicly available via a web-interface and part of the ProteinsPlus (Fährrolfes *et al.*, 2017; Schöning-Stierand *et al.*, 2020) server (https://proteins.plus). A search of geometrical, textual and numerical queries can be easily performed on protein–ligand interfaces derived from complexes contained in the PDB in a reasonably short time. In the following, we will illustrate the features of GeoMine by different use cases. Detailed descriptions of the methods are available in the PELIKAN (Inhester *et al.*, 2017) publication.

2 Usage and output

According to the ProteinsPlus workflow, GeoMine is started with only a PDB structure as input. 3D query design is guided by a precalculated pocket with selectable features in the embedded NGL viewer (Rose *et al.*, 2018). The query generation process allows the user to select amongst others atoms and aromatic ring centers or to place such points at unoccupied positions. Point–point constraints, i.e. interatomic distance ranges or interactions, and angle constraints between any pair of connected point–point constraints or aromatic ring normals allow the definition of any geometric arrangement.

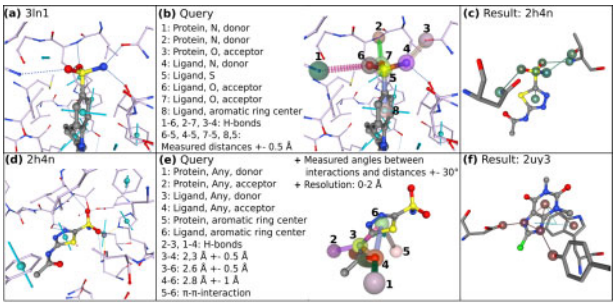

Fig 1. Query construction for the search of celecoxib (**a**, **b**) and acetazolamide (**d**, **e**) using GeoMine. (**c**, **f**) Example results from the searches of the queries defined in (b) and (e), respectively. The structures matching the query are highlighted in the results.

The resulting query is shown in the NGL viewer and simultaneously in tables for further modification by a variety of geometrical constraints, e.g. the range of a distance, and chemical properties. Main properties like the molecule type of an atom are automatically set while more discriminative ones, like the functional group of a ligand atom, can be explicitly defined by the user. Additionally, it is possible to query textual and numerical properties of the protein–ligand complex and its components, i.e. the depth of a pocket or the EC number. All search types can be used separately or together either on a PDB subselection or on the complete dataset. The first 150 matches are listed in a table and can be superimposed for visual analysis onto the 3D query in the viewer. An extensive statistics report, which allows a more sophisticated analysis of all results, as well as the pockets of the first 150 matches can be downloaded. The complete result set can be filtered continuing to search in it with the current query as a new starting point.

3 Applications

Since GeoMine is able to find structural similarities between binding sites of unrelated proteins, it is a valuable tool for off-target studies, e.g. with the aim of lead optimization, drug repurposing or explaining side effects. In the following, we will describe two different off-target searches showcasing the comprehensiveness of GeoMine results. Additional application examples are available in the PELIKAN (Inhester *et al.*, 2017) paper. The GeoMine database searches are performed using up to 30 cores of a 2× Intel Xeon Gold 6248 processor (20 cores/2.5 GHz), 200 GB of main memory and a Dell 1.6TB NVMe HHHL AIC PM1725b solid state drive with an xfs file system.

The identification of protein–ligand complexes with similar interaction patterns like a given query complex can generate ideas about potential off-target proteins. For our first example application, we choose the COX-2 selective inhibitor celecoxib (PDB code: 3LN1;Fig. 1a). In the precalculated pocket, the unsubstituted arylsulfonamide moiety of celecoxib interacts with the protein environment via 4 hydrogen bonds (Fig. 1a). A query describing partially this interacting moiety (Fig. 1b) took 45 s and resulted in 43 matches (see statistics report in Supplementary Material S1). A variety of different protein classes emerged by this search, for example carbonic anhydrase (CA II) complexed with the inhibitor acetazolamide (PDB code: 2H4N; Fig. 1c). According to studies, CA II is an off-target for unsubstituted sulfonamides like celecoxib (Weber *et al.*, 2004) implicating the enzyme in celecoxib side effects. Validation results for this query are 3LN1, 5JW1 (celecoxib cocrystallized with COX-2) and 1OQ5 (celecoxib cocrystallized with CA II). 1OQ5 was found removing one hydrogen bond from the query.

For the illustration of the second off-target search, we choose the previously found CA II complexed with acetazolamide (PDB code: 2H4N; Fig. 1d). Dependent on the arrangement of available functional groups in a protein pocket, ligands may form different interaction patterns. To find similar geometric arrangements, we constructed a query from the complex 2H4N with parts of the ligand and hypothetical alternative interactions. It includes the ligands' thiadiazole ring center and the donor and acceptor of its acetamide fragment (Fig. 1e). Potential interaction directions are defined by angles. Geometrical flexibility is achieved by relatively large tolerance values for angles and distances. Keeping the atom elements unspecified by describing only their interaction and molecule types ensures chemical fuzziness. Low-quality results are prevented by a numerical filter. The search took 52 s and resulted in 57 matches (see statistics report in Supplementary Material S2), for instance, chitinase that binds the inhibitor theophylline (PDB code: 2UY3; Fig. 1f). According to a study, chitinase is an off-target for acetazolamide, which results as a promising lead for antifungal drug development (Schüttelkopf *et al.*, 2010). A validation result can be found in PDB file 2UY4 (acetazolamide cocrystallized with chitinase).

4 Conclusions

GeoMine addresses the computational challenge of efficient geometrical data mining of protein–ligand binding sites. Reasonable queries can be answered by GeoMine within seconds up to a few minutes. All structures that match the query are found and presented in a comprehensive manner. The search infrastructure of GeoMine is easy to use and publicly available as part of the ProteinsPlus web service.

Funding

This work was supported by the German Federal Ministry of Education and Research as part of the German Network for Bioinformatics Infrastructure – de.NBI [031L0172, 031L0105].

Conflict of Interest: ProteinsPlus and in the NAOMI ChemBio Suite use some methods that are jointly owned and/or licensed to BioSolveIT GmbH, Germany, M.R. is a shareholder of BioSolveIT GmbH.

References

Angles,R. *et al.* (2020) GSP4PDB: a web tool to visualize, search and explore protein–ligand structural patterns. *BMC Bioinformatics*, **21**, 85.
Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
Fährrolfes,R. *et al.* (2017) ProteinsPlus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res.*, **45**, 337–343.
Hendlich,M. *et al.* (2003) Relibase: design and Development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.*, **326**, 607–620.
Inhester,T. *et al.* (2017) Index-based searching of interaction patterns in large collections of protein–ligand interfaces. *J. Chem. Inf. Model.*, **57**, 148–158.
Korb,O. *et al.* (2016) Interactive and versatile navigation of structural databases. *J. Med. Chem.*, **59**, 4257–4266.
Mobilio,D. *et al.* (2010) A protein relational database and protein family knowledge bases to facilitate structure-based design analyses. *Chem. Biol. Drug Des.*, **76**, 142–153.
Rose,A.S. *et al.* (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
Schöning-Stierand,K. *et al.* (2020) ProteinsPlus: interactive analysis of protein–ligand binding interfaces. *Nucleic Acids Res.*, 48, 48-53.
Schüttelkopf,A.W. *et al.* (2010) Acetazolamide-based fungal chitinase inhibitors. *Bioorg. Med. Chem.*, **18**, 8334–8340.
Weber,A. *et al.* (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.*, **47**, 550–557.
Weisel,M. *et al.* (2012) PROLIX: rapid mining of protein–ligand interactions in large crystal structure databases. *J. Chem. Inf. Model.*, **52**, 1450–1461.