BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 10 2014, pages 1481–1483
doi:10.1093/bioinformatics/btu052

Structural bioinformatics

Advance Access publication January 27, 2014

ChEpiMod: a knowledgebase for chemical modulators of epigenome reader domains

Jamel Meslamani, Steven G. Smith, Roberto Sanchez and Ming-Ming Zhou*

Department of Structural and Chemical Biology, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY, 10029, USA

Associate editor: Anna Tramontano

ABSTRACT

Context: Epigenome reader domains are rapidly emerging as a new class of drug targets for a wide array of human diseases. To facilitate study of structure–activity relationship and small-molecule ligand design for these domains, we have created ChEpiMod. ChEpiMod is a free knowledgebase of chemical modulators with documented modulatory activity for epigenome reader domains.

Methods: ChEpiMod organizes information about chemical modulators and their associated binding-affinity data, as well as available structures of epigenome readers from the Protein Data Bank. The data are gathered from the literature and patents. Entries are supplemented by annotation. The current version of ChEpiMod covers six epigenome reader domain families (Bromodomain, PHD finger, Chromodomain, MBT, PWWP and Tudor). The database can be used to browse existing chemical modulators and bioactivity data, as well as, all available structures of readers and their molecular interactions. The database is updated weekly.

Availability: ChEpiMod is freely available at http://chepimod.org

Contact: ming-ming.zhou@mssm.edu

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

Received on November 21, 2013; revised on January 14, 2014; accepted on January 20, 2014

1 INTRODUCTION

Epigenetic control of gene expression plays a fundamental role in numerous biological processes including development, cell proliferation and differentiation, and its dysregulation has been linked to multiple disorders such as cancer, inflammation and infection (Kouzarides, 2007). The evolutionarily conserved modular domains that are embedded in transcription-associated proteins and recognize post-translational modifications such as lysine acetylation or methylation are called 'Epigenome Readers' (Yap and Zhou, 2010). The desire to better understand and modulate the molecular functions of these readers has encouraged the development of small-molecule compounds (or 'chemical modulators') that serve as research tools and as potential therapeutic agents for a wide array of human diseases (Arrowsmith, *et al*. 2012). The ever-increasing number of compounds with documented modulatory activity against epigenome readers has prompted us to develop ChEpiMod, a comprehensive knowledgebase of structure–activity relationship data for these domains and their chemical modulators. The goal of ChEpiMod is to organize, in a single place, all publicly available and relevant research data to facilitate the structure-guided development and use of epigenome reader modulators as tools for chemical epigenomics research. General purpose bioactivity databases such as ChEMBL (Gaulton *et al*., 2012) and PubChem BioAssay (Wang *et al*., 2012) contain information about some ligands for epigenome readers, but are far from comprehensive for these domains and do not provide structural information on protein–ligand interactions. Additionally, ligands stored in these general databases have their affinities listed for proteins, rather than domains. ChEpiMod overcomes these limitations by providing comprehensive coverage of epigenome reader bioactivities, coupled with structural information and advanced analysis tools.

2 DATABASE CONTENT

ChEpiMod currently contains 130 682 compounds (2042 with affinity $<10\,\mu M$) (Supplementary Fig. S1). A total of 133 target proteins are covered matching 224 different domains. The database contains 132 524 bioactivities for which 44% is assigned to a defined domain (2375 with affinity $<10\,\mu M$). More than 99% of the bioactivities that are assigned to identified domains are reported for Bromodomains (Supplementary Table S1). Structures collected in ChEpiMod correspond to Bromodomain (44%), PHD finger (22%), Chromodomains (12%), MBT (10%), PWWP (7%) and Tudor domains (5%) (Supplementary Table S1). Ligand-interaction data derived from these structures is also stored (Figs 1 and 2).

3 USER INTERFACE AND FUNCTIONALITY

Users can access ChEpiMod via a web interface. Search results are displayed in a table reporting compound, protein, domain, structure and binding affinity, which can be exported in various formats (Fig. 1). Within the table, users can access summaries about a specific compound, protein or domain. The 'domain summary' page displays the total number of compounds, structures and bioactivities available for that domain. A first tab on this page enables the user to browse compounds clustered based on their structural similarity, allowing the chemical diversity of all known ligands for a domain to be explored. A second tab provides a comparative view of the molecular interactions of all available compounds with the target domain. Interactions can be displayed either for the complete domain sequence or for the binding site residues alone (Fig. 2). The binding site level view provides information on interaction types (hydrophobic,

*To whom correspondence should be addressed.

Fig. 1. Sample ChEpiMod search results. Partial screenshot of the results table for bioactivities of Bromodomain-containing protein 4 (BRD4). Each row in the table contains one bioactivity. Default columns show the compound structure, target protein and domain, affinity values and reference. Compound summary (1) and domain summary (2) pages are accessible from here. A display button (3) can be used to select multiple protein structures, align and visualize them. Selected bioactivities, compounds and structures can be downloaded in various formats (4). Buttons available under each PDB code to access the structure in the PDB, to visualize the 3D protein structure, and to search for similar binding sites in other domain structures

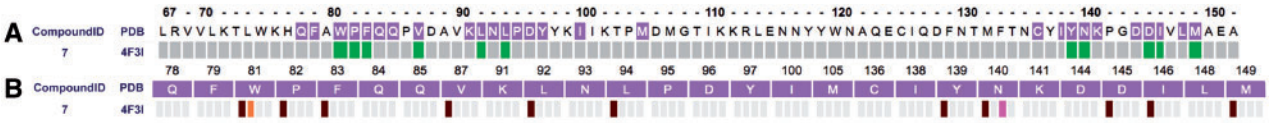

Fig. 2. Domain–ligand interactions. (**A**) Residues interacting with compound MS417 (PDB: 4F3I) are shown in green in the context of the full BRD4-BrD1 sequence. Consensus binding site residues are colored in purple. (**B**) Consensus acetyl-lysine binding site for BRD4-BrD1 and contacts with MS417. Each contact is represented by four types of interactions: hydrophobic (maroon), aromatic (orange), hydrogen bond (purple) and ionic (cyan)

aromatic, polar and ionic). These features are also mapped on the structure for visualization (Supplementary Fig. S2). The 'compound summary' page contains physicochemical properties, number of targeted proteins, domains, bioactivities, and protein structures for a compound. Affinity values expressed in K_i, K_d and IC50 (μM) are mapped on phylogenetic trees which provide a family wide view of the coverage and selectivity of a specific chemical modulator. ChEpiMod provides several other functionalities. For example, the binding site similarity search identifies domains that have similar binding sites, which might be useful for discovering potential new interactions for existing modulators (Klabunde, 2007) and to assess their selectivity.

4 DATA COLLECTION AND PROCESSING

ChEpiMod is organized around bioactivities between epigenome readers and chemical modulators. Bioactivities from journals and patents are manually extracted and supplemented by data from the ChEMBL_17 database (Gaulton *et al.*, 2012). Affinity values expressed as binding constants K_i, K_d and IC_{50} are stored in units of micromolar. Compounds are represented by their 2D structures. Targets are annotated according to the Uniprot database (Consortium, 2013), and the recommended Uniprot name is used for each target name. Other protein names and aliases are also extracted and stored, providing the user the ability to search the database by protein name synonyms. Domains are annotated according to the PFAM database (Punta *et al.*, 2012). Since some targets have more than one domain of the same type the domain number is specified for any bioactivity. When the target domain is missing (e.g. in patents) no target domain is specified and the bioactivity is reported for the protein. Each bioactivity is linked to its original source by the Pubmed identifier for journals, and the patent identifier for patents.

All crystallographic structures of domains are collected from the PDB (Berman *et al.*, 2003). For each downloaded structure a binding site is defined as all residues with at least one heavy atom within 6Å of any ligand heavy atom; and a 'Consensus binding' site for each domain is also defined (cf. Supplementary Materials for definition). Molecular interactions between ligands and each residue of the consensus binding site are computed using the Interaction Fingerprints program (Marcou and Rognan, 2007) and grouped in four categories according to the interaction type (hydrophobic, aromatic, polar and ionic).

5 CONCLUSION

ChEpiMod has been developed to facilitate structure-based rational design of new chemical modulators for epigenetic reader domains. The molecular interaction data available in the knowledge-database can guide to further improve compound affinity and selectivity, as well as identify protein residues important for ligand recognition.

Funding: This work was funded in part by the research grants from the National Institutes of Health (HG004508 and CA87658 to M.-M.Z.)

Conflict of Interest: none declared.

REFERENCES

Arrowsmith,C.H. *et al.* (2012) Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discov.*, **11**, 384–400.

Berman,H. *et al.* (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.

Consortium,U. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.

Gaulton,A.A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
Klabunde,T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *British J. Pharmacol.*, **152**, 5–7.
Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
Marcou,G. and Rognan,D. (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inform. Model.*, **47**, 195–207.
Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
Wang,Y. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.
Yap,K.L. and Zhou,M.-M. (2010) Keeping it in the family: diverse histone recognition by conserved structural folds. *Critic. Rev. Biochem. Mol. Biol.*, **45**, 488–505.