IChem:AVersatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions

Franck da Silva, Jérémy Desaphy, Didier Rognan

To cite this version:

Franck da Silva, Jérémy Desaphy, Didier Rognan. IChem:AVersatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. ChemMedChem, Wiley-VCH Verlag, 2017, 10.1002/cmdc.201700505 . hal-02158151

HAL Id: hal-02158151

https://hal.archives-ouvertes.fr/hal-02158151

Submitted on 17 Jun 2019

SPECIAL ISSUE

IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions

Franck Da Silva,[a] Jeremy Desaphy,[a, b] and Didier Rognan*[a]

Structure-based ligand design requires an exact description of the topology of molecular entities under scrutiny. IChem is a software package that reflects the many contributions of our research group in this area over the last decade. It facilitates and automates many tasks (e.g., ligand/cofactor atom typing, identification of key water molecules) usually left to the modeler's choice. It therefore permits the detection of molecular interactions between two molecules in a very precise and flexible manner. Moreover, IChem enables the conversion of intricate three-dimensional (3D) molecular objects into simple representations (fingerprints, graphs) that facilitate knowledge acquisition at very high throughput. The toolkit is an ideal companion for setting up and performing many structure-based design computations.

IChem is a suite of tools consisting of about 50 000 lines of computer code written in C++, decomposed in nine modules, for detecting and comparing molecular objects (proteins, ligand-binding cavities, ligands, protein–ligand and protein–protein complexes) frequently manipulated in structure-based computational chemistry (Table 1). Herein we describe four of the most frequent and important uses. For deeper investigation of all modules, the reader should refer to the publicly available user guide.[1]

Setting the scene: pdbconv. A reasonable start to any structure-based design project is the retrieval of experimentally determined protein structures from the Protein Data Bank (PDB),[2] a web resource that currently stores over 134 000 entries. Unfortunately, PDB structures cannot be used directly, as many important features (e.g., protonation and ionization states, atom types and bond orders for organic molecules) are missing. The *pdbconv* module of the IChem toolkit automates the preparation of ready-to-use protein–ligand structures. The process first assigns a specific class to each residue name (Table 2).

Table 1. IChem modules.

Module	Purpose
pdbconv	Post-process raw PDB files
realign	Structural alignment of two molecules
sims	Fingerprint comparison
utils	Miscellaneous
Volsite	Cavity detection and druggability prediction
IFP	Fingerprinting protein–ligand interactions
Ints	Fingerprinting protein–ligand interaction patterns
Grim	Converts protein–ligand interaction pattern in graphs
DetectPPI	predict biologically relevant PPIs[a]

[a] Protein–protein interfaces.

Table 2. HET residue classes in IChem, exemplified by one residue.

HET[a]	Template[b]	Updated[c]	Class[d]
ACO	1	/	COFACTOR
3NI	1	/	ION
001	1	/	LIGAND
3CO	1	/	METAL
004	1	/	MOD_AA
02I	1	/	NUCLEIC
01L	1	/	ORGANOMET
BH1	0	BPH	PROSTHETIC
ALA	1	/	STD_AA
045	1	/	SUGAR
000	1	/	UNWANTED
HOH	1	/	WATER

[a] Three-character alphanumeric code of each chemical component. [b] Template present (1) or absent (0). [c] Updated (/) or deprecated and replaced (e.g., BH1 updated into BPH). [d] Residue class.

It then applies a correct atom type to every heavy atom, generates the corresponding covalent bonds, and selects strongly bound water molecules while removing bulk water. The process relies on a predefined list of all possible residues with the corresponding templates for every HET record of the PDB file. Correct atom types and 3D coordinates are provided for every template by converting, with Corina,[3] PDB SMILES strings into the corresponding MOL2 file.

The residue list (Table 2) assigns the encountered HET record to one of the 12 possible residue classes (cofactor, ion, ligand, metal, modified amino acid, nucleic acid, organometallic, prosthetic, standard amino acid, sugar, unwanted, water). Please note that molecules originating from crystallization buffer ("un-

[a] Dr. F. Da Silva, Dr. J. Desaphy, Dr. D. Rognan
Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS—Université de Strasbourg, 74 route du Rhin, 67400 Illkirch (France)
E-mail: rognan@unistra.fr

[b] Dr. J. Desaphy
Discovery Chemistry Research and Technologies, Eli Lilly and Company, Lilly Corporate Center, Indianapolis IN 46285 (USA)

The ORCID identification number(s) for the author(s) of this article can be found under:
https://doi.org/10.1002/cmdc.201700505.

SPECIAL ISSUE This article is part of a Special Issue on Cheminformatics in Drug Discovery. To view the complete issue, visit:
http://onlinelibrary.wiley.com/doi/10.1002/cmdc.v13.6/issuetoc.

wanted" class) are automatically identified and discarded. Once atom types have been properly defined for every molecule type (protein and accessory molecules, solvent, ligand), any third-party tool (e.g., Protoss)[4] can be used to finally add the missing hydrogen atoms while optimizing both the ionization and tautomeric state of each molecule of the PDB entry.

Working with predefined residue lists and molecular templates provides both advantages and drawbacks. The main advantage is a uniform treatment of all chemical components of a PDB entry with a presumably correct atom typing. As a main drawback, the procedure requires an updated residue list and thus fails in case of a newly released PDB entry. We therefore propose regular updates along every new release of the in-house developed sc-PDB database of druggable protein–ligand complexes.[5]

Detecting ligandable cavities: Volsite. *Volsite* is a tool to automatically detect cavities at the surface of a macromolecule of interest, and predict its structural druggability.[6] It can be run in two modes depending on whether coordinates of a bound ligand are given (ligand-restricted mode) or not (full unrestricted mode). In any case, the target is first placed in a 2 Å resolution grid lattice and each voxel is assigned a state as whether its accessibility exceeds a user-defined threshold. Accessible voxels are then assigned a pharmacophoric property (hydrophobic, aromatic, hydrogen bond donor, hydrogen bond acceptor, positive ionizable, negative ionizable) complementary to that of the nearest protein atom according to a set of topological rules.[7] The pharmacophoric properties of all atoms are detected on the fly by the general IChem atom parser thereby enabling to consider accessory molecules (Table 2) or not during the cavity detection. Because every voxel has a fixed volume, the total number of pharmacophore-annotated voxels approximates the overall cavity volume. The method is fast (a few seconds) and precisely delineates the cavity borders at a very high precision (Figure 1).

In addition, a set of 73 cavity descriptors are computed for each cavity and used as input to a support vector machine (SVM) classifier to predict the structural druggability (or ligandability) of the inspected cavity. In a standard benchmarking exercise consisting of 113 cavities of known druggability, Volsite presented the highest accuracy when compared with state-of-the-art tools.[6] In case of multiple cavities, all druggable cavities are saved as readable MOL2 files, along with their predicted druggability score.

Interestingly, the similarity of two Volsite cavities can be estimated by analogy to classical ligand similarity measurements, using a companion tool (*Shaper*)[6] that uses a smooth Gaussian function to maximize the overlap of their volume and pharmacophoric properties. High-throughput cavity comparisons are increasingly used in computational chemistry notably to identify ligands for novel cavities, design inhibitors with precise selectivity patterns and predict their possible side effects.[8]

Converting protein–ligand complexes into fingerprints and graphs: IFP, GRIM. A major feature of IChem is the possibility to generate diverse simplified representations (fingerprints, graphs) of protein–ligand interactions. For example, the *IFP* module enables to list all protein–ligand interactions occurring in a complex and to output an interaction fingerprint as a bit string (Figure 2).

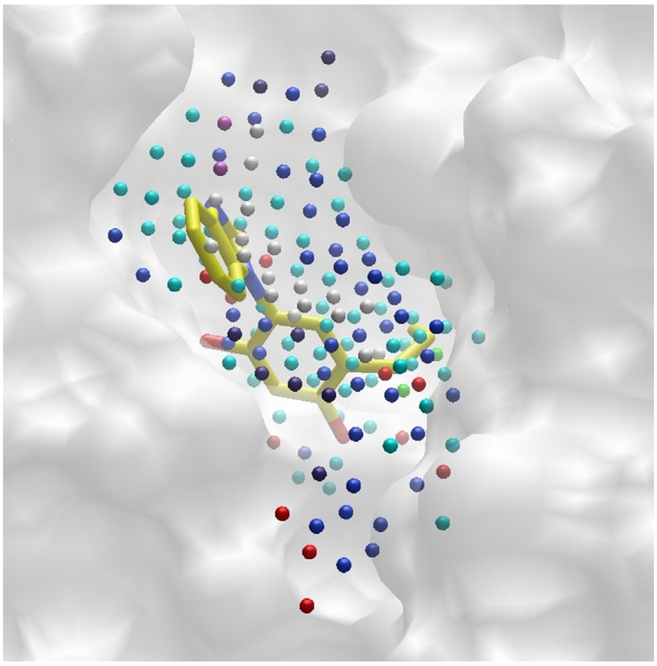

Figure 1. Cavity detection at the surface of the heat shock protein HSP90α (PDB ID: 4YKR). The Volsite cavity (volume = 502.87 Å^3, druggability = 1.11) is represented by pharmacophoric points (hydrophobic, cyan; aromatic, green; hydrogen-bond donor, blue; hydrogen bond acceptor, red; positive ionizable, dark slate blue; negative ionizable, orchid; dummy, white). Volsite points nicely encompass the bound inhibitor (HET code 4ep, yellow sticks) ignored during the cavity detection procedure.

```
HBond LIG     OD1|31   |ASP 113-A  O17|1   |CAU  2.60 134.67
HBond PROT    ND2|99   |ASN 312-A  O17|1   |CAU  2.76 159.23
Hydrophobic   CZ3|60   |TRP 286-A  C16|2   |CAU  4.05
Hydrophobic   CZ |82   |PHE 289-A  C16|2   |CAU  3.74
Ionic LIG     OD2|32   |ASP 113-A  N19|4   |CAU  2.94
Ionic LIG     OD1|31   |ASP 113-A  N19|4   |CAU  3.60
Hydrophobic   CZ3|13   |TRP 109-A  C21|6   |CAU  4.43
Hydrophobic   CH2|14   |TRP 109-A  C22|7   |CAU  3.72
Hydrophobic   CZ |116  |TYR 316-A  C22|7   |CAU  4.46
Hydrophobic   CZ3|60   |TRP 286-A  C15|8   |CAU  4.06
Hydrophobic   CE2|81   |PHE 289-A  C15|8   |CAU  4.03
HBond LIG     OG |42   |SER 203-A  N7 |16  |CAU  3.31 127.77

 |A W109|A D113|A S203|A W286|A F289|A N312|A Y316
1000000000010100001001000000100000000010001000000
```

Figure 2. Table of protein–ligand interactions and interaction fingerprint (IFP) generated by IChem. For every ligand-binding residue, seven bits are switched either on (1) or off (0) as whether a particular interaction is detected or not with the ligand. Interactions are registered in a precise order (hydrophobic, aromatic face-to-face, aromatic edge-to-face, hydrogen bond accepted by ligand, hydrogen bond donated by ligand, ionic bond with ligand negatively charged, ionic bond with ligand positively charged).

Several years ago, we[7] and other groups[9] proposed the use of IFPs to post-process docking data and pick poses producing IFPs similar to that of known actives. Computing interaction fingerprints (IFPs) from docking poses is a robust and very efficient manner to predict ligand binding modes,[10] propose reliable scaffold hops,[11] and enrich virtual hits in true actives.[12] The success of this post-processing approach is based on the idea that true ligands of a same target often share key interactions with key anchoring residues and thereby produce relatively similar IFPs. However, a clear limitation is the strict dependence to the number of active site residues, preventing to compare interaction fingerprints across binding sites of different sizes.

We therefore recently designed size-invariants descriptors conceptualized by a graph describing the exact protein–ligand interaction pattern.[13] The method called GRIM (Graph Interaction Matching) defines three interaction pseudoatoms (IPAs) for every detected protein–ligand interaction: one on the ligand-interacting atom, one on the protein-interacting atom and one at the barycenter of the latter two atoms (Figure 3).

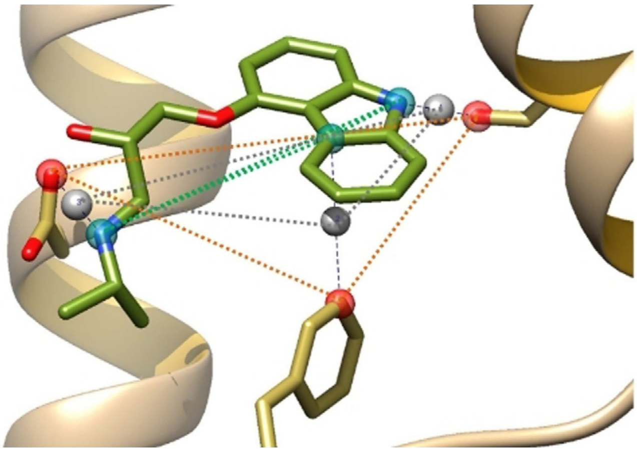

Figure 3. Example of a set of interaction pseudoatoms (ligand-based, red; protein-based, cyan; centered, gray) for the β2 adrenergic receptor in complex with carazolol (PDB ID: 2RH1).

The full set of IPAs defines an interaction pattern that is unique to every protein–ligand complex and that can be converted into a graph where IPAs will define nodes.[13]

A particular interaction pattern can be easily compared and aligned to another one by a simple graph matching technique aiming at identifying the maximal common subgraph (clique).[13] The similarity of two interaction pattern graphs is measured by an empirically derived score (GRIMscore) that can be used for example to post-process docking poses and reward those corresponding to interaction patterns already visited in reference X-ray structures (Figure 4). In three consecutive international docking contests aimed at predicting ligand binding modes prior to the release of the corresponding X-ray structures, GRIM rescoring was always quoted as one of the very best methods for generating near-native docking poses.[14] The same advantage over fast scoring functions was reported in virtual screening against diverse target families (e.g., G protein-coupled receptors, nuclear hormone receptors, protein kinases).[13]

GRIM presents several advantages over alternative knowledge-based rescoring strategies: 1) it can be coupled to any docking algorithm, 2) it does not constrain ligand docking but rewards interaction patterns already present among PDB templates, 3) it takes advantage of ligands with similar binding modes and not necessarily similar chemical structures, 4) it can be applied in a target family-biased pose selection process in which PDB templates from the same protein but also from similar targets can be used to store reference interaction patterns, and 5) it permits to directly quantify binding mode similarity between a predicted protein–ligand complex and any PDB template at a very high throughput.

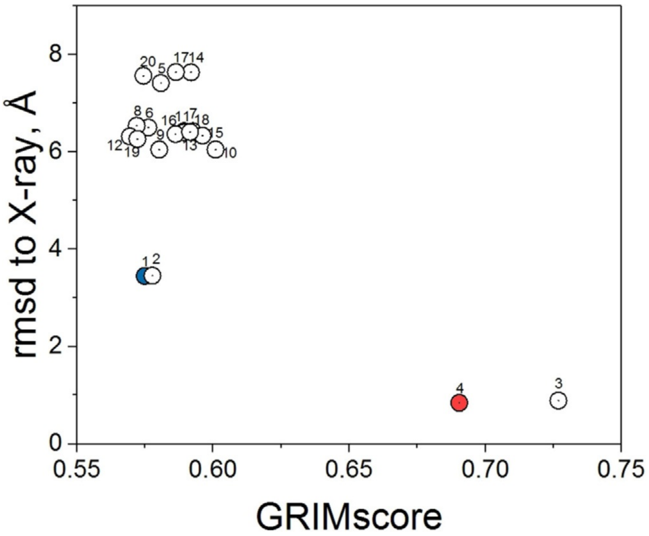

Figure 4. Surflex-Dock rescoring 20 docking poses of the MAP27 inhibitor[14b] to the human MAP4K4 by graph similarity (GRIMscore) to the X-ray structure of the same kinase with the inhibitor GNE-495 (PDB ID: 4ZK5). The top-ranked pose according to Surflex-Dock (cyan circle) is irrelevant. The two best poses according to GRIM (red circles) are < 1.0 Å RMSD away from the true X-ray pose. All poses are numbered from 1 to 20 according to the Surflex-Dock score.

Detecting biologically relevant protein–protein interfaces: detectPPI. Protein–protein interfaces (PPIs) represent challenging but very promising targets for drug discovery.[15] Hence, PPIs describe a vast unexplored biological space for which small molecular weight modulators[16] are expected to offer very high potency and selectivity profiles. Although mainly discovered by biophysics-driven fragment-based approaches, computational chemistry is expected to play a major role in designing the future PPI modulators,[17] notably upon relying on the huge structural information already available in the Protein Data Bank. To discriminate biologically relevant from crystallographic artifacts, computational methods are needed to rapidly detect PPIs and predict their biological relevance from a structural point of view. IChemPIC[18] was designed to address this need. The *detectPPI* module uses the general IChem functions (molecule reader, interaction detection) to detect the interface, identifies the corresponding IPAs and generate a fixed-length property vector (Figure 5) as input for a Random Forest classifier previously trained on a set of 400 PPIs (200 biologically relevant, 200 irrelevant interfaces). IChemPIC is equally robust to detect both classes with the same accuracy, independently on the size of the PPI.[18]

Interestingly, this new IChem module can be used at a high throughout to detect biologically relevant PPIs at the PDB

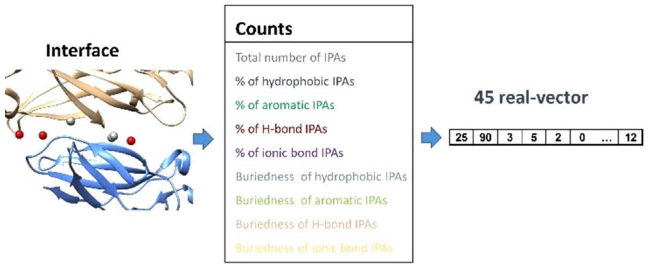

Figure 5. Prediction of biologically relevant protein–protein interfaces by IChem. Interaction pseudoatoms (IPAs) are first computed from the 3D structure of the protein–protein complex and converted into a 45 real vector, featuring important physicochemical properties (interface size, pharmacophoric properties, buriedness), read by a Random Forest classifier for relevance prediction.

scale. Alternatively, the method can be used on-line (http://bioinfo-pharma.u-strasbg.fr/IChemPIC) by just specifying the PDB three-letter code.

In conclusion, IChem is a suite of software dedicated to the analysis and comparison of three-dimensional molecular objects. It converts an intricate three-dimensional information into much simpler fingerprints or graphs, thereby enabling high-throughput comparisons and fueling machine learning models for predicting important features like protein–protein interfaces, druggable cavities, interaction patterns and binding poses. IChem is available for nonprofit academic research at http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html.

Acknowledgements

We thank the Centre National de la Recherche Scientifique (CNRS), the Alsace Region, and the Institut de Recherches Servier for PhD grants to F.D.S. and J.D. Part of this work was supported by the MEDALIS Drug Discovery Center (LABEX ANR-10-LABX-0034_Medalis). Although J.D.'s affiliation is currently Eli Lilly, all presented results pertain to work prior to such affiliation.

Conflict of interest

The authors declare no conflict of interest.

Keywords: cavity · drug design · fingerprints · interactions · virtual screening

[1] The software and user guide can be downloaded at http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html.

[2] P. W. Rose, A. Prlic, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y. P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman, S. K. Burley, *Nucleic Acids Res.* **2017**, *45*, D271 – D281.

[3] Corina v.3.40, Molecular Networks GmbH, Erlangen (Germany).

[4] S. Bietz, S. Urbaczek, B. Schulz, M. Rarey, *J. Cheminf.* **2014**, *6*, 12.

[5] J. Desaphy, G. Bret, D. Rognan, E. Kellenberger, *Nucleic Acids Res.* **2015**, *43*, D399 – D404.

[6] J. Desaphy, K. Azdimousa, E. Kellenberger, D. Rognan, *J. Chem. Inf. Model.* **2012**, *52*, 2287 – 2299.

[7] G. Marcou, D. Rognan, *J. Chem. Inf. Model.* **2007**, *47*, 195 – 207.

[8] C. Ehrt, T. Brinkjost, O. Koch, *J. Med. Chem.* **2016**, *59*, 4121 – 4151.

[9] a) Z. Deng, C. Chuaqui, J. Singh, *J. Med. Chem.* **2004**, *47*, 337 – 344; b) M. D. Kelly, R. L. Mancera, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942 – 1951; c) C. P. Mpamhanga, B. Chen, I. M. McLay, P. Willett, *J. Chem. Inf. Model.* **2006**, *46*, 686 – 698.

[10] M. Chalopin, A. Tesse, M. C. Martinez, D. Rognan, J. F. Arnal, R. Andriantsitohaina, *PLoS One* **2010**, *5*, e8554.

[11] J. Venhorst, S. Nunez, J. W. Terpstra, C. G. Kruse, *J. Med. Chem.* **2008**, *51*, 3222 – 3229.

[12] C. de Graaf, A. J. Kooistra, H. F. Vischer, V. Katritch, M. Kuijer, M. Shiroishi, S. Iwata, T. Shimamura, R. C. Stevens, I. J. de Esch, R. Leurs, *J. Med. Chem.* **2011**, *54*, 8195 – 8206.

[13] J. Desaphy, E. Raimbaud, P. Ducrot, D. Rognan, *J. Chem. Inf. Model.* **2013**, *53*, 623 – 637.

[14] a) P. D. S. F. C. Gomes, F. Da Silva, G. Bret, D. Rognan, *J. Comput.-Aided. Mol. Des.* **2017**, DOI: https://doi.org/10.1007/s10822-017-0046-1; b) I. Slynko, F. Da Silva, G. Bret, D. Rognan, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 669 – 683.

[15] M. R. Arkin, Y. Tang, J. A. Wells, *Chem. Biol.* **2014**, *21*, 1102 – 1114.

[16] M. A. Kuenemann, C. M. Labbe, A. H. Cerdan, O. Sperandio, *Sci. Rep.* **2016**, *6*, 23815.

[17] L. Laraia, G. McKenzie, D. R. Spring, A. R. Venkitaraman, D. J. Huggins, *Chem. Biol.* **2015**, *22*, 689 – 703.

[18] F. Da Silva, J. Desaphy, G. Bret, D. Rognan, *J. Chem. Inf. Model.* **2015**, *55*, 2005 – 2014.

Manuscript received: August 24, 2017
Revised manuscript received: October 6, 2017
Accepted manuscript online: October 10, 2017
Version of record online: November 7, 2017