EXERCISE: PYTHON + STRUCTURAL BIOINFORMATICS

Introduction

Protein-protein interactions (PPIs) are fundamental for all the functions in the cell. Knowing the structure of PPIs can be very useful to understand protein function or the effect of amino acid substitutions. However, determining the structure of PPIs is expensive and time consuming. Therefore, the number of PPIs with available structures is scarce. In contrast, high-throughput methods for the assessment of PPIs provide abundant experimental data that is available in public databases [1]. In this scenario, computational methods can be used to fill the gap between experimental and structural data.

Many computational strategies have been developed to model PPIs. We can classify them in three groups: template based modeling, protein-protein docking and hybrid/integrative modeling. Template based methods work by applying the structural features from a template structure to the PPI of interest. These methods require the template to be a PPI structure whose members are homologous to the proteins we want to model. They sample a small conformational space and have a low computational cost. Protein-protein docking methods work by sampling and scoring lots of possible interacting conformations. They don't require any template, they sample a large conformational space and have a high computational cost. Hybrid/integrative modeling methods are based on combining experimental data with computational approaches. Thanks to the use of experimental data, the sampling of the conformational space and the computational cost are reduced [2,3]. The identification and classification of interfaces plays a relevant role for all these approaches (i.e. by the classification of PRISM [4] or 3DiD [5]).

Interactions of proteins are only components of large complexes that will finally perform the biochemical function, final purpose of protein expression. Approaches for modeling large complexes require the integration of different type of experimental data [6,7]. Even if we know all individual interaction pairs of a complex, modelling the complete complex is a challenge [8] and requires often the guide of the enveloping surface [9,10]. More specifically, some algorithms were developed to fit protein complexes within a surface derived from electron-microscopy data [11]. A particular group of complexes with biological relevance are the transcription complexes. Despite many experiments that determine the capacity of some proteins to bind a specific site of DNA (i.e. in vivo experiments, such as bacterial and yeast one hybrid assays [12,13] and ChIP-Seq [14]; and in vitro experiments, such as SELEX [15,16], SMiLE-SEQ [17], protein-binding microarrays known as PBM [18-20], MPRA [21,22] and DAP-Seq[23]), this is not necessarily in line with the regulation of a gene, requiring the conjunction of many factors (phenomenon known as epistasis) to finally produce its expression or repression.

Project description

Here, **we propose the development of a method to model the macro-complex structure of biomolecules, formed by proteins and DNA, without the explicit requirement of EM data, but with information of each pairing interaction of a complex** (protein-protein, protein-DNA, and also by partial structures of protein-domains and DNA binding-sites: domain-domain, domain-binding). The input for the experiment will be the protein and DNA sequences, and the structures of protein-protein and protein-DNA interactions. The classical examples to train the approach are the structures of ribosomes, nucleosomes, the complex of transcription with polymeraseIII [24] or the complex of interferon-β enhanceosome [25]. The team is free to select the approach: using modeller, IMP, X3DNA [26], or superposition of structures to complete the macro-molecular complex. The training can easily be done by: 1) splitting a known complex in its chains, 2) then joining pairs of interacting molecules as a single PDB file, and 3) using your program to reconstruct the original complex. Some complexes may involve the formation of homo-dimers or the implication of a protein sequence (or even the same DNA binding site) several times. Therefore, it is also acceptable to request as an input the stoichiometry of the complex. Some structures, in fact the majority, characterize only a partial fragment of a protein sequence. The team is free to leave the final protein incomplete, or try to finalize the full complex with the use of MODELLER or iTASSER [27]. The team is also free to derive amino-acid pair interactions or amino-acid -nucleotide interactions from the original complex to infer the final model. This will imply a requisite in the input of residue-pair interactions. However, this is only recommended if it may be derived experimentally (i.e. for residues in the surface) and it is not encouraged the use of other theoretical approaches to obtain them due to the low ratio of success.

Input

Mandatory arguments:

- Directory with all the base pdb files to use.
- Fasta file with the sequences of the complex
- Output directory where store the results

Optional arguments:

- Stechiometry
- Others

Output

Mandatory output:

- PDB file(s) with the suggested files
- Others

Examples for training

Protein examples from PDB: 6GMH, 6OM3, 5FJ8, 4G83, 3T72, 5NSS
Incomplete (real) examples. Human Progesterone Receptor, Human Thyroid

receptor

Requisites and extras

Each team has to have 2-3 members. (3 max. 2 min.)

The team has to have at least one member registered in the course of "structural bioinformatics" and one in "Introduction to Python".

Besides the program, there must be a written tutorial with full explanations on the background theory, the examples selected and the results of the analyses (i.e. comparison of real structure with models)

The program must run also for complexes of only proteins.

Handling small compounds, such as hormones, peptides, metabolites or drugs is a plus, but it is not requested.

Allowing backbone flexibility and conformational changes (i.e. using normal modes or MD) is also a plus, but it is not requested.

DNA conformations can be preselected with X3DNA or constructed via combination of PDBs

Grades of the exercise PYTHON

The evaluation of the project is independent for structural biology and Introduction to Python.

The evaluation will consider:

- Documentation/README
  - How to execute
  - How to install
  - Requirements
  - Code documentation that allows reusability of the package
  - Extras (README.md, examples, pip3/github)
- Structure:
  - Package structure
  - Standalone application
- Installation:
  - complete setup.py
- Arguments/config:
  - Usage of a standard module for arguments/help management
  - Default values
  - config file if necessary
  - help option
- Program model:

    - Definition of classes (if necessary)
    - Reusability
- Usage of third-party packages
- Management of different types of exceptions
- Extras:
    - Graphical output
    - GUI
    - …

References

1 Garcia-Garcia, J. *et al.* Networks of ProteinProtein Interactions: From Uncertainty to Molecular Details. *Mol Inform* **31**, 342-362, doi:10.1002/minf.201200005 (2012).

2 Segura, J., Marin-Lopez, M. A., Jones, P. F., Oliva, B. & Fernandez-Fuentes, N. VORFFIP-driven dock: V-D2OCK, a fast and accurate protein docking strategy. *PLoS One* **10**, e0118107, doi:10.1371/journal.pone.0118107 (2015).

3 Lensink, M. F. *et al.* Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. *Proteins*, doi:10.1002/prot.25007 (2016).

4 Tuncbag, N., Gursoy, A., Nussinov, R. & Keskin, O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* **6**, 1341-1354, doi:nprot.2011.367 [pii]
10.1038/nprot.2011.367 (2011).

5 Stein, A., Panjkovich, A. & Aloy, P. 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res* **37**, D300-304, doi:gkn690 [pii]
10.1093/nar/gkn690 (2009).

6 Saltzberg, D. *et al.* Modeling Biological Complexes Using Integrative Modeling Platform. *Methods Mol Biol* **2022**, 353-377, doi:10.1007/978-1-4939-9608-7_15 (2019).

7 Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384-1403, doi:10.1016/j.cell.2019.05.016 (2019).

8 Russell, R. B. *et al.* A structural perspective on protein-protein interactions. *Curr Opin Struct Biol* **14**, 313-324, doi:10.1016/j.sbi.2004.04.006
S0959440X04000739 [pii] (2004).

9 Lasker, K., Topf, M., Sali, A. & Wolfson, H. J. Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J Mol Biol* **388**, 180-194, doi:10.1016/j.jmb.2009.02.031 (2009).

10 Topf, M. *et al.* Protein structure fitting and refinement guided by cryo-EM density. *Structure* **16**, 295-307, doi:10.1016/j.str.2007.11.016 (2008).

11 Tjioe, E., Lasker, K., Webb, B., Wolfson, H. J. & Sali, A. MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map. *Nucleic Acids Res* **39**, W167-170, doi:10.1093/nar/gkr490 (2011).

12 Meng, X. & Wolfe, S. A. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat Protoc* **1**, 30-45, doi:10.1038/nprot.2006.6 (2006).

13 Deplancke, B., Dupuy, D., Vidal, M. & Walhout, A. J. A gateway-compatible yeast one-hybrid system. *Genome Res* **14**, 2093-2101, doi:10.1101/gr.2445504 (2004).
14 Ambrosini, G., Dreos, R., Kumar, S. & Bucher, P. The ChIP-Seq tools and web server: a resource for analyzing ChIP-seq and other types of genomic data. *BMC Genomics* **17**, 938, doi:10.1186/s12864-016-3288-8 (2016).
15 Hallikas, O. & Taipale, J. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. *Nat Protoc* **1**, 215-222, doi:10.1038/nprot.2006.33 (2006).
16 Roulet, E. *et al.* High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* **20**, 831-835, doi:10.1038/nbt718 (2002).
17 Isakova, A. *et al.* SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat Methods* **14**, 316-322, doi:10.1038/nmeth.4143 (2017).
18 Berger, M. F. & Bulyk, M. L. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* **338**, 245-260, doi:10.1385/1-59745-097-9:245 (2006).
19 Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723, doi:10.1126/science.1162327 (2009).
20 Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **43**, D117-122, doi:10.1093/nar/gku1045 (2015).
21 Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173-1175, doi:10.1038/nbt.1589 (2009).
22 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271-277, doi:10.1038/nbt.2137 (2012).
23 O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280-1292, doi:10.1016/j.cell.2016.04.038 (2016).
24 Han, Y., Yan, C., Fishbain, S., Ivanov, I. & He, Y. Structural visualization of RNA polymerase III transcription machineries. *Cell Discov* **4**, 40, doi:10.1038/s41421-018-0044-z (2018).
25 Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111-1123, doi:10.1016/j.cell.2007.05.019 (2007).
26 Li, S., Olson, W. K. & Lu, X. J. Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res* **47**, W26-W34, doi:10.1093/nar/gkz394 (2019).
27 Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-738, doi:10.1038/nprot.2010.5 (2010).