Computational challenges

The ongoing proliferation of analytical tools and imaging methods places an ever-growing load on the public repositories who make the resulting data available to the public. The orchestration of experiments involving multiple types of data increasingly depend on the capacity of the underlying infrastracture to make the relevant data amenable to computation.

In the case 3D models, cryoEM in particular, the resulting structures are information-rich at multiple scales: atomic, residue, protein, topological. Given that comparison is necessarily involved in the study of heterogeneity, a consistent method to compare these structures' facets precludes any fruitful analysis whether it be sequences conservation, protein clustering or conformational states. (Ex. RPs can be compared across strucutres on the basis of their homology, conservation signatures, spatial position, interfaces with other RPs/rRNA, function and other). By and large for historical reasons data associated with each of these assays belong to different repositories which introduces significant fragmentation into experimental design. Insufficient integration and lack of common ontologies among these repositories become acute as the dataset of interest becomes more heterogenous. Though the challenges to integration are not of any inherent scientific interest, they are an ever-growing impediment to computer-aided investigation and a barrier to certain investigations altogether. The challenges associated with ontology and infrastracture are among significant ones.

Regarding RPs clustering

Infrastracture

There is a number of limitations to the current infrastracture, both between and within repositories:

Integration:

Referring to a particular `.pdb`/`.mmcif` file as a point of departure and given that none of the API's from Uniprot-KB, RCSB PDB and PFAM simultaneously contain a protein's *chain id*, *uniprot accession* and *pfam family*, it appears neccessary to request each of these data from the corresponding API sequentially to profile each protein further.

This is a cumbersome process that can be somewhat ameliorated with the migration to graph-databases and introduction of GraphQL API to query highly-connected and loosely-structured biological data in a more programmatic way. These technologies are seing adoption[1,2,3], especially in the ELIXIR ecosystem, but are still far from being the go-to model.

This can be viewed as a part of a wider aspiration of integrating bioinformatic tools and resources across the ecosystem.

Ambiguity regarding the source of truth:

In small-scale effort like this one it would seem nominally easier to just download the databases of interest locally to eschew dynamic web-requests, especially given that the PDB's SIFTS service aspires to provide just that. Unfortunately, there is a question of consistency, wherein https://www.ebi.ac.uk/pdbe/docs/sifts/quick.html derived via the UniProt mapping does not provide quite the same data as manual querying. (ex. https://www.uniprot.org/uniprot/P0A0F8 is not present in the `.csv` version of the database, could find more).

Given that both PDB Europe's API and that of RCSB PDB provide separate endpoints and serve data in slightly different manner, there is a natural question of what resource to refer to and whether this duplication is necessary.

It is also sometimes ambiguous which endpoint to use within a single resource wether due to documentation or complexity. (Ex. RCSB's XML endpoint[5] provides "RNA" tags whereas the JSON endpoint[6] doesn't). Some of these legacy APIs are being deprecated [7] in 2020 in favor of the more capable ones, but it still leaves many questions of intra-resources integration open.

Bandwidth:

Complete lack or low capacity of programmable interfaces is likewise a major cap on automation of experiments.

(Ex. Uniprot provides a mapping service via a GUI that is capable of handling a large job, but no bulk-processing API endpoint which limits the construction of pipelines that depend on Uniprot-derived mappings to single-datum requests. Ex. InterPro provides a custom InterProScan tool, but for more fine-grained access to the database EMBL-EBI [8] urges not to exceed 30 data per request. Ex. PFAM provides no bulk-processing capabilities.)

Ontologies

Although the nomenclature for ribosomal proteins suggested by Ban et al. has been adopted in most recent structural studies and offers a provisional standard for further depositions, its main goal was to reconcile multiple historical conventions of naming RPs with each other. Consequently, the issue of bringing old nomenclatures into the fold with the ontologies of the existing and proliferating databases remains unadressed. It is important to recognize that families of proteins which are constructed programmatically from sequence and homology data are themselves entities evolving based on the new input and are therefore agnostic to teleologically-inspired top layer of nomenclature. As a result, the correspondence between PFAM families and Ban's classes is still ambiguous and programmatic protein classification presents a challenge because of one-to-many mappings. At present, UniprotKB deals with this issue by providing links to *all* known matches(hits) derived from multile databases and their corresponding

methods of protein-signature generation (*patterns*, *profiles*,*fingerprints*, *HMMs*). Most of these contributing databases are being pooled by the ELXIR consortium at http://www.ebi.ac.uk/interpro/.

Data-driven vs application driven

A lot of obstacls to data-integration both in infrastructure and ontology seem to be borne out of historical contingency where a tool or a database is constructed specifically for the purposes of a single experiment or investigation. Extensibility and interoperability with existing databases and tools always confers additional costs on development and is rarely accounted for in the grants [9]. Hence the mounting heterogeneity and systemic fragmentation.

Perhaps with the advent of graph-databases, HPC and high-dimensional models mega-hubs like InterPro and PDB proper will be enable more research that takes advantage of multiple modalities of data simultaneously and thus embraces overparametrization that biology is rife with. If this is the case, the focus in bioinformatics shifts from "developing databases and tools" to contributing to a data fabric of high connectivity. The question of how to encapsulate this dynamic in the infrastracture and ontologies that are being constructed at present is a pressing one.

1:(https://www.rcsb.org/news?year=2020&article=5eb18ccfd62245129947212a&feature=true)

2:(https://www.ebi.ac.uk/training/events/2020/mining-pdbe-and-pdbe-kb-using-graph-database)

3: @article{cook2019european, title={The European Bioinformatics Institute in 2020: building a global infrastructure of interconnected data resources for the life sciences}, author={Cook, Charles E and Stroe, Oana and Cochrane, Guy and Birney, Ewan and Apweiler, Rolf}, journal={Nucleic Acids Research}, year={2019} }

4:@article{orengo2020community, title={A community proposal to integrate structural bioinformatics activities in ELIXIR (3D-Bioinfo Community)}, author={Orengo, Christine and Velankar, Sameer and Wodak, Shoshana and Zoete, Vincent and Bonvin, Alexandre MJJ and Elofsson, Arne and Feenstra, K Anton and Gerloff, Dietland L and Hamelryck, Thomas and Hancock, John M and others}, journal={F1000Research}, volume={9}, number={278}, pages={278}, year={2020}, publisher={F1000 Research Limited} }

5:(https://www.rcsb.org/pdb/rest/describeMol?structureId=3j9m)

6:(https://www.rcsb.org/pdb/json/describeMol?structureId=3j9m)

7:(https://www.rcsb.org/pages/webservices)

8:(https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/Job+Dispatcher+Sequence+Analysis+Tools+Hom

9:Guzey, A. How Life Sciences Actually Work: Findings of a Year-Long Investigation. Guzey.com. 2019 August. Available from https://guzey.com/how-life-sciences-actually-work