M7.7 / M30 "Big Data software introduced”

Contents

Big Data approaches and Structural Biology

Big Data is often described in terms of the "four Vs": Volume, Velocity, Variety and Veracity. With the latest generation of detectors for crystallography and EM, datasets are now typically in the TBs (volume) and generated in hours (velocity). This suggests the need for hardware solutions, in terms of processing power, storage and network. Some operations are limited by the size of data; for example, it is not so easy for a structural biologist to move images around, and they may not be available in some computer environments where (s)he typically works.

Nevertheless, raw data from individual structural biology techniques can be considered structured data, in that they are well-understood and conform to established formats and data models. The main exception is textual information, recorded in publications or laboratory notebooks and reports. This contrasts with many areas of Big Data science which deal with unstructured data. In moving to interdisciplinary scientific studies, researchers increasingly have to integrate datasets from completely separate experiments (variety), which are sometimes of unknown provenance (veracity).

In this report, we survey some of the technologies being applied in modern Big Data science, and discuss their applicability to Structural Biology. We then propose four small prototypes to demonstrate real use cases. These prototypes are being developed within the West-Life consortium, with input from IBM and the STFC Hartree Centre. Progress is disseminated via the West-Life wiki pages.

Techniques

Map Reduce

The MapReduce formulation [1] was developed to distribute the analysis of large datasets over many compute nodes. The MapReduce approach was popularized by Google for handling massively distributed queries, but has since been applied in a wide range of domains. A typical MapReduce implementation is based on map() and reduce() operations that work on a local subset of the data, but the power of the approach comes from an intermediate step called shuffle() or collate() which is responsible for re-distributing the data across the compute nodes.

The basic data structures for MapReduce operations are key/value (KV) pairs, and key/multivalue (KMV) pairs that consist of a unique key and a set of associated values. In map(), KV pairs are generated by reading data from files or processing existing KV pairs to create new ones. The collate() operation extracts unique keys and maps all the values associated with these keys to create KMV pairs. The reduce() operation processes KMV pairs to produce new KV pairs as input to the following steps of the algorithm. In a parallel environment, the map() and reduce() operations work on local data, while the collate() operation builds KMV pairs using values stored on all processors.

Apache Hadoop is a popular implementation of the MapReduce programming model. In addition, Hadoop provides the Hadoop Distributed File System (HDFS) for storing data on commodity machines, and Hadoop YARN responsible for managing computing resources in clusters. There are a number of additional software packages that extend Hadoop, of which Apache Spark is one of the most popular (Spark also supports other filesystems and cluster management systems).

Machine Learning

The term machine learning covers a range of analytical techniques, such as Decision Trees, Clustering and Neural Networks. In general, each technique has a complex functional form which can be adjusted automatically to explain a set of training data. The aim is that the trained machine learning model is then able to make accurate predictions about similar data. There is always the danger, however, of over-fitting to the training data, in which case predictions for new data will be poor. Thus, various techniques exist to clean and balance training data, and to test for over-fitting.

Machine learning is already used in some structural biology applications. For example, PSIPRED is a simple and accurate secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST. Machine learning, and in particular deep learning, are at the basis of the recent improvement in residue-residue contact predictions for protein structures as seen in the latest CASP12 experiment [2]. The SuRVoS workbench [3] trains a machine learning model from initial user annotations to segment volumes from cryogenic soft X-ray tomography.

Aspects of machine learning are incorporated into Natural Language Processing and Image recognition, described below.

Natural Language Processing

Natural Language Processing is concerned with automated processing and interpretation of natural language corpora, such as scientific articles or patents. Such approaches are likely to be useful in structural biology to extract metadata on the samples used and on non-structural experiments which provide additional information on the macromolecules studied. By using the literature to extract additional metadata on the context of the structural biology experiment, this approach contributes to the area of data provenance addressed in Work Package 6.

An NLP approach needs to be able to recognise significant phrases from within larger texts. Automated annotation of relevant texts (e.g. abstracts or articles) can provide a starting point, but annotation by experts is usually required for better accuracy. The STFC Hartree Centre has developed a user interface which presents experts with sample sentences, and allows quick annotation from a list of possible categories. For example, annotating phrases as gene, protein, residue or metabolite allows the NLP system to build a model of these concepts. Where ontologies are available, the meaning and inter-relation between phrases can be formalised.

Image recognition

Another major application area of machine learning is image/object recognition. While these techniques are widespread in sectors such as business and security, they have also been adopted by the medical image analysis community. A survey of 308 recent publications in medical image analysis is presented in [4]. Many medical imaging techniques are intrinsically 2-dimensional, for example chest X-rays, and image recognition techniques can be adopted straightforwardly. Other techniques produce 3-dimensional data, for example CT scans, and several approaches have been used. Pseudo-2D approaches use independent slices through the 2D data, or multiple projection views, while other studies have used a true 3D analysis. In structural biology, we have a mix of 2D (e.g. diffraction images or electron micrographs) and 3D (e.g. electron densities) datasets.

The most successful type of models for image/object analysis to date are convolutional neural networks (CNNs). In CNNs, weights in the network are shared such that the network performs convolutions of the images. This reduces the number of free parameters in the model, and also makes the feature extraction translationally invariant. In a typical deep learning application, the network has many layers, and convolution layers are mixed with other kinds of layers. CNNs have been developed to recognise specific types of features in electron tomograms [5].

Proposed prototypes

Prototype 1: NLP

Macromolecular structures in the Protein Data Bank are associated with a publication describing the structure. These papers typically describe the wider study, and are a potential source of valuable metadata for the structure. We will Natural Language Processing techniques to annotate a corpus of publications associated with the PDB, and begin to extract some useful metadata.

A meeting was held on 20th Sept 2017 between Chris Morris, Rebecca Mackenzie, Andrew Gargett (STFC), Geeth De Mel (IBM), Sameer Velankar (EMBL-EBI, PDBe) and Johanna McEntyre (EMBL-EBI, Europe PMC) to discuss this prototype. The prototype will work with the corpus of abstracts linked to PDB entries, for example http://www.ebi.ac.uk/pdbe/entry/pdb/1cbs/citations. These are also available as JSON files, and can be obtained via the PDBe REST API

```
from pdbe import pyPDBeREST
p = pyPDBeREST()
print(p.PDB.getPublications(pdbid='1cbs'))
print(p.PDB.getRelatedPublications(pdbid='1cbs'))
```

Abstracts are available for a large number of PDB entries, but not all. Where they are available, the prototype tool will output the following:

- A Named Entity Recogniser for specific residues within the chain, e.g. "ARG132".
- Code to verify that the identified residue exists within the protein of interest.
- Validation of the above by comparing different papers on the same protein
- Code to generate an annotation in the PDB, linking to a specific sentence in the paper.

For the example of entry 1CBS, the abstract contains the sentence "*The carboxylate of the ligand interacts with the expected trio of residues (Arg132, Tyr134 and Arg111; CRABP II numbering)*". The residue numbering from the literature abstract agrees with the residue numbering in the PDB file, but may be offset in other papers / structures of this protein.

The prototype will demonstrate the ability to extract metadata on important residues from the abstracts of related papers. Future work will extend this to full papers (where available). Other future work could include:

- Integration of the new Named Entity Recogniser into EuropePubMed ingest system.
- Integration into IUCR authorship system.
- Interface with Uniprot and OMIM residue anotations.
- Additions to Sequence Types and Features Ontology

Prototype 2: Image recognition

In electron cryo-microscopy, maps are often sharpened to enhance high resolution features and help with model building. However, map sharpening also enhances noise, and can hinder map interpretation by breaking the continuity of the macromolecular chain. It would be useful to have an automated assessment of map interpretability as a function of degree of sharpening. As a second example, attempts to phase diffraction data from crystallography result in electron density maps which can be poor or even wrong. It would be useful to have a quick assessment to decide if the maps are correct, and if so whether the quality is good enough to take forward to model building.

We will explore machine learning techniques, related to image and object recognition, to classify regions of maps from cryoEM or crystallography. Such techniques are gaining widespread acceptance in the medical imaging community [4], and we wish to investigate whether they are useful for structural biology.

We have generated two test datasets from a set of sharpened or blurred EM maps. The first consists of 2D slices extracted from the maps, along the 3 axes and using a set window size. The code for extracting slices makes use of the mrcfile library from CCP-EM (STFC partner). The second consists of screenshots of the maps in different orientations. The latter are essentially 2D projections with depth cueing, and are closer to what a scientist would work with.

A convolutional neural network (CNN) to analyse these datasets is being set up using the Keras package (https://keras.io/) with Theano (http://deeplearning.net/software/theano/) as the backend. The prototype will explore the optimum architecture of the CNN, as well as the best way to present the cryoEM and crystallographic data.

Prototype 3: Map Reduce

The MapReduce paradigm allows for the parallel processing of large datasets. Within structural biology, there are many applications where a scientist wishes to query all structures in the PDB, and such operations can be reformulated as a MapReduce problem. We will use a Hadoop/Spark platform available at the STFC Hartree Centre to trial such a setup.

A possible framework for parallel analysis of PDB structures is provided by PDB-Hadoop https://zenodo.org/record/16030#.WfD6JHZryM9 This consists of a set of bash scripts. One script prepares the set of PDB files for deployment on HDFS. Another script provides the mapper in a call of hadoop:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
       -Dmapred.sort.avoidance=0 -Dmapred.reduce.tasks=0 \
       -D stream.non.zero.exit.status.is.failure=false \
       -input /user/hduser/pdb-text-full.txt \
       -output /user/hduser/pdb-legacy-output \
       -mapper "/path/to/local/pdb-hadoop/pdb-hadoop.sh"
```

We have deployed this package on our test platform, and tried some example applications. Anna Paola Carrieri (IBM) is working on applications of Hadoop for genomics, and is providing advice.

Prototype 4: large datasets

Identifying and quantifying structural similarities between macromolecules is a common but complex operation often performed on the Protein Data Bank. The West-Life partner Masaryk University is investigating how an all-against-all structural superposition can be performed efficiently on the current holdings of the PDB. This would likely make use of the CCP4 program Gesamt [6] which implements a fast and general graph-based matching algorithm. The result of pre-calculating the structural features would be an indexing of the PDB which would enable a quick structural search with any new structures.

References

[1] Dean J, Ghemawat S: (2009) “MapReduce: Simplified data processing on large clusters.” *Computing in Science and Engineering*, **11**, 29-41

[2] J. Schaarschmidt, B. Monastyrskyy, A.y Kryshtafovych and A.M.J.J. Bonvin. (2017) “Assessment of Contact Predictions in CASP12: Co-Evolution and Deep Learning Coming of Age.” *Proteins: Struc. Funct. & Bioinformatics.* Advanced Online Publication

[3] I. Luengo et al., (2017) “SuRVoS: Super-Region Volume Segmentation workbench.” *J Struct Biol*. **198**, 43-53

[4] Geert Litjens et al. (2017) "A Survey on Deep Learning in Medical Image Analysis" arXiv:1702.05747

[5] Chen M et al. (2017) “Convolutional Neural Networks for Automated Annotation of Cellular Cryo-Electron Tomograms” arXiv:1701.05567

[6] Krissinel, E & Uski, V (2017) "Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB" *J Comp Sci Appl Inform Technol*. **2**, 1-7.