Protein Classificaiton Project Update

Christopher Mertin Samuel Leventhal

October 28, 2015

- The progress you have made towards your goal. (This can not be just "We collected data".) (50%)

  While proteins can be somewhat classified based on their domains and structures it is not a trivial task since overlap exists between families in which domains they contain. As a result not all families are solely classified based on the domains that they contain. For example, the protein family known as "Regulator of G-protein signalling" (RGS) is a protein group that was defined as having the RGS domain in its protein chain. However, there are other families that contain this protein as well, so simply labeling the protein based on the domain that it contains is not ideal [1].

  Even though it is not ideal, it can be used to classify the proteins. There are alternative methods that can be used to classify the proteins, however they're more laborious. The four different approaches that can be used are:

  - **Patterns:** Many important sequence features are represented by only a few amino acids, which can be modeled quite easily with *patterns*.
  - **Profiles:** Used to model protein families and domains by looking at the frequency of which certain aspects occur.
  - **Fingerprints:** Used for multiple motifs for being able to distinguish individual *sub families* in proteins as it can represent small variations.
  - **Hidden Markov Models (HMMs):** Adept at representing amino acid insertions and deletions by using a score of the position of the sequence. These are the most advanced type of data sets that are used to represent these proteins, as they represent the most amount of data.

  One interesting feature that we thought we might implement was using multiple machine learning algorithms to test their efficiency and accuracy on the same data. For example, the following algorithms can be tried to be used to represent the data: ID3, Markov Chaining, Support Vector Machines (SMVs), Logistic Regression, and Boosted Decision Trees. Using these, it will be interesting to see how each of these perform. It would also be interesting to try the different sets of data as well, since they represent different features.

  Another novel approach we are considering is to exploit the large amount of visual models of proteins. Such models, specifically 3D representations, are abundant in the RCSB protein data bank [2]. We predict it may be possible to use the highly developed methods in image analysis such as Convolution Neural Networks in order to construct fixed-length vector representations of proteins which may be more insightful than current classification techniques such as domains or other previously mentioned approaches. It may then be possible to employ another learning algorithm which uses the last hidden layer as input.

  It was decided that the best types of datasets to look at were those from *Profiles* and *Hidden Markov Models*. The data from the profiles are known as HAMAP [3] and PROSITE [4]. There are a bit more databases with data about HMMs which include Pfam [5], SMART [6], TIGRFAM [7], PIRSF [8], PANTHER [9], Superfamily [10], and Gene3D [11].

  After implementing a successful classification algorithm we then hope to use generative machine learning models to build theoretically possible proteins. One possibility is Markov chaining. However, a more successful approach would be Recurrent Neural Networks (RNN), and specifically Long-Short Term

Memory (LSTM) due to its use of memory and addressing better the issue of exploding/vanishing gradients. Our choice of using LSTM is natural due to it's recent success in sequence generation [12]. Once a generative model is in place it would then be interesting to attempt to improve our hyperparamters and model selection by feeding generated sequences into our classifier.

One specific method of layering learning models for improved training would be to use the image approach mentioned followed by a recurrent neural network. Specifically we could first pre-train a CNN for the protein image classification task and use the finite vector representation as input to a LSTM decoder that generates potential proteins. Similar work along this vein has been done by Vinyals *et al.* for generating properly formed descriptions of images [13].

We have not finished going through these data bases yet, but we will need to find the best ones for each set of data in what we want to use to classify them. More than likely the databases cannot work between one another due to the way that they implement their classification algorithms is most likely different than the others. So, even though we downloaded the data for the proteins, we need to look at the data to see which one would be the best to use.

- Details of your plan for rest of the semester (30%)

  The rest of the semester will be done by learning about the alternative machine learning models that we'd like to implement and test with. We're going to be using at least two data sets to classify the data, as more than likely the Markov Chaining will do better with the data produced by the Markov Models.

  The protein data itself needs to be analyzed before we can use it for this project, so for a good portion of the semester we will be looking at the data and figuring out the best way to interact with it.

  Thankfully, these alternative machine learning methods will be implemented in future homework assignments, so we will have to learn about the anyways and have to implement them for our assignments, so the algorithms won't need to be changed much between the homework and the project that we're working on.

References

[1] European Bioinformatics Institute, "Protein Classification," `https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification`

[2] "RCSB Protein Data Bank," `http://www.rcsb.org/pdb/home/home.do`

[3] T. Lima *et al.*, 2009. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res. 37:D471-8 `http://europepmc.org/abstract/MED/18849571`

[4] CJ Sigrist *et al.*, 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38:D161-6 `http://europepmc.org/abstract/MED/19858104`

[5] European Bioinformatics Institute, "Pfam," `http://pfam.xfam.org/`

[6] European Molecular Biology Lab, "SMART," `http://smart.embl-heidelberg.de/`

[7] J. Craig Venter Institute, "TIGRFAM," `http://blast.jcvi.org/web-hmm/`

[8] Georgetown University Medical Center, "PIRSF," `http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml`

[9] Paul Thomas, "PANTHER," `http://pantherdb.org/`

[10] "Superfamily," `http://supfam.org/SUPERFAMILY/`

[11] "Gene3D," `http://gene3d.biochem.ucl.ac.uk/Gene3D/`

[12] A. Graves. Generating sequences with recurrent neural net- works. arXiv:1308.0850, 2013.

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555, 2014.