Problem statement: Deep learning on protein sequences

Goal: Using a large corpus of protein sequences and their annotated protein families, we learn dense vector representations for amino acid sequences using the co-occurrence statistics of short fragments.

Database: UniProt

Background:

Previous work:

1. searching for common motifs or through phylogenetic comparison against known protein sequences. Classifying new proteins using this technique used a lot of feature engineering using prior knowledge and exports for annotations.
2. word2vec skip-gram was used to create vectors representing trigrams of amino acids could be trained on large amounts of protein sequence data. In Skip-gram, the model takes a target word as input and tries to predict the surrounding context words within a fixed-sized window. This was successful.

What is being tried:

1. The paper tries to evaluate if Global Vectors for Word Represetation (GloVe) can generate improve representations using the full co-occurrence statistics available in our corpus.

   Trigram : a trigram refers to a sequence of three consecutive amino acids.

   The paper talks about “Createing Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics”
   Each trigram is transformed into a point in a multi-dimensional space. This "distributed representation" means that each dimension of this point represents some aspect of the trigram's properties, such as its role in the protein structure, its chemical characteristics, or its evolutionary conservation.
   They chunk a sequence into a group of 3 AA(a trigram) and create a distributed representation using GloVe

   Differences between GloVe and word2vec: Word2Vec focuses on local context, learning embeddings based on nearby words in sentences whereas GloVe takes a global approach, considering word co-occurrences across the entire corpus.

   Why did GloVe work better?
   Using full co-occurrence statistics available in our corpus must have something to do with it.

GloVe uses a co-occurrence matrix that represents how often words appear together in a corpus. It then factorizes this matrix to obtain word embeddings. capture the ratios of word co-occurrence probabilities.

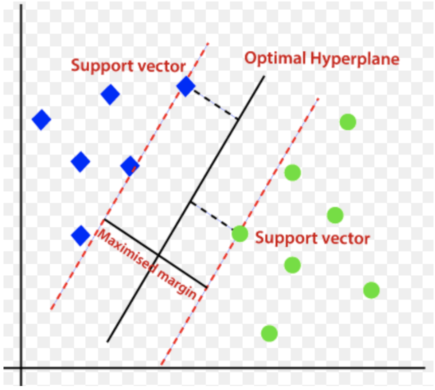
For the GloVe embeddings this is how the co-occurence matric is created:
They compute the trigram co-occurrence matrix for all sequences using a symmetric window of 12 on each side. With non-overlapping trigrams in all three shifts resulting from shifting the reference location by 1.
This means they're taking sets of three items from the sequence in three different ways, without overlapping, by slightly changing the starting point each time.
For example: Seq: ABCDEFG -> ABC DEF
Then shifting Shift by one: GABCDEF and computing non overlapping trigrams again

This resulted in 10,311 trigram represented and 62 million non-zero entries in the cooccurrence matrix. Based off of the histogram of co-coccurrence counts (Fig. 4a), we selected a co-coccurence cap of 20 for the GloVe model.

How can we improve this? BioBERT, Bio-Vec, ProtVec

SVM:
soft margin classifier
For the SVMs a protein sequence is represented by the sum of all its trigram representations
In SVM all we are trying to do is find a plane which will help us separate out data points along with maximizing the margins.

What do we mean by that?



Our job is to make sure the margin is maximized for a fair classification of data even if it means allowing some misclassficiations.
What data points matter to us? The support vectors.

NNs:
Neural network is the swiggle that fits all classes
Parameters estimates are analogous to the slope and intercept value that we solve when we fit a straight line to the data

GRUs
We talked about NN now we need to understand RNNS to move onto GRU.
RNNs are a type of neural network that are specialized for processing sequences of data.
They are used when the context or the order of the data is important in our case protein sequences and the corresponding trigrams. (24 trigrams)

What is Special about RNNS
RNNs have memory that captures information about what has been calculated so far. This is implemented through hidden states that retain some information about a sequence up to the current point.

Looks great whats the catch?

Challenges with Basic RNNs:

Vanishing Gradient Problem: When training RNNs using backpropagation, the gradients can tend to become very small, effectively preventing the network from learning long-distance relationships within the data.

Exploding Gradient Problem: Conversely, gradients can also grow exponentially, which can cause wildly fluctuating weights and unstable training.

Gates: GRUs have gates that decide what information should be passed along to the output and what should continue to be remembered for future use.
They have two types of gates:
Reset Gate: Determines how to combine the new input with the previous memory.

Update Gate: Decides how much of the past information needs to be passed along to the future.

The writer does mention more details about how they selected nodes, layers or other hyperparameters for the model in the results table

Activation function is softmax

How was the model improved?
By maxpooling over the entire sequence of hidden states and our final models concatenated the maxpool and final hidden state.
With the maxpooled output, we could overfit smaller datasets easily and on the full dataset our

models were able to reach validation F1 scores above 0.9 within the first 7 epochs.
What else ?
GRU model was training 10% faster than the LSTM models.

LSTM
Instead of using the same feed back loop connections for events that happened long ago and events that happen yesterday to make predictions about tomorrow
it breaks it into two loops one for long memory back and one for short memory back

How do we avoid the issues?
The long term memories are allowed to be passed in a way without any weights associated with them
Whereas the short term memory has weights associated to it

Slide 1:
DNA:

1. DNA contains all the information but the functional way to grab that information is protein also knows and the polypeptides
2. **Polypeptide** is just another word for a chain of amino acids. Although many proteins consist of a single polypeptide, some are made up of multiple polypeptides(hemoglobin). Genes that specify polypeptides are called **protein-coding** genes.

3. First step is dna duplication process and after that the non-coding strand of the dna also known as the template strand acts as the template and the process of transcription occurs on it.

4. Promoter Regions: Specific DNA sequences that act as molecular 'on-switches' where transcription happens and RNA polymerase initiate the transcription of genes.

This raw rna is known as the primary rna

TRANSCRIPTION

Transcription Initiation: RNA polymerase binds to the promoter region, unwinding the DNA to transcribe the gene into pre-mRNA which eventuall becomes mRNA.

Before we talk about mRNA we need to know some after terms like:

Regulatory RNA: For example lets think of a cell as a factory and genes as the workers. Most of the workers are involved in making the toys(which for us are the protien) while some are the managers which control the start, stop, speed and series of production. These managers/RNA are giving instructions to the genes which for us is what regulatory RNA does. In more scientific terms: Regulatory RNAs don't build the proteins; they regulate how genes are expressed, if a gene is turned on or off, how much of a gene's product is made.

Transfer RNA (tRNA): Lets compare a restaurant to a cell and protein to a meal. The wait stuff or tRNA is responsible for bring the ingredients(which are the amino acids) to the chef(which in this case are the ribosomes) in the correct order based on the recipe(mRNA). Each waiter is responsible for a specific ingredient, and they must ensure they deliver it precisely when the chef needs it to add to your dish.

In more scientific terms: Transfer RNAs (tRNAs) are the delivery workers of the cell. They carry the building blocks (amino acids) that are used to make proteins. Each tRNA has a specific matching code (anticodon) that pairs with the recipe on the mRNA. This ensures that the amino acids are added in the correct order to make the protein exactly as it's supposed to be.

RBS:

Ribosome Binding Site is a specific parking site on the street that is messenger RNA (mRNA) molecule. The ribosome being the car for that specific parking site. the RBS is a specific sequence of nucleotides (the building blocks of RNA) where the ribosome knows to "park" itself to start its work. This site is located within the upfront section of the mRNA (5' UTR) that sets the stage for this translation but doesn't actually code for the protein itself.

mRNA Processing: Introns removed, exons spliced together, addition of a 5’ cap and a 3’ poly-A tail(for stability) to form mature mRNA.

TRANSLATION:
After mRNA is created it gets translated in amino acids where a group of 3 neucleotides form one codon and each codon has a corresponding amino acids. Amino acids are linked together, forming a growing polypeptide chain as the ribosome reads mRNA codons.

Termination of translation occurs when a stop codon (UAA, UAG, UGA) is reached.

Protein folding & Post translational modification
Newly synthesized polypeptides fold into unique 3D structures, often aided by chaperone proteins. Chaperone proteins are like the helpful guides and quality control inspectors of the cellular world. chaperone proteins assist other proteins within the cell to fold into their correct three-dimensional shapes as incorrect shape of a protein can lead to it not functioning and even cause harm

Proteomics: beyond translation

- The large-scale study of proteomes—entire sets of proteins produced by organisms.
- Proteomics examines the structure, function, and interactions of proteins, categorizing them into families based on sequence and structural similarities.

SLIDE 2:

Protein families:
A protein family is a group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure.
Protein families are often arranged into hierarchies, with proteins that share a common ancestor subdivided into smaller, more closely related groups. The terms superfamily (describing a large group of distantly related proteins) and subfamily (describing a small group of closely related proteins)

WHY STUDY PROTIEN FAMILIES
Studying protein families helps us understand how proteins evolve, predict their functions,

Evolutionary Basis of Protein Families:

We talk about grouping families of proteins together but we need to understand how are these families formed what happened before that led to creating of these groups:

Gene Duplication: The origin of protein families is often due to gene duplication events, where a gene is accidentally copied in an organism's DNA.
Divergence: Over time, these duplicated genes can accumulate mutations leading to slightly different proteins that may perform related, but distinct functions.
Conserved Sequences: Members of a protein family often have conserved sequences, known as motifs or domains, that have specific structural or functional roles.

How do we classify these Protein Families

Sequence Alignment: Proteins are classified into families by aligning sequences to find conserved patterns.
Structural Similarity: Proteins with low sequence similarity but similar structure and function may also be classified in the same family.
Functional Classification: Enzymes are often grouped into families based on the reactions they catalyze, irrespective of their sequence or structural similarities