Res2Vec: Amino acid vector embeddings from 3d-protein structure

Scott A. Longwell[1] (longwell@stanford.edu),
Tyler C. Shimko[2] (tshimko@stanford.edu)

[1]Department of Bioengineering, Stanford University, Stanford, CA, 94305; [2]Department of Genetics, Stanford University, Stanford, CA, 94305

Abstract

The twenty naturally-occurring amino acid residues are the building blocks of all proteins. These residues confer distinct physicochemical properties to proteins based on their chemical composition, size, charge, and other properties. Here, inspired by recent work in the field of natural language processing [1], we propose a vector embedding model for amino acids based on their neighbors in 3-dimensional space within folded, active protein structures. We then test the utility of such a model by using the embedded vectors to predict the effect of mutation on protein activity.

Data

The data for this project were acquired from the RCSB PDB [2] and split in training, validation, and test datasets using the methodology described by ProteinNet (`https://github.com/aqlaboratory/proteinnet`). The number of example proteins by sequence similarity/dataset is shown in the table to the right. Note that each protein is composed of hundreds of amino acids.

Seq. sim. threshold	Proteins
30%	22,344
50%	29,936
70%	36,005
90%	42,507
95%	43,544
100%	87,573
VAL	224
TEST	78

Features

To provide a uniform coordinate basis for each focus residue, we centered the coordinate system at the alpha carbon of the residue of interest and transformed the coordinates such as to adhere to the coordinate system outlined below. This change of basis allowed us to implicitly correct for rotational invariance.

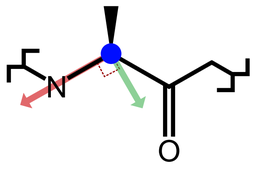


Figure 1: The transformed coordinate system shown with respect to the target amino acid. The red arrow indicates the positive direction of the x-axis and the green arrow denotes the positive direction of the y-axis, both in the plane of the page. The blue circle indicates an arrow in the positive direction along the z-axis perpendicular to the plane of the page.

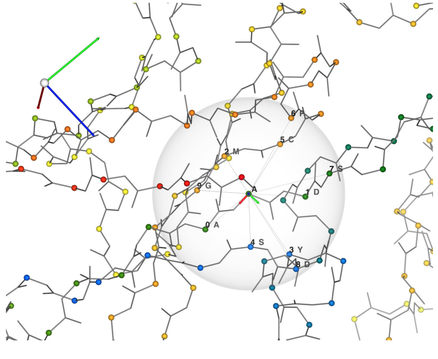


Figure 2: The transformed coordinate system shown in place at a specific focus residue. The x, y, and z axes are shown in red, green, and blue, respectively, and a line is drawn between the alpha carbon of the focus residue and the alpha carbon of each of the 10 closest residues. Original PDB coordinate system shown in upper left.

Models

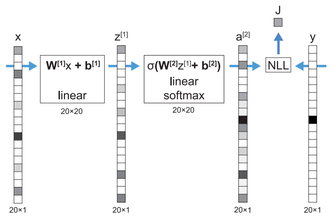


Figure 3: A diagram of the model architecture. The following vector is used to weight the 10 context residues to generate "x":

```
ReLU(linear(ReLU(linear(coords))))
```

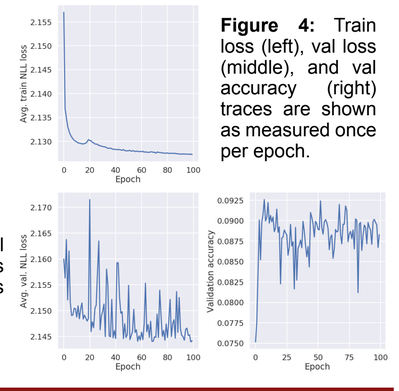


Figure 4: Train loss (left), val loss (middle), and val accuracy (right) traces are shown as measured once per epoch.

Results

Many different model architectures were trained, but we ultimately decided on an architecture with 20 hidden units and a fully-connected coordinate combination layer. This model was trained on the 30% sequence similarity dataset and ultimately achieved a training set average loss of 2.12, a validation set average loss of 2.14, and a validation set accuracy of 8.7%. On the test set, the model had an accuracy of 8.9%.

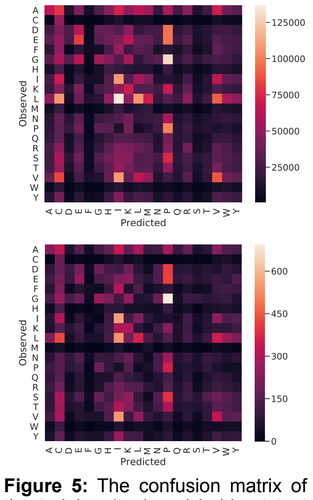


Figure 5: The confusion matrix of the training (top) and held-out test set (bottom).

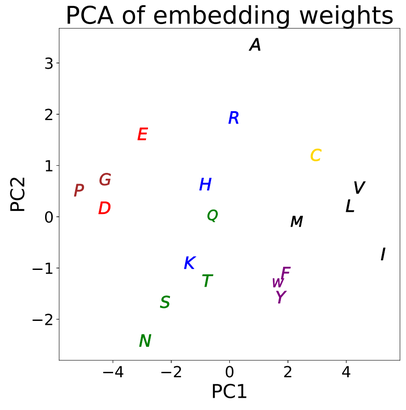
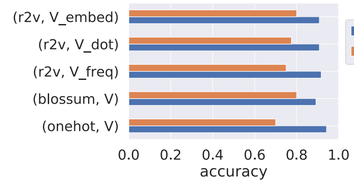


Figure 6: The first 2 principal components of the embedding vectors for each residue are shown (44.5% and 9.4% V.E., respectively). Residues are identified by single letter code and colored by predominant chemical property.

Discussion

We assessed the performance of the embeddings by training an SVM to predict the effect of mutation on the T4 lysozyme protein [3]. We compared the performance of our embeddings to one-hot encoding vectors and BLOSUM empirical substitution matrices [4].



V_dot

$$N_{row}[i,j] = M[i,j] / \sum_j M[i,j]$$
$$N_{col}[i,j] = M[i,j] / \sum_j M[i,j]$$
$$Row_i = N_{row}[i,:] / \sqrt{\sum_k (N_{row}[i,k])^2}$$
$$Row_j = N_{row}[j,:] / \sqrt{\sum_k (N_{row}[j,k])^2}$$
$$Col_i = N_{col}[:,i] / \sqrt{\sum_k (N_{col}[k,i])^2}$$
$$Col_j = N_{col}[:,j] / \sqrt{\sum_k (N_{col}[k,j])^2}$$
$$S_{dot}[i,j] = \log\{dot(Row_i, Row_j) + dot(Col_i, Col_j)\}$$

V_freq

$$p(i,j) = M[i,j] / \sum_i \sum_j M[i,j]$$
$$q_{row}(i) = \sum_j M[i,j] / \sum_i \sum_j M[i,j]$$
$$q_{col}(j) = \sum_i M[i,j] / \sum_i \sum_j M[i,j]$$
$$S_{freq} = \log\{p(i,j) / q_{row}(i) * q_{col}(j)\}$$
$$S_{freq} = (S_{freq} + S_{freq}{}^T)/2$$

Figure 7: SVM model accuracy for T4 lysozyme mutation effect prediction is shown. Four-fold cross validation was employed and mean accuracy is shown. The formulae for the V_dot and V_freq scores (adapted from [3]) are shown at right.

Conclusions/Future Directions

The vector embeddings we created capture a sizable portion of the variation in amino acid properties. We applied these embeddings to predict the (categorical) effect of mutations on the T4 lysozyme protein and our embeddings matched performance of the current standard in the field, BLOSUM empirical substitution matrices, without the same need for computationally expensive sequence alignment. Given further time to improve model performance we would implement the following changes:

To improve training:

- Incorporate information about angles of context amino acid side chains
- Use distance to any atom in side chain as the context distance metric

To evaluate utility:

- Train a regression model for mutation effect prediction in addition to the SVM described in the Discussion section above

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv.org, vol. cs.CL. 16-Jan-2013.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank,” Nucl. Acids Res., vol. 28, no. 1, pp. 235–242, Jan. 2000.

[3] W. Torng and R. B. Altman, “3D deep convolutional neural networks for amino acid environment similarity analysis,” BMC bioinformatics, vol. 18, no. 1, p. 302, Dec. 2017.

[4] S. Henikoff and J. G. Henikoff, “Amino acid substitution matrices from protein blocks,” PNAS, vol. 89, no. 22, pp. 10915–10919, Nov. 1992.