Predicting Correctness of Protein Binding Orientations

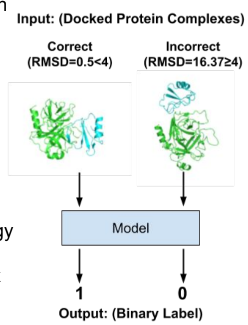
Sarah Gurev
sgurev@stanford.edu

Nidhi Manoj
nmanoj@stanford.edu

Kaylie Zhu
kayliez@stanford.edu

Motivation

- Most protein complexes do not have an experimentally determined structure.
- Scientists would like to use individual protein structures to model a 2 protein complex, but many orientations can result from the docking simulation.
- We built a SVM and ResNext 3D CNN to predict whether a docked protein structure has the correct binding orientation.
- **Input:** Positions of all atoms and energy values for docked complex (PDB File) and RMSD of complex to true complex (label)
- **Output:** RMSD (regression) or Correct/Incorrect (classification)



Data and Features

Get Dockings

- Obtain 10,000 Haddock dockings for each complex in Docking Benchmark 5 as PDB files (from Levitt Lab) and RMSD as labels for each docking

Split Data

- Split 135 complexes amongst train, validation, test set
- Undersample dockings used so equal number of positives and negatives

Preprocess

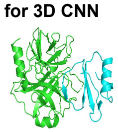
- Normalize by mean and standard deviation of training data
- Threshold RMSD at 4.0 for classification

Extract Features

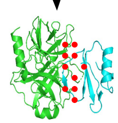
- For SVM: electrostatics, van der waals, buried surface area, and solvation energy (all provided as raw data)
- For CNN: Make cubic inputs (derived feature)

Make cubic inputs for 3D CNN

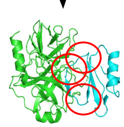
Randomly rotate atom positions of docked protein complex



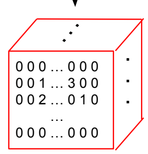
Find center of each interaction between proteins



Cluster interaction centers and find cluster center and all atoms within radius



For each cluster, make cubes with 1 cubic angstrom voxels that are 0 if no atom and number for atom type otherwise



Models

Support Vector Machine (SVM):

- Equation parameterized with w, b

$$\mathrm{h}_{w,b}(x) = g(w^T x + b)$$
where $g(z) = 1$ if $z \geq 0$ and $g(z) = -1$ otherwise

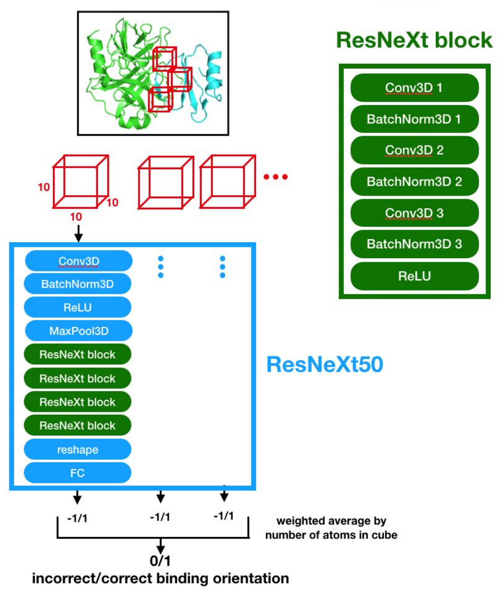
- Radial Basis Function (RBF) Kernel
- Grid search over optimal $2^{-15} <= C <= 2^{10}$ and $2^{-10} <= \gamma <= 2^{10}$
- Optimizing problem:

$$min_{w,b} \frac{1}{2} ||w||^2$$
$$\text{such that } y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, ..., m$$

ResNeXt 3D CNN:

- Experimented with various learning rates, batch sizes, and ResNeXt model depths
- Aggregated model output decisions for each docking
- Used ADAM optimizer, learning rate reduced on plateau, and binary cross entropy with logits loss function:

$$\mathrm{L}(\mathrm{X}, \mathrm{y}) = -\, \mathrm{w}_+ \cdot y log p(Y = 1|X)$$
$$-\, \mathrm{w}_- \cdot (1 - y) log p(Y = 0|X)$$
where w_+ and w_- are the fraction of correct and incorrect orientations



Results

Model	Hyperparameter values	Train F1 Score	Test F1 Score	Train ROC-AUC	Test ROC-AUC	Train R^2 (5-fold cross validation)	Test R^2
SVM Regression	C=4, γ=32					0.445	0.171
SVM Classification	C=2, γ=32	0.888	0.870	0.950	0.501		
3D CNN ResNeXt101	lr = 0.5e-2, bs = 32	0.851	0.825	0.903	0.864		
3D CNN ResNeXt50	lr = 0.5e-2, bs = 32	0.956	0.929	0.981	0.954		

- Train has 7812 samples and Test has 1472 samples
- bs = batch size, lr = learning rate

Discussion

- First we formulated the problem of predicting the correctness of a protein binding orientation as a regression task (predicting RMSD values) which did not perform well.
- In order to improve performance, we reframed the problem as a classification task, which achieved more promising results as expected.
- We then used a ResNeXt 3D CNN that accepts 3D cubes of atom position data from the protein interface region and computed a weighted average over all cubes in the protein.
- Our best model is a ResNeXt50, which has a F1 score of 0.929 and, as expected, outperforms our SVM results.

Future Work

- Determine the efficacy of our model on the original, unbalanced dataset (more negative than positive simulated examples) with precision, recall, and the average rank of the top true positive
- Add more attributes (atom charge and specific atom type) and make a multichannel 3D CNN

References

F. Amir, A. Minhas, B. J. Geiss, and A. B. hur (corresponding. Pairpred: Partner-specific prediction of interacting residues from sequence and structure, 2013.

S. Basu and B. Wallner. Finding correct proteinrotein docking models using proqdock. *Bioinformatics,* 32(12):i262–i270, 2016.

S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie, and P. E. Bourne. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J Chem Inf Model*, 51(2):408–419, Feb 2011.

R. Sanchez-Garcia, C. O. S. Sorzano, J. M. Carazo, and J. Segura. Bipspi: a method for the prediction of partner-specific protein–protein interfaces. *Bioinformatics,* page bty647,2018.

T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kastritis, M. Torchala, R. Chaleil, B. Jimenez-Garcıa, P. A. Bates, J. Fernandez-Recio, A. M. Bonvin, and Z. Weng. Updates to the integrated protein–protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. Journal of Molecular Biology, 427(19):3031– 3041, 2015