Modeling the Structure of SH3-Peptide Complex

Mor Vazana, Magal Gafni, Hilla Asida and Gabriella Levy

SH3 is a small protein domain that has a key role in mediation of signal regulation in the cell and is regulated by binding peptides. Due to its important role, understanding its structure as a domain-peptide complex is a challenge we would like to concur. Here we introduce SH3-Net , a neural network that, given a sequence of SH3 domain and peptide, predicts the structure of the complex.

Introduction:

The SH3 (SRC Homology 3) is a small protein domain of ~60 amino acid residues. These domains are involved in the regulation of important cellular pathways, such as cell proliferation, migration and cytoskeletal modifications and have been identified as protein modules that recognize proline-rich sequences.

The structure of SH3 domain comprises five to eight β-strands arranged into two antiparallel β-sheets or in a β-barrel. This structure conforms a PxxP-binding site, adapted for the recognition of left-handed proline-rich type II, a binding region formed by the RT loop between strands β1 and β2, and the n-Src loop between strands β2 and β3, which is named the specificity site.

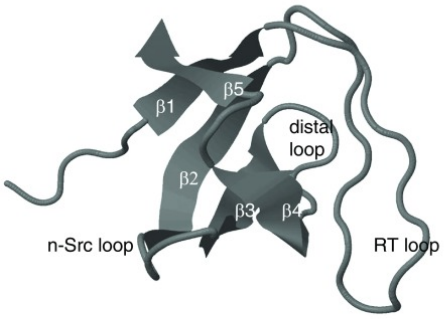

Fig 1. Topology of secondary structure of the Src homology 3 (SH3) domain

SH3 domains are regulated by association with other proteins, phosphorylation, dimerization, and *cis–trans* isomerization. High-affinity peptides that bind SH3 domains are used in drug development as candidates for anticancer treatment.

To summarize, the importance of SH3 in cell biology is demonstrated by its key role as a mediator of signal regulation. Such vital processes take a central role when new treatments rely on an understanding of cellular control mechanisms.
In this project, our goal was to model the structure of the SH3-peptide complex.

Results

We created a Neural Network, SH3-Net, that predicts the structure of SH3-peptide complex.

In order to evaluate our model's performance, we compared its predictions for the test set to AlphaFold2's predictions, based on the RMSD measurement.

For both AlphaFold2 and SH3-Net's predictions, we aligned the SH3 domain in the reference structure to the SH3 domain of the predicted structure. The output transformations were used to transform the total reference structure. Finally, we calculated the RMSD between the transformed reference structure and the predicted structure, for both the SH3 domain and peptide.

As observed in Fig 2., SH3-Net yielded lower variance in the RMSD of the peptides, although the mean RMSD of the AlphaFold2 prediction is lower.

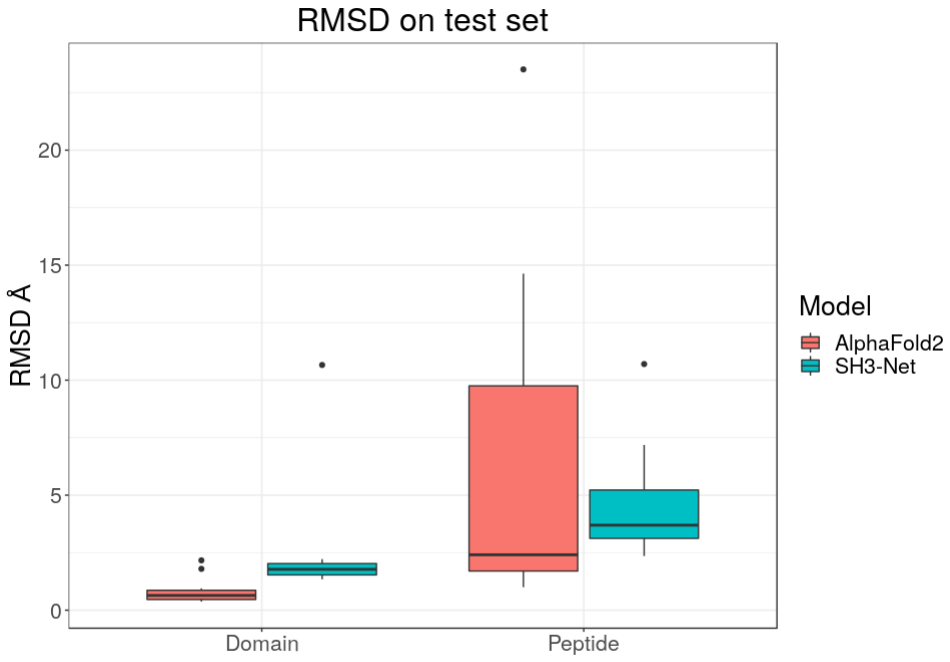

Fig 2. RMSD for both SH3-net and AlphFold2 on test set

In Fig 3, we see that SH3-Net prediction for the peptide was much more accurate than AlphaFold2 prediction.

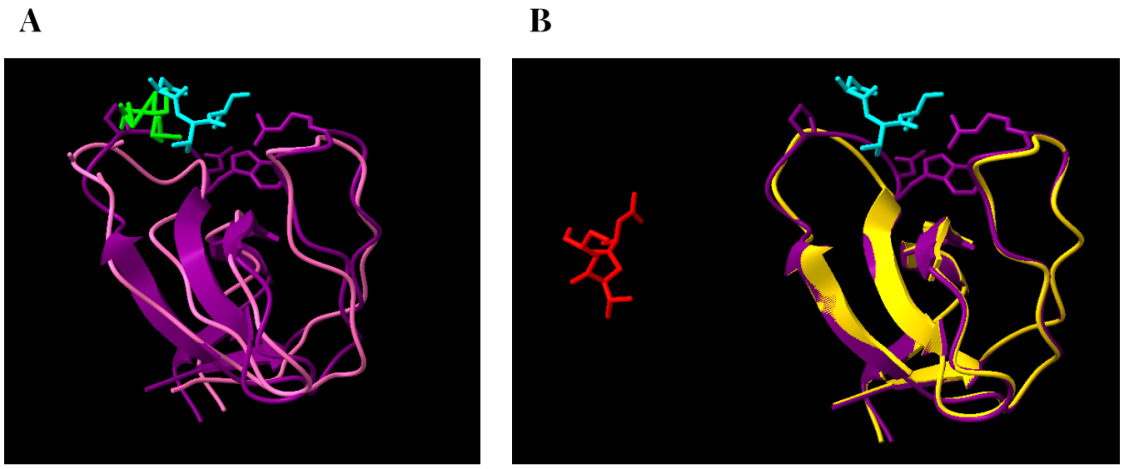

Fig 2. Alignment between predicted structure to the reference structure
(A). Alignment between SH3-Net prediction to the reference structure. Purple - reference sH3 domain, Cyan - reference peptide, Pink - predicted SH3 domain, Green - predicted peptide.
(B). Alignment between AlphaFold2 prediction to the reference structure. Purple - reference sH3 domain, Cyan - reference peptide, Yellow- predicted SH3 domain,Red - predicted peptide.

To further analyze our results, the RMSDs calculated above were used in order to visualize the success of our predictions in comparison to AlphaFold2's prediction, as seen in Fig 3. AlphaFold2's prediction was more accurate than the SH3-Net model's prediction in most cases, but SH3-Net predicted the peptide structure more accurately in 33% of the test set.

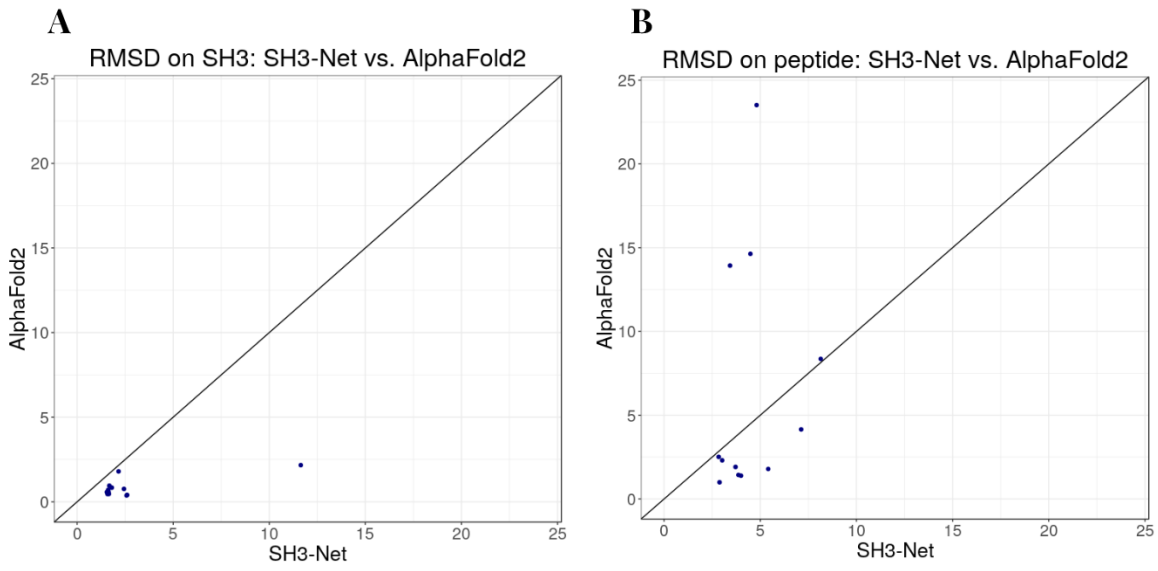

Fig 3. SH3-Net's RMSD vs. AlphaFold2's RMSD on test set
(A). RMSD on SH3, (B). RMSD on peptides

Methods:

Data and processing

Initially, our data consisted of 1,237 PDB files of the SH3 domain.

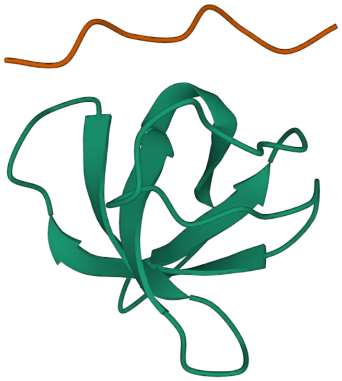

Fig 3. 3D structure of the model complex (1CKA)

We chose the 1CKA structure as our model structure since it contains only SH3 and a peptide and has high resolution. This structure was used as our reference structure throughout the data processing stage.

In order to identify the SH3 domain, we aligned each chain from each structure to the SH3 domain from the reference structure and chose the chain with the best alignment to be the SH3 domain of this structure. The alignment was evaluated by fixed thresholds - the number of atoms that were matched (at least 40 atoms) and to the rmsd (at least 2.0 Å) . During this process, we filtered out the unsuitable structures - structures that didn't contain any chain that fulfilled these requirements.
Next, we filtered the remaining PDBs by finding only the PDB files that contain a peptide (in addition to SH3 domain found in the previous step). In this step, we also found the start and end residues of the SH3 domain and the peptide. In order to do so, we used GeomHash - a geometric hash table class that enables efficient hashing . We used this tool to map all coordinates of Cα atoms of all chains of each of the PDB files. Then for each Cα of the reference we extracted all the IDs of the residues lying within a 5 Å radius, and chose the closest one to each Cα.
Finally, we chose the peptide to be the chain with the most residues that were closest to the Cαs of the reference and found the exact residues ID of that chain. The purpose of this step was to filter out the most appropriate peptide for each PDB even if it is a part of a larger protein, that way we could make noisy PDBs suitable for our neural network.

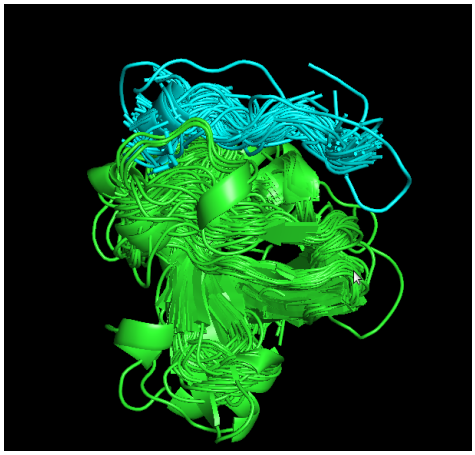

Fig 4. Pymol view of the 146 PDBs
Green- SH3 domain, Cyan - peptide

We used the same program to clear not only the peptide, but also to find the start and end residue ID of the SH3 of each PDB that fits the reference. Finally, we remained with 146 PDBs out of the original 1237.

Training Neural Network

As we learned in class, due to the recent success of the tool AlphaFold2, neural networks have become a common tool for modeling proteins. Therefore we created a SH3-Net , a neural network that receives as input equal length sequences that if necessary were "padded" . We defined the max length of the SH3 to be 130 amino acids and of the peptide to be 30 amino acids. The SH3-net output is a prediction of the structure.

The input is a one-hot encoding matrix of the sequences of both the domain and peptide in this order. Each row of the matrix represents an amino acid in the sequence, and each column represents an amino acid (20 real amino acids, 1 column for unknown amino acid and one column for padding).
In addition we added two columns describing to which protein each amino acid belongs (SH3 domain or peptide). In total, the input matrix is of size (160,24).

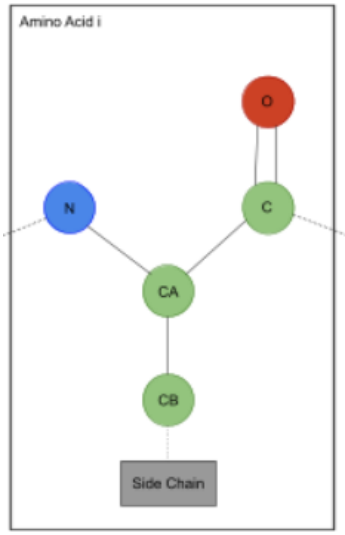

Fig 5. Amino acid backbone and Cβ

The output matrix contains the predicted x,y,x coordinates of the backbone (N,Cα,C,O) and the Cβ of the given complex (see fig 5.). In total, the output matrix is of size (160,15).

Similar to the neural network we built in exercise 4, SH3-Net consists of the following layers: 1D convolution layer, First ResNet Layer, 1D convolution, Second ResNet Layer, Dropout Layer, 1D Convolution Layer, Dense Layer (see fig 6.). For the first resnet block we defined 64 kernels of size 15 and 3 blocks, for the second resnet block we defined 64 kernels of size 5 . For the dropout layer we chose a percentage of 0.2. Our batch size is 16 and our learning rate is 0.001. We did 200 epochs and saved the best model we got from all.

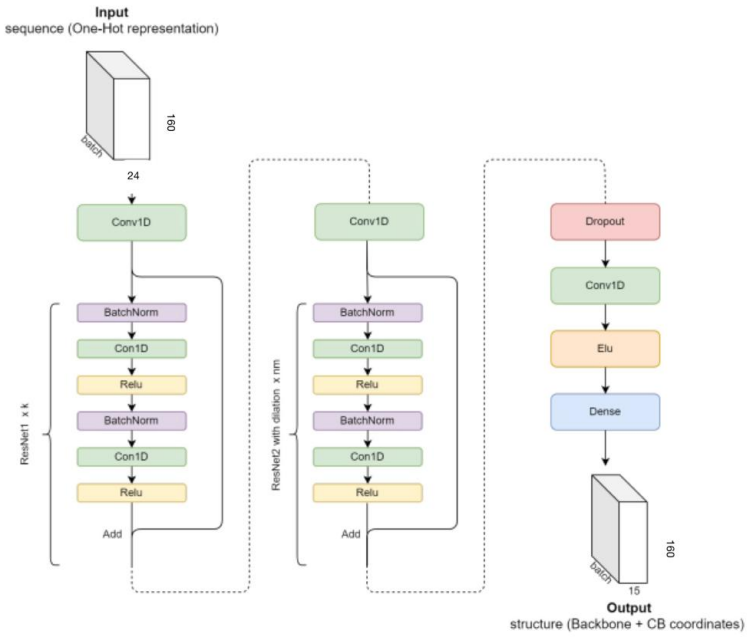

Fig 6. SH3-Net architecture

Discussion:

To conclude, our predictions were not found to be better than AlphaFold2's predictions, but were not very far from them.

In some cases the prediction of the peptide achieved a better RMSD score. As the peptide structure is less known to us (in comparison to the SH3 structure), the prediction of it can be more relevant to us.

We believe our Achilles' heel is the lack of training- data. As mentioned before,due to the processing and filtration of the raw PDBs, we were left with only 146 PDBs. This is obviously a small amount for training a neural network, with more data we believe we could achieve better results that are closer to AlphaFold's results and work towards getting even better results because of the specificity of our model to SH3.

In the process we also created a pipeline to filter out a domain and a peptide from raw PDBs. The same method can also be used with other proteins. We also created a database of clean SH3-peptide PDBs that can be used for future works.

References:

Kurochkina N, Guha U. SH3 domains: modules of protein-protein interactions. *Biophys Rev*. 2013;5(1):29-39. doi:10.1007/s12551-012-0081-z

Teyra J, Huang H, Jain S, Guan X, Dong A, Liu Y, Tempel W, Min J, Tong Y, Kim PM, Bader GD, Sidhu SS. Comprehensive Analysis of the Human SH3 Domain Family Reveals a Wide Variety of Non-canonical Specificities. Structure. 2017 Oct 3;25(10):1598-1610.e3. doi: 10.1016/j.str.2017.07.017. Epub 2017 Sep 7. PMID: 28890361.