Technical Updates

Topic : Large-Scale prediction of Disulphide Bond Connectivity

Name : Prince Gupta

Roll No. : B19BB033

Background :-

Predicting protein residue-residue contacts is an important 2D prediction task. It is useful for *ab initio* structure prediction and understanding protein folding. Several algorithms for reconstructing 3D structure from contacts have been developed in both the structure prediction and determination (NMR) literature. Contact map prediction is also useful for inferring protein folding rates and pathways. In spite of steady progress over the past decade, contact prediction remains still largely unsolved.

We can use a new contact map predictor (SVMcon) that uses support vector machines to predict medium range and long range contacts. SVMcon integrates profiles, secondary structure, relative solvent accessibility, contact potentials, and other useful features. On the same test data set, we can achieve a higher accuracy with SVMcon. SVMcon recently participated in the seventh edition of the Critical Assessment of Techniques for Protein Structure Prediction (CASP7) experiment and was evaluated along with seven other contact map predictors. SVMcon was ranked as one of the top predictors, yielding the second best coverage and accuracy for contacts with sequence separation >= 12 on 13 *de novo* domains.

Update : -

Some of the technical updates that we can do in this paper regarding the large scale prediction of disulphide bonds in protein sequences could be related to the feature selection and using improved classifier.

1. Feature Selection :

Feature selection is useful to improve the performance of machine learning methods, particularly when there is a large number of features as in this

study. However, since there are more than 310,000 training data points, it would take many days to conduct a round of training and testing on a computer. Thus a thorough feature selection is currently not feasible. So we can try to remove only some features (pairwise profile correlation, pairwise mutual information, residue type, and protein information features) once a time to test how they affect the prediction accuracy. We may get some improved accuracy through feature selection. However, we cannot clearly state if the improvement is due to the random variation or due to the removal of the features. But at least, these features were not essential or being compensated by other similar features. Thus, a more thorough feature selection should be conducted to improve the performance when more computing power is available.

2. SVM learning :

For an input feature vector associated with a pair of residues, we can use Support Vector Machines (SVMs) to predict if the two residues are in contact (positive) or not (negative). SVMs provide a non-linear classifier model by non-linearly mapping the input vectors into a feature space and using linear methods for classification in the feature space. Thus SVMs, and more generally kernel methods, attempt to combine the advantages of both linear and nonlinear methods by first embedding the data into a feature space equipped with a dot product and then using linear methods in the feature space to perform classification or regression tasks based on the Gram matrix of dot products between data points. A key property of kernel methods is that the embedding does not need to be given in explicit form, the Gram matrix of dot products $K(x, y) = \varphi(x) \cdot \varphi(y)$ between data points is all is needed to proceed with classification or regression.

We can experiment with several common kernels including linear kernels, Gaussian radial basis kernels (RBF), polynomial kernels, and sigmoidal kernels. Depending on the data here, RBF kernel could provide the best results. Using the RBF kernel, $f(x)$ is actually a weighted sum of Gaussians centred on the support vectors. Almost any separating boundary or regression function can be obtained with such a kernel, thus it is important to tune the SVM parameters carefully in order to achieve good generalization performance and avoid overfitting.

We may adjust the width parameter γ of the RBF kernel, leaving all other parameters to their default value. γ is the inverse of the variance (σ^2) of the

RBF and controls the width of the Gaussian functions centred on the support vectors. The bigger is γ, the more peaked are the Gaussians, and the more complex are the resulting decision boundaries.

Conclusion :-

Here, we have described a new contact map predictor (SVMcon) that uses support vector machines to integrate a large number of useful information including profiles, secondary structure, solvent accessibility, contact potentials, residue types, segment window information, and protein-level information. SVMcon can yield good performance on medium range to long range contact predictions and can be modularly incorporated into a structure prediction pipeline. In the blind CASP7 experiment, SVMcon is ranked as one of the top contact predictors. The method represents an effort towards a good 2D structure prediction. It can be used to improve *ab initio* structure prediction and analogous fold recognition.