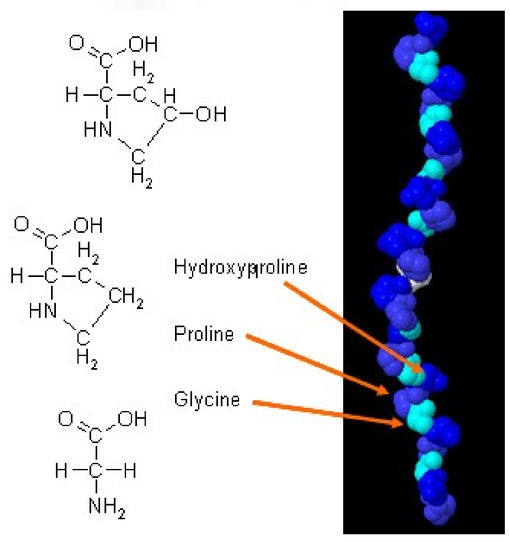
Rule For Predicting Proline Hydroxylation

Chuan Hock Koh (▪泉福) and Limsoon Wong (黄任祥)

BACKGROUND

Proline hydroxylation is an important posttranslational modification as hydroxyproline plays a crucial role in the folding and stabilizing of collagen protein. Collagen, the most abundant protein in higher vertebrates, plays several critical roles such as tissue development and cell structure. Low quality collagen proteins would lead to severe diseases such as scurvy and osteogenesis imperfecta. Therefore, in this paper, we aim to better understand the proline hydroxylation process, in particular, to predict which prolines would be hydroxylated when given a sequence.



SIRIUS PREDICTION SYSTEM BUILDER

We recently developed a software tool, Sirius Prediction System Builder (PSB). Sirius PSB exudes userfriendliness in that it is equipped with a nice Graphical User Interface that will allows anyone with just some basic knowledge in data mining to be able to build a prediction system without any programming involved. Sirius PSB uses a general computational approach of feature generation, feature selection, feature integration and where applicable, the construction of cascade classifiers to build prediction models.

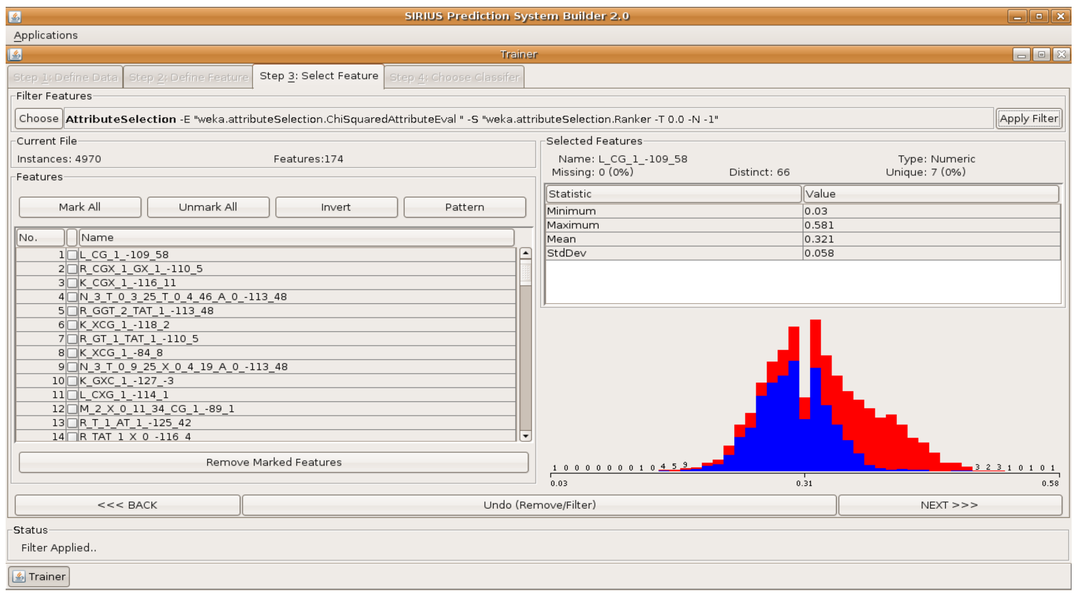


Figure 1. Snapshot of Sirius PSB

RESULTS

Sirius PSB

Table 1. Prediction Performance based on 10-fold cross-validation using Sirius PSB approach

	Equal Error Rate (Closest)			
Flank Size	Sensitivity	Specificity	Threshold	Area Under ROC
6	0.811	0.813	0.329	0.885
8	0.856	0.863	0.208	0.935
10	0.838	0.842	0.253	0.902
12	0.804	0.802	0.274	0.866
14	0.818	0.814	0.219	0.892
16	0.801	0.798	0.310	0.869

Positional Weight Matrix

Table 2. Prediction Performance based on 10-fold cross-validation using PWM approach

	Equal Error Rate (Closest)			
Flank Size	Sensitivity	Specificity	Threshold	Area Under ROC
6	0.710	0.707	0.285	0.807
8	0.727	0.734	0.276	0.810
10	0.746	0.742	0.255	0.829
12	0.750	0.747	0.266	0.829
14	0.765	0.761	0.244	0.842
16	0.778	0.781	0.263	0.840

DISCUSSION

Using Sirius PSB, we generated 1gram, 2gram and a series of features based on biochemistry properties. Using chi-square to select a subset of features that are statistically useful, coupled with support vector machine as our learning algorithm, we were able to achieve an Area Under ROC Curve (AUC) of 0.935 (10-fold cross-validation) from a set of 295 extracellular protein sequences downloaded from public databases that are experimentally known to contain hydroxyproline. As a comparison, conventional methods such as positional weight matrix (PWM) with all positions equally weighted achieved an AUC of only 0.842 (10-fold cross-validation).

CONCLUSION

Both PWM and Sirius PSB approaches suggest that having a glycine at every 3rd position indicates a higher probability of hydroxylation of that particular proline. Incorporating that information into PWM allows it to achieve a higher AUC of 0.866 (10-fold cross-validation). Therefore, we concluded that in extracellular protein sequences, there exists the unique characteristic of X and Y residues in GlyXY, with X occurring as any amino acid, and Y as hydroxyproline. To achieve higher confidence on our results, we are currently working on increasing our dataset size.

REFERENCES

[1] Simone, A.D., Vitagliano, L. and Berisio, R. (2008). Role of hydration in collagen triple helix stabilization. Biochemical and Biophysical Research Communications, Vol.372, May 2008, pp. 121-125.

[2] Koh, C.H. and Wong, L. (2007). Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences. In Proceedings of 18th International Conference on Genome Informatics, (Singapore, December 3-5, 2007), pp. 73-82.

[3] Vitagliano, L., Berisio, R., Mazzarella, L. and Zagari, A. (2001). Structural Bases of Collagen Stabilization Induced by Proline Hydroxylation. Biopolymers, Vol.58, 2001, pp. 459-464.

[4] Witten, I.H. and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann, San Francisco, 2005.