173. Finding Signal Peptides in Human Protein Sequences Using Recurrent Neural Networks

Petko Fiziev [1 2 3] , **Artemis Hatzigeorgiou** [1 3 4] ,
Eike Staub [3], **Martin Reczko** [4 5]

Keywords: signal peptides , recurrent neural networks

1 Introduction

A new approach called Sigfind for the prediction of signal peptides in human protein sequences is introduced. The method is based on a novel neural network architecture (BRNN) that was first presented by [Baldi et al., 2000]. The BRNN architecture is used for sequential learning problems with sequences of finite lengths. Other recurent NNs can only model causal dynamical system, where the output of the system at a certain time does not depend on future inputs.The BRNN model has a symmetric memory for storing influences of past and future inputs to an output at time t. The modifications to this architecture and a better learning algorithm result in a very accurate identification of signal peptides (99.5% correct in fivefold crossvalidation) [Reczko et al., 2002].

The output $Y_t \in \mathbb{R}^s$ is calculated after the calculation of F_t and B_t has finished using

$$Y_t = \eta(F_t, B_t, U_t) \tag{1}$$

In our modified BRNN, the state units are not connected directly to the output, but to the hidden layer. It is reasonable to assume that several activation patterns occurring on the combination of the forward and backward state units are not linearly separable and thus cannot be distinguished without the additional transformation performed by this hidden layer. There is an implicit asymmetry in the amount of context information accumulated in the state vectors. At the start of the sequence ($t = 1$), the forward state vector F_1 contains information about only one input vector U_1 whereas the backward state vector B_1 has accumulated all information of the complete sequence. At the end of the sequence ($t = T$), the opposite situation occurs. The processing of the state vectors with a hidden layer can help detecting these asymmetric situations and process the more relevant activations.

2 Software and files

The Sigfind system is avaliable on the WWW for predictions (http://www.stepc.gr/ synaptic/sigfind.html).

[1]Center for Bioinformatics, Department of Genetics, University of Pennsylvania, School of Medicine, Philadelphia, PA 19104-6145, USA
[2]Free University Berlin, Kaiserswerther Str. 16-18, 14195 Berlin,Germany
[3]metaGen Pharmaceuticals GmbH, Oudenader Str.16 D-13347 Berlin, Germany
[4]Synaptic Ltd., Science and Technology Park of Crete, P.O. Box 1447, 711 10 Voutes Heraklion, Greece
[5]email: reczko@web.de

3 Figures and tables

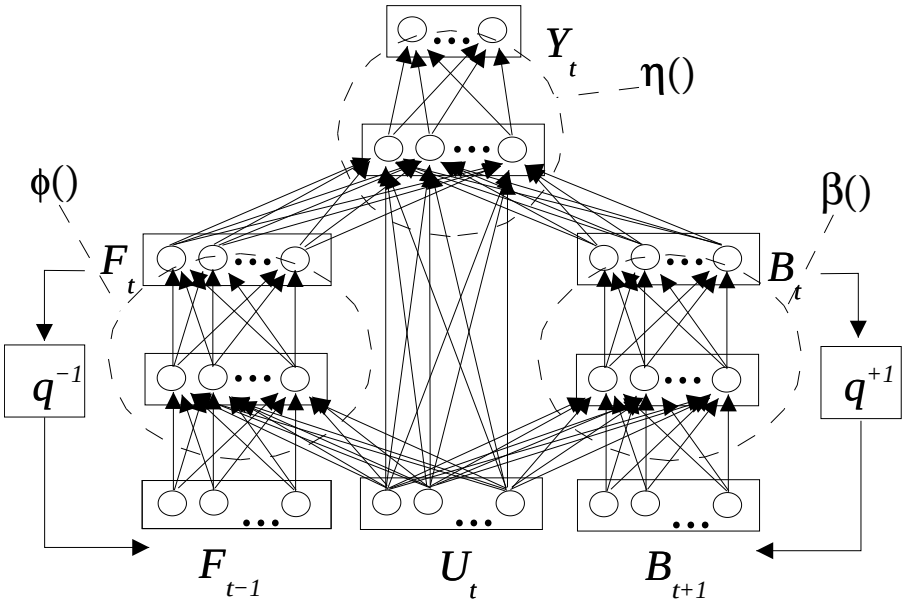

Figure 1: The modified bidirectional recurrent neural network architecture.

References

[Baldi et al., 2000] Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., and Soda, G. (2000). Bidirectional dynamics for protein secondary structure prediction. In Sun, R. and Giles, L., editors, *Sequence Learning: Paradigms, Algorithms, and Applications*. Springer Verlag.

[Reczko et al., 2002] Reczko, M., Fiziev, P., Staub, E., and Hatzigeorgiou, A. (2002). Finding signal peptides in human protein sequences using recurrent neural networks. In *Algorithms in Bioinformatics WABI 2002*, pages 60–67. Springer.