

Predicting protein localization using recurrent neural networks

Background

The subcellular translocation of proteins usually relies on a short N-terminal targeting peptide (sequence of amino acid residues). The peptide directs the protein to its appropriate subcellular location. After delivery the targeting peptide is cleaved from the mature protein and is presumed to be of no further use and digested. Some properties that correlate with the subcellular sorting are known, but statistical indicators are very weak. As a group of biological sequences, targeting peptides evolve quickly and exhibit great diversity.

Foreground

A series of feed forward neural network (FNN) based predictors have shown potential in predicting localization and cleavage sites of various proteins. **TargetP** distinguishes between proteins destined for mitochondria [mTP], for chloroplast [cTP], for the secretory pathway [SP], and other (Emanuelsson & von Heijne, 2001).

A neural network is trained from example data to find a solution (which is later evaluated by presenting novel data). In contrast to conventional statistical tools, the network architecture imposes a bias (or constraint) on the search for the solution. The vast number of possible combinations of amino acids that may signal transportation and the relatively few examples available, require us to be careful in selecting an appropriate architecture.

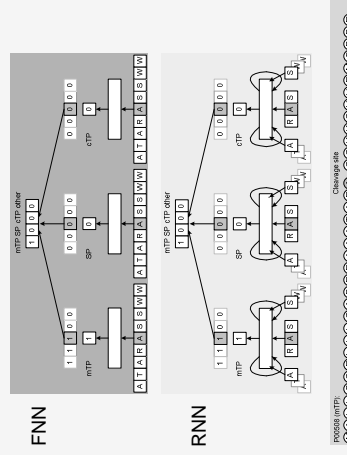
We explore (bi-directional) recurrent neural networks (RNN) and their ability to help in predicting localization of proteins. We use the same data, learning task and evaluation methods as TargetP to objectively assess the usefulness of a range of recurrent neural networks. The recurrent neural networks are used to spatially scan and detect target sequences. By recursively creating an upstream and downstream sequence state from the residues next to each position in the sequence, the middle residue is classified as being part of the target sequence or not (cf. Baldi et al., 1999). The detection output is then fed through a feed forward neural network which identifies the destination of the protein.

The recurrent state can incorporate variable ranges of dependencies between events located both upstream and downstream of sequences. The recursion reduces the number of weights in a network, and consequently reduces the risk of overfitting to data. But how well does it work in practice? See results...

References

- Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999). Exploiting the Past and the Future in Protein Secondary Structure Prediction. *Bioinformatics*, 15, 937-946.
- Bodén, M. Using evolutionary noise to improve prediction of rapidly evolving targeting peptides, submitted to CEC 2003.
- Emanuelsson, O., & von Heijne, G. (2001). Prediction of organellar targeting signals. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1541(1-2), 114-119.

Neural network architectures



Results

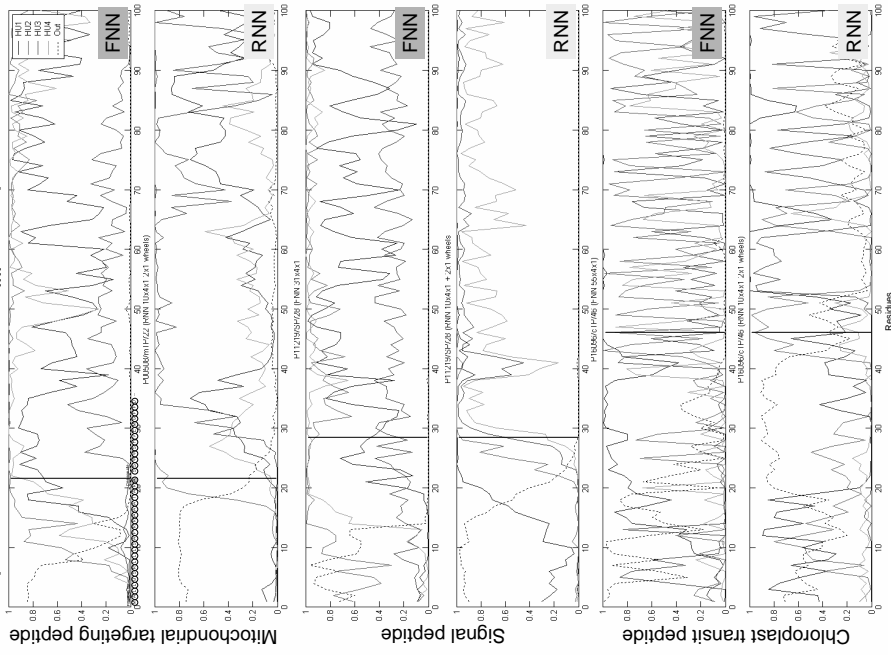
Version	Network	Class:	other	mTP	SP	cTP	Correct	All
Plant	FNN 1 (TargetP replica)	other	139	10	3	10	0.858	
		mTP	32	293	5	38	0.796	
		SP	18	6	245	0	0.911	
		cTP	13	30	0	98	0.695	0.824
	RNN 1 (as above + 1-residue wheels)	other	145	7	2	8	0.895	
		mTP	46	290	3	29	0.788	
		SP	19	5	244	1	0.907	
		cTP	15	28	0	98	0.695	0.827
	RNN 2 (10 residue window, 4 hidden, 1-residue wheels)	other	139	9	3	11	0.858	
		mTP	25	307	7	29	0.834	
		SP	14	7	248	0	0.922	
		cTP	13	25	2	101	0.716	0.846
RNN 3 (-10 residue window, 8 hidden, 2-residue wheels)	other	138	7	8	9	0.852		
	mTP	14	326	8	20	0.886		
	SP	11	6	250	2	0.929		
	cTP	9	19	0	113	0.901	0.880	
Non-plant FNN 1 (TargetP replica)	other	1513	105	34	916	0.916		
	mTP	17	206	5	9	0.865		
	SP	26	10	679	0	0.880	0.910	
	other	1508	102	42	913	0.913		
RNN 1 (15 residue window, 10 hidden, 1-residue wheels)	other	36	327	8	881	0.881		
	mTP	32	12	671	0	0.938		
	SP							

Conclusions

Recurrent neural networks demonstrate advantages compared with feed forward neural networks for predicting protein subcellular localization. Specifically, for plant data, RNNs consistently perform better overall. For non-plant data, the two approaches are equally successful. However, the performance profile is changed. Mitochondrial targeting and chloroplast transit peptides are predicted with higher precision using RNN whereas signal peptides are better handled using FNNs. On a related note, we have used "mutated" training samples to improve the prediction of the rapidly evolving signal peptides (Bodén, submitted).

Mikael Bodén (mikael@itee.uq.edu.au)

Examples of network internal states ||| and outputs



Look for...

State changes indicate what changes in the input the network is sensitive to. FNN output fluctuates a lot. Why? Each output is a result of single window residues. Two neighbouring predictions are independent. RNN output and state variables exhibit slack. Why? The RNN state is partially determined from the flanking predictions (an upstream and a downstream trace).