

Using recurrent neural networks to predict subcellular localisation

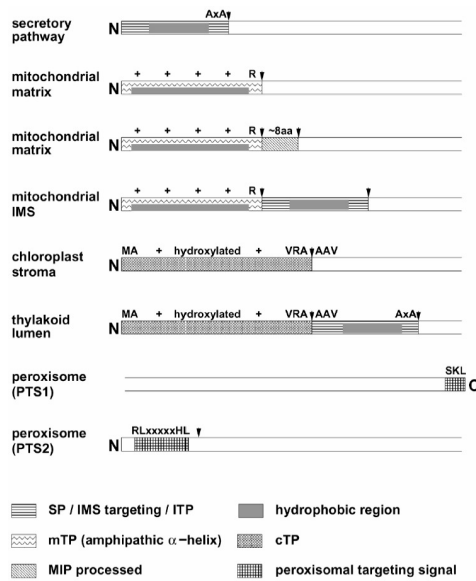
- In eukaryotic cells, membrane-bound organellar compartments hold specific protein and lipid content (functional integrity)
- Proteins are translocated from ribosomes to specific compartments using complex, dynamic processes (difficult to model)
- Predicting localisation for nascent proteins
 - helps determining their function, and possible interactions
 - provides knowledge of sorting signals to be used for drug design etc
 - BUT is difficult since sorting signals exhibit little sequence similarity

Using recurrent neural networks to predict subcellular localisation

- Conventional *feed forward* neural networks are used successfully for biological sequence analysis, but
 - process sequential data statically through a bounded input window, and
 - a large window creates a large input space (which may hide relevant data dependencies)
- *Recurrent* neural networks
 - process sequential data dynamically (using *states*)
 - are biased towards finding sequential patterns

Targeting sequences

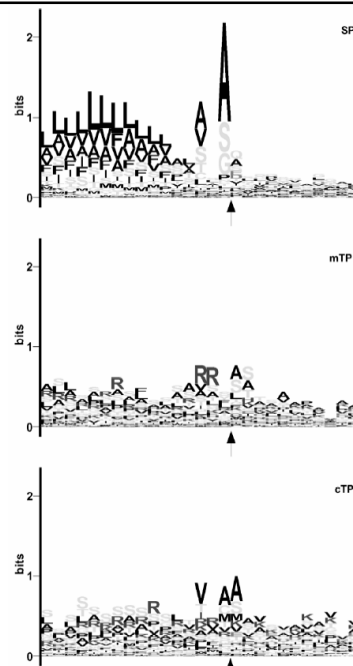
- Signal peptides (SP)
 - Target proteins to the ER, the secretory pathway
 - 15-30 residues long.
 - N-terminal domain positively charged, central hydrophobic region, small apolar residues -3, -1 relative cleavage site.
- Mitochondrial targeting peptides (mTP)
 - Target proteins to mitochondria (outer and/or inner membrane)
 - 20-40 residues long.
 - Arg, Ala and Ser overrepresented, while negative residues scarce.
 - Weak conservation around cleavage site(s).
- Chloroplast transit peptides (cTP)
 - Target proteins to chloroplast.
 - 20-120 residues long.
 - Hydroxylated residues common, acidic uncommon.
 - Semi-conserved motif around cleavage site.



From Emanuelsson, *Brief. Bioinformatics* 3(4) 2002.

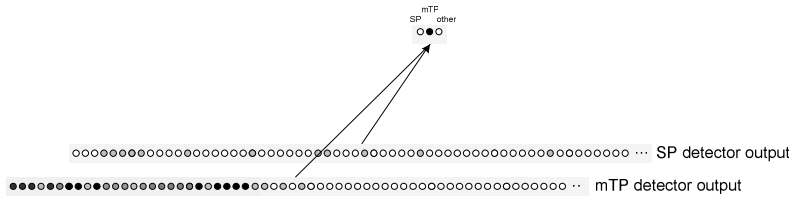
Targeting sequences

- Signal peptides (SP)
 - Target proteins to the ER, the secretory pathway
 - 15-30 residues long.
 - N-terminal domain positively charged, central hydrophobic region, small apolar residues -3, -1 relative cleavage site.
- Mitochondrial targeting peptides (mTP)
 - Target proteins to mitochondria (outer and/or inner membrane)
 - 20-40 residues long.
 - Arg, Ala and Ser overrepresented, while negative residues scarce.
 - Weak conservation around cleavage site(s).
- Chloroplast transit peptides (cTP)
 - Target proteins to chloroplast.
 - 20-120 residues long.
 - Hydroxylated residues common, acidic uncommon.
 - Semi-conserved motif around cleavage site.

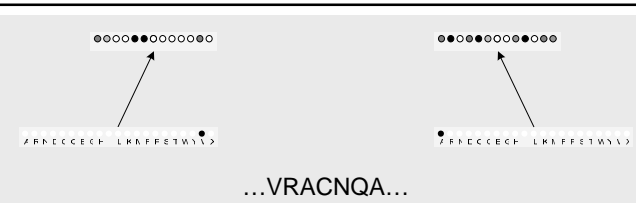
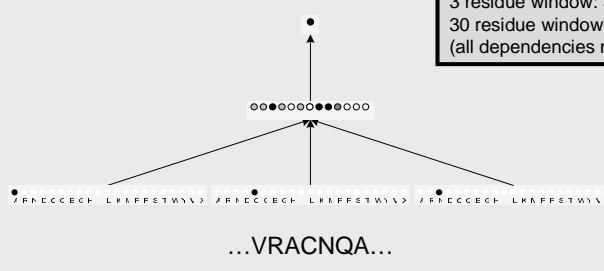


From Emanuelsson, *Brief. Bioinformatics* 3(4) 2002.

TargetP (cf. SignalP)



3 residue window: $3 \cdot 21 = 63$ bits
 30 residue window: $30 \cdot 21 = 630$ bits
 (all dependencies must be within)

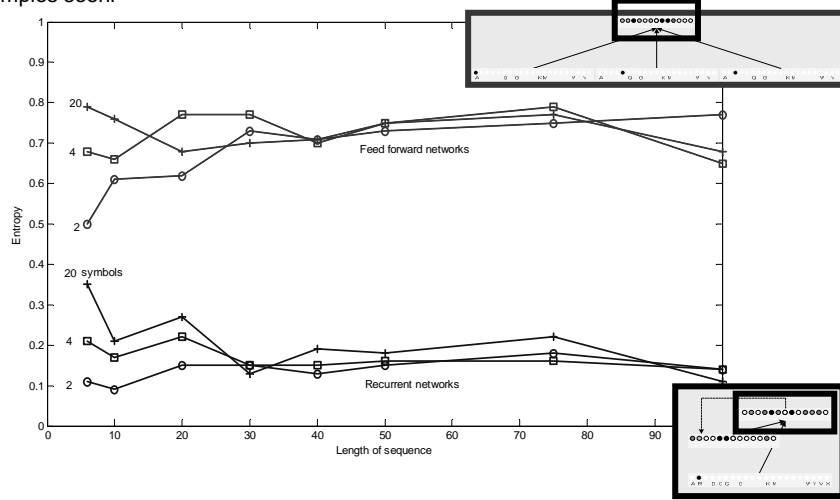


1 residue wheels: $(1+1) \cdot 21 = 63$ bits
 10 residue wheels: $(10+10+1) \cdot 21 = 441$ bits
 (dependencies can range outside)

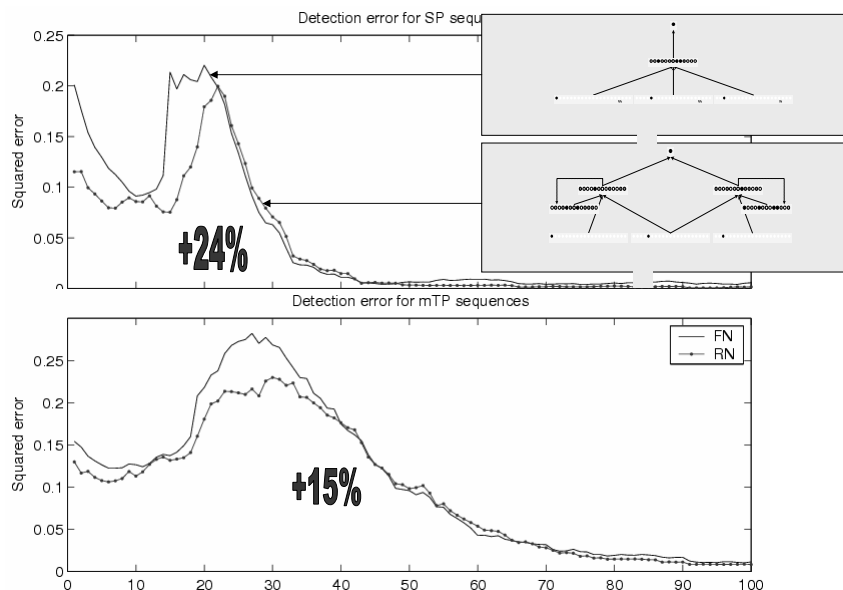


Separation in state space

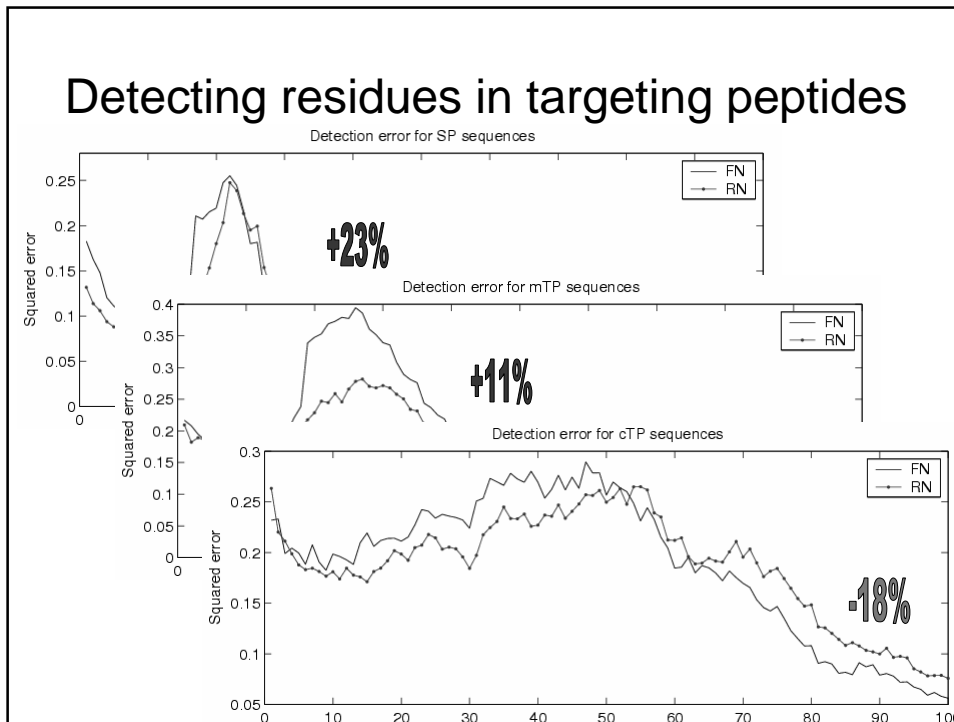
between sequences with and without a randomly chosen motif (1/3 of sequence length) before training - no examples seen!



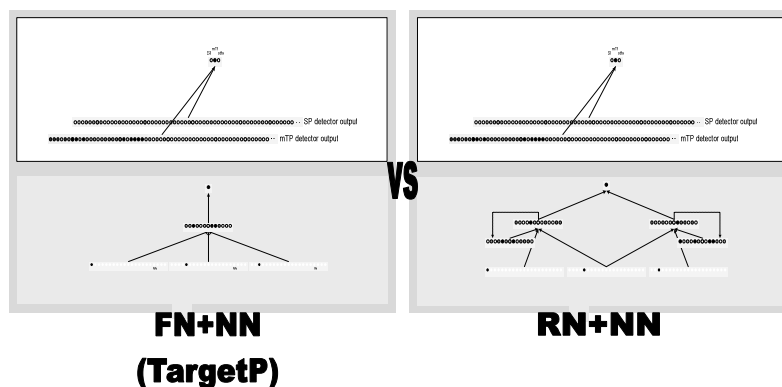
Detecting residues in targeting peptides



Detecting residues in targeting peptides



Sorting the detection signals



Accuracy

	Target	Sensitivity	Specificity	MCC
TargetP	SP	0.960	0.920	0.920
	mTP	0.890	0.670	0.730
	Other	0.880	0.970	0.820
	Mean	0.910	0.853	0.823
FN+	SP	0.936 (0.005)	0.943 (0.009)	0.918 (0.006)
KNN	mTP	0.813 (0.010)	0.856 (0.017)	0.809 (0.015)
sort	Other	0.952 (0.003)	0.939 (0.004)	0.861 (0.007)
	Mean	0.900	0.912	0.862
FN+	SP	0.945 (0.005)	0.938 (0.005)	0.921 (0.006)
NN	mTP	0.883 (0.010)	0.768 (0.013)	0.794 (0.012)
sort	Other	0.923 (0.004)	0.959 (0.002)	0.855 (0.007)
	Mean	0.917	0.888	0.856
RN+	SP	0.946 (0.007)	0.922 (0.008)	0.911 (0.005)
KNN	mTP	0.818 (0.015)	0.837 (0.019)	0.801 (0.011)
sort	Other	0.936 (0.006)	0.941 (0.003)	0.846 (0.008)
	Mean	0.900	0.900	0.852
RN+	SP	0.948 (0.006)	0.923 (0.009)	0.912 (0.004)
NN	mTP	0.892 (0.009)	0.768 (0.018)	0.798 (0.009)
sort	Other	0.914 (0.003)	0.961 (0.003)	0.849 (0.006)
	Mean	0.918	0.884	0.853
Combined	SP	0.954	0.936	0.925
	mTP	0.865	0.834	0.825
	Other	0.938	0.954	0.866
	Mean	0.919	0.908	0.872

+6%

	Target	Sensitivity	Specificity	MCC
TargetP	SP	0.910	0.950	0.900
	mTP	0.820	0.900	0.770
	cTP	0.850	0.690	0.720
	Other	0.850	0.780	0.770
	Mean	0.858	0.830	0.790
FN+	SP	0.924 (0.016)	0.957 (0.013)	0.917 (0.013)
KNN	mTP	0.849 (0.016)	0.905 (0.018)	0.801 (0.027)
sort	cTP	0.796 (0.020)	0.783 (0.047)	0.752 (0.035)
	Other	0.862 (0.022)	0.731 (0.006)	0.747 (0.012)
	Mean	0.858	0.844	0.804
FN+	SP	0.926 (0.006)	0.966 (0.010)	0.925 (0.009)
NN	mTP	0.857 (0.025)	0.915 (0.018)	0.816 (0.030)
sort	cTP	0.812 (0.038)	0.772 (0.028)	0.754 (0.026)
	Other	0.877 (0.017)	0.750 (0.011)	0.768 (0.015)
	Mean	0.868	0.851	0.815
RN+	SP	0.943 (0.008)	0.965 (0.007)	0.936 (0.003)
KNN	mTP	0.876 (0.014)	0.899 (0.020)	0.816 (0.020)
sort	cTP	0.779 (0.031)	0.801 (0.041)	0.753 (0.031)
	Other	0.863 (0.042)	0.774 (0.023)	0.776 (0.020)
	Mean	0.865	0.860	0.820
RN+	SP	0.938 (0.009)	0.975 (0.005)	0.939 (0.007)
NN	mTP	0.857 (0.011)	0.925 (0.012)	0.825 (0.013)
sort	cTP	0.811 (0.022)	0.772 (0.019)	0.753 (0.019)
	Other	0.892 (0.022)	0.752 (0.028)	0.778 (0.027)
	Mean	0.874	0.856	0.824
Combined	SP	0.944	0.985	0.950
	mTP	0.870	0.917	0.827
	cTP	0.800	0.789	0.758
	Other	0.907	0.774	0.801
	Mean	0.880	0.866	0.834

+5%

Conclusions

- Combined model achieves
 - sensitivity 0.919 and specificity 0.908 on non-plant proteins,
 - sensitivity 0.880 and specificity 0.866 on plant proteins
- Recurrent networks demonstrate ability to
 - make accessible, and
 - find sequence patterns
- TargetP sequence-target mapping is not optimal: sorting system fails to fully exploit detection improvement offered by recurrent networks
- The Protein Prowler server, incorporating recurrent networks for subcellular localisation prediction, is under development

Simulation details

- All models are subject to five-fold cross-validation, each repeated six times
- Data consists of 2738 non-plant and 940 plant proteins (same as TargetP), sequences presented with a uniform distribution over classes
- Detection networks
 - window sizes as TargetP for FNs, 10+10 for all RNs, 4 hidden nodes (as TargetP),
 - use binomial outputs, and
 - are trained to optimise cross-entropy (learning rate 0.01, 30,000 sequences)
- Sorting networks
 - no hidden nodes (as TargetP),
 - use multinomial outputs (softmax), and
 - are trained to optimise the log-likelihood (learning rate 0.01, 30,000 sequences)
 - K=3 for k-nearest neighbour