

Convolutional LSTM Networks for Subcellular Localization of Proteins

Søren Kaae Sønderby
Casper Kaae Sønderby
Henrik Nielsen
Ole Winther

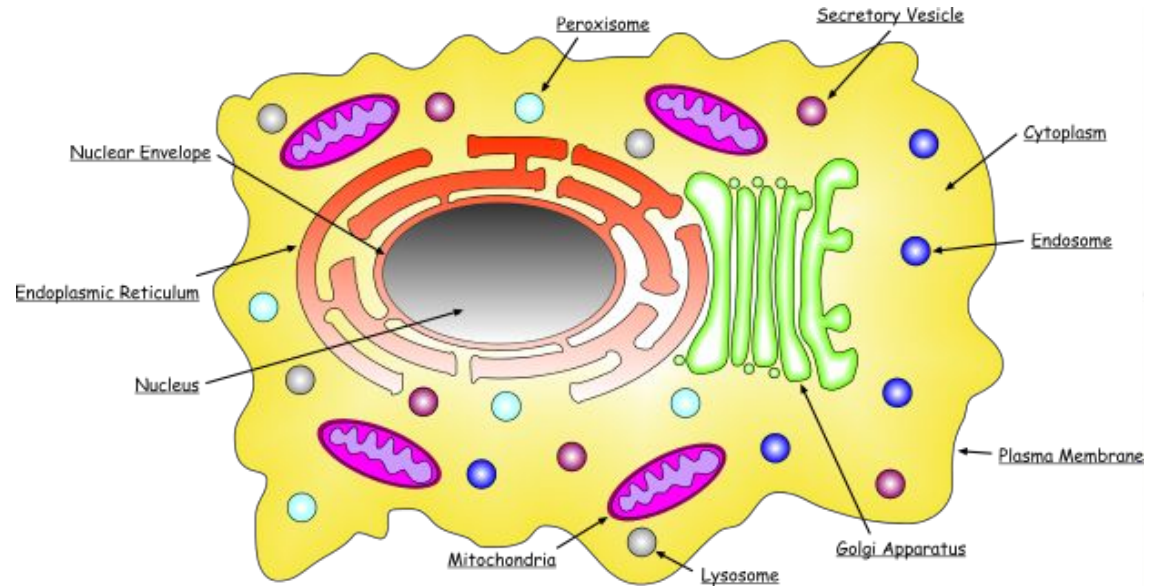
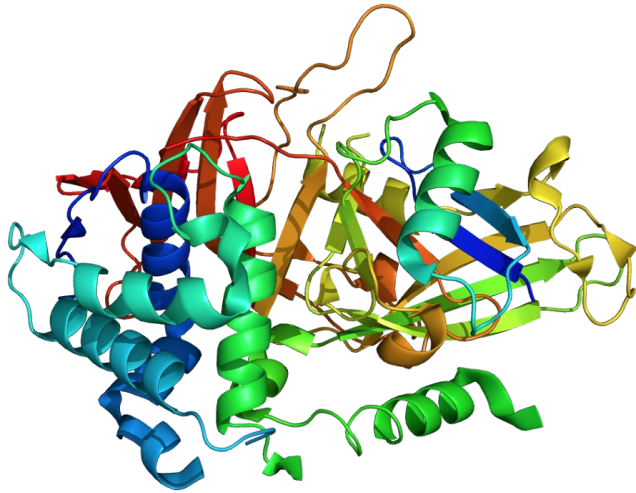
2015

2016.11.09

DeepRegret

Salil Bhate
Bosh Liu
Scott Longwell
Tyler Shimko
Daniel Thirman
Shashwat Udit

Subcellular Localization



Goal












Protein Sequence

```
>ExTopoDBID:10[Uniprot_AC:O34653]  
MTEQTIAHKQKQLTKQVAFAQFETKNSLIQLLNTFIPFFGLWFLAYLSLDVSYLLTLAL  
TVIAAGFLTRIFIIFHDCCHQSFFKQKRYNHILGFLTGVLTLPYQLQWQSHSIHHATSS  
NLDKRGTDGIWMLTVNEYKAASRRTKLAYRLYRNPFIIMLGPIYVFLITNRFNKKGARR  
KERVNTYLTNLAIVALAAACCLIFGWQSFLLVQGPFIPLISGSIGVWLFYVQHTFEDSYFE  
ADENWSYVQAAVEGSSFYKLPKLLQWLTGNIGYHHVHLSPKVPNYKLEVAHEHHEPLEKN  
VPTITLKTSLQSLAFRLWDEDNKQFVSFRAIKHIPVSLPPDSPEKQKLRKNA
```

predict



Subcellular Localization

-  ER
-  Golgi
-  Extracellular
-  Lysosomal
-  Plasma membrane
-  Vacuolar
-  Chloroplast
-  Mitochondrial
-  Cytoplasmic
-  Nuclear
-  Peroxisomal

Motivation

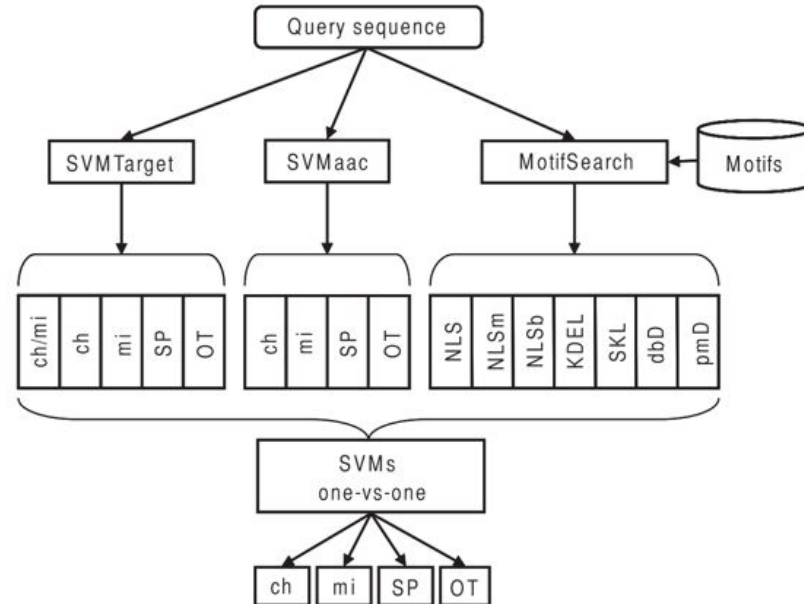
Predicting subcellular localization can:

- Imply protein function
- Suggest interacting proteins
- Identify therapeutic and diagnostic targets
- Help understand disease
- Inform protein engineering

Existing Algorithms

MultiLoc2, SherLoc2 (2009)

- SVM-based
- Several classifiers:
 - Sequence
 - Phylogenetic profile
 - Gene-Ontology (GO)
 - PubMed Abstracts
- *Lots of feature engineering*



Claims

1. We show that LSTM networks combined with convolutions are efficient for predicting subcellular localization of proteins from sequence.
2. We show that convolutional filters can be used for amino acid sequence analysis and introduce a visualization technique.
3. We investigate an attention mechanism that lets us visualize where the LSTM network focuses.
4. We show that the LSTM network effectively extracts a fixed length representation of variable length proteins.

Claims

1. We show that LSTM networks combined with convolutions are efficient for predicting subcellular localization of proteins from sequence.

LSTMs predict subcellular localization

2. We show that convolutional filters can be used for amino acid sequence analysis and introduce a visualization technique.

Convolutional filters provide interpretable sequence motifs

3. We investigate an attention mechanism that lets us visualize where the LSTM network focuses.

A-LSTMs highlight key parts of protein sequences

4. We show that the LSTM network effectively extracts a fixed length representation of variable length proteins.

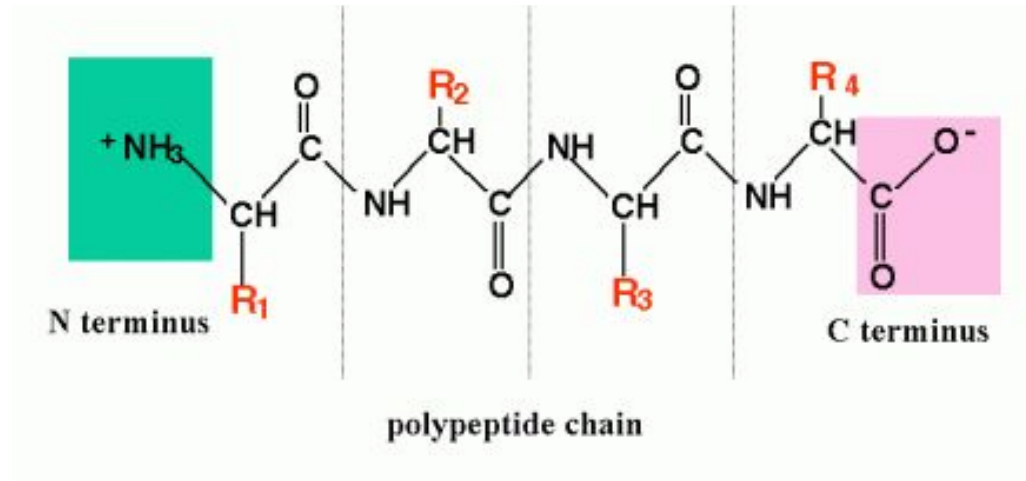
LSTMs generate fixed-length protein representations

Data

- **X**: 5959 protein sequences
- **Y**: 11 subcellular locations
- 80-20 Training-Test split

Data

- Truncated/padded to 1,000 AA residues
- Keep N- and C-terminus



Data

Each AA encoded by:

- 1-of-K encoding
- BLOSUM80
- HSDM
- Sequence profiles from ProfilPro

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7	-3	-3	-3	-1	-2	-2	0	-3	-3	-3	-1	-2	-4	-1	2	0	-5	-4	-1
R	-3	9	-1	-3	-6	1	-1	-4	0	-5	-4	3	-3	-5	-3	-2	-2	-5	-4	-4
N	-3	-1	9	2	-5	0	-1	-1	1	-6	-6	0	-4	-6	-4	1	0	-7	-4	-5
D	-3	-3	2	10	-7	-1	2	-3	-2	-7	-7	-2	-6	-6	-3	-1	-2	-8	-6	-6
C	-1	-6	-5	-7	13	-5	-7	-6	-7	-2	-3	-6	-3	-4	-6	-2	-2	-5	-5	-2
Q	-2	1	0	-1	-5	9	3	-4	1	-5	-4	2	-1	-5	-3	-1	-1	-4	-3	-4
E	-2	-1	-1	2	-7	3	8	-4	0	-6	-6	1	-4	-6	-2	-1	-2	-6	-5	-4
G	0	-4	-1	-3	-6	-4	-4	9	-4	-7	-7	-3	-5	-6	-5	-1	-3	-6	-6	-6
H	-3	0	1	-2	-7	1	0	-4	12	-6	-5	-1	-4	-2	-4	-2	-3	-4	3	-5
I	-3	-5	-6	-7	-2	-5	-6	-7	-6	7	2	-5	2	-1	-5	-4	-2	-5	-3	4
L	-3	-4	-6	-7	-3	-4	-6	-7	-5	2	6	-4	3	0	-5	-4	-3	-4	-2	1
K	-1	3	0	-2	-6	2	1	-3	-1	-5	-4	8	-3	-5	-2	-1	-1	-6	-4	-4
M	-2	-3	-4	-6	-3	-1	-4	-5	-4	2	3	-3	9	0	-4	-3	-1	-3	-3	1
F	-4	-5	-6	-6	-4	-5	-6	-6	-2	-1	0	-5	0	10	-6	-4	-4	0	4	-2
P	-1	-3	-4	-3	-6	-3	-2	-5	-4	-5	-5	-2	-4	-6	12	-2	-3	-7	-6	-4
S	2	-2	1	-1	-2	-1	-1	-1	-2	-4	-4	-1	-3	-4	-2	7	2	-6	-3	-3
T	0	-2	0	-2	-2	-1	-2	-3	-3	-2	-3	-1	-1	-4	-3	2	8	-5	-3	0
W	-5	-5	-7	-8	-5	-4	-6	-6	-4	-5	-4	-6	-3	0	-7	-6	-5	16	3	-5
Y	-4	-4	-4	-6	-5	-3	-5	-6	3	-3	-2	-4	-3	4	-6	-3	-3	3	11	-3
V	-1	-4	-5	-6	-2	-4	-4	-6	-5	4	1	-4	1	-2	-4	-3	0	-5	-3	7

1 of K

20

BLOSUM80

20

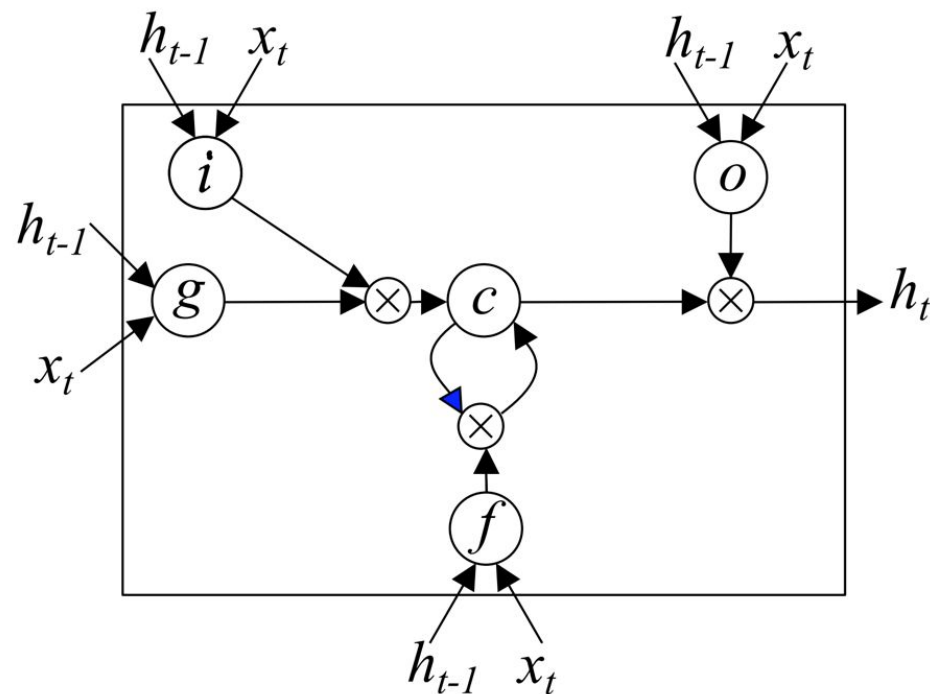
HSDM

20

ProfilPro

20

LSTM cell



$$i_t = \sigma(D(x_t)W_{xi} + h_{t-1}W_{hi} + b_i)$$

$$f_t = \sigma(D(x_t)W_{xf} + h_{t-1}W_{hf} + b_f)$$

$$g_t = \tanh(D(x_t)W_{xg} + h_{t-1}W_{hg} + b_g)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$o_t = \sigma(D(x_t)W_{xo} + h_{t-1}W_{ho} + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

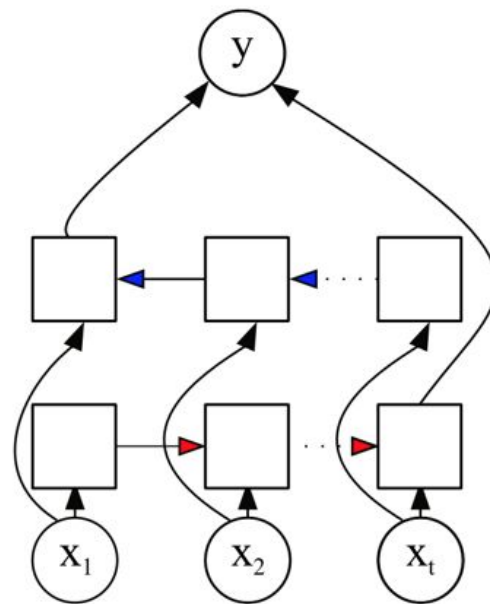
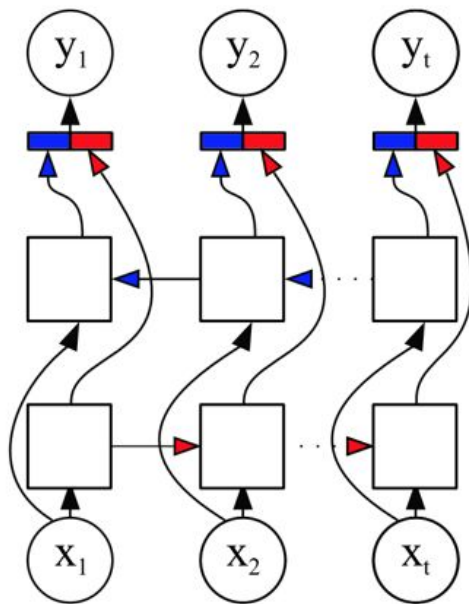
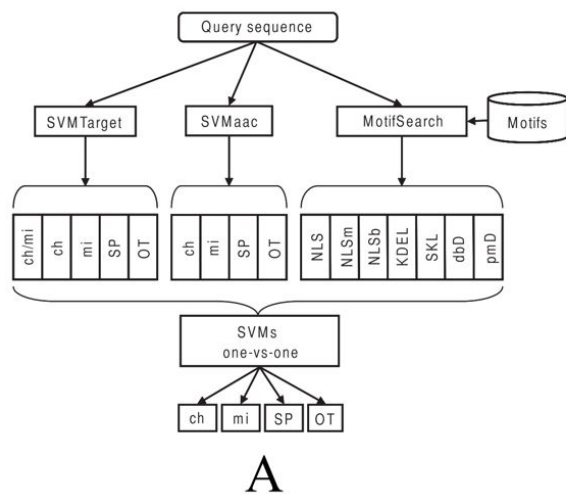
$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

\odot : Elementwise multiplication

D : Dropout, set values to zero with probability p

x_t : input from the previous layer: h_t^{l-1}

Regular LSTM Networks for Predicting Single Targets

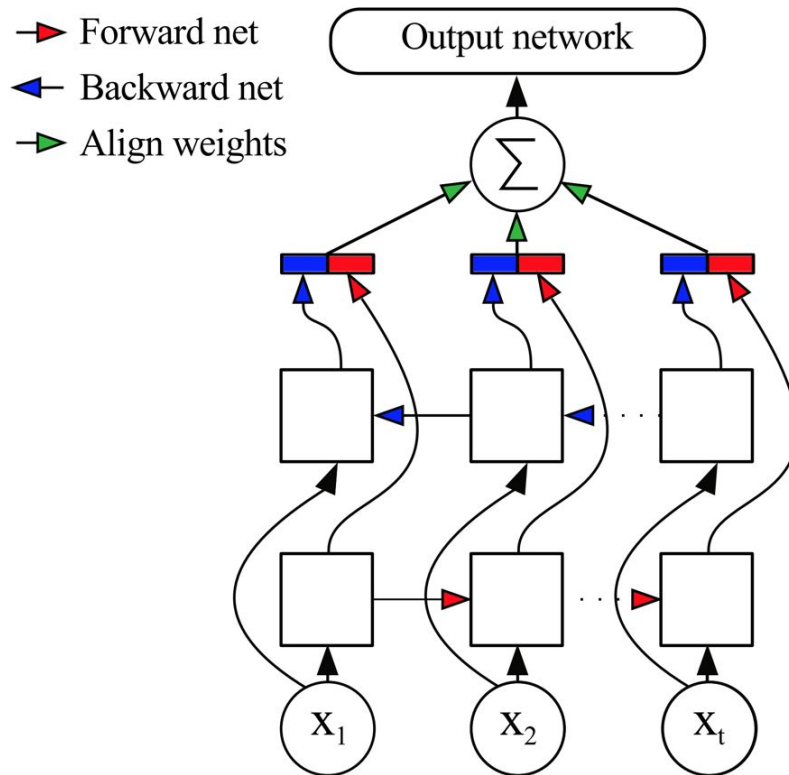


Attention Mechanism LSTM

$$a_t = \tanh(h_t W_a) v_a^T$$

$$\alpha_t = \frac{\exp(a_t)}{\sum_{t'=1}^T \exp(a_{t'})}$$

$$c = \sum_{t=1}^T h_t \alpha_t$$



Model Implementation

- 1D Convolutional layer between inputs and LSTM
- ADAM optimizer
- 50% dropout
- Convolution: ReLU
- LSTM: tanh
- 100-epochs
- 3 models:
 - A-LSTM
 - R-LSTM
 - 10-ensemble of R-LSTM

Results

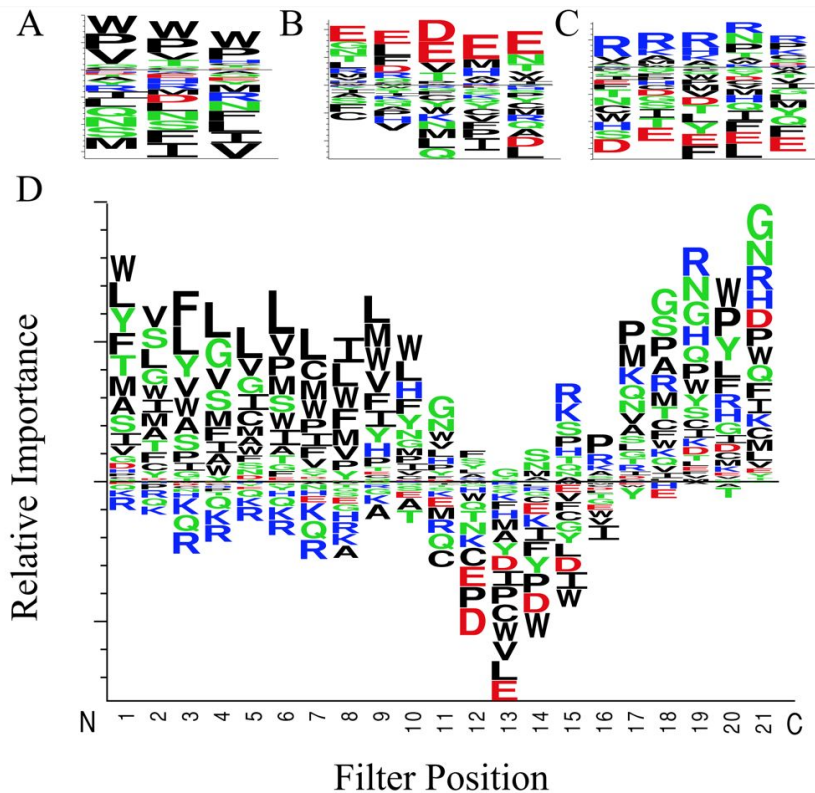
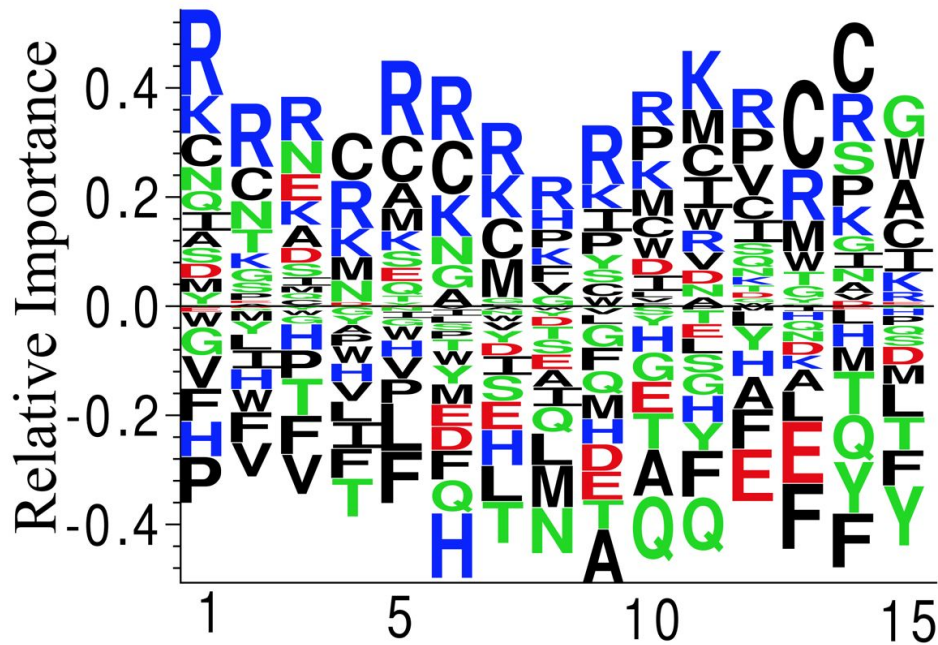
1. LSTMs predict subcellular localization

Model	Accuracy
Input: Protein Sequence	
R-LSTM	0.879
A-LSTM	0.854
R-LSTM ensemble	0.902
MultiLoc	0.767
Input: Protein Sequence + Metadata	
MultiLoc + PhyloLoc	0.842
MultiLoc + PhyloLoc + GOLoc	0.871
MultiLoc2	0.887
SherLoc2	0.930

Confusion Matrix											
ER	26	1	0	0	8	1	0	0	0	3	0
Golgi	1	28	0	0	0	0	0	0	0	1	0
Chloroplast	0	0	82	3	0	0	5	0	0	0	0
Cytoplasmic	0	0	1	266	0	0	3	12	0	0	0
Extracellular	0	0	0	1	166	0	0	0	0	1	0
Lysosomal	0	0	0	0	5	12	0	0	0	3	0
Mitochondrial	0	0	2	5	0	0	94	1	0	0	0
Nuclear	0	0	0	27	1	0	3	137	0	0	0
Peroxisomal	0	1	0	10	0	0	0	1	18	2	0
Plasma membrane	0	0	0	0	5	0	1	1	0	241	0
Vacuolar	0	0	0	0	7	0	0	0	0	1	5

Results

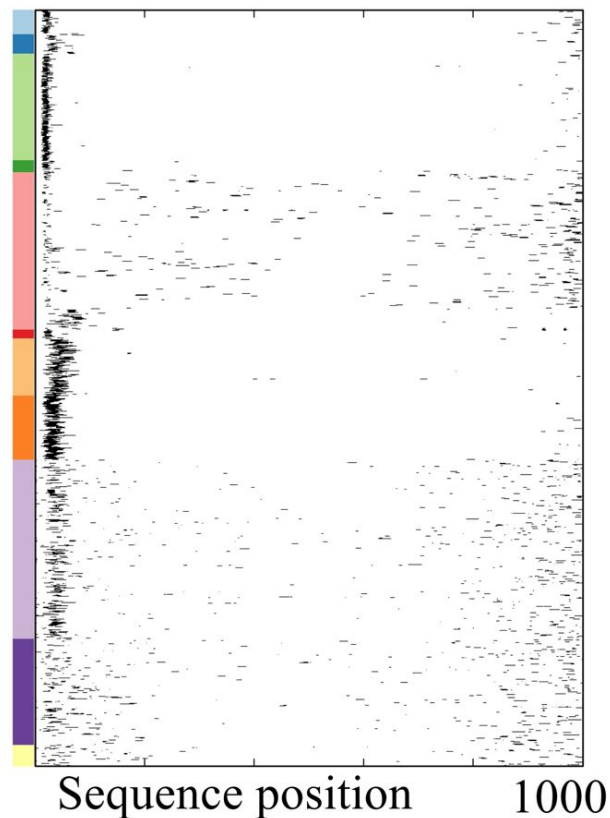
2. Convolutional filters provide interpretable sequence motifs



Results

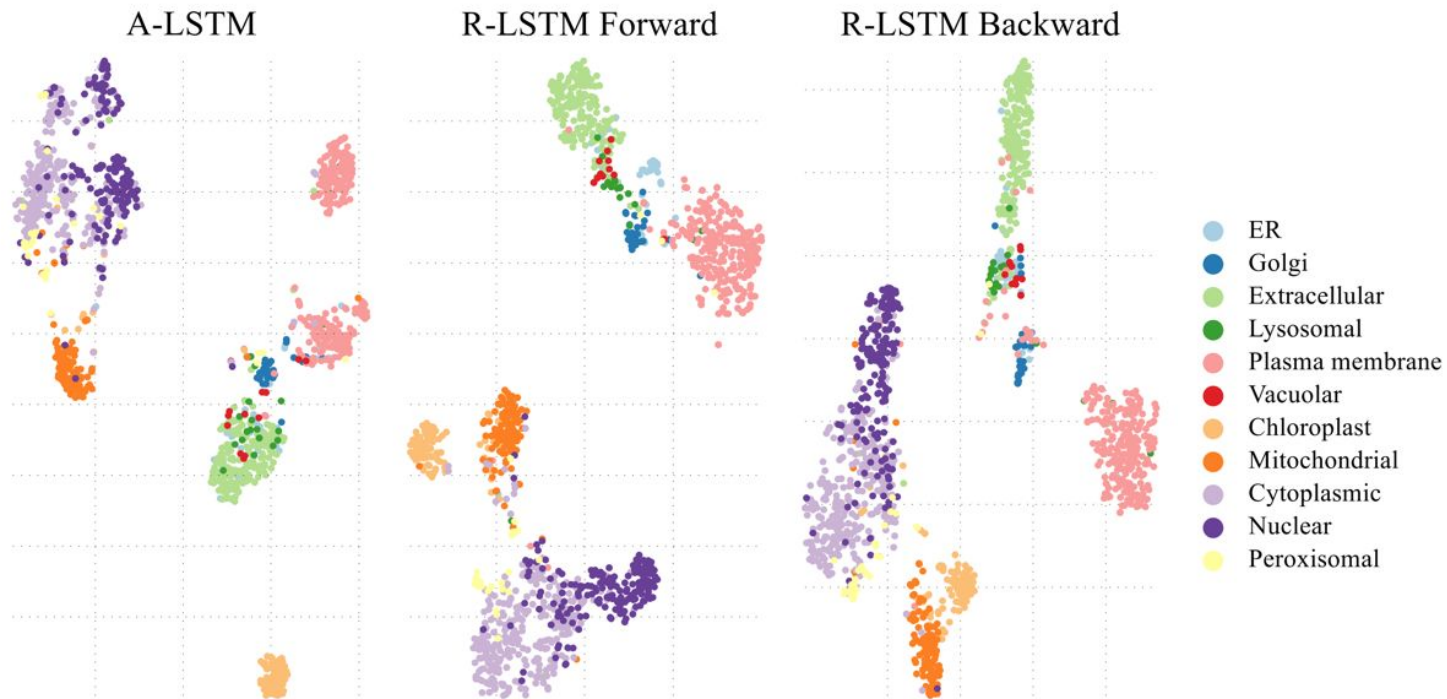
3. A-LSTMs highlight key parts of protein sequences

- ER
- Golgi
- Extracellular
- Lysosomal
- Plasma membrane
- Vacuolar
- Chloroplast
- Mitochondrial
- Cytoplasmic
- Nuclear
- Peroxisomal



Results

4. LSTMs generate fixed-length protein representations



Criticism

- 6000 proteins
- In general, could have characterized mis-classified proteins better (i.e. which proteins trip up the model)
- Is randomly selecting train/test appropriate?
- No indication how well their method generalizes

Criticism

- Errors appear somewhat correlated with class size: how well would a naive classifier do?
- More in silico experiments required for attention (could this be an artefact of initialization)?

Extensions

Input data:

2007). Each amino acid was encoded using 1-of-K encoding, the BLOSUM80 (Henikoff & Henikoff, 1992) and HSDM (Prlić et al., 2000) substitution matrices and sequence profiles, yielding 80 features per amino acids. Sequence profiles were created with ProfilePro² using 3 blastpgp³ iterations on UNIREF50 (Magrane et al., 2011).

- How much is the network relying on sequence data vs. substitution frequency to determine localization?
 - Some interpretation scheme
 - Visualizing weights of inputs
 - Looking for important combinations of sequence and substitution frequency

Extensions

*D358–D363 Nucleic Acids Research, 2014, Vol. 42, Database issue
doi:10.1093/nar/gkt1115*

Published online 13 November 2013

The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases

Sandra Orchard^{1,*}, Mais Ammari², Bruno Aranda¹, Lionel Breuza³, Leonardo Briganti⁴, Fiona Broackes-Carter⁵, Nancy H. Campbell⁶, Gayatri Chavali¹, Carol Chen⁷, Noemi del-Toro¹, Margaret Duesbury¹, Marine Dumousseau¹, Eugenia Galeota⁴, Ursula Hinz³, Marta Iannuccelli⁴, Sruthi Jagannathan⁸, Rafael Jimenez¹, Jyoti Khadake¹, Astrid Lagreid⁹, Luana Licata⁴, Ruth C. Lovering⁶, Birgit Meldal¹, Anna N. Melidoni⁶, Mila Milagros¹, Daniele Peluso¹⁰, Livia Perfetto⁴, Pablo Porras¹, Arathi Raghunath¹¹, Sylvie Ricard-Blum¹², Bernd Roechert³, Andre Stutz³, Michael Tognolli³, Kim van Roey¹³, Gianni Cesareni^{4,10,*} and Henning Hermjakob^{1,*}

- Using protein-protein interaction databases to further improve accuracy

Extensions

- Proteins can change localization during their “lifetime”

9310–9324 *Nucleic Acids Research*, 2013, Vol. 41, No. 20
doi:10.1093/nar/gkt715

Published online 9 August 2013

Genome-wide single-cell-level screen for protein abundance and localization changes in response to DNA damage in *S. cerevisiae*

Aprotim Mazumder^{1,2,3}, Laia Quiros Pesudo^{1,2}, Siobhan McRee^{1,2}, Mark Bathe^{1,2,3} and Leona D. Samson^{1,2,4,5,*}

- Probabilistic classification may be more representative of actual localization

Thanks!

All models were implemented in Theano (Bastien et al., 2012) using a modified version of the Lasagne library⁴ and trained with gradient descent. The learning rate was controlled with ADAM ($\alpha = 0.0002$, $\beta_1 = 0.1$, $\beta_2 = 0.001$, $\epsilon = 10^8$ and $\lambda = 10^{-8}$) (Kingma & Ba, 2014). Initial weights were sampled uniformly from the interval $[-0.05, 0.05]$. The network architecture is a 1D convolutional layer followed by an LSTM layer, a fully connected layer and a final softmax layer. All layers use 50% dropout. The 1D convolutional layer uses convolutions of sizes 1, 3, 5, 9, 15 and 21 with 10 filters of each size. Dense and convolutional layers use ReLU activation (Nair & Hinton, 2010) and the LSTM layer uses hyperbolic tangent. For the A-LSTM model the size of the first dimension of W_a was 400. Based on previous experiments we trained