Fast model-based protein homology detection without alignment

Aleksander Sudakov

29.01.08

Source

Hochreiter, S., Heusel, M. And Obermeyer, K. 2007. Fast model-based protein homology detection without alignment. Bioinformatics 23:1728-1736.

Tasks

- To analyze protein sequences from newly sequenced genomes
- Detect protein homology to other proteins
- Identify protein function, class, 3D structure

Alignment-based similarity methods

- Pairwise alignments (BLAST)
- Support vector machine (SVM)
- Position-specific scoring matrices (PSSM)

Model- based methods

- Account for relevant patterns or chemical properties
- Interpretation of classification results
- Various input

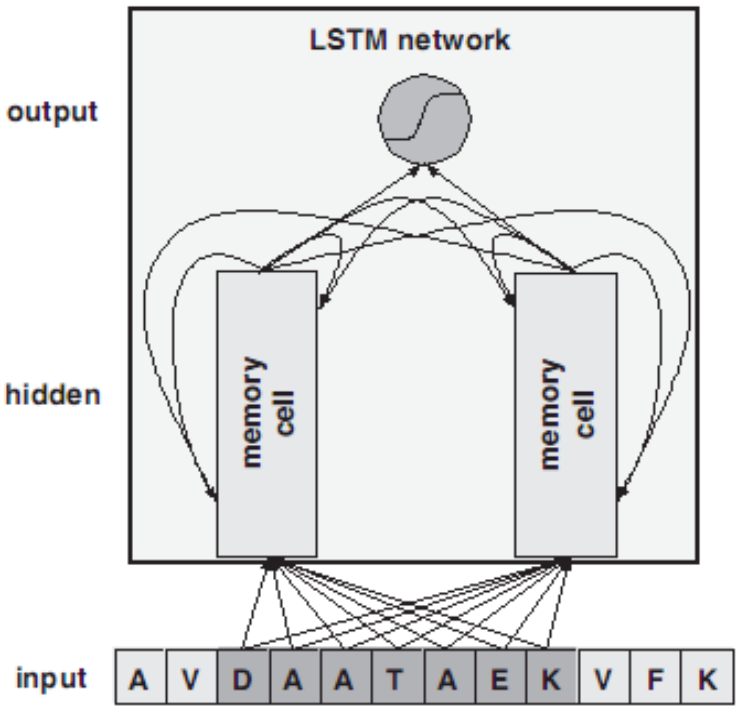
Recurrent neural networks

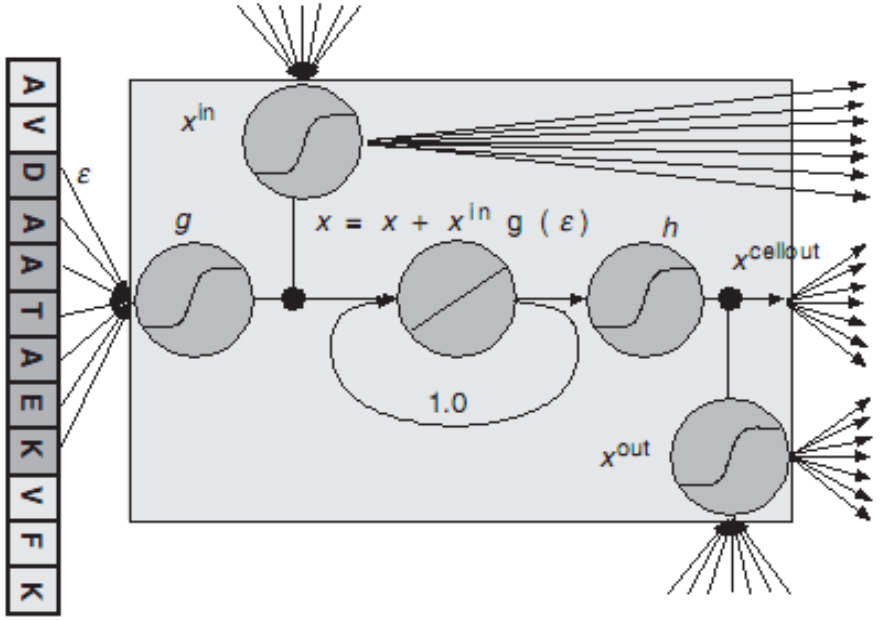
- Can extract dependencies between subsequences (AB is only indicative if followed by CD)
- Can extract correlation within subsequences (AB and CD are indicative, but AD or BC may be not indicative)
- Can extract local and global sequence characteristics (hydrophobicity, atomic weight etc.)
- Can extract dependencies between amino acids over a long interval in the sequence

Long Short-Term Memory

- Classical RNN – exponential decay of previously seen information
- LSTM – memory cell architecture
- Stores patterns from previously scanned regions

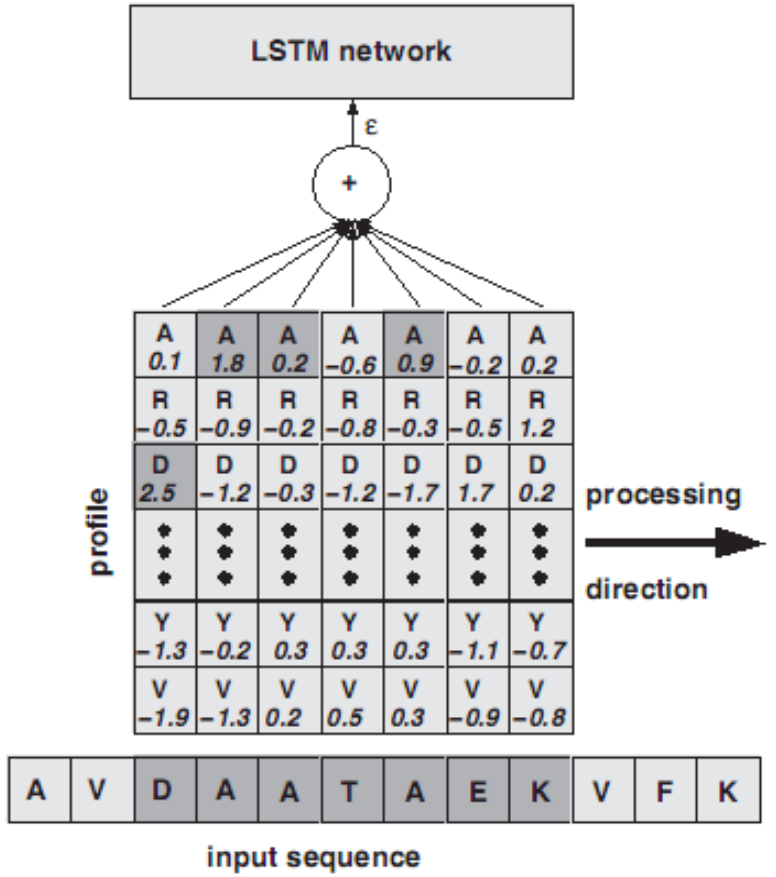
Memory cell


LSTM network
output
hidden
memory cell
memory cell
input
A V D A A T A E K V F K


A V D A A T A E K V F K
ε
x^{in}
g
$x = x + x^{in}$ g (ε)
h
$x^{cellout}$
1.0
x^{out}

Profile

- Weigted sum of amino acids in window



Computational complexity

- O(L) to classify a new sequence L
- Alignment - $O(L^2)$
- SVM - $O(N_{sv} L^2)$

SCOP

Method	M	P	V	S	ROC	ROC50	Time
(a) PSI-BLAST	–	–	–	–	0.693	0.264	5.5 s
(b) FPS	–	–	–	–	0.596	–	6800 s
(c) SAM-T98	+	–	–	–	0.674	0.374	200 s
(d) Fisher	–	–	–	+	0.887	0.250	>200 s
(e) Mismatch	–	–	–	+	0.872	0.400	380 s
(f) Pairwise	–	–	–	+	0.896	0.464	>700 s
(g) SW	–	–	–	+	0.916	0.585	>470 s
(h) LA	–	–	–	+	0.923	0.661	550 h
(i) Oligomer	–	–	–	+	0.919	0.508	2000 s
(j) HMMSTR	–	+	+	+	–	0.640	>500 h
(j) Mismatch-PSSM	–	+	+	+	0.980	0.794	>500 h
(j) SW-PSSM	–	+	+	+	0.982	0.904	>620 h
(k) LSTM	+	–	+	–	0.932	0.652	20 s

M – model based; P – profile input;

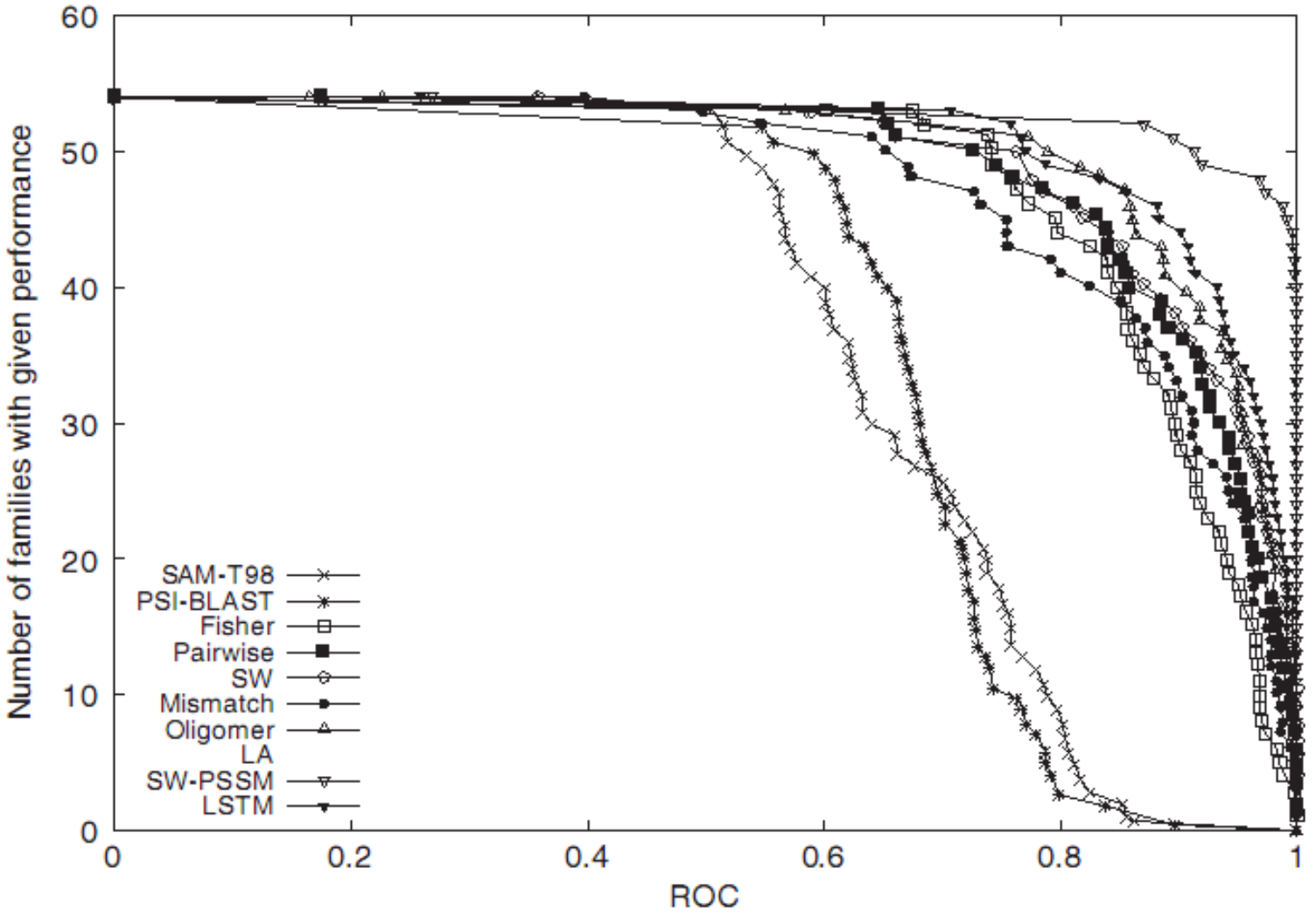
V – semi- supervised; S – SVM

ROC – area under curve; ROC50 – are under top 50 false positives

SVM methods training set creation – 110h

LSTM training set creation – 117h

SCOP 1.53 ROC


Number of families with given performance
ROC
0
10
20
30
40
50
60
0
0.2
0.4
0.6
0.8
1
SAM-T98
PSI-BLAST
Fisher
Pairwise
SW
Mismatch
Oligomer
LA
SW-PSSM
LSTM

SCOP 1.53 ROC50

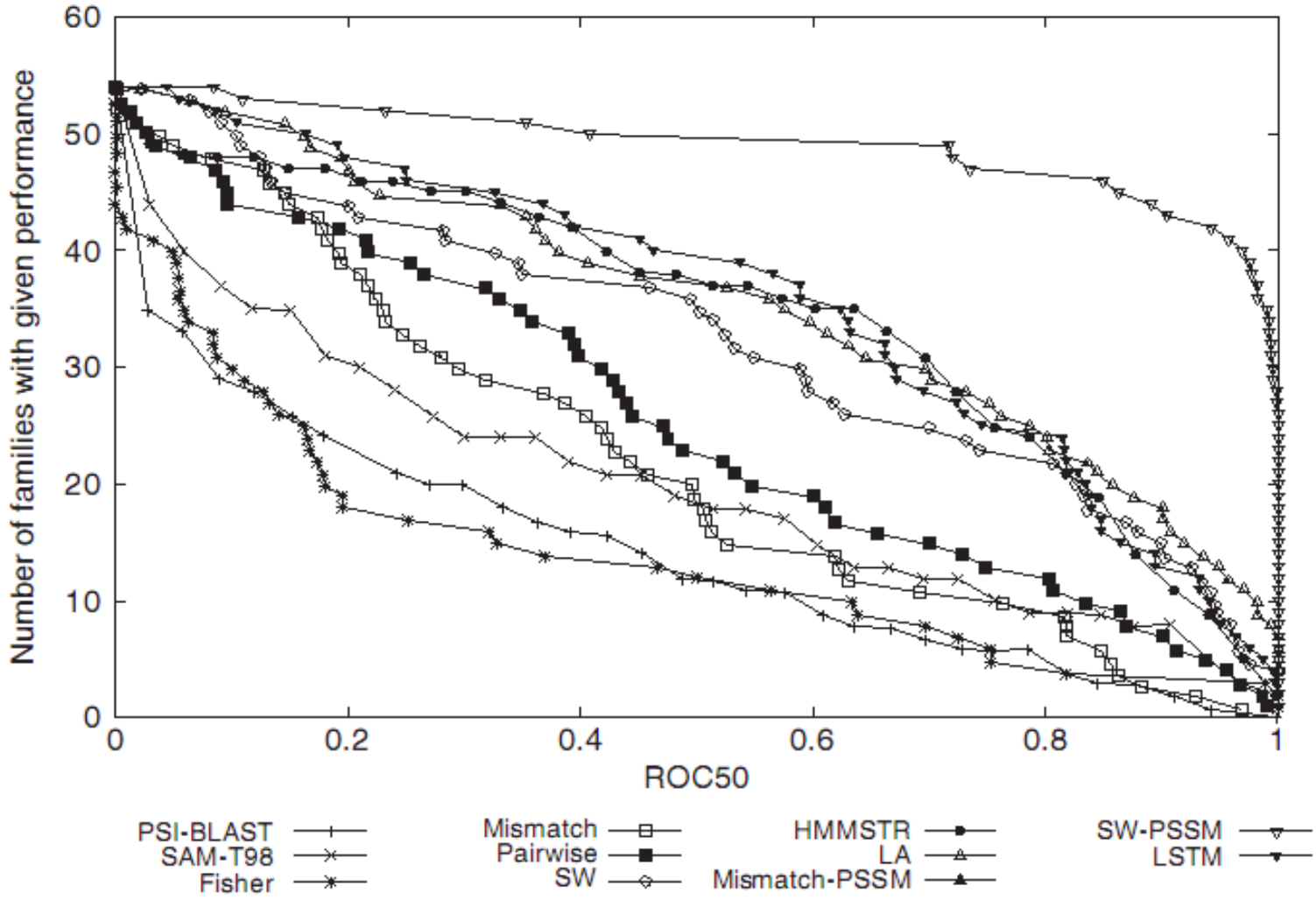


Table 3. Results on the data set from Ding and Dubchak (2001) for different machine-learning methods

Method	Q	Method	Q
NN	41.8	SVM	45.2
LSTM	51.7		

where 'NN' means neural network and 'SVM' support vector machine. LSTM yields the highest accuracy.

Features AA composition, predicted secondary structure, hydrophobicity, polarity etc.

Table 4. Results of PROSITE protein classification tested on the SwissProt database

Method/motif	Sensitivity	Specificity	Balanced Error
PROSITE	85.91 (15.62)	99.94 (0.15)	7.08 (7.79)
LSTM	98.24 (3.55)	99.79 (0.19)	0.99 (1.82)
Motif	86.82 (9.2)	99.93 (0.16)	6.63 (4.59)

All numbers are averaged over given 15 classes, the SD of the results is given in brackets. Results are reported for the PROSITE motif ('PROSITE'), for LSTM ('LSTM'), and for the motif extracted from LSTM ('motif'). The columns show (left to right): method, sensitivity (true positives divided by all positives in percent, 'sens.'), specificity (true negatives divided by all negatives in percent, 'spec.'), and the balanced error in percent ('bal. err.'). The balanced error is the mean of the class 1 and the class 2 error rate and is an appropriate measure for classification with unbalanced class sizes.

• 4FE4S_ FERREDOXIN (385)

PROSITE C-x(2)-C-x(2)-C-x(3)-C-[PEG]

LSTM (≈) C-x(2)-C-x(2)-C-x(2)-{C}-[AC]-[PEG]
•AA_ TRNA_ LIGASE_ I (913)

PROSITE P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

LSTM (≈) [ACFILMPV]-H-[ILMVFY]-G-[HGNT]-{DEHNPQR}-{DEP}-{CHKRY}-{DER}-[AILMSTVY]-{EGHPW}
•ATPASE_ ALPHA_ BETA (376)

PROSITE P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S

LSTM (≉) [ILV]-G-[CELR]-x(0,2)-[DGNV]-x-[ILRSV]-[AGS]-[DEKNQRV]-[AEGPV]-[DILMV]-[ADRT]-[DEGLNV]
•CITRATE_ SYNTHASE (76)

PROSITE G-[FYA]-[GA]-H-x-[IV]-x(1,2)-[RKT]-x(2)-D-[PS]-R

LSTM (≉) [ASG]-R-x(2)-G-W-x-A-H-x(2)-E OR
[ASG]-[QK]-x-P-x-[LIVM]-[AV]-A-x(2)-Y
•CYTOCHROME_ C (388)

PROSITE C-{CPWHF}-{CPWR}-C-H-{CFYW}

LSTM (≈) C-{CFP}-{CRWY}-C-H-{CFHWY}
•DEHYDROQUINASE_ I (44)

PROSITE D-[LIVM]-[DE]-[LIVMN]-x(18,20)-[LIVM](2)-x-[SC]-[NHY]-H-[DN]

LSTM (≉) D-[LIVA]-[LIVAY]-E-[LIVFW]-R-[LIVA]-D
•HISTONE_ H3_ 1 (44)

PROSITE K-A-P-R-K-Q-L

LSTM (≈) T-G-x-K-A-P-R
•INSULIN (194)

PROSITE C-C-{P}-x(2)-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C

LSTM (≈) C-C-{CDW}-x(2)-C-[DEIKNPSTB]-x(3)-[FILMV]-x(3)-C
•INVOLUCRIN (14)

PROSITE M-S-[QH]-Q-x-T-[LV]-P-V-T-[LV]

LSTM (≉) L-E-L-P-E-Q-Q OR Q-Q-E-S-x-E-x-E-L
•PHOSPHOFRUCTOKINASE (97)

PROSITE [RK]-x(4)-G-H-x-Q-[QR]-G-G-x(5)-D-R

LSTM (≉) [ILV]-E-V-M-G-[HR]-x(2)-[GS]
•PHOSPHOPANTETHEINE (198)

PROSITE [DEQGSTALMKRH]-...-[DNEKHS]-S-[LIVMST]-PCFY-...-[LIVMWSTA]-[LIVGSTACR]-x(2)-[LIVMFA]

LSTM (≈) [LFT]-x(1,2)-[DEQSTAK]-...-[DEHQSN]-S-[LIVMA]-x(4)-[LIVMSTA]-x(3)-[LIVMAF]-[DEHQSTAR]
•SERPIN (156)

PROSITE [LIVMFY]-x-[LIVMFYAC]-[DNQ]-[RKHQS]-[PST]-F-[LIVMFY]-[LIVMFYC]-x-[LIVMFAH]

LSTM (≉) F-{ADEGINP}-[IKLMNSV]-x(6,7)-V-x-M-M
•UPF0011 (26)

PROSITE S-D-A-G-x-P-x-[LIV]-[SN]-D-P-G

LSTM (≉) R-x(4)-[LF]-x(5)-[LIVF]-x(2)-E-D-T-R
•ZINC_ FINGER_ C2H2_ 1 (792)

PROSITE C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

LSTM (≈) [CFA]-x(2)-C-x(3)-[CFY]-x(5)-[LFQ]-x(2)-H-x(3)-H
•ZINC_ PROTEASE (546)

PROSITE [GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ]

LSTM (≈) [GILNSTV]-[AFILMTVY]-x-H-E-[AFILMTVY]-[AGILMSTV]-H

Conclusion

LSTM

- Novel method for protein classification and motif extraction
- 3 orders of magnitude faster than best performing SVM
- „State-of-art“ results
- Complementary to alignment-based approaches