nature biotechnology

SUPPLEMENTARY INFORMATION

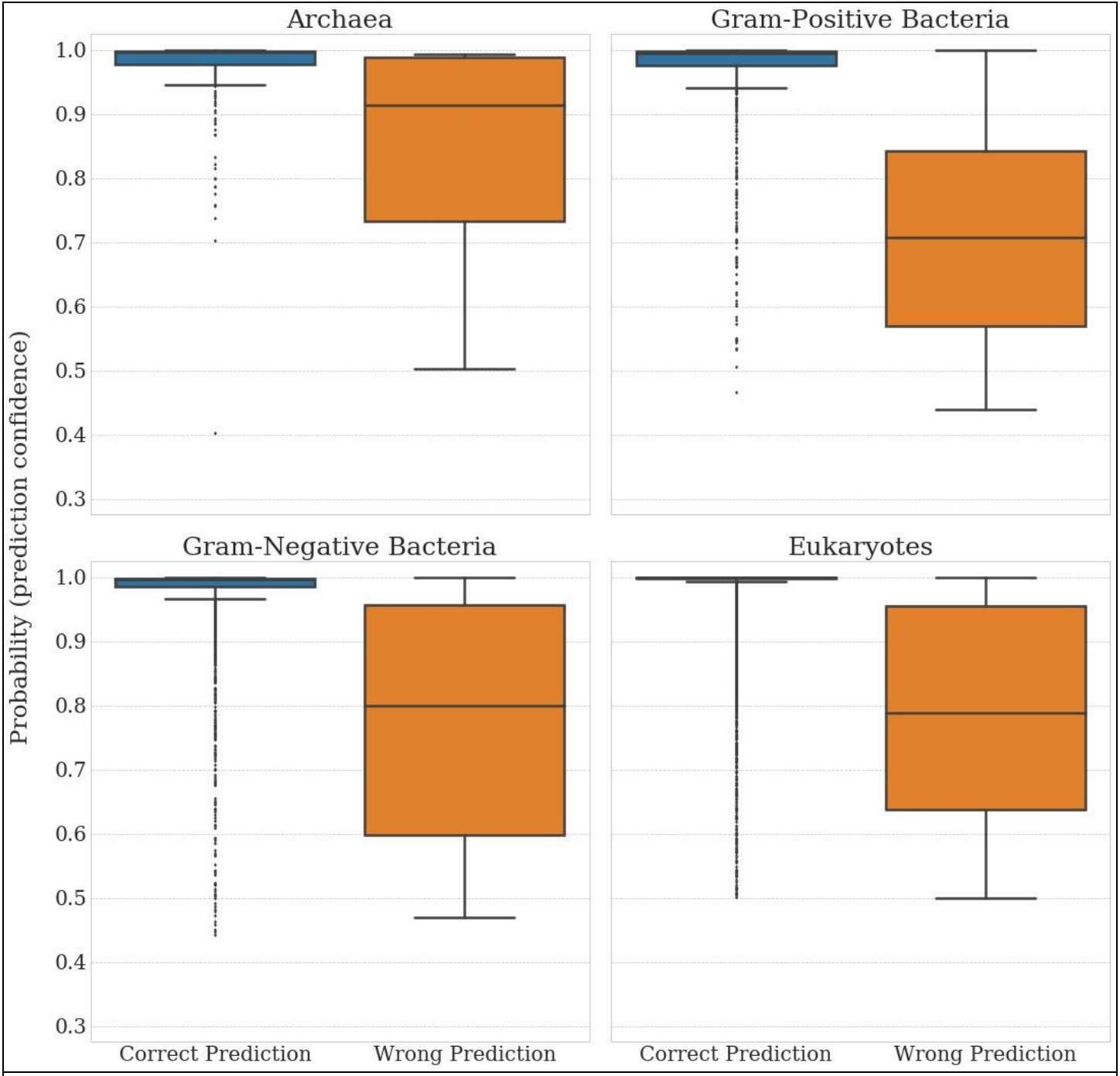
https://doi.org/10.1038/s41587-019-0036-z

In the format provided by the authors and unedited.

SignalP 5.0 improves signal peptide predictions using deep neural networks

José Juan Almagro Armenteros[1,9], Konstantinos D. Tsirigos[1,2,3,4,9], Casper Kaae Sønderby[5], Thomas Nordahl Petersen[6], Ole Winther[5,7], Søren Brunak[1,8], Gunnar von Heijne[2,3] and Henrik Nielsen[1]*

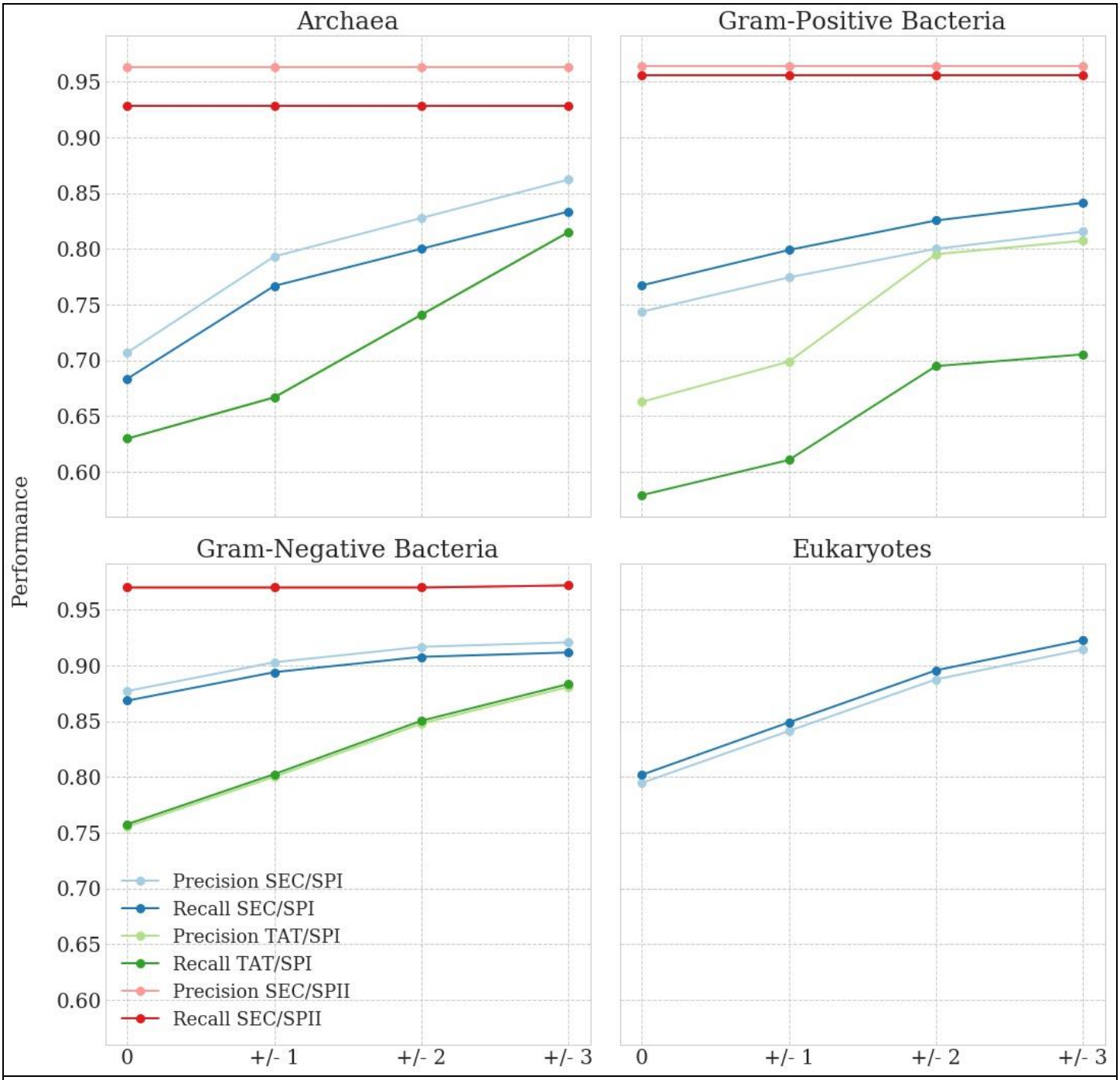
[1]Department of Bio and Health Informatics, Technical University of Denmark, Kgs Lyngby, Denmark. [2]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden. [3]Science for Life Laboratory, Stockholm University, Solna, Sweden. [4]Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany. [5]Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [6]National Food Institute, Technical University of Denmark, Kgs Lyngby, Denmark. [7]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs Lyngby, Denmark. [8]Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [9]These authors contributed equally: José Juan Almagro Armenteros, Konstantinos D. Tsirigos. *e-mail: hnielsen@bioinformatics.dtu.dk



Supplementary Figure 1

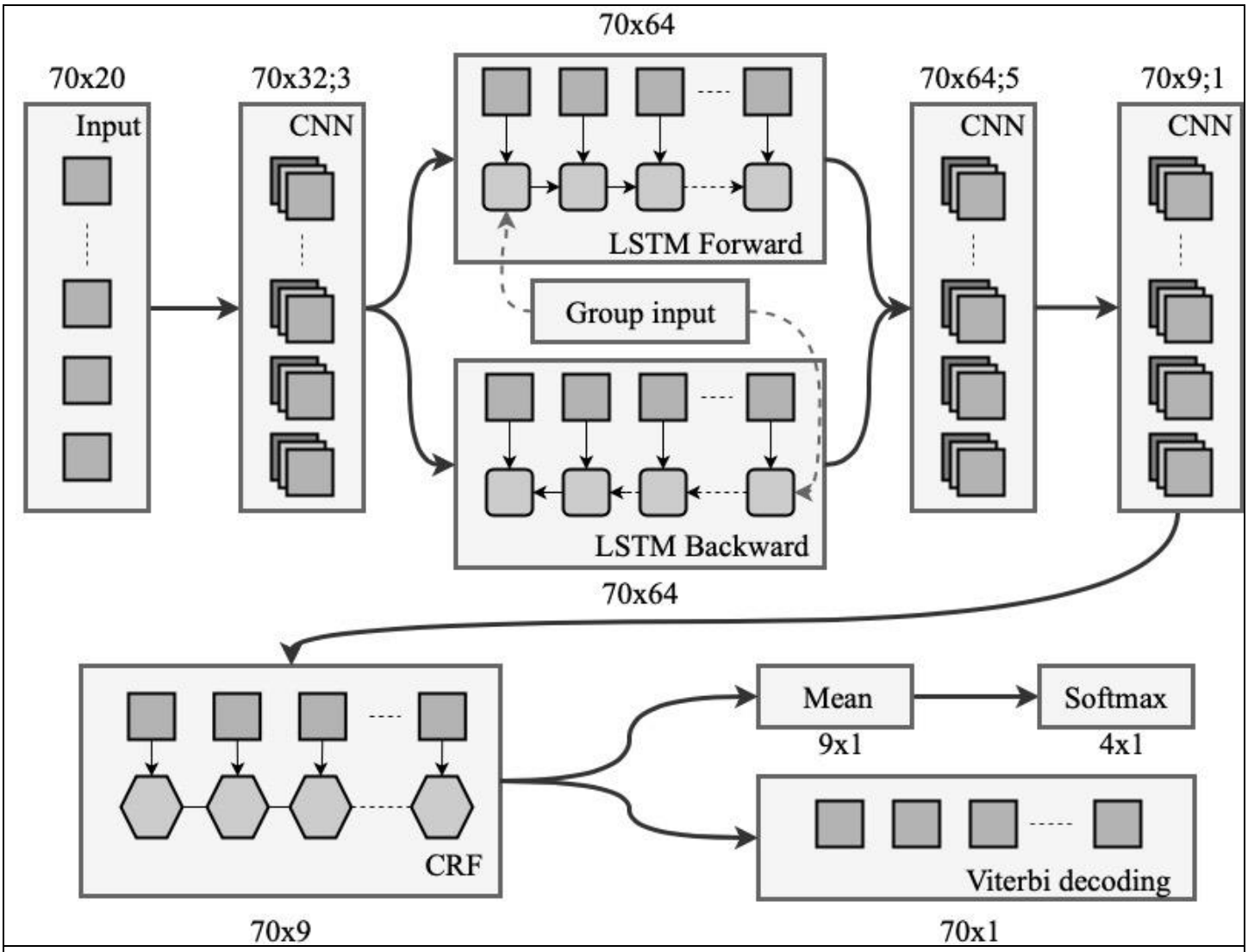
Box plot of the probability of the predicted class for correct and incorrect predictions

A probability close to 1 means a highly reliable prediction. For Archaea, Gram-Positive and Gram-Negative the probability threshold is 0.25, as there are four possible classes (*Sec/SPI*, *Tat/SPI*, *Sec/SPII* and *Other*). For Eukarya this threshold is 0.5, as it has only two classes (*Sec/SPI* and *Other*). A probability close to this threshold means a very unreliable prediction. All classes, namely *Sec/SPI*, *Tat/SPI*, *Sec/SPII* and *Other* are combined in this plot.



Supplementary Figure 2

Performance of SignalP 5.0 on cleavage site detection when considering a window of 0, 1, 2 and 3 amino acids centered on the real cleavage site.



Supplementary Figure 3

The SignalP 5.0 neural network architecture.

Supplementary Material

Supplementary Note 1

The special case of Signal-BLAST

Signal-BLAST runs BLAST against a pre-constructed database of signal peptides, and, if it finds a hit with high similarity, it essentially conducts a database look-up (current version of Signal-BLAST uses Uniprot 2017_6 as reference). Because of that, we sought to find out how many of the proteins in our datasets fall under the "look-up" category. Interestingly, out of the 7,457 proteins in Eukaryotes, modes (parameter sets) 1, 2 and 3 of Signal-BLAST find 7,444 /7,457 (99.8%) with 100% identical sequence in the reference database, and only mode 4 is different with 3,160 /7,457 (42.4%). In this benchmark, the overall MCC of modes 1, 2 and 3 was almost equal to 1.0, while for the mode 4 it merely reached 0.145. In Archaea, out of 182 proteins, there are 126 (for modes 1, 2 and 3) or 109 (from mode 4) proteins with a 100% identity, which translates into 69.2% or 59.8% of the total proteins in this set. In Gram-negative bacteria, out of 783 proteins, there are 723 (for modes 1, 2 and 3) or 684 (from mode 4) proteins with a 100% identity in the reference database (constituting 92.3% or 87.4% of them). Finally, in Gram-positive bacteria, out of 389 proteins, there are 343 (for modes 1, 2 and 3) or 324 (from mode 4) proteins with 100% identity to one of the proteins in the reference database, i.e. 88.2% or 83.2% of them. Based on this finding, the Signal-BLAST method was not included in our benchmark, since its predictive performance would be artificially high. Signal-BLAST further exemplifies that the idea of using sequence similarity alone is probably not suited for the task of SP prediction, since, if no homology is found, then the prediction performance is very low.

Supplementary Note 2

Signal peptide classification and Cleavage Site prediction reliability

We observed that correctly classified proteins have a prediction confidence close to 1, even though some outliers were present. The median of these probabilities for Archaea, Gram-positive

bacteria, Gram-negative bacteria and Eukaryotes is 0.996, 0.995, 0.997 and 0.999 respectively. In contrast, incorrectly classified proteins have wider probability distributions, ranging between 0.4 and 1.0. The median of these probabilities for Archaea, Gram-Positive, Gram-Negative and Eukaryotes is 0.914, 0.707, 0.800 and 0.788, respectively, highlighting the lower prediction confidence of incorrect predictions.

CS prediction performance for Tat/SPI and Sec/SPI proteins increased across all domains in relation to window size (**Supplementary Figure 2**). For Eukaryotes, recall increases from 0.802 to 0.923 when considering a window of three amino acids, indicating that a high proportion of the erroneous predictions are within 3 AAs of the annotated CS position. We can therefore conclude that, when predicting the SPase I CS with SignalP 5.0, we can be highly confident that the real CS will be located in a window of three AAs around the predicted CS. For Sec/SPII SPs, the cleavage site performance is very high and is not improved by considering a wider window, reflecting that the SPase II CS is dependent on the presence of a cysteine residue.

Supplementary Note 3

Proteomic analysis using SignalP 5.0

For *Escherichia coli,* UniProt reports 498 proteins with a SP, 137 of which are experimentally verified and 361 predicted by various ways. Out of the 137 SPs, 122 are Sec/SPI, 11 are Sec/SPII and 4 are Tat/SPI SPs. SignalP 5.0 predicts 612 SPs in total, out of which 414 Sec/SPI, 161 Sec/SPII and 37 Tat/SPI. When we compare the UniProt annotated SPs to the SignalP 5.0 predicted ones, we observe that SignalP 5.0 detects 136/137 experimental SPs, together with

their respective type (misses one Tat/SPI SP). Regarding CS prediction, 131/137 are predicted to have precisely the same CS, whereas, 5/137 are within the ±3 window.

The respective analysis for *Saccharomyces cerevisiae* is as follows: UniProt annotation reports 297 proteins, 37 of which are experimentally annotated and 260 are predicted by various ways. SignalP 5.0 predicts 314 Sec/SPI SPs in total. When we compare the UniProt annotated SPs to the SignalP 5.0 predicted ones, we observe that SignalP 5.0 detects 36/37 experimental SPs. Regarding CS prediction, 29/37 are predicted to have precisely the same CS, while, a further of 2/37 are within the ±3 window.

These two analyses show us that SignalP 5.0's confident predictions can potentially help towards a better annotation of proteomes.

Method	URL	Organism	Prediction type
SignalP 5.0	http://www.cbs.dtu.dk/services/SignalP/	A | E | P | N	S | L | T | O
SignalP 4.1[12]	http://www.cbs.dtu.dk/services/SignalP-4.1/	E | P | N	S | O
DeepSig[19]	https://deepsig.biocomp.unibo.it/deepsig/	E | P | N	S | O
LipoP[47]	http://www.cbs.dtu.dk/services/LipoP/	A | P | N	S | L | O
Philius[21]	http://www.yeastrc.org/philius/pages/philius/runPhilius.jsp	A | E | P | N	S | O
Phobius[22]	http://phobius.sbc.su.se/	A | E | P | N	S | O
PolyPhobius[48]	http://phobius.sbc.su.se/poly.html	A | E | P | N	S | O
PRED-LIPO[30]	http://bioinformatics.biol.uoa.gr/PRED-LIPO/	A | P | N	S | L | O
PRED-SIGNAL[20]	http://bioinformatics.biol.uoa.gr/PRED-SIGNAL/	A | E | P | N	S | O
PRED-TAT[25]	http://www.compgen.org/tools/PRED-TAT/	A | E | P | N	S | T | O
PrediSi[49]	http://www.predisi.de/	E | P | N	S | O
Signal-3L 2.0[33]	http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/	E | P | N	S | O
Signal-BLAST[17]	http://sigpep.services.came.sbg.ac.at/signalblast.html	A | E | P | N	S | O
Signal-CF[34]	http://www.csbio.sjtu.edu.cn/bioinf/Signal-CF/	E | P | N	S | O
SOSUIsignal[50]	http://harrier.nagahama-i-bio.ac.jp/sosui/sosuisignal/sosuisignal_submit.html	E | P | N	S | O
SPEPlip[35]	http://gpcr.biocomp.unibo.it/cgi/predictors/spep/pred_spepcgi.cgi	E | P | N	S | L | O
SPOCTOPUS[23]	http://octopus.cbr.su.se/	A | E | P | N	S | O
TATFIND[26]	http://signalfind.org/tatfind.html	A | E | P | N	T | O

TatP[51]	http://www.cbs.dtu.dk/services/TatP/	A | P | N	T | O
TOPCONS2[24]	http://www.topcons.net/	A | E | P | N	S | O

Supplementary Table 1. The full list of methods used for benchmarking of SignalP 5.0. In the 'Organism' column, where we specify the type of organisms each method was designed for or can work with, 'A' stands for Archaea, 'E' for Eukaryotes, 'N' for Gram-negative bacteria and 'P' for Gram-positive bacteria. In the 'Prediction type' column, where we specify which type of predictions each method can produce, 'S' stands for Sec/SPI signal peptides, 'L' for Lipoprotein (Sec/SPII) signal peptides, 'T' for Tat/SPI signal peptides and 'O' for other (globular and/or transmembrane proteins).

Type	Archaea	Eukaryotes	Gram-negative bacteria	Gram-positive bacteria
Sec/SPI signals	60 (50)	2,614 (210)	509 (90)	189 (25)
Sec/SPII signals	28 (19)	--	1,063 (442)	449 (201)
Tat/SPI signals	27 (22)	--	334 (98)	95 (74)
Globular	78 (63)	13,612 (6,929)	202 (103)	140 (64)
TM	44 (28)	1,044 (318)	220 (50)	50 (25)
Total	237 (182)	17,270 (7,457)	2,328 (783)	923 (389)

Supplementary Table 2. The composition of the training and test datasets used during the development of SignalP 5.0 with the benchmark dataset numbers in parentheses

SignalP 5.0 (cross-val)	Archaea		Gram-negative bacteria		Gram-positive bacteria		Eukaryotes
	MCC^1	MCC^2	MCC^1	MCC^2	MCC^1	MCC^2	MCC
Sec/SPI	0.938	0.933	0.907	0.918	0.890	0.882	0.966
Sec/SPII	0.956	0.938	0.960	0.960	0.957	0.957	-
Tat/SPI	0.977	0.958	0.981	0.981	0.868	0.866	-

Supplementary Table 3. Cross-validated performance of SignalP 5.0 on SP detection over the whole training dataset (20,758 proteins). 'MCC^1' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by TM+Globular proteins only; 'MCC^2' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by Sec/SPII SPs,Tat/SPI SPs, TM and Globular proteins.

SignalP 5.0 (cross-val)	Sec/SPI	Sec/SPII	Tat/SPI	Other
Total dataset (real/predicted)				
Sec/SPI	3,254 (339)	22 (7)	8 (1)	88 (28)
Sec/SPII	24 (14)	1,503 (637)	2 (2)	11 (9)
Tat/SPI	13 (9)	8 (6)	433 (177)	2 (2)
Globular	40 (21)	1 (1)	0 (0)	13,991 (7,137)

TM	64 (17)	1 (0)	2 (1)	1,291 (403)
Archaea (real/predicted)				
Sec/SPI	56 (46)	1 (1)	0 (0)	3 (3)
Sec/SPII	1 (1)	26 (17)	1 (1)	0 (0)
Tat/SPI	0 (0)	0 (0)	26 (21)	1 (1)
Globular	0 (0)	0 (0)	0 (0)	78 (63)
TM	1 (1)	0 (0)	0 (0)	43 (27)
Eukaryotes (real/predicted)				
Sec/SPI	2,550 (194)	-	-	64 (16)
Globular	39 (21)	-	-	13,573 (6,908)
TM	49 (13)	-	-	995 (305)
Gram-negative bacteria (real/predicted)				
Sec/SPI	474 (76)	18 (6)	3 (0)	14 (8)
Sec/SPII	19 (9)	1,040 (430)	1 (1)	3 (2)
Tat/SPI	2 (2)	3 (2)	329 (94)	0 (0)
Globular	0 (0)	1 (1)	0 (0)	201 (102)
TM	9 (2)	1 (0)	2 (1)	208 (47)
Gram-positive bacteria (real/predicted)				
Sec/SPI	174 (23)	3 (0)	5 (1)	7 (1)
Sec/SPII	4 (4)	437 (190)	0 (0)	8 (7)
Tat/SPI	11 (7)	5 (4)	78 (62)	1 (1)
Globular	1 (0)	0 (0)	0 (0)	139 (64)
TM	5 (1)	0 (0)	0 (0)	45 (24)

Supplementary Table 4 . Confusion matrices for the different type of predictions that SignalP 5.0 makes on the training and the benchmark (in brackets) dataset.

SignalP 5.0 (cross-val)	Archaea				Gram-negative bacteria				Gram-positive bacteria				Eukaryotes			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall																
Sec/SPI	0.683	0.767	0.800	0.833	0.868	0.894	0.908	0.912	0.767	0.799	0.825	0.841	0.802	0.849	0.896	0.923
Sec/SPII	0.929	0.929	0.929	0.929	0.970	0.970	0.970	0.972	0.955	0.955	0.955	0.955	-	-	-	-
Tat/SPI	0.630	0.667	0.741	0.815	0.757	0.802	0.850	0.883	0.579	0.611	0.695	0.705	-	-	-	-
CS precision																
Sec/SPI	0.707	0.793	0.828	0.862	0.877	0.903	0.917	0.921	0.744	0.744	0.800	0.815	0.795	0.841	0.887	0.914
Sec/SPII	0.963	0.963	0.963	0.963	0.970	0.970	0.970	0.972	0.964	0.964	0.964	0.964	-	-	-	-
Tat/SPI	0.630	0.667	0.741	0.815	0.755	0.800	0.848	0.881	0.663	0.699	0.795	0.807	-	-	-	-

Supplementary Table 5. Cross-validated performance of SignalP 5.0 on CS recall and precision over the whole training dataset (20,758 proteins).

	CRF and transfer learning	Only CRF	Only transfer learning	No CRF nor transfer learning
Signal peptide accuracy	0.986	0.981	0.984	0.981
Cleavage site accuracy	0.808	0.790	0.637	0.626

Supplementary Table 6. Comparison of the model performance with and without CRF and transfer learning. Since all taxonomic groups were combined for this analysis, we report just the accuracy of SignalP 5.0 (i.e. the fraction of correctly identified signal peptides and the fraction of correctly predicted cleavage sites)

Method	Archaea		Eukaryotes	Gram-negative bacteria		Gram-positive bacteria	
	MCC^1	MCC^2	MCC	MCC^1	MCC^2	MCC^1	MCC^2
SignalP 5.0	**0.922**	**0.917**	**0.883**	**0.860**	**0.830**	0.922	**0.760**
SignalP 4.1	n.d.	n.d.	0.808	0.851	0.248	**0.949**	0.148
DeepSig	n.d.	n.d.	0.819	0.792	0.166	0.870	0.115
LipoP	0.829	0.604	0.363	0.787	0.483	0.922	0.403
Philius	0.765	0.447	0.421	0.802	0.127	0.843	0.075
Phobius	0.813	0.514	0.510	0.818	0.132	0.818	0.074
PolyPhobius	0.752	0.453	0.456	0.844	0.144	0.852	0.111
PrediSi	n.d.	n.d.	0.553	0.802	0.244	0.781	0.121
PRED-LIPO	0.796	0.586	0.234	0.775	0.398	0.896	0.410
PRED-SIGNAL	0.897	0.584	0.272	0.724	0.098	0.830	0.114
PRED-TAT	0.788	0.626	0.326	0.769	0.187	0.830	0.189
Signal-3L 2.0	n.d.	n.d.	0.597	0.806	0.110	0.922	0.106
Signal-CF	n.d.	n.d.	0.326	0.561	0.106	0.558	0.084
SOSUIsignal	n.d.	n.d.	0.375	0.693	0.108	0.722	0.047
SPEPlip	n.d.	n.d.	0.655	0.746	0.498	0.646	0.350
SPOCTOPUS	0.765	0.408	0.492	**0.860**	0.127	0.922	0.109
TOPCONS2	0.765	0.432	0.477	0.860	0.131	0.897	0.071

Supplementary Table 7. Benchmarking of Sec/SPI signal peptide detection predictions. The highest performance values have been highlighted in bold. 'MCC^1' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by TM+Globular proteins only; 'MCC^2' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPI SPs and the negative dataset by Sec/SPII SPs,Tat/SPI SPs, TM and Globular proteins.

Method	Archaea				Eukaryotes				Gram-negative bacteria				Gram-positive bacteria			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall																
SignalP 5.0	0.660	0.740	0.780	0.820	**0.729**	**0.762**	**0.795**	**0.833**	**0.733**	**0.767**	0.800	0.800	**0.840**	**0.840**	**0.880**	**0.880**
SignalP 4.1	n.d.	n.d.	n.d.	n.d.	0.695	0.729	0.762	0.786	0.644	0.711	0.733	0.744	**0.840**	**0.840**	0.840	0.840
DeepSig	n.d.	n.d.	n.d.	n.d.	0.624	0.652	0.690	0.724	0.600	0.656	0.667	0.678	0.760	0.760	0.840	0.840
LipoP	0.480	0.620	0.660	0.720	0.343	0.386	0.419	0.448	0.733	**0.767**	0.789	0.789	0.600	0.600	0.640	0.640
Philius	0.580	0.680	0.700	0.700	0.619	0.686	0.743	0.781	0.700	0.744	0.789	0.811	0.600	0.600	0.600	0.600
Phobius	0.540	0.640	0.660	0.700	0.667	0.700	0.738	0.786	0.644	0.722	0.789	0.811	0.600	0.600	0.600	0.600
PolyPhobius	0.560	0.680	0.680	0.700	0.681	0.733	0.776	**0.833**	0.644	0.733	0.811	0.822	0.680	0.680	0.720	0.720
PrediSi	n.d.	n.d.	n.d.	n.d.	0.652	0.695	0.719	0.767	0.722	0.789	0.811	0.822	0.640	0.640	0.760	0.800
PRED-LIPO	0.480	0.600	0.660	0.680	0.095	0.114	0.152	0.181	0.467	0.522	0.567	0.600	0.760	0.760	0.760	0.760
PRED-SIGNAL	**0.800**	**0.900**	**0.900**	**0.900**	0.224	0.290	0.329	0.362	0.444	0.522	0.622	0.644	0.680	0.680	0.720	0.720
PRED-TAT	0.580	0.720	0.800	0.820	0.410	0.510	0.571	0.614	0.711	0.767	0.800	0.822	0.720	0.720	0.760	0.760
Signal-3L 2.0	n.d.	n.d.	n.d.	n.d.	0.648	0.686	0.733	0.762	0.644	0.700	0.722	0.733	0.800	0.800	0.840	0.840
Signal-CF	n.d.	n.d.	n.d.	n.d.	0.652	0.676	0.724	0.762	0.689	0.711	0.744	0.778	0.720	0.720	0.800	0.800
SOSUIsignal	n.d.	n.d.	n.d.	n.d.	0.176	0.329	0.467	0.576	0.267	0.367	0.567	0.622	0.200	0.240	0.280	0.440
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.710	0.733	0.771	0.810	0.611	0.678	0.722	0.733	0.680	0.680	0.720	0.720
SPOCTOPUS	0.340	0.480	0.520	0.560	0.390	0.533	0.686	0.757	0.467	0.689	**0.833**	**0.867**	0.640	0.760	0.800	0.880
TOPCONS2	0.480	0.600	0.620	0.640	0.371	0.505	0.638	0.729	0.544	0.622	0.733	0.767	0.240	0.320	0.400	0.440
CS precision																
SignalP 5.0	**0.771**	**0.688**	**0.812**	**0.812**	**0.671**	**0.702**	**0.732**	**0.732**	**0.742**	**0.775**	**0.809**	**0.809**	**0.600**	**0.600**	**0.629**	**0.629**
SignalP 4.1	n.d.	n.d.	n.d.	n.d.	0.613	0.643	0.672	0.693	0.151	0.167	0.172	0.175	0.083	0.083	0.083	0.083
DeepSig	n.d.	n.d.	n.d.	n.d.	0.604	0.631	0.668	0.700	0.131	0.144	0.146	0.148	0.073	0.073	0.080	0.080
LipoP	0.484	0.375	0.516	0.562	0.159	0.178	0.194	0.207	0.327	0.342	0.351	0.351	0.153	0.153	0.163	0.163
Philius	0.425	0.362	0.438	0.438	0.151	0.168	0.182	0.191	0.106	0.112	0.119	0.122	0.054	0.054	0.054	0.054
Phobius	0.395	0.333	0.407	0.432	0.226	0.237	0.250	0.267	0.098	0.110	0.120	0.124	0.054	0.054	0.054	0.054
PolyPhobius	0.395	0.326	0.395	0.407	0.176	0.190	0.201	0.216	0.097	0.110	0.122	0.124	0.060	0.060	0.063	0.063
PrediSi	n.d.	n.d.	n.d.	n.d.	0.273	0.291	0.301	0.321	0.144	0.157	0.162	0.164	0.062	0.062	0.074	0.078
PRED-LIPO	0.455	0.364	0.5	0.515	0.069	0.083	0.110	0.131	0.212	0.237	0.258	0.273	0.216	0.216	0.216	0.216
PRED-SIGNAL	0.489	0.435	0.489	0.489	0.066	0.085	0.096	0.106	0.076	0.089	0.106	0.110	0.060	0.060	0.064	0.064
PRED-TAT	0.493	0.397	0.548	0.562	0.080	0.099	0.111	0.119	0.125	0.135	0.141	0.145	0.082	0.082	0.087	0.087
Signal-3L 2.0	n.d.	n.d.	n.d.	n.d.	0.322	0.341	0.365	0.379	0.113	0.123	0.127	0.129	0.074	0.074	0.078	0.078
Signal-CF	n.d.	n.d.	n.d.	n.d.	0.105	0.109	0.117	0.123	0.102	0.105	0.110	0.115	0.059	0.059	0.065	0.065
SOSUIsignal	n.d.	n.d.	n.d.	n.d.	0.037	0.069	0.098	0.121	0.040	0.055	0.086	0.094	0.018	0.021	0.025	0.039
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.366	0.378	0.398	0.418	0.276	0.307	0.327	0.332	0.187	0.187	0.198	0.198

SPOCTOPUS	0.293	0.207	0.317	0.341	0.120	0.164	0.211	0.233	0.067	0.098	0.119	0.124	0.056	0.066	0.070	0.077
TOPCONS2	0.366	0.293	0.378	0.390	0.107	0.146	0.184	0.210	0.081	0.093	0.110	0.115	0.022	0.029	0.036	0.039

Supplementary Table 8. Benchmarking of Sec/SPI signal peptide cleavage site predictions, measured as recall and precision. The highest performance values have been highlighted in bold.

Method	Archaea		Gram-negative bacteria		Gram-positive bacteria	
	MCC[1]	MCC[2]	MCC[1]	MCC[2]	MCC[1]	MCC[2]
SignalP 5.0	**0.936**	**0.910**	**0.945**	**0.946**	**0.917**	**0.923**
LipoP	0.836	0.755	0.791	0.833	0.798	0.822
PRED-LIPO	0.766	0.743	0.629	0.707	0.775	0.775
SPEPlip	n.d.	n.d.	0.857	0.884	0.845	0.843

Supplementary Table 9. Benchmarking of Sec/SPII signal peptide detection predictions. The highest performance values have been highlighted in bold. 'MCC[1]' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPII SPs and the negative dataset by TM+Globular proteins only; 'MCC[2]' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Sec/SPII SPs and the negative dataset by Sec/SPI SPs,Tat/SPI SPs, TM and Globular proteins.

Method	Archaea				Gram-negative bacteria				Gram-positive bacteria			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall												
SignalP 5.0	**0.895**	**0.895**	**0.895**	**0.895**	**0.964**	**0.964**	**0.964**	**0.968**	**0.925**	**0.925**	**0.925**	**0.925**
LipoP	0.684	0.684	0.737	0.737	0.860	0.860	0.860	0.862	0.831	0.831	0.831	0.831
PRED-LIPO	0.632	0.632	0.632	0.632	0.717	0.717	0.717	0.719	0.816	0.816	0.816	0.816
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.912	0.912	0.914	0.914	0.876	0.876	0.876	0.876
CS precision												
SignalP 5.0	**0.944**	**0.944**	**0.944**	**0.944**	**0.970**	**0.970**	0.970	**0.975**	**0.959**	**0.959**	**0.959**	**0.959**
LipoP	0.765	0.765	0.824	0.824	0.969	0.969	0.969	0.972	0.944	0.944	0.944	0.944
PRED-LIPO	0.923	0.923	0.923	0.923	0.969	0.969	0.969	0.972	0.921	0.921	0.921	0.921
SPEPlip	n.d.	n.d.	n.d.	n.d.	0.969	0.969	**0.971**	0.971	0.936	0.936	0.936	0.936

Supplementary Table 10. Benchmarking of Sec/SPII signal peptide cleavage site predictions, measured as recall and precision. The highest performance values have been highlighted in bold.

Method	Archaea		Gram-negative bacteria		Gram-positive bacteria	
	MCC[1]	MCC[2]	MCC[1]	MCC[2]	MCC[1]	MCC[2]
SignalP 5.0	**0.972**	**0.948**	**0.958**	**0.965**	0.859	**0.889**
PRED-TAT	**0.972**	**0.948**	**0.958**	0.948	**0.903**	0.853

TatP	0.824	0.667	0.787	0.689	0.752	0.680
TATFIND	**0.972**	0.902	0.893	0.910	0.793	0.800

Supplementary Table 11. Benchmarking of Tat/SPI signal peptide detection predictions. The highest performance values have been highlighted in bold. 'MCC[1]' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Tat/SPI SPs and the negative dataset by TM+Globular proteins only; 'MCC[2]' refers to signal peptide vs non-signal peptide detection when the positive dataset is comprised by Tat/SPI SPs and the negative dataset by Sec/SPI SPs, Sec/SPII SPs, TM and Globular proteins.

Method	Archaea				Gram-negative bacteria				Gram-positive bacteria			
	0	±1	±2	±3	0	±1	±2	±3	0	±1	±2	±3
CS recall												
SignalP 5.0	**0.591**	**0.636**	**0.727**	**0.773**	0.684	0.724	0.745	0.776	0.595	**0.622**	**0.676**	**0.689**
PRED-TAT	0.500	0.545	0.636	0.636	**0.735**	**0.755**	**0.776**	**0.806**	**0.622**	**0.622**	0.635	**0.689**
TatP	0.318	0.409	0.500	0.500	0.653	0.673	0.694	0.704	0.446	0.473	0.514	0.581
TATFIND	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
CS precision												
SignalP 5.0	**0.591**	**0.636**	**0.727**	**0.727**	0.698	**0.740**	**0.760**	0.760	**0.698**	**0.730**	**0.794**	**0.794**
PRED-TAT	0.500	0.545	0.636	0.636	**0.713**	0.733	0.752	**0.782**	0.590	0.590	0.603	0.654
TatP	0.269	0.346	0.423	0.423	0.427	0.440	0.453	0.460	0.355	0.376	0.409	0.462
TATFIND	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.

Supplementary Table 12. Benchmarking of Tat/SPI signal peptide cleavage site predictions, measured as recall and precision. The highest performance values have been highlighted in bold.