Supplementary Data

A list of plastid assigned, manually curated gene models was prepared from the first release of the *Phaeodactylum tricornutum* genome v1.0[1] (Bowler et al., in preparation). Predicted signal peptides and signal peptide cleavage sites were displayed according to the results of SignalP[2] (Bendtsen et al., 2004), taking into account the different prediction methods Neuronal networks (NN) (Nielsen et al., 1997) or Hidden Markov models (HMM) (Nielsen and Krogh, 1998) (supplementary Fig. 6). The results were used to prepare sequence logos (Schneider and Stephens, 1990) using the WebLogo server[3] (Crooks et al., 2004) to illustrate the predictions of the different prediction algorithms with predictions combining computation and manual correction (Fig. 1).

References

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–95.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–90.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 8: 581–99.

Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 6: 122–30.

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097–100.

[1] `http://genome.jgi-psf.org/Phatr1/Phatr1.home.html`
[2] `http://www.cbs.dtu.dk/services/SignalP/`
[3] `http://weblogo.berkeley.edu/`

Name	Protein Id	Sequence	NN Ymax	HMM Cmax
(a)		01 05 10 15 20 25 30 35 40 45 50		
Albino3	28289	MARTRHGARLARCAFVWLWVASSTTAFTTTSPRLAAHFRSASRTQRTTTT	**0.758***	0.720*
AtpC	21239	MRSFCIAALLAVASAFTTQPTSFTVKTANVGERASGVFPEQSSAHRTRKA	**0.701***	0.595*
FbaC1	22933	MKLSTAALFFIPAVVAFAPPQAAFRSNPALFATETAAEKTTFSKMPASVK	**0.692***	0.549*
FCP	48997	MAKFSLAILAALVATASAFVASPTTSSASTALRATQEGVWDPLGLMTLGT	**0.640***	0.637*
FCP	50582	MKSSAVLALAMAGSTAAFAPTSSTQASTSSTSLQAAMPDRLWNTMVDKTE	**0.436***	0.259*
FCP	51914	MKITALCLTALVSASHAFAPSTPSSASSSARALSTESTSDVPLLQIKEKV	**0.517***	0.429*
FCP	49699	MKYRTSILFSALATSATAFAPTQSITRTLTQTNMADFNKDDFLSFKEKDK	**0.533***	0.528*
FCP	62104	MKLFTIFLPLVLVGTAAGFASGPFSKKASPSPEVSIESMPGIVAPTGFFD	**0.605***	0.555*
FCP	47005	MMKLALIASLVAGAAAFAPASKQASSSALKAFENEAGVIQPTGFFDPFGL	**0.664***	0.632*
Hcf136	48926	MKFSLTLVAVFVSGSAAFVPNSFSATARTDALRMTSPGYEIESESSGARR	**0.717***	0.690*
OEE3	48703	MKLALVFSLFATAAAFSQEASRREALTKGAAAFGAAFLPVAANAAVGESP	**0.575***	0.444*
petC	47867	MKIIPTVTSLALLAISIRAFTPTLAPRHTHASSKTWAASLEEPAYEGTID	**0.641***	0.529*
PsbM	50581	MKSFQLLTLFALIAASLAFAPNQAPQQVAKAAFTKAAGALPVAALAAPAF	**0.618***	0.586*
putative Tic62	35796	MTVSTYFIFFFTLGRGTAAFSPSLGASLTRSTAVPRFALSPSSNDDGIHS	**0.728***	0.533*
recA	61939	MMHRKIALVAGLWFCLLGSSCHGFGLLGGAAHSRTVRWPQQWAVSSPIGP	**0.610***	0.527*
Sep	62042	MHHFAKVILLVCVAMLAAVYTEAFATPSRSVQSSAVSITNSMPFRTSALN	**0.688***	0.682*
Trx-f	36974	MMQQQQQQQHAAQRRAGVGTILLSLWFVPAATAFAPSTPLAFRTQTPVVT	**0.739***	0.738*
Trx-m	62165	MYRAQQYCRSRTLFIHYAIVLVLTRYCSAFCSLEPVLRPSRWISNRSKSL	**0.706***	0.472*
GLNA2	14218	MKLNIAAIALFAASASAFAPRFASPRSHATVLSAVLEERTGQSQLDPAVI	**0.515***	**0.515***
ADK	45920	MLRSLAFTASVALLFSLDPLLAFAPIRTTTSVASPSGVSIRAQTGDQLFA	0.520*	**0.767***
CPE	46734	MKLPWLGPSAAALLLSSQTMAFLPSSLPSQSARNAGVTLQEKPSASDSSF	0.624*	**0.793***
Elip	18985	MAPLRTTFALLLSLVSASAFAPVQNVARKQTSVSAFKIDPQLYDDAVSDW	0.599*	**0.725***
Enolase	44169	MLFKPSTLLALFAVAGTTLAFAPRSTTTPLTSTTRGSASSSVTTLAMSGI	0.697*	**0.704***
FbaC2	14856	MKIAVVAFFVIAQCGAFAPAHYSRTVTSSTLLGAKEKGGTSKELDLPCAD	0.488*	**0.631***
FBPC1	45363	MEKWGFSRGQVPLLLSVIALSFVLLPTTNSFQTSTGQSQPQRLPASSASL	0.762*	**0.884***
FBPC2	41361	MKIALLPVVFSAISVRAFLPTRPSPATQYFRYGPRLATASLSQAAGAAVS	0.491*	**0.897***
FCP	22723	MKTSAIVAILAVSGASAFTPNTNAPQQLTKVGATAELDNMLGVDIETGKK	0.724*	**0.961***
FCP	46231	MKCIAAIALLATTASAFNAFGAAKKAAPKKPVFSIETIPGALAPVGIFDP	0.654*	**0.894***
Fcp related	62176	MKKILKRSTICCCLGLYTGSPCSYAFRPPPISEEIGCSPTLAIFLKEDCH	0.328*	**0.923***
Flavodoxin	23541	MNTQFVSALLLASAAITNGFAFVNTHRYTASTTALEAGVKIYYSSSTGNT	0.608*	**0.623***
FtrB	43703	MKVFLSFVALLVLSLTQAFMPVSKPSFGRVGGTVYMAKEMTPEEEEIAVE	0.799*	**0.985***
FtsZ	31286	MRISSKLMAVWLTTTVGIASAFQPATPSRIRQSPRTTVVLAMSDDWKPPA	0.729*	**0.798***
GLRXC2	43651	MTTNRVLFLAILALTCLRTRAFLPVTTFVSRTVAGSHGVHSLSMADMADD	0.715*	**0.927***
GLRXC3	41758	MAILSRTLVGVLTATFACIPHETNAFLKPHLGTVSSLVSIGEPSSTRLHI	0.448*	**0.767***
GLRXC4	41850	MKSFIACAFILASAAAFAPQPGSPVLRSTSALSASPSDDRSPNPIIKVMA	0.488*	**0.648***
Hmox	45923	MRNITKLRLLAVLAVAASPIRTSQAFAPSVSRKQTNCHKTPLTTSVSIKN	0.688*	**0.820***
HSP70D	44604	MRSQILLCFRILFLFLLPTTTIAFSTHPKFRARRKQPCASSQLAVTDKHE	0.917*	**0.923***
NiR	48100	MVQTLPVFRSLSVMAIGSFLLALGPSPSADAFAFSTRLPTQSLSLRSTST	0.363*	**0.442***
NTT1	50541	MIPTTSVASHRSVVLVGLACTTLLLAVLSPSSTEAFAPSAHRYSANQAVP	0.453*	**0.744***
PAP fibrillin II	45766	MPGRVVPLSLLAVVCASDLFRSTLAFHPKVSTSSRTIPPSTTQLHAFGFL	0.632*	**0.894***
PetJ	43773	MKLAVIATLLATASAFSIQAEFSKVAKGAAAVGVGAVIAAAPALAGDVGA	0.789*	**0.839***
PPdK	49184	MKFSSAAATTVGLLLSGHAPMIFSFVTPPSRFASGHQASGSERSIISHST	0.433*	**0.823***
Prk	43777	MKFAVFASLTATAAAFAPTAFVPSNLRGVAPSASSLNMALKEGQTPIIIG	0.502*	**0.539***
PtCA1	45670	MKFLSASIALLACATSVEAFNANKAFRFGAKAMPEVSSESATSALSAGGA	0.792*	**0.951***
Rpe	24020	MKFTIVSLAAVVASASAFAPATKSVRSVSALNVWGDKDYLIAPSILSADF	0.484*	**0.602***
RPI	47776	MRLTAGASILLASSAHAFTNPAFFPRTATFASTSSYTSALILKMGVDQDE	0.550*	**0.890***
		01 05 10 15 20 25 30 35 40 45 50	highest value **bold**	

(continuation on next page)

BOLD: prediction by NN, UNDERLINED: prediction by HMM, GREY: conserved motif at signal peptide cleavage site, BLACK: hydrophobic (ACFGILMPVWY), GREEN: hydrophilic (NQST), BLUE: basic (HKR), RED: acidic (DE)
* predicted signal peptide cleavage site coincides with conserved motif

supplementary **Fig. 6**, caption on page 4

Name	Protein Id	Sequence	NN Ymax	HMM Cmax
(a, continuation**)**		01 05 10 15 20 25 30 35 40 45 50		
secA	43680	MRLTLTIQVALTLLLLPSSTVWAFRTSPATQVFSRSRPMRSVASSSSSFS	0.803*	**0.978***
SRP54	46102	MRLQSGCVLTLLAATFYPSTQAFSIFSVGPSFASSSFASRERSDVSRKSY	0.624*	**0.947***
sufD	50465	MKFTSVTLACFFFVESSGVSAFTFSTPTQRQRSIPLPIVSSSRRTRSALH	0.690*	**0.802***
TAL	51830	MKSLVPLAFVLATSSGFAPREHHYYRPNQPARTNALVLQAESTAAVLASA	0.385*	**0.795***
Tic55	62146	MALRRSISRLAMVYLVTLCLKTAFVSAFSSTNTPTTPTSSVQQRSANQQA	0.674*	**0.702***
Tkt	43792	MKFSSIALATVFFAAKAQGFTGSSIKPRFGVQAHSVLKMSTATETDKMAA	0.595*	**0.782***
Tpt3	18575	MMMKRALVVLTLSVGVSARASAFAPGAAVKNHAGATQSAIHKQTPFPTTE	0.601*	**0.779***
VDE-like2	62162	MKLHRKGRYRLLVTAVLLGTVCSFVPENLRSGSVRIPRKNANAGSVPGTH	0.748*	**0.858***
Tic32	11808	MKYSILGNILLSLGLVLETTKAWSVPPPPQTSPAPRESSGDSGTSTPVSP	0.741*	**0.538***
FBPC4	27440	MGRGVIIFCVKNFAVWLLIITSAVSIQAWIPLPLSATVKARIDSTTLFFS	0.523*	**0.556***
FSA	48116	MAISRSRRRSNGLGIVLVWTIFISAVWGWTPPLRGSSRLFLDSADPNVWN	0.825*	**0.886***
GLRXC1	44755	MKLVSRSFCTLAWTCGAAQAWSVAPRTVARNRPAWVAARHPHTPTACAFS	0.537*	**0.659***
Hlip2	50908	MRWTCAFLWCVVVPTLHAWVPSTTNPASRIGTRRWEALGDRELEEPRMNP	0.847*	**0.921***
PAP fibrillin I	49299	MMREQRMLAILWGAGLWFGGSGVHAWQFPNLFTVPIQPSQKFSQGSTAKS	0.601*	**0.951***
GAPDH	29119	MKFSAATFAALVGSAAAYSSSSFTGSALKSSASNDASMSMATGMGVNGFG	0.439*	**0.390***
VDE-like1	62158	MRFAWVVAAGVVLTTTTQALVPLDCTGMGETRTSGIRPIRGLESNMARYA	0.760*	**0.858***
		01 05 10 15 20 25 30 35 40 45 50		
(b)		01 05 10 15 20 25 30 35 40 45 50		
FbaC5	44147	MRFSLQSSLAVLLVLQASHAAAFSAPVSSSNGKNGIRSFAPLSMSLDKYA	**0.709***	0.347
FCP	43697	MKFAATILALIGSAAAFAPAQTSRASTSLQYAKEDLVGAIPPVGFFDPLG	**0.561***	0.312
FCP	24872	MKTAVIASLIAGAAAFAPAKNAARTSVATNMAFEDELGAQPPLGFFDPLG	**0.448***	0.330
PGM	43607	MKGYLFATWACLTISSNASTEAFAHRGPRAPCGLHASKLKTMDSGKLVDV	**0.380***	0.375
DDE	62131	MKFLGVTSLLCLWSVVNRENVSEAFAPRHQSLSRPSSRTTSAFSRAPILS	0.614	**0.737***
PsbO	28739	MKFTAACSIALAASASAFAPIPSVSRTTDLSMSLQKDLANVGKVAAAGAL	0.322	**0.535***
FCP	22033	MKSIIFASLLTSAAAFAPASSSTTRTATPTALNEEFCRGYVGGESVEPMF	0.458*	**0.506**
FCP	18759	MKFAILASMLSAAAAFAPASQGAGKASVALNAEKSPAMPFLPYPENLKGY	0.400*	**0.563**
FCP	18195	MMRSTILAALLASAAAFAPASMQSQRAGSVSLNAEEMSKSIPFLVKPDKL	0.459*	**0.629**
Hlip1	62173	MLTLLILMTRLSLSESFGVTTPRIFRPAPCRHTHRPLKTTLRHSTLPSET	0.542*	**0.611**
		01 05 10 15 20 25 30 35 40 45 50		
(c)		01 05 10 15 20 25 30 35 40 45 50		
FCP	47183	MKYAVFASLLASAAAFAPAAKPAASTSALNAEMSKSMPFLTAPKNTGGYV	**0.467**	0.430
FCP	44994	MKLSLAILALCASTNAAFAPSVSQRTPRDLAGVVAPTGFFDPAGFAARAD	**0.676**	0.672
FCP	24752	MKFAVFAFLLASAAAFAPAQQSARTSVATNMAFENEIGAQQPLGYWDPLG	**0.585**	0.330
Tpt1	25658	MKVATTLTLAFICCASAFGLNGQTTSVMKKVGFDAGSKPMVQAIDVQGNR	**0.673**	0.511
FBPC3	16659	MFILKSPALWLLLYPVVAFTAARANSIRPAAALSVFDLSSVEAVPSRKTK	0.587	**0.832**
HY2	34372	MKLVPAWTMMTRNAIFSARNPSHFLEVTALFCILIASGRGARTNTAFVNP	0.492	**0.758**
Tic110	62144	MKFTGVALAVSLAQQQALSNTGGGIVGAYTVSSPSFFTPKSFGSSFVRGP	0.424	**0.535**
FCP	44082	MKLFGVSVVIVVVATAVSLSDSVAAFAMAHGGSHKLSNTALRVTGFEDEL	**0.500**	0.231
chlD	43587	MVSSSKSTWMMAGACLILMAFQVQSFTFVPATRATSTVKRVAPAFMSAVA	**0.635**	0.486
		01 05 10 15 20 25 30 35 40 45 50	highest value **bold**	

BOLD: prediction by NN, UNDERLINED: prediction by HMM, GREY: conserved motif at signal peptide cleavage site, BLACK: hydrophobic (ACFGILMPVWY), GREEN: hydrophilic (NQST), BLUE: basic (HKR), RED: acidic (DE)
* predicted signal peptide cleavage site coincides with conserved motif

supplementary **Fig. 6**, caption on page 4

supplementary **Fig. 6** (pages 2 and 3): List of 81 plastid assigned, manually curated gene models from the *Phaeodactylum tricornutum* genome used to construct the sequence logos (Fig. 1) comparing SignalP's Neuronal networks (NN) and Hidden Markov models (HMM). **(a)** sequences with identical predictions between NN and HMM and coincidence with an "ASAFAP"-motif (grey) in both prediction models (62 sequences). **(b)** sequences with differing predictions from NN and HMM but coincidence with an "ASAFAP"-motif in one of the prediction models (10 sequences). **(c)** sequences with no coincidence between the predicted cleavage sites of both models and an "ASAFAP"-motif (9 sequences). Protein IDs refer to the first release of the *P. tricornutum* genome v1.0.