PREDALGO, a new multi-subcellular localization prediction tool dedicated to Algae

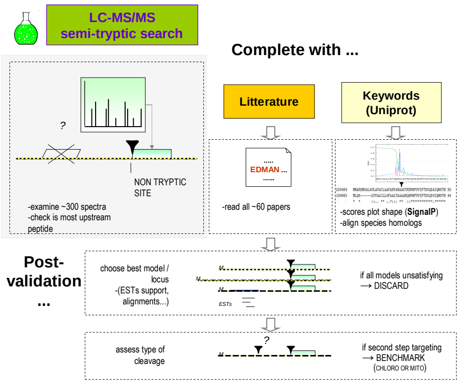
Tardif, M., Atteia, A., Vallon, O., Specht, M., Cogne, G., Rolland, N., Brugière, S., Hippler, M., Ferro, M., Bruley, C., Peltier, G. and Cournac, L.*

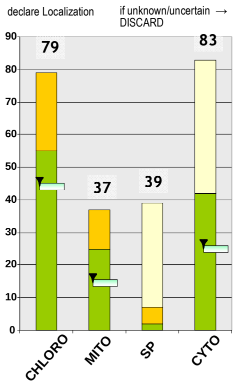
EDyP / CEA Grenoble

Introduction

Despite of organella-specific proteomics efforts, assigning a subcellular localization to proteins still need to be assisted by **localization-predicting softwares**. However, the existing Plant-dedicated tools tend to mispredict algal chloroplast-localized proteins to mitochondria. We thus developped a tool **adapted to Algae**, using the feature of a cleavable N-terminal peptide present in proteins targeted to either of the mitochondria, chloroplast or endoplastic reticulum compartments. PredAlgo was implemented using training sets of ***Chlamydomonas reinhardtii*** proteins, which principaly consisted in **new N-terminal peptides of mature proteins identified from screening MS/MS data** of the recently published mitochondria and chloroplast proteomics surveys (1,2).

1 Identify maturation sites → Training sets



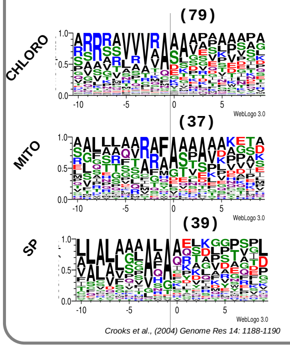


CRITERIA for peptide match

- **non tryptic cleavage at Nter**
- **most upstream valid match on the protein**
- **within the first 150 amino acids**

The training Sets were finally constituted of candidates with MS/MS (green), Edman (orange) or SignalP prediction (yellow) **evidence of cleavage site**. The MS/MS fraction of the "CYTO" set was constituted by proteins for which the N-terminal of the sequence in the database was matched by a proteomic peptide, supporting the **absence of cleavage.**

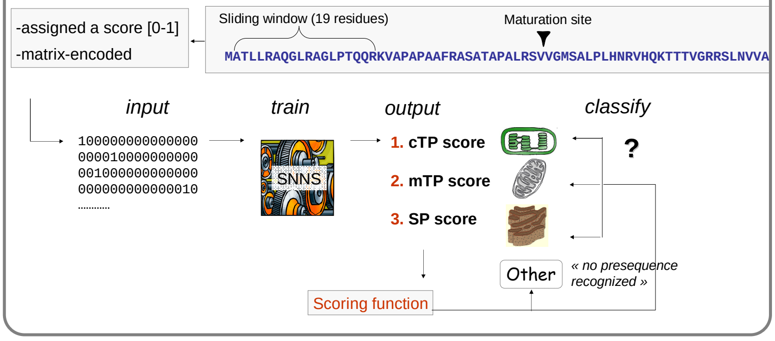
2 Sequences around maturation sites



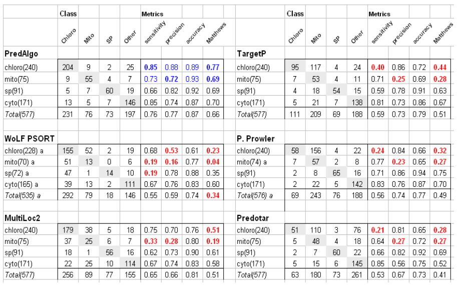
Crooks et al., (2004) Genome Res 14: 1188-1190

The sequences surrounding the cleavage site were converted into WebLogos. In the "chloro" and "mito" logos, negative charges predominated upstream of cleavage while practically no positively charged (acidic) residues were present. Too weak differential patterns between chloro and mito, make the application of **Artificial Neural Networks** a solution of choice.

3 Implementation of PredAlgo (Neural Network)



4 Performance evaluation in *C. reinhardtii*



	Class				Metrics			
	Chloro	Mito	SP	Other	sensitivity	precision	accuracy	Matthews
PredAlgo								
chloro(240)	204	9	2	25	**0.85**	0.88	0.89	**0.77**
mito(75)	9	55	4	7	0.73	**0.72**	0.93	**0.69**
sp(91)	5	7	60	19	0.66	0.82	0.92	0.69
cyto(171)	13	5	7	146	0.85	0.74	0.87	0.70
Total(577)	231	76	73	197	0.76	0.77	0.87	0.66
WoLF PSORT								
chloro(228) a	155	52	2	19	0.68	**0.53**	0.61	**0.23**
mito(70) a	51	13	0	6	**0.19**	**0.16**	0.77	**0.04**
sp(72) a	47	1	14	10	**0.19**	0.78	0.88	0.35
cyto(165) a	39	13	2	111	0.67	0.76	0.83	0.60
Total(535) a	292	79	18	146	0.55	0.59	0.74	**0.34**
MultiLoc2								
chloro(240)	179	38	5	18	0.75	0.70	0.76	**0.51**
mito(75)	37	25	6	7	**0.33**	**0.28**	0.80	**0.19**
sp(91)	18	1	56	16	0.62	0.73	0.90	0.61
cyto(171)	22	25	10	114	0.67	0.74	0.83	0.58
Total(577)	256	89	77	155	0.65	0.66	0.81	0.51

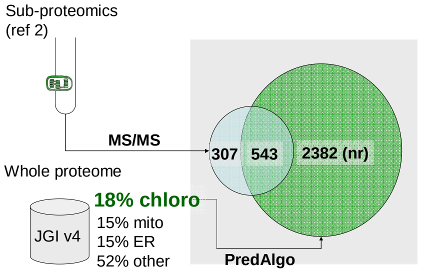
	Class				Metrics			
	Chloro	Mito	SP	Other	sensitivity	precision	accuracy	Matthews
TargetP								
chloro(240)	95	117	4	24	**0.40**	0.86	0.72	**0.44**
mito(75)	7	53	4	11	0.71	**0.25**	0.69	**0.28**
sp(91)	4	18	54	15	0.59	0.78	0.91	0.63
cyto(171)	5	21	7	138	0.81	0.73	0.86	0.67
Total(577)	111	209	69	188	0.59	0.73	0.79	0.51
P. Prowler								
chloro(240)	58	156	4	22	**0.24**	0.84	0.66	**0.32**
mito(74) a	7	57	2	8	0.77	**0.23**	0.65	**0.27**
sp(91)	2	8	65	16	0.71	0.86	0.94	0.75
cyto(171)	2	22	5	142	0.83	0.76	0.87	0.70
Total(576) a	69	243	76	188	0.56	0.74	0.77	0.49
Predotar								
chloro(240)	51	110	3	76	**0.21**	0.81	0.65	**0.28**
mito(75)	5	48	4	18	0.64	**0.27**	0.72	**0.27**
sp(91)	2	7	60	22	0.66	0.82	0.92	0.69
cyto(171)	5	15	6	145	0.85	0.56	0.75	0.52
Total(577)	63	180	73	261	0.53	0.67	0.73	0.41

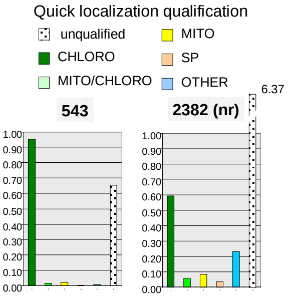
$$\text{Accuracy} = \frac{TP+TN}{(TP+FN)+(TN+FP)}$$

$$\text{Matthews Corr. Coeff.} = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

PredAlgo was assessed on independant **Benchmark sets** of *C. reinhardtii* sequences, and compared to existing softwares. PredAlgo circumvented the weakness of most current tools as it generated outputs with **highly improved discrimination between the chloroplast and mitochondria**, resulting in 85% sensitivity for the chloroplast, 72% precision for the mitochondria. The precision calculated for the mitochondrion is not to be taken as an absolute criteria due to the small size of the "mito" set relatively to other sets. The discriminating power between compartments is better reflected by the MCC values which were fairly improved again for "Chloro" and "Mito" outputs (0.77 and 0.69).

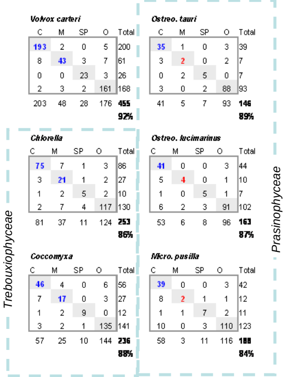
5 First application : the *C. reinhardtii* chloroplast proteome





PredAlgo predicted a chloroplast localization for 64% (543 / 850) of proteins previously identifed as chloroplastic by MS/MS (2) (while TargetP recovered only 19%, not shown). The core of a *C. reinhardtii* chloroplast could be delimitated by the common set of 543 proteins. In addition, PredAlgo predicted 2382 (non redundant) proteins the vast majority has no obvious localization but which represent **potentially "new" chloroplastic proteins**.

6 Performances in other Green Algae



Trebouxiophyceae

Volvox carteri

C	M	SP	O	Total
193	2	0	5	200
8	43	3	7	61
0	0	23	3	26
2	3	2	161	168
203	48	28	176	**455**
				92%

Chlorella

C	M	SP	O	Total
75	7	1	3	86
3	21	1	2	27
1	2	5	2	10
2	7	4	117	130
81	37	11	124	**253**
				86%

Coccomyxa

C	M	SP	O	Total
46	4	0	6	56
7	17	0	3	27
1	2	9	0	12
3	2	1	135	141
57	25	10	144	**236**
				88%

Prasinophyceae

Ostreo. tauri

C	M	SP	O	Total
35	1	0	3	39
3	2	0	2	7
0	2	5	0	7
3	0	2	88	93
41	5	7	93	**146**
				89%

Ostreo. lucimarinus

C	M	SP	O	Total
41	0	0	3	44
5	4	0	1	10
1	0	5	1	7
6	2	3	91	102
53	6	8	96	**163**
				87%

Micro. pusilla

C	M	SP	O	Total
39	0	0	3	42
8	2	1	1	12
1	1	7	2	11
10	0	3	110	123
58	3	11	116	**188**
				84%

To assess whether PredAlgo is **a suitable tool for green algae in general**, orthologs pairs from whole protomes were computed as Blast Best Reciprocal Hits between *C. reinhardtii* and each of six Chlorophyta algal species. PredAlgo was then run on the algal proteomes and the predicted localization was compared to that of the *Chlamydomonas* ortholog. The counts shown here were restricted to sets of pairs for which the localization of the Chlamydomonas protein is known with certainty, i.e. to the Training and Benchmark sets.

i) Predictions in *Volvox carteri* correlated very well
ii) 'chloroplast' recall of at least 80% in all six algae
iii) Except for *Volvox*, consistency in 'mito' predictions was moderate (*Chlorella*, *Coccomyxa*) or not evaluable (*Ostreoccocus tauri*, *O. lucimarinus*, *Micromonas pusilla*)

Conclusion

The identification of N-terminal peptides of mainly mitochondria- and chloroplast-targeted proteins via **MS/MS "semi-trypsic" strategies** brought up sufficient experimental data to implement a localization predictor specific to Algae. **PredAlgo** appeared as a **first position software to be used in *C. reinhardtii*** and its closest homolog ***V. carteri*** with a great discrimination capacity between chloroplast and mitochondria. It is also well-suited in **more distant algal species** if dedicated at chloroplast prediction recovery.

Noticeably, in organisms tested, a fair number of predictions errors were due to erroneous gene models. We assume that PredAlgo would perform even better as predicted models will become more accurate.

Perspectives

❑ **Localization of *C. reinhardtii* Central metabolic enzymes** (G. Cogne & co)
❑ **Mitochondria proteome refining in *C. reinhardtii*** (A. Atteia & co)
❑ ...

References

1) Atteia A, et al. N (2009) Mol Biol Evol 26: 1533-1548
2) Terashima M, et al. (2010) Mol Cell Proteomics 9: 1514-1532

Abbreviations cTP, chloroplast transit peptide; mTP, mitochondria transit peptide; SP, signal peptide
JGI v4, Joint Genome Institute, protein models version 4

We **acknowledge** Annie Adrait, Laurent Kellner for spectrometric and biochemical expertise. We thanks Charles Hauser, Erik Hom and Mario Stanke for having made available Chlamydomonas-related ESTs and annotations files.