

On the road towards the annotated *Physcomitrella patens* genome

Andreas Zimmer*, Daniel Lang, Jan Mitschke, Mark von Stackelberg, Ralf Reski and Stefan A. Rensing

Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaezlestr. 1, 79104 Freiburg, Germany,

*andreas.zimmer@biologie.uni-freiburg.de, fon +49 761 203 6974 fax +49 761 203 6945

<http://www.plant-biotech.net>, <http://www.cosmoss.org>

Physcomitrella subcellular protein localization prediction

We have collected and curated 102 *Physcomitrella patens* (Uniprot) proteins to be used as a starting point to compare protein target prediction tools with reference to *Physcomitrella*. Our initial analysis revealed that e.g. the commonly used tool targetP predicts the localization for only ~40% of the *Physcomitrella* proteins correctly. In order to improve the localization prediction for the forthcoming annotation of the *Physcomitrella* genome, we are cooperating with the Kohlbacher group to “mossify” their localization prediction tool MultiLoc.

Dually targeted proteins are proteins, which are transcribed and translated from one gene and targeted to more than one location. The underlying mechanisms are not well understood and only few proteins have been demonstrated to be dually targeted. For *Physcomitrella patens* three dually targeted proteins are known. Whereas subcellular prediction for proteins targeted to a single localization is available, there is no reliable tool yet to predict dually targeted proteins. Therefore we are developing methods to predict dually targeted proteins.

Genetic map

Genetic maps are crucial for the scaffolding of the genome assembly following sequencing. At present we are developing a SSR based genetic map for *Physcomitrella patens*. To select informative markers we have analyzed more than 1,200 EST-derived (from non-redundant 48,944 virtual transcripts) microsatellites or simple sequence repeats (SSR) till now. A linkage map has been established with MAPMAKER as well as with JoinMap. 228 markers were grouped into 45, respectively 46, linkage groups spanning 1342 cM, respectively 1697 cM. We have developed a SSR marker mapping pipeline, which extracts the analyzed microsatellite markers from the transcript sequences and reliably maps them onto genome scaffolds using spliced alignments. We are currently developing and establishing visualization of marker localizations along the genome scaffolds.

Gene Structure Prediction

An important prerequisite for gene structure prediction is a reliable, organism-specific trainingset. We have collected and manually curated all available *Physcomitrella patens* genes to derive coding, non-coding and intergenic regions that can be used for statistical parameter estimation to filter the EST evidence which is used for *ab initio* gene prediction. We will present the typical prototype of a *Physcomitrella* gene as it can be derived from the training set. To evaluate the dataset and find suitable spliced alignment software for the annotation of the genome, we compared several spliced alignment tools. In addition we are building species-specific Bayesian Splice Site Models (BSSM) to be used with the GenomeThreader software. For the gene structure prediction itself we are cooperating with the van de Peer group who are using the Eugène pipeline.

Transposons

To further support iterative assembly and genome annotation we are detecting LTR retrotransposons in the *Physcomitrella* genome *in silico*. We will present initial results on this topic.

Re-Launch of cosmoss.org

The *Physcomitrella* transcriptome resource, www.cosmoss.org, has been expanded to envelop new and future tools.

Financial support by the DFG (Re 837/10-1) ist gratefully acknowledged.