JOURNAL OF PROTEOMICS XX (2011) XXX–XXX

available at www.sciencedirect.com

www.elsevier.com/locate/jprot

Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods

Noor Remmerie[a,b], *Thomas De Vijlder*[a,b], *Dirk Valkenborg*[a,c,d], *Kris Laukens*[e,f], *Koen Smets*[g], *Jilles Vreeken*[g], *Inge Mertens*[a,c], *Sebastien Carpentier*[h], *Bart Panis*[h], *Geert De Jaeger*[i,j], *Ronny Blust*[k], *Els Prinsen*[b], *Erwin Witters*[a,c,k,*]

[a]*Center for Proteomics (CFP), Groenenborgerlaan 171, B-2020 Antwerp, Belgium*
[b]*Laboratory of Plant Growth and Development, Department of Biology, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium*
[c]*Vlaamse Instelling voor Technologisch Onderzoek (VITO), Boeretang 200, B-2400 Mol, Belgium*
[d]*Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Agoralaan 1, B-3590 Diepenbeek, Belgium*
[e]*Intelligent Systems Laboratory, Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium*
[f]*Biomedical Informatics Research Center Antwerp (Biomina), Wilrijkstraat 10, B-2650 Edegem, Belgium*
[g]*Advanced Database Research and Modelling, Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium*
[h]*Division of Crop Biotechnics, Department of Biosystems, K.U.Leuven, Kasteelpark Arenberg 13, B-3001 Heverlee, Belgium*
[i]*Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium*
[j]*Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium*
[k]*Laboratory for Ecophysiology, Biochemistry and Toxicology, Department of Biology, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium*

ARTICLE INFO

Keywords:
Blue native gel electrophoresis
Liquid chromatography
Clustering
Data mining
Nicotiana tabacum cv. Bright Yellow-2
Protein complexes

ABSTRACT

To understand physiological processes, insight into protein complexes is very important. Through a combination of blue native gel electrophoresis and LC–MS/MS, we were able to isolate protein complexes and identify their potential subunits from *Nicotiana tabacum* cv. Bright Yellow-2. For this purpose, a bioanalytical approach was used that works without *a priori* knowledge of the interacting proteins. Different clustering methods (e.g., k-means and hierarchical clustering) and a biclustering approach were evaluated according to their ability to group proteins by their migration profile and to correlate the proteins to a specific complex. The biclustering approach was identified as a very powerful tool for the exploration of protein complexes of whole cell lysates since it allows for the promiscuous nature of proteins. Furthermore, it searches for associations between proteins that co-occur frequently throughout the BN gel, which increases the confidence of the putative associations between co-migrating proteins. The statistical significance and biological relevance of the profile clusters were verified using functional gene ontology annotation.

Abbreviations: ATP, adenosine-5′-triphosphate; BiFC, bimolecular fluorescence complementation; BN, blue native; BY-2, *Nicotiana tabacum* cv. Bright Yellow-2; COS1, coronatine insensitive1 suppressor; FDR, false discovery rate; GO, gene ontology; HRC, hierarchical clustering; HMW, high molecular weight; KMC, k-means clustering; LMW, low molecular weight; ME, malic enzyme; MW, molecular weight; NAD(P), nicotinamide adenine dinucleotide (phosphate); SM, similarity metric; UDP, uridine diphosphate.

* *Corresponding author at:* University of Antwerp, Center for Proteomics, Groenenborgerlaan 171, B-2020 Antwerp, Belgium. Tel.: +32 32653594; fax: +32 32653697.
E-mail address: Erwin.witters@ua.ac.be (E. Witters).

1874-3919/$ – see front matter
doi:10.1016/j.jprot.2011.03.023

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

The proof of concept for identifying protein complexes by our BN PAGE/LC–MS/MS approach is provided through the analysis of known protein complexes. Both well characterized long-lived protein complexes as well as potential temporary sequential multi-enzyme complexes were characterized.

1. Introduction

Protein complexes play a critical role in many biological processes. Most proteins are, at some time point in the lifespan of the cell, involved in complex formation with multiple protein interaction partners [1]. Identifying the component proteins in a protein complex is an important step towards the understanding of the complex and in elucidating the related biological activities. Complex formation is of utmost importance in plants, as illustrated for their most typical physiological processes such as photosynthesis [2], cell wall growth [3], and phytohormone sensing [4]. To date protein–protein interactions are widely studied by techniques such as tandem affinity purification [5,6], yeast two hybrid studies [7], co-immunoprecipitation [8], BiFC [9], and through *in silico* prediction [10]. An alternate way to study protein–protein interactions, is to define all protein complexes within the cell. To this end, protein complexes within plant models have been studied through biochemical approaches, including zone gradient centrifugal sedimentation [11], and native chromatography, or combinations thereof [12]. In this study, blue native gel electrophoresis (BN PAGE) was used, since it allows for medium to high throughput screening of protein complexes within whole plant cell lysates [13]. The technique is well established for the separation of both soluble and membrane-bound protein complexes [14–16], and both direct and indirect protein–protein interactions can be elucidated in one single experiment. It thus continues to gain interest from the proteomics community [17,18]. The general workflow comprises native separation by BN PAGE followed by a denaturing second dimension SDS PAGE—in which each complex is dissected in its individual components. An alternative for the denaturing SDS PAGE step is LC–MS/MS, which allows for a rapid identification of all proteins within each gel slice and for protein profiling across the BN gel [19]. Wessels et al. [20] suggested that potentially interacting proteins can be identified by searching for similar protein profiles after BN PAGE separation, which has also been shown by Helbig et al. [21]. Protein correlation profiling permits the analysis of multiprotein complexes that can be enriched by fractionation but not purified to homogeneity [22]. A promising group of methods to find correlations between co-migration proteins is cluster analysis. These are statistical methods, which have been successfully applied on gene expression data [23–25], and reports of their implementation to reveal putative protein complexes are rapidly emerging [26–30]. Several clustering methods (e.g., hierarchical clustering, k-means) can be used to analyze protein interaction data. To detect co-migrating unrelated proteins, a functional gene ontology (GO) annotation is often performed [31]. Since proteins within the same protein complex are generally aggregated to take part within a similar biological process, the functional coherence of a cluster can be used to indicate its tendency to be a genuine complex.

Previously two-dimensional BN/SDS PAGE has been used by the authors to unravel protein complexes from whole cell lysates of *Nicotiana tabacum* cv. Bright Yellow-2 cell cultures (BY-2). While not a model plant for genome study purposes, yet it is an important model system to study cell physiology, hormone signaling, cell cycle, cell growth and stress situations [32]. Here, BN PAGE gave indications about the oligomerization state of several tobacco proteins and revealed potentially novel protein–protein interactions [13]. In this follow-up study a combination between 1D BN PAGE and LC–MS/MS was used as a semi-high throughput strategy to create a 'complexome' of BY-2 cells. First, the mass spectrometric identification method was optimized. Since tobacco BY-2 is still a mostly unsequenced and badly annotated plant system, identification of the proteins and their interactions relies on cross-species identification based on homology and orthology [33]. To enhance peptide identification, multiple search engines were employed [34]. Secondly, to reveal candidate interacting proteins, proteins were clustered according to their migration profile and functional annotation. In this study, popular clustering methods like hierarchical clustering and k-means were compared to a modern biclustering technique based on itemset mining. In what follows, we denote the k-means and hierarchical clustering methods as the classical approaches. To evaluate and compare the outcome of these clustering methods, known protein complexes such as the 20S proteasome and 26S proteasome were used as benchmarks.

2. Materials and methods

2.1. Chemicals and material

Unless otherwise indicated, all biochemicals and kits were from Sigma (St Louis, MO, USA)) or GE Healthcare (Uppsala, Sweden).

2.2. Isolation of protein complexes from BY-2 cell suspension

BY-2 cell suspensions were cultivated as published [32]. The stationary culture was sampled 7 days after subcultivation. The exponential culture was sampled 3 days after a 50 fold dilution of a stationary culture in fresh media. Throughout, approximately 340 mg cells were collected by vacuum paper filtration (Whatman nr.1) and ground in liquid nitrogen. To these samples, 1 mL ice cold extraction buffer containing 50 mM HEPES (pH 7.4), 30 mM potassium acetate, 5 mM EGTA, 2% (w/v) digitonin (high purity; Calbiochem, San Diego, CA, USA), 1% (v/v) plant protein protease inhibitor cocktail (P-9599), 10 μL Nuclease Mix, 1% (w/v) polyvinylpolypyrrolidone was

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

added during grinding in liquid nitrogen. To remove debris, lysates were centrifuged at 20,000 *g* at 4 °C for 45 min and supernatant was passed through a 0.22 μm filter. Protein concentration was determined using the Bradford Protein Assay (Bio-Rad Laboratories, Hercules, CA, USA).

2.3. BN PAGE

Precast NativePAGE™ NOVEX® BIS TRIS gels were used (Invitrogen Life Technologies, Carlsbad, CA, USA) (4%–16%, 8 cm × 8 cm × 0.1 mm) and after addition of 0.2% Coomassie and 1% digitonin to the samples, 15 μg/well of each sample was loaded into the gel. Gels were run at 4 °C (60 min at 150 V constant; 35 min at 250 V). Contrary to the anode buffer (2.5 mM BisTris, 2.5 mM Tricine, pH 6.8), the cathode buffer (2.5 mM BisTris, 2.5 mM Tricine, pH 6.8) contained 0.002% (w/v) Coomassie G-250. The molecular size of the complexes was estimated using a HMW native marker kit (Invitrogen Life Technologies, Carlsbad, CA, USA). BN gels were stained with colloidal CBB (SERVA Electrophoresis GmbH, Heidelberg, Germany) [35].

2.4. LC–MS/MS analysis

Above 70 kDa, the whole gel lane was cut in several (on average 47) equally sized slices of 1 mm and in-gel protein digestion was performed according to Shevchenko et al. [36]. To remove the Coomassie stain, gel slices were first washed in a fixing solution (50% methanol, 10% acetic acid) [37]. Subsequently, gel slices were extensively washed with water and shrunk with acetonitrile until gel plugs were completely white. Each gel slice was subjected to reduction/alkylation and the proteins digested with trypsin (Promega MS Gold, Madison, WI, USA). Trypsin was added to each gel slice (1:50 trypsin/proteins) and proteins were digested overnight at 37 °C. Supernatant was collected, vacuum-dried and resolubilized in water containing 0.5% formic acid. Nanoflow LC–MS/MS was performed on a NanoLC Ultra 2D system (Eksigent, Dublin, California, USA) connected to a LTQ Orbitrap Velos mass spectrometer (Thermo Electron, Bremen, Germany), equipped with a Triversa chip based electrospray source (Advion Biosystems, Ithaca, NY, USA) operating at 1.8 kV. A volume of 25 μL of each sample was loaded on a C_{18} precolumn (PepMap 100, 5 μm particles, 20 mm × 200 μm ID; Dionex, Sunnyvale, California, USA) at a flow rate of 6 μL/min in solvent A (2% acetonitrile and 0.1% formic acid in water). This trapping column was connected to an analytical C_{18} column (Acclaim Pepmap 100, 3 μm particles, 150 mm × 75 μm ID) (Dionex, Sunnyvale, California, USA) via a column switching setup. Swift elution of peptides was accomplished using an isocratic flow of solvent B (30% acetonitrile and 0.04% formic acid in water) at a flow rate of 500 nL/min for 30 min. MS spectra were recorded in the Orbitrap with a resolution of 60,000 (at m/z 400) to an AGC target setting of 500,000. The maximum injection time was set to 500 ms and lock mass was enabled (polysiloxane ion at m/z 445.12024). Collision induced dissociation MS/MS spectra were acquired in the LTQ Velos ion trap in data dependent mode selecting the 20 most abundant multiply charged precursor ions from the MS spectrum. The maximum injection time was set to 50 ms and AGC was set to 7500. Fragmentation was accomplished by CID wideband activation at a normalized collision energy of 35 and with an activation time of 30 ms. After MS/MS the precursor m/z's were excluded for 60 s. A permanent exclusion list containing m/z values for abundant trypsin and keratin peptides was used.

2.5. Protein identification

For protein identification, database searches were performed with Mascot (Version 2.2; Matrix science, London, England), Sequest (version 1.0.43. embedded in Proteome Discoverer 1.0; Thermo Fisher Scientific, San Jose, CA, USA) and Phenyx (Version 2.6; (GeneBio SA, Geneva, Switzerland) against the NCBInr database (version 7 July 2009; taxonomy: Viridiplantae; number of sequences: 700843). The following settings were used: the enzyme was trypsin and one miscleavage was allowed, cystein-carbamidomethylation was chosen as a fixed modification and methionine-oxidation as a variable one. The peptide tolerance was set at 3 ppm and the MS/MS tolerance at 0.8 Da. The Phenyx database has the ability of finding post-translational modifications (phosphorylation, biotinylation, …) in an extended search of the spectra. The results of all searches were combined by Scaffold (version Scaffold 3.00.03; Proteome Software Inc., Portland, OR, USA) with the following settings: a peptide confidence level of 95% as specified by the Peptide Prophet algorithm [38], a protein confidence level of 95% and the thresholds of each search engines separately. Mascot identifications required at least ion scores greater than 31. Phenyx identifications required at least z-scores greater than 5.0. Sequest identifications required at least deltaCn scores greater than 0.10 and XCorr scores greater than 2.8 for doubly, triply and quadruply charged peptides. Protein identifications were accepted if they were established at greater than 95% probability and contained at least 1 identified peptide. Protein probabilities were assigned by the Protein Prophet algorithm [39]. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped according to the principles of parsimony.

2.6. Data analysis

The HMW native marker (Invitrogen Life Technologies, Carlsbad, CA, USA), used during BN PAGE, was used as a benchmark to determine the molecular weight (MW) boundaries of each BN gel slice. Based on the known molecular mass of these protein standards, the migration distance of the marker proteins and the number of slices between two sequential protein standards, the migration distance of each gel slice was estimated. Protein candidates were excluded from the final list if the MW boundaries of their corresponding gel slice did not exceed the theoretical mass of the protein by more than a 1.5 fold granted 'mobility mismatch'. These proteins were considered to be monomeric.

2.7. BLAST search

Since identification was based upon cross-species identification, all identified proteins were blasted in batch against the *Arabidopsis thaliana* TAIR9 protein sequence database by using

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

command-line BLAST version blastall 2.2.17 [40] with the following arguments: –p blastp –m 8 –P 1 –A 0. To collapse redundant matches into a single entry, only the *A. thaliana* hit with the largest bitscore was retained for each identified protein (see supplemental data).

2.8. Protein complex clustering

To identify putative protein complexes, a comparison of protein profiles was performed by classical clustering methods throughout the whole BN lane with the statistical program MATLAB (version 7, Mathworks, Inc., USA). The premise is that proteins belonging to a particular complex occur in the same BN slice. The spectral count and the migration distance for each protein within the BN gel were used as input parameters for the profile clustering. In order to account for possible uncertainty in the data, cluster analysis was performed on normalized and non-normalized spectrum counts. This uncertainty is due to the cross-species setting that hampers protein identification. Poorly annotated peptides in the non-model organism influence the spectrum count of each protein and consequently affect the clustering that is based on this spectrum count. Normalization was done throughout the whole BN gel for each protein identification. Only protein groups that clustered together through all clustering methods or in multiple clusters in the biclustering approach were maintained. Two different classical clustering approaches were applied. We employed hierarchical clustering using Spearman rank correlation [41] to calculate the dissimilarity matrix. The unweighted average distance was used to calculate the linkage between two clusters in the agglomerative hierarchical cluster tree. The cut-off was empirically determined through visual inspection of the dendrogram, making a biologically relevant compromise between size and specificity of the clusters, which yielded 16 clusters. In addition, we employed a single-run of k-means clustering, using the same settings as for hierarchical clustering. In order to facilitate comparison with the results obtained through hierarchical clustering, k was chosen to partition the data into 16 clusters.

2.9. Protein tiling

In addition to classical clustering methods, we also applied a biclustering or "tiling" approach. Tiling is a data-mining technique that discovers that smallest group of itemsets, or clusters, which together optimally describe the data. To this end, we regard the data as a binary matrix, where a 1 indicates an interaction between the row and the column. In our setting, each row in the dataset represents a protein, whereas the columns correspond to the BN slices. The process of finding the best biclustering can be easily explained as trying to 'color' all the cells of the matrix containing a 1 as efficiently as possible. Starting off with the situation where none of the 1s have a color assigned, we iteratively find that itemset (combination of rows and columns) with which we can 'color' the largest 'uncolored' area of the data set, with the restriction that we may only color the selected 1s in a row if it contains 1s for all the selected columns (e.g., if we have selected proteins A, B and C, we may only color the corresponding 1s for lane X if proteins A, B, and C all occur in lane X). As such, the data set is being 'colored' with as large as possible patches of the same color and as few as possible colors in total. The area that such a tile colors (or, better, covers) is calculated by multiplying the number of elements in the itemset or cluster, by the number of rows in which all the elements of the itemset co-occur.

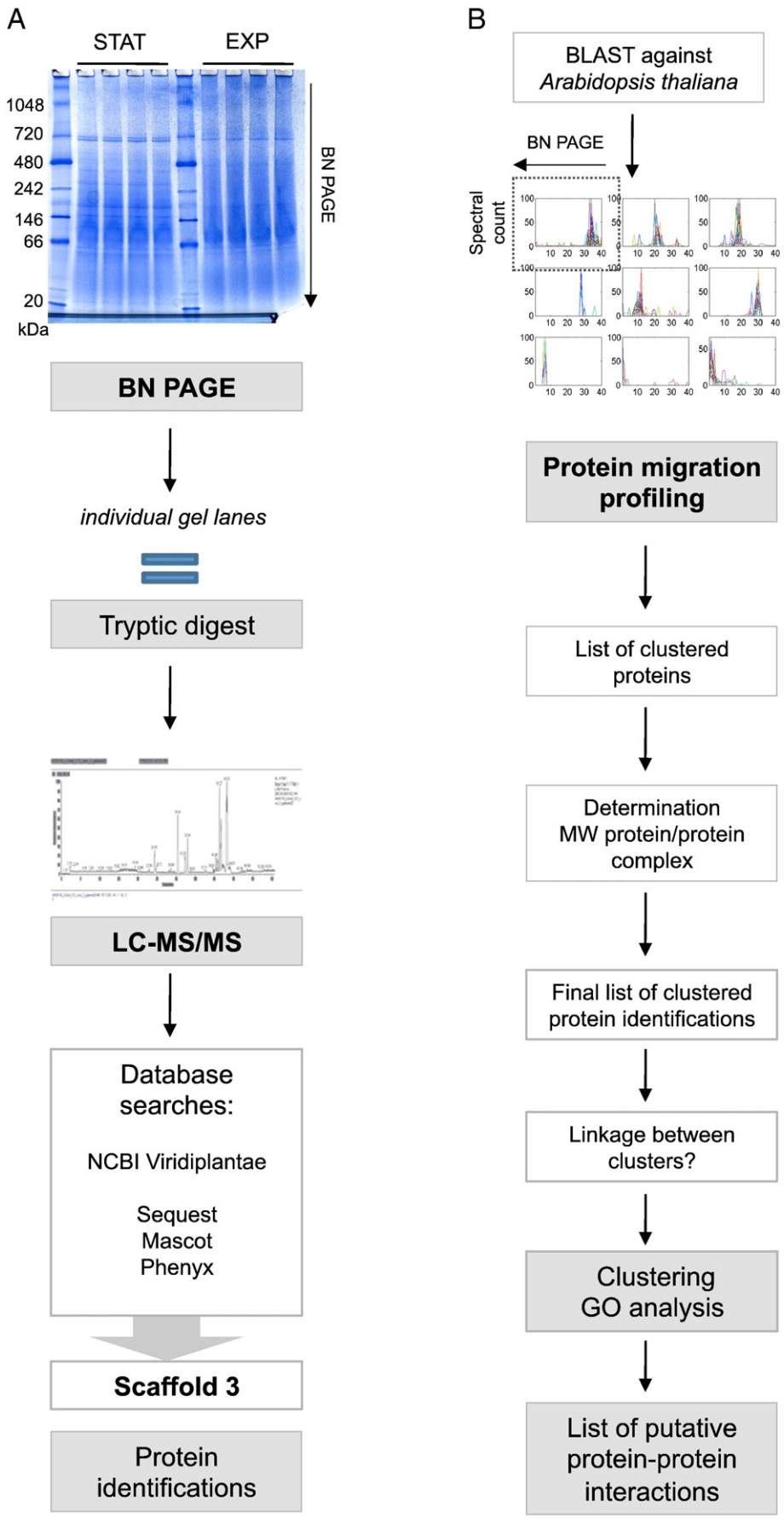
To apply the tiling method, the data matrix was discretized to binary values, where a value of 1 represents the presence of a protein, and a value of 0 corresponds to its absence. Due to this binary dataset, biclustering does not take quantitative information into account. The algorithm of Geerts et al. [42] was used in an implementation publically available for research purposes (http://www.adrem.ua.ac.be/tiling). We mined tilings in which the individual tiles were allowed to overlap. Besides the input (a binary data matrix, in sparse format), and the choice of allowing overlap or not, there are no further parameters.

2.10. Functional annotation of clustered proteins

To evaluate the enrichment of functional annotations in the set of proteins within each cluster we used the BINGO plugin [43] within Cytoscape (version 2.7.0) [44]. A hypergeometric test and the Benjamini–Hochberg correction was used to test the statistical significance of the enrichment of each given GO term in a list of proteins with respect to a reference list. As a reference list, the full gene ontology annotation of *A. thaliana* was downloaded from the gene ontology website (www.geneontology.org). The "biological process" category of the

Q5 **Fig. 1 – Representation of the workflow. A. From BN gel to protein identification within one gel lane. Stationary samples are indicated with STAT and exponential BY-2 with EXP. Above 70 kDa, the BN gel was cut into equal parts of 1 mm. After tryptic in-gel digestion, LC–MS/MS was performed on each of these gel slices. Identification data of 3 different search algorithms (Mascot, Sequest, and Phenyx) were combined within Scaffold 3.00.03. B. Computational approaches to reveal putative protein–protein interactions based upon orthology in *Arabidopsis thaliana*. The redundancy of protein identifications due to the cross-species approach was limited by protein grouping (Scaffold 3.00.03) and by peptide homology searching against *Arabidopsis thaliana*. Proteins were grouped according to their migration profiles within the BN gel by two different classical clustering methods (hierarchical clustering and k-means clustering) and a biclustering approach. Here, protein migration profiles of the classical clustering approaches are shown. The dashed box corresponds to a single cluster that represents a group of proteins that have a similar migration profile over the whole BN lane (X-axis). The Y-axis represents the spectral count values of each protein throughout the whole BN gel lane. The molecular weight of a protein and its migration position on the BN gel is used as a criterion of its possible involvement in a protein complex. The significance of the outcome of these clusters was tested by functional GO-annotation (biological process) and database-searching.**

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023



Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

GO ontology was used to assign a biological meaning to each cluster. Categories with a p-value less than 0.01 were considered statistically significantly overrepresented in a cluster compared to the whole *Arabidopsis* annotation.

To find putative interacting proteins, a hierarchical clustering with Pearson correlation was done for each protein within a complex, according to GO annotations. For this purpose, only annotations that were shared between at least two proteins were taken into account for a given cluster. For the evaluation of known protein complexes, interaction data was retrieved from the STRING (http://string-db.org) [45], AtPIN release 10 (http://bioinfo.esalq.usp.br/atpin/atpin.pl) [46] or the IntAct database (http://www.ebi.ac.uk/intact/main.xhtml) [47].

3. Results and discussion

3.1. Workflow selection and protein selection boundaries

BN PAGE was applied to separate protein complexes from whole plant cell lysates. Fig. 1 gives an overview of the entire workflow presented here and further explained in the next paragraphs. First a list of co-migrating proteins was obtained after BN PAGE/LC–MS/MS (Fig. 1A). The BN gel lanes were cut into an average 47 equally sized pieces between 70 kDa and 1300 kDa. After proteolysis, the extracted peptides were separated by reversed-phase nano-LC and analyzed by tandem MS. The resulting spectra were used for peptide based homology protein identification and subsequently putative interaction partners between the identified proteins were searched by three clustering methods (the classical clustering approaches and biclustering) (Fig. 1B). In order to reduce the complexity of the protein–protein interaction analysis, the low MW complexes (<70 kDa) co-migrating with the bulk of monomeric proteins were omitted from the analysis. This threshold was selected since the MW distribution of *A. thaliana* proteins showed that 81% of all proteins fall beneath this MW boundary (see Supplementary data Fig. 1).

3.2. Cross species identification of proteins by multiple search engines

Since the tobacco genome is still not fully sequenced, protein identification had to rely on cross-species identification based on a peptide homology search against the NCBInr database (taxonomy Viridiplantae). In order to improve the reliability and sensitivity of the protein identification [34], the database was searched by three different search algorithms (Mascot, Phenyx and Sequest). By combining data from the multiple search engines using integrating software (Scaffold version 3.00.03, Proteome Software Inc., Portland, OR), the peptide false discovery rate decreased from 1.2% (for the best performing single search engine) to 0.6% (for the 3 search engines combined) and the number of identified unique peptides increased by 20%, which together lead to an increased confidence of the protein identification, as well as an increased number of identified proteins. This analysis produced an initial list of 191 and 185 protein identifications, respectively for the stationary and exponential samples. For both datasets, 81% of these identifications were corroborated by two or more peptides (0.1% Protein FDR, 0.6% Peptide FDR). All mass spectrometry data are available in the supplemental data section. Redundancy in the protein identification, caused by scattering across various orthologues due to the cross-species approach, was reduced at two levels: first, all data were combined within the Scaffold software to allow proteins that share the same pool of peptides to be merged into a single protein group. Secondly, redundancy was further reduced by performing a sequence similarity search against the *A. thaliana* protein sequences, which led to the collapse of several similar protein hits into a single entry and resulted in a final list of 165 (reduction of 11%) and 180 (reduction of 3%) non-redundant proteins for respectively the stationary and the exponential sample. The BLAST results (see supplemental data) show that all proteins obtained within this *N. tabacum* dataset are well conserved in *A. thaliana*. More than 50% of the BLAST hits had a perfect match (E-value=0). The other hits had an E-value between 1.00*E–185 and 1.00*E–45 (see Supplemental Fig. 2). GO categorization of the orthologues of the BY-2 dataset showed that the largest part of the proteins were involved in metabolic processes, protein metabolism and stress response (see supplemental data). These proteins are indeed known to be highly conserved across species [48].

3.3. Detecting protein complexes by protein migration profiling and clustering

Besides the classical clustering methods to cluster the proteins upon their migration profile in the BN gel, a “biclustering” or “tiling” approach used in data mining [42,49] was evaluated.

3.3.1. Evaluation of clustering methods

First, we compared the commonly used hierarchical clustering and k-means clustering methods. Both clustering methods were applied on normalized and non-normalized data. Normalization of the data did produce an artifact by amplifying the noise, i.e., the proteins with low spectral counts (data not shown). As a result, the low-abundance proteins were grouped as additional clusters. Since these clusters are not biologically relevant, non-normalized spectral count data were used instead (Fig. 2) to allow comparison between the classical clustering approaches and the biclustering technique. For both classical clustering techniques the number of clusters discovered is essentially a user-defined choice. Furthermore, interpretation of k-means clustering results should take into account the stochastic nature of this method, yielding possibly non-coherent clusters between different clustering rounds on the same data set. This particularly hampers the comparison between cluster analyses of different samples or different technical replicates (data not shown).

It should be noted that hierarchical and k-means clustering both allow each protein to be assigned to only one cluster. In contrast, the biclustering method allows for overlap between clusters, and may thus reveal (a likely) participation of certain individual proteins in more than one different protein complex. As a result, the clusters obtained with the latter method are smaller than the clusters of the classical approach.

In order to compare the results of the classical clustering approaches to the results of the biclustering method, a similarity metric (SM) was calculated (Fig. 3). The similarity

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

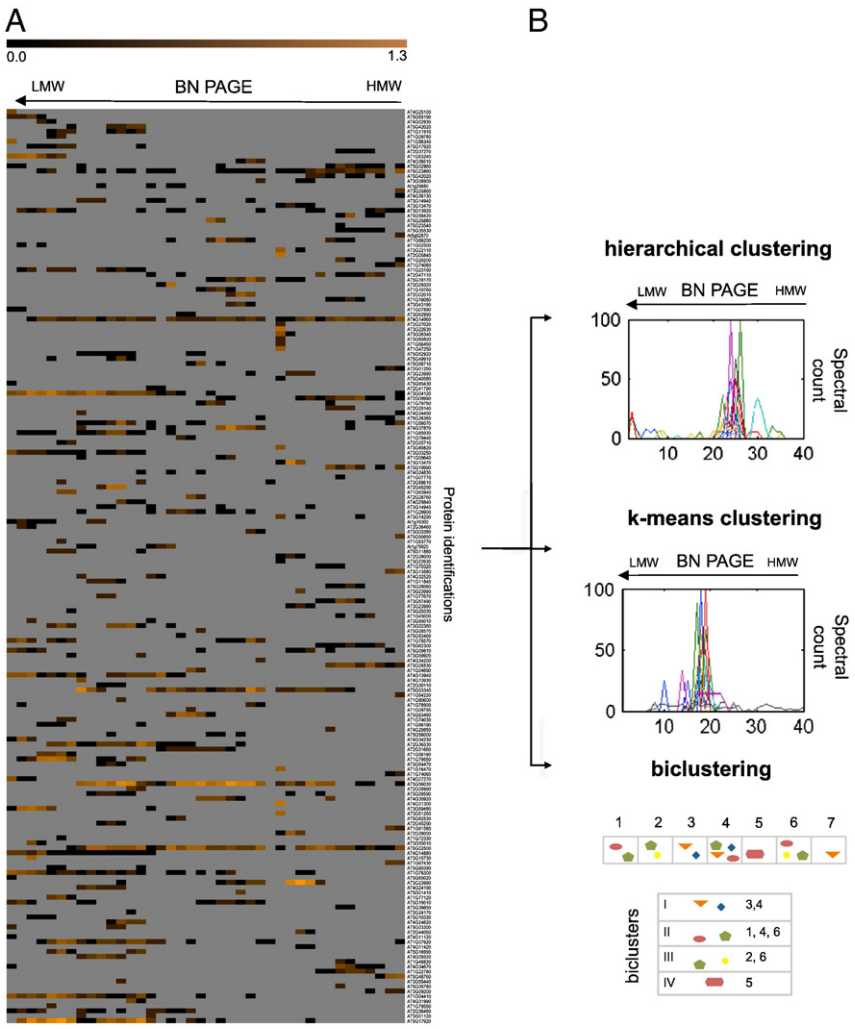


Fig. 2 – Reconstruction of protein complexes from *Nicotiana tabacum* by BN PAGE and clustering strategies. A. Non-normalized spectral count data from all proteins *versus* their migration within the BN gel (heatmap) (X-axis) were used. Within the heatmap representation, the spectral count values are given as a $\log_{10}$ and ranges from 0 (black) to 1.3 (yellow). Each column represents a BN gel slice and each individual row presents an individual protein (Y-axis). B. Three different clustering methods were applied. The premise is that proteins belonging to a particular complex occur in the same BN slice. Two classical approaches (hierarchical clustering and k-means clustering) only look at the migration of proteins (X-as) to cluster them while the biclustering approach groups proteins that co-occur frequently together throughout the whole BN lane. For the classical clustering approaches, the spectral count values of each protein are needed to obtain the clustering of proteins while for biclustering, the data matrix was discretized to binary values (present or not present in a gel slice). The Arabic numbers represent BN gel slices and Roman numbers represent biclusters.

metric calculates the number of protein identifications in the intersection of both clusters, divided by the number of protein identifications in both clusters minus the intersection. Fig. 3A graphically illustrates the calculation of this metric. In theory, the metric presents the odds of the Jaccard index and is defined in the range, [0, ∞].

First, both classical approaches were compared (Fig. 3B) and thereafter, the classical approaches were, in turn, individually compared to the biclustering method (Fig. 4A). The larger SM, the more protein identifications both clusters have in common. When SM equals 1 (or $\log_{10}(SM)=0$), the number of protein identifications in the intersection equals the number of protein identifications not in common. When SM is smaller than 1 (or $\log_{10}(SM)<0$), there is an overlap between the identifications, but both clusters contain more protein identifications not in common.

Fig. 4B shows that for the dataset of exponentially growing BY-2 cells, there exists an overlap of 16% between both classical clustering methods and that 79% of the compared clusters only share very few protein identifications. Within 21% of the matched clusters, both clusters contained more protein identifications in common than uncommon. No completely identical clusters among the results of hierarchical clustering and the results of k-means clustering were found. Rather, each cluster can be mapped to an average of two other clusters within the matrix. To test our similarity score, we

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

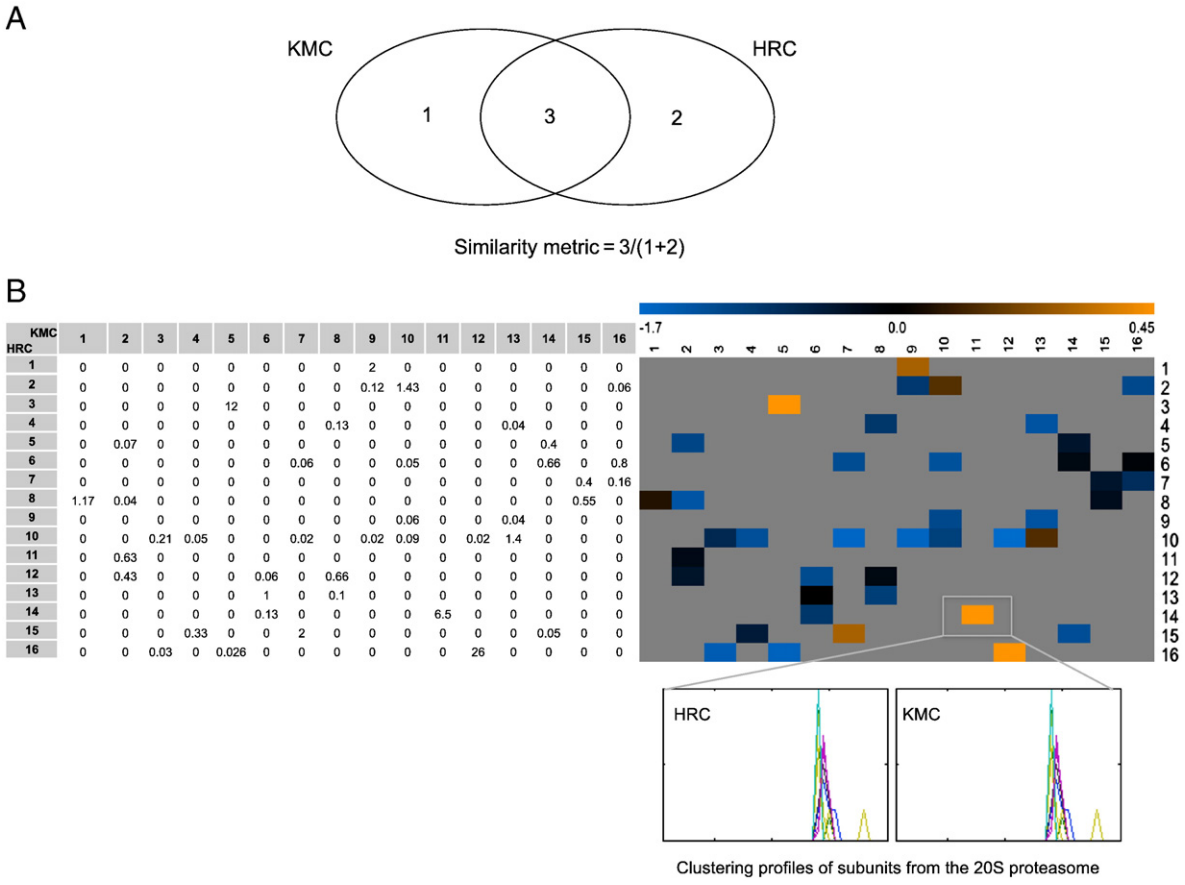


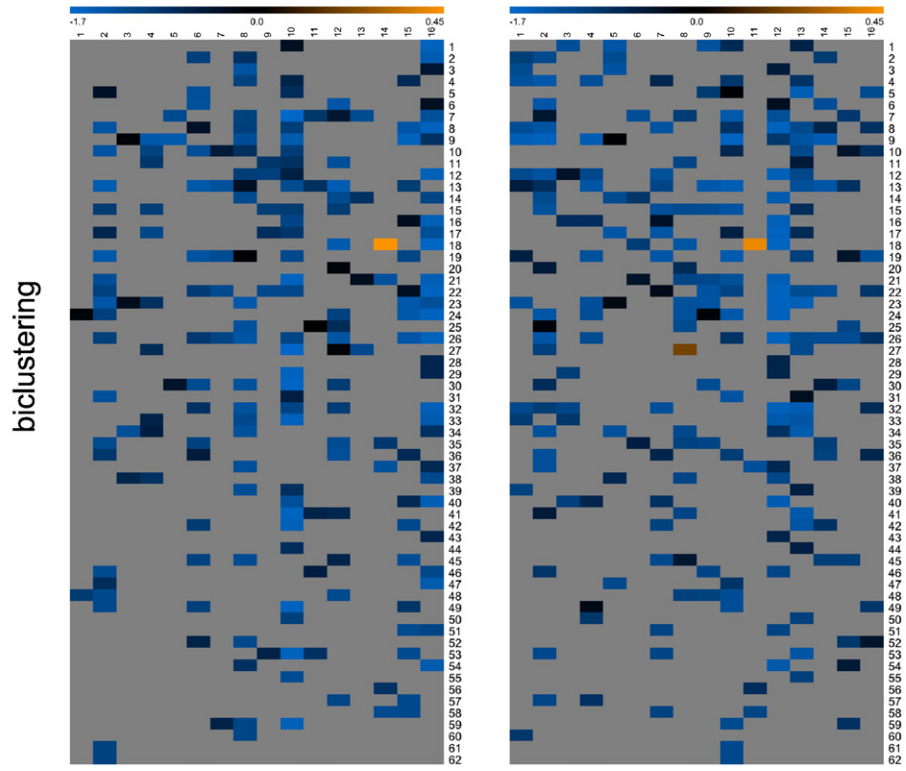
Fig. 3 – Definition of a similarity metric to compare clustering methods. A. Protein clusters are compared by dividing the number of the shared protein identifications (3) by the sum of all distinct protein identifications within both clusters (1+2). B. Comparison both classical clustering approaches (hierarchical clustering and k-means clustering) of the exponential dataset. Within the heatmap representation, the similarity values are given as a $\log_{10}$. The color scale bar ranges from −1.7 (blue) to 0.45 (yellow). When $\log_{10}$ (SM) < 0, overlap exist between the identifications of the compared clusters but they contain more protein identifications not in common. If $\log_{10}$(SM) ≥ 0, both clusters have more than half of their protein identifications in common. The column and row headers represent the cluster numbers of each clustering method, given in the supplemental data. The cluster profiles of subunits of the 20S proteasome illustrate that compared clusters with a high similarity score show a high overlap in their protein migration profile.

evaluated the most similar match for each cluster (note that the other matches can present links to subcomplexes thereof). For example the most similar match for cluster 14_{HRC} (SM = 6.5) was cluster 11_{KMC} (Fig. 3), both of which represent the 20S proteasome (Fig. 3B).

A larger number of matches were observed between the respective classical clustering methods and the tiling approach, 21% for hierarchical clustering and 23% for k-means clustering (Fig. 4A and B). However, in 99% of the cases the observed similarity between a classical cluster and a bicluster are relative small or medium ($\log_{10}$(SM) < 0) (Fig. 4B). This is explained by the relative small size of the biclusters compared to the size of the classical clusters. For example, the aforementioned clusters 14_{HRC} and 11_{KMC} showed a high similarity against protein bicluster 18, respectively SM = 5 and SM = 2.6. Each of these three clusters corresponds to the same biologically important entity that we previously isolated from BY-2 cells by 2-dimensional BN/SDS PAGE [13]: the 20S proteasome. As a valid benchmark complex, it demonstrates that each of the clustering methods used here correctly assigns all members to the complex. For a distinct protein complex, such as the 20S proteasome, all clustering methods performed very well. For less distinct protein complexes however, like for example multi-enzyme complexes that are known to be involved in several metabolic pathways through a variety of interactions, the biclustering method is more appropriate to correctly reveal their promiscuous nature, as it searches and retrieves several possible combinations of interacting proteins.

Fig. 5A shows the comparison between two highly similar clusters (SM = 26) obtained by hierarchical clustering (cluster 16_{HRC}) and k-means (cluster 12_{KMC}) with all biclusters of a dataset of exponentially growing BY-2 cells. Each of the classical clusters can be matched against multiple biclusters. Reconstruction of protein complexes through their shared and distinct components shows that the biclustering method separates co-migrating protein complexes more efficiently than the classical clustering approaches (Fig. 5B). Complexes

A



B

	no match	match	large overlap[a]	medium to small overlap[b]
vs. k-means				
hierarchical	84%	16%	21%	79%
vs. biclustering				
hierarchical	79%	21%	0.90%	99%
k-means	76%	23%	0.80%	99%

[a] log10(SM)≥0

[b] log10(SM)<0

Fig. 4 – Comparison of both classical clustering methods to the biclustering method by a calculated similarity metric. A. Heatmap representation of the similarity between both classical clustering method (X) and the biclustering approach (Y). The classical clustering approach was set to partition the data in 16 clusters while biclustering yielded 62 biclusters. The similarity values are given as $\log_{10}$. The color scale bar ranges from –1.7 (blue) to 0.45 (yellow). B. Comparison of all clustering methods by a similarity metric. About 20% of the classical approaches matched against 1 or more biclusters but the majority of these overlaps are small. This means that the compared clusters share only a few proteins but they both contain more distinct proteins.

that were grouped in a single classical cluster are returned as multiple biclusters. This tiling method searches across the whole gel lane for reiterating protein sets. As a consequence, the more frequently a set of proteins co-occurs throughout multiple gel slices, the stronger their association will be. Therefore the biclustering method is less vulnerable to grouping (unrelated) co-migrating protein complexes together, compared to the more classical approaches. This is the case for the dimeric tubulin complex (TUA6 and TUB8), the oligomeric chaperonin 60, the oligomeric TPPII and pyruvate dehydrogenase (LTA2 and MAB1). Furthermore, the biclustering results in Fig. 5B show an association between the 40S ribosome (light green) and 19S proteasome (red) by e.g., the RTP5A subunit, between the 40S (light green) and the 60S ribosome (dark green)

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

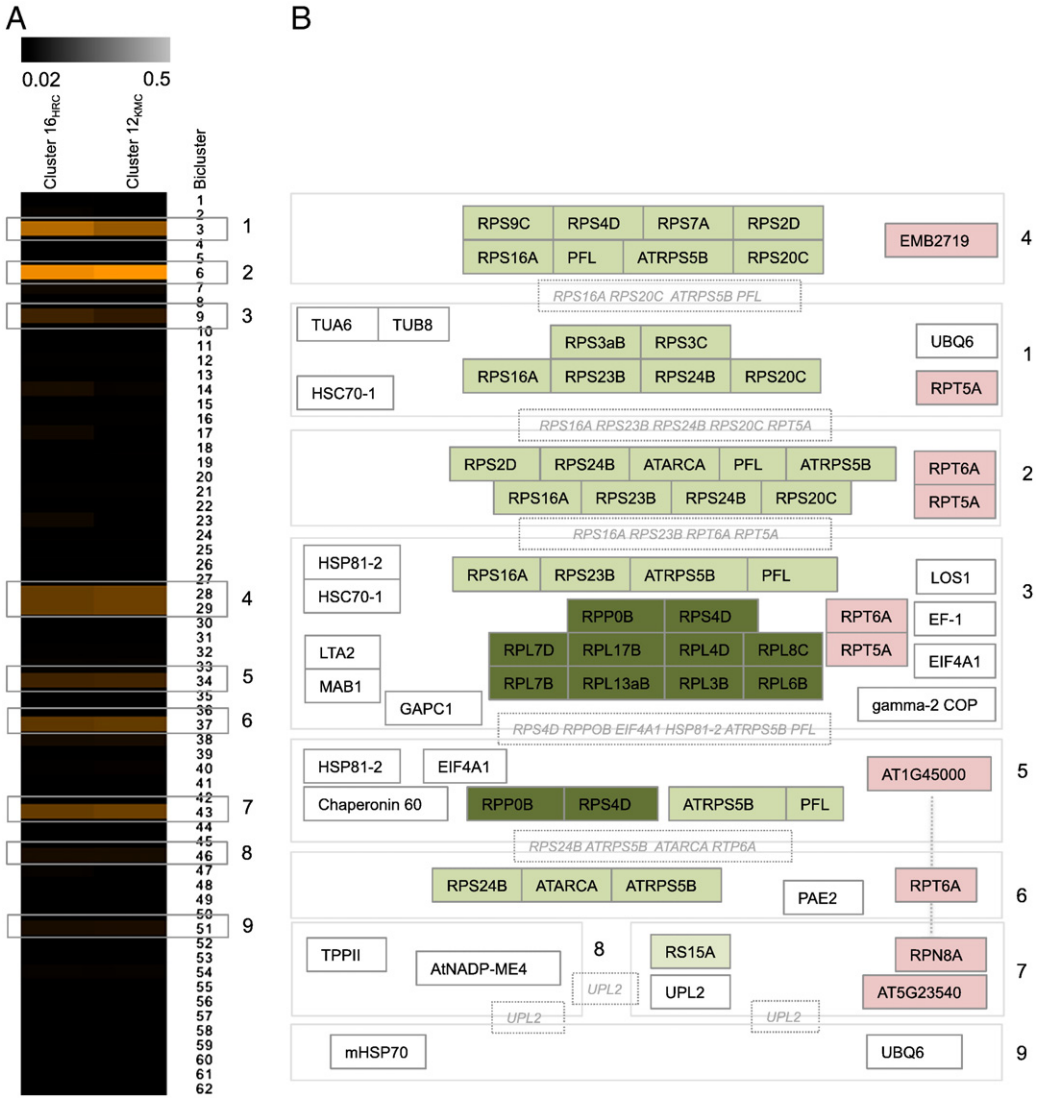


Fig. 5 – Reconstruction of protein complexes by linking protein tiles through shared components. A. Comparison between two matched (SM=26) classical clusters (cluster16$_{HRC}$ and cluster12$_{KMC}$) to all biclusters of an exponential BY-2 sample. A single classic cluster shares multiple small overlaps with several biclusters. B. Biclustering allows for the participation of a protein in multiple biclusters. The numbers represent each bicluster that has a similarity with the classic clusters (boxes in A). By linking biclusters through their shared components (proteins in overlapping dashed boxes in B), protein complexes can be reconstructed and this reveals an interesting link between protein complexes. The 40S ribosome is colored light green, the 60S ribosome dark green and subunits of the 19S are colored red. Evidence of the interaction between the LOS1, EF-1a, EIF4A1 and gamma-2 COP proteins and the 60S ribosome was found within the STRING database (see Fig. 7B).

through their RPP0B subunit. The co-occurrence of certain sets of proteins within several biclusters may be interpreted as core complexes, whereas unique members of a certain bicluster could be seen as attachments [50,51].

3.3.2. *Gene ontology analysis of the purified protein complexes*

Protein migration profiling is possible in the presence of many background proteins but co-localization of unrelated proteins cannot be excluded. To further distinguish between related and unrelated co-migrating proteins, a functional characterization was performed under the assumption that proteins within a given protein complex are likely to be involved in a similar biological process. For this functional annotation, the gene ontology (GO) is the *de facto* standard and can be used for functional association analysis of proteomic data [31]. The statistical overrepresentation of GO categories within each cluster was analyzed. A typical result of such BiNGO analysis for bicluster 9 (from exponential BY-2 cells) is shown in Fig. 6. It contains multiple large protein complexes (60S ribosome, 40S ribosome, 26S proteasome, and pyruvate dehydrogenase). The analysis shows that the majority of the proteins are functionally related due to their involvement in metabolic and cellular processes, and herein a further distinction can be made between complex members involved in translation, proteolysis, and pyruvate metabolism. A small group was labeled as responsive to stimuli. For 3 proteins (RPL7D, RPL4D

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

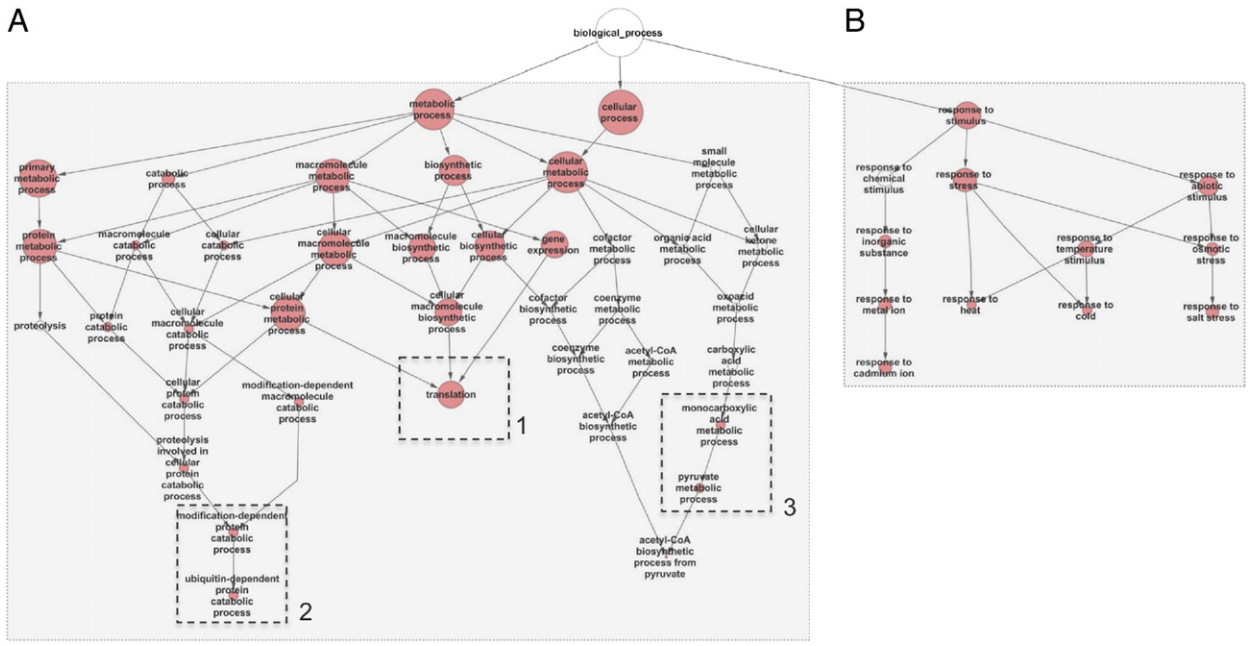


Fig. 6 – Functional annotation analysis of clustered proteins. Red colored dots represent significantly overrepresented (p = 0.01) functional groups within bicluster 9 (exponential dataset). The size of the spots depends on the number of proteins within each GO category. GO categories with a single entry are not taken into account for further complex analysis. Two main groups can be discriminated: A. Proteins involved within metabolic and cellular processes. Here, proteins are grouped within 3 functional biological processes: translation (1), ubiquitin-dependent protein catabolic process (2) and pyruvate metabolic process (3) B. A distinct group of all proteins are also annotated as responsive to stimuli.

and EF-1alpha), although known to be involved in translation, no GO annotation was found.

The gene ontology annotations can possibly be used to reveal groups of functionally associated proteins within a cluster, as well as potentially unrelated co-migrating groups of proteins. To this end, GO annotations were used to hierarchically partition clusters according to the functions of their members. Fig. 7A illustrates that this method allows to discriminate between functionally (un)related protein complexes that co-migrate, such as the 26S proteasome and the ribosome. Due to the incompleteness and varying depth of available GO annotations for *Arabidopsis*, this clustering should nevertheless be carefully interpreted.

The approach was further evaluated by retrieving interaction data of these proteins from the STRING database. Fig. 7B shows that the proteins indeed assemble into protein complexes, and that some of these protein complexes are even related. Some of these relationships were also found by the GO-driven cluster partitioning approach described above, e.g., between the ribosomal subunit RPPOB and the translation elongation factor LOS1, or between the 60S ribosome and the translation initiation factor EIFA4. Although functional relationships between members of a cluster can readily explain the fact that its members co-migrate, functional differences within a cluster can also be biologically interesting, if they represent putative links between different functions within a cell.

3.3.3. *Finding protein complexes throughout multiple samples*

The biclustering method can both be used to analyze patterns over the whole BN gel dataset, as well as within each one of the two (stationary and exponential) cell culture stages. Within an experiment for comparative analysis, proteins that cluster persistently throughout multiple samples can be considered stable protein complexes and those proteins that cluster intermittently with this stable core can be regarded as sample-dependent associations, as shown for the EIFA4 interaction with the ribosome in proliferating cells as described below. The protein biclustering method allows finding patterns over multiple samples simultaneously and then returns both protein sets that co-occur frequently throughout all samples but, as well, those that are condition dependent. Within both samples (stationary and exponential BY-2 cells) a protein set of 5 proteins (ATGSR2, RHM1, GDH1, GAD4 and SUS4) was found at ~556 kDa for the exponential cells and at ~660 kDa in the stationary cells. Since ATGSR2 (glutamate-ammonia ligase; glutamine synthase), GDH1 (glutamate dehydrogenase) and GAD4 (glutamate decarboxylase) all share glutamate as a substrate these enzymes are possibly associated. Within the stationary cell samples, the multimeric RSR4 (REDUCED SUGAR RESPONSE 4) was biclustered with these proteins. This protein is part of the glutamine amidotransferase complex and its presence in higher MW protein complex was reported [52] and explains the MW shift within the stationary cells.

In plants, sucrose synthase occurs as a tetramer of ~92 kDa subunits (368 kDa) [53] and so it interacts with other proteins within our BN gel. Matic et al. [54] showed that sucrose synthase has a high affinity for UDP-glucose in BY-2 cells. RHM-1 (RHAMNOSE BIOSYNTHESIS 1, UDP-glucose 4,6-dehydratase/catalytic) uses UDP-glucose as a substrate in the UDP-

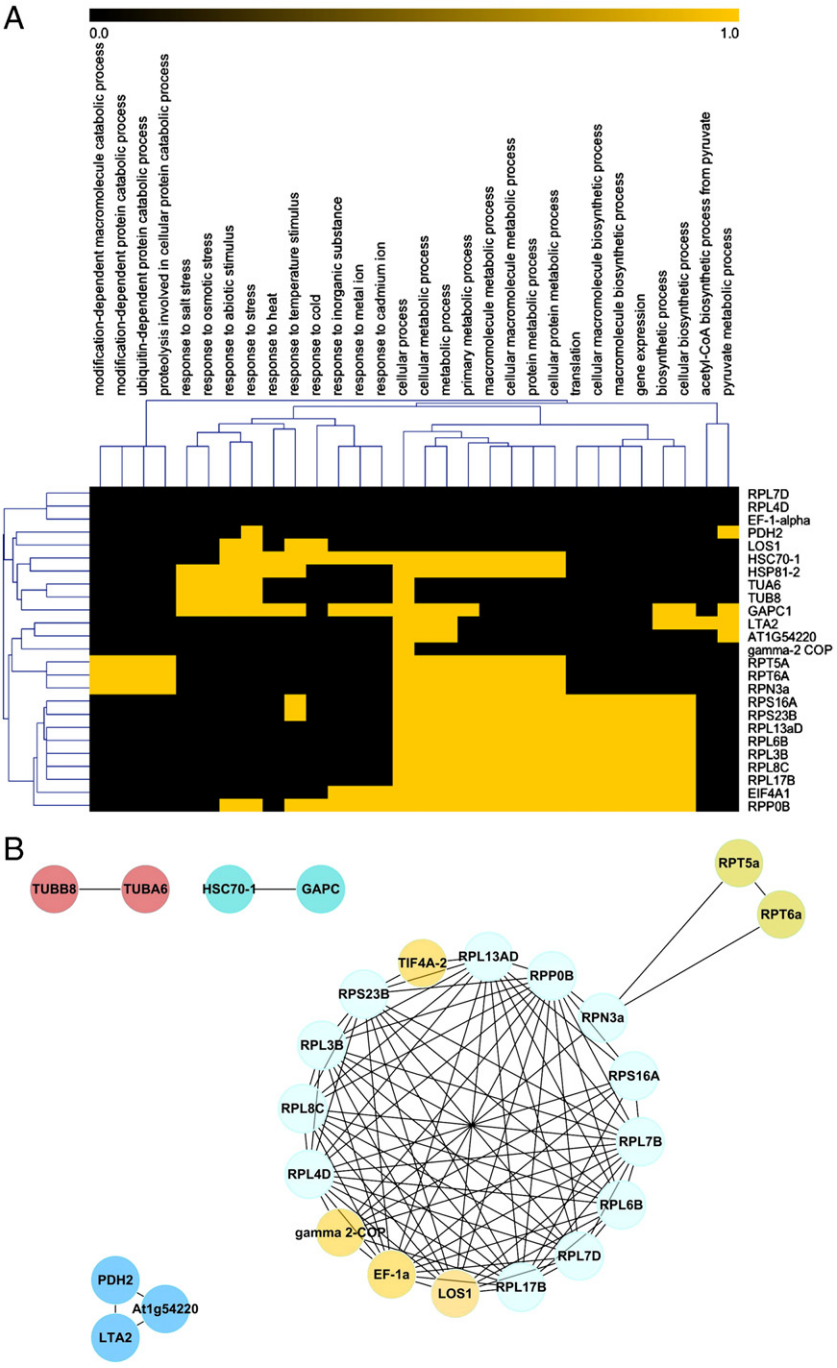


Fig. 7 – Hierarchical clustering of the GO annotation (biological process) within a bicluster of exponential BY-2 cells. Functional annotation can be used to sort out co-migrating (un)related protein complexes. A. Hierarchical clustering of a BiNGO analysis of a single bicluster (bicluster 9) from exponential BY-2 cells. This bicluster (bicluster 9) corresponds to box 3 in Fig. 5B. B. Interactions within the cluster were retrieved from the STRING database and visualized within a protein network. The different colors used represent (un-)related protein–protein interactions/protein complexes that were grouped together in bicluster 9 but sorted out by functional annotation.

rhamnose biosynthesis pathway [55] and was frequently clustered with the sucrose synthase in our dataset, both for stationary and exponentially growing BY-2 cells (see supplemental data, bicluster $10_{STAT+EXP}$).

Subunits of the 60S ribosome were biclustered in both samples. An interaction between the 60S ribosome and eIF4A, a DEAD-box RNA helicase was found. This eIF4A is a highly dynamic subunit of the translation initiation complex eIF4F that unwinds the mRNA prior to translation in proliferating plant cells [56]. In our dataset, this interaction between the ribosome and this DEAD BOX helicase was only found in BY-2 cells that are actively proliferating in the exponential phase.

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

The consequences of this interaction for protein translation are not understood yet.

3.3.4. *Analysis of BY-2 protein complexes by clustering methods*

Within this section some examples of protein complexes are described. Multiprotein complexes of the proteasome pathway were found by the BN PAGE/LC–MS/MS approach (see Table 1). The most prominent protein complexes that were found are the subcomplexes of the 26S proteasome. These large protein complexes function in controlled proteolysis and were previously identified by 2-dimensional BN/SDS PAGE [13,57]. The 20S proteasome is an ATP and ubiquitin-independent protease, consisting of 14 different subunit in an $\alpha_{1-7}/\beta_{1-7}/\alpha_{1-7}/\beta_{1-7}$ configuration. This complex was clustered together with the tetradecameric protein chaperonin 60 by two clustering methods (HRC and biclustering). Although they cluster together both in their migration profiles as in their functional annotations, these protein complexes are non-interacting co-migrating protein complexes since they are both identified in the BN gel around 850 kDa, a molecular weight too small to harbor both complexes. An association between the 20S proteasome and lumazine synthase, COS1, was reported previously after BN SDS PAGE [13]. Within this dataset, this protein clustered with the 20S proteasome and chaperonin 60 by both classical clustering methods. The biclustering method only linked the COS1 protein to chaperonin 60. When searching for similar protein migration profiles across the BN gel, the classical clustering approaches are not sensitive to the exact place in the BN gel and can cluster over several BN slices. Since the COS1 was found within a gel slice of slightly different molecular weight (~957 kDa) and was not present in the slice of the 20S proteasome (~850 kDa), they were not grouped by biclustering, which primarily retrieves proteins that co-occur within multiple gel slices. In *A. thaliana* and spinach, this oligomeric lumazine synthase was also found at 738 kDa [58] and 991 kDa [59], corresponding quite well to the data presented here (~957 kDa). Since these co-migrating protein complexes all play a role within plant defense [59,60], an association between them is not rejected but both protein complexes are known to form large oligomers (up to 60-mer) [60]. The existence of their large oligomeric states and their migration within the BN gel (~850–950 kDa) rules out a direct interaction between both large protein complexes (>1600 kDa), unless smaller subcomplexes of both proteins exist that can associate at this molecular weight. The presence of such subcomplexes is not reported yet. At ~1100 kDa, the 19S regulatory particle of the 26S proteasome was found through clustering. A large number of subunits were found (see Table 1). Previously, the RPN1a subunit of this protein complex was used as a bait in an *A. thaliana* affinity purification experiment [6]. This protein interacted with several of the proteins that co-migrated in our experiment.

The serine protease protein complex, tripeptidyl peptidase II (TPPII), is often seen as contaminant of the 26S proteasome because of its similar size and function [61]. Both classical clusterings grouped this oligomeric protein complex together with the 26S proteasome and 40S ribosome while protein tiling dedicated this TPPII protein to a single separate bicluster. The similarity of this bicluster to the classic clusters was close to zero. This means that there is a small overlap but that the classic clusters contain multiple distinct protein identifications. Biclustering showed that no link is present with the 26S proteasome (Fig. 5B). Within the BN gel, the TPPII complex migrated at ~1000 kDa as an oligomer of 140–150 kDa subunits. Normally it forms a large protein complex of more than 5 MDa but the existence of active TPPII subcomplexes was already reported within *A. thaliana* [61]. The large E3 ubiquitin protein ligase 2 (upl2) (405 kDa) is clustered with TPPII and with an ubiquitin extension protein, a constituent of the ribosome. Mass spectrometry data showed that E3 protein ligase contained an N-terminal biotin (see supplemental data). Biotin acts as a covalently-bound cofactor on a family of enzymes that catalyze reactions in a variety of crucial metabolic processes and are mainly found on (de)carboxylases [62]. Here, a first lead is present of a possible biotin-containing ubiquitin ligase but further validation is required. In general, little is known of this HECT (Homologous to E6AP C-Terminus) domain-containing ligase protein in plants and it is thought to act as a single component within the ubiquitin-mediated protein degradation pathway [63]. Since they are all clustered as active within the proteolytic pathway, a direct or in-direct link between these proteins cannot be ruled out. Within the gel slice (~1000 kDa), other subunits of the 40S ribosome were found (see Table 1) together with these proteolytic proteins.

Another protease complex (DegP7) was found around 750 kDa. These protein complexes are known to form large homo-oligomers with a trimer as their functional unit [64]. In our approach the hexameric form of this protein complex clustered with the membrane peripheral part of the vacuolar ATP synthase complex but functional annotation and their place of migration ruled out that these protein complexes interact with each other. V-ATPase in plants are large heteromeric protein complexes from >700 kDa but free subcomplexes of the vacuolar ATPase were reported previously [65]. Furthermore, the V-ATPase amount, its subunit composition and their stoichiometry seem to vary in different types of tissue, in response to environmental factors and due to the developmental state of the plant. Between 690 kDa and 810 kDa only two subunits (A and B subunits) of the membrane peripheral V-ATPase were found and no potential interactors could be retrieved. The tobacco V-ATPase was extensively studied by Drobny et al. [66] and they showed that 8 subunits of the tobacco V-ATPase could be identified after enrichment by immunoprecipitation but that the assignment of the polypeptides to specific V-ATPase subunits was not straightforward. This difficult identification of all subunits could explain the lack of other subunits within our dataset. *De novo* sequencing could be necessary to reveal other subunits. The subunits found here are the two ubiquitous major subunits, A and B, present in three copies per functioning enzyme forming the catalytic hexameric (A_3B_3) cylinder of the V-ATPase and carries the catalytic nucleotide-binding site [67].

Microtubuli subunits are found at multiple places across the whole BN lane. It is known that microtubuli are assembled from dimers from α-tubulin and β-tubulin [68]. These dimers are stable protein complexes that interact with a large amount of proteins within the plant cell [69]. With the BN PAGE/LC–MS/MS approach, interactions were found with the 40S ribosome and some elongation factors (e.g., EF1α and EIF4A).

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

Table 1 – Examples of protein complexes identified by clustering approaches after BN PAGE/LC–MS/MS.

Protein complex	TAIR ID	NCBI ID	Protein description	References
20S proteasome (~850 kDa)				
Core components	AT3G22110	gi|12229904	PAC1	[13,57]
	AT2G05840	gi|12229948	PAA2	
	AT2G27020	gi|14594925	PAG1	
	AT1G21720	gi|14594927	PBC1	
	AT3G22630	gi|14594929	PBD1	
	AT3G26340	gi|14594931	20S proteasome beta subunit E	
	AT3G60820	gi|14594933	PBF1	
	AT1G56450	gi|14594935	PBG1	
	AT1G47250	gi|147856362	PAF2	
	AT5G40580	gi|15237451	PBB2	
	AT1G53850	gi|159478278	PAE1	
	AT5G66140	gi|162458131	PAD2	
	AT3G60820	gi|14594933	PBF1	
	AT3G14290	gi|217071540	PAE2	
	AT1G16470	gi|255634578	PAB1	
	AT4G31300	gi|3024440	PBA1	
Potential interactors	AT5G60160	gi|223550330	Aspartyl aminopeptidase, putative	
	AT3G07110	gi|76573371	60S ribosomal protein L13A	
	AT4G16260	gi|119004	Catalytic/cation binding/hydrolase	
19S proteasome (~1100 kDa)				
Core components	AT5G19990	gi|18420092	RPT6A	[57]
	AT1G53750	gi|115449095	RPT1A	
	AT4G29040	gi|6652880	RPT2a	
	AT4G24820	gi|117607065	RPN7	
	AT1G09100	gi|15217431	RPT5B	
	AT5G58290	gi|1709798	RPT3	
	AT2G32730	gi|171854677	Rpn2	
	AT5G09900	gi|225438483	EMB2107/RPN5	
	AT2G20580	gi|225446449	RPN1A	
	AT1G20200	gi|12230970	26S proteasome regulatory subunit S3	
	AT5G05780	gi|77745499	RPN8A	
	AT4G19006	gi|78059502	26S proteasome regulatory subunit	
	AT2G20140	gi|168002980	26S protease regulatory complex subunit 4	
	AT1G45000	gi|224062085	EMB2719/26S proteasome regulatory complex subunit p42D	
	AT3G05530	gi|225432252	RPT5A	
Potential interactors			Multiple partners of the 40S ribosome	
TPP2 (tripeptidyl peptidase II) (~1000 kDa)				
Core component	AT4G20850	gi|225470769	TPP2	[61]
Potential interactors	AT1G70320	gi|223533281	UPL2	
	AT2G47110	gi|126038342	UBQ6	
	AT4G30920	gi|27463709	Cytosol aminopeptidase family protein	
UPL2 (ubiquitin-protein ligase 2) (~450 kDa)				
Core component	AT1G70320	gi|223533281	UPL2	[63]
Potential interactor	AT2G09990	gi|159138763	RPS16A	
	AT4G20850	gi|225470769	TPP2	
40S ribosome (>850 kDa)				
Core components	AT2G09990	gi|159138763	RPS16A	
	AT5G62300	gi|224134518	RPS20C	
	AT5G02960	gi|115440881	RPS23B	
	AT3G57490	gi|223547389	RPS2D	
	AT1G22780	gi|76573321	PFL	
	AT5G28060	gi|223542604	RPS24B	
	AT2G37270	gi|115433960	ATRPS5B	
	AT1G18080	gi|1346110	ATARCA	
	AT5G35530	gi|118488288	RPS3C	

These interactions were previously described [70]. Throughout the whole BN gel, the chaperonins HSC70 and cytosolic Hsp80-1 (member of HSP90 family) co-migrated frequently. These two heat shock proteins function together with Hsp70, they may be considered as parts of a larger multi-chaperone system. The Hsc70 also clusters with multiple subunits of the

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

ribosome (Fig. 5B), which is in agreement with its role as a regulator of the folding of nascent polypeptides [71].

3.3.5. *Detection of temporary sequential multi-enzyme complexes*

In this study, not only long-lived protein complexes were detected. Indications of temporary associations between sequential enzymes of a metabolic pathway (referred to as metabolons) [72] were found as well. Several of such enzyme complexes were found throughout the BN lane but further experimentation is needed to determine the biological significance of these suggested multi-enzyme complexes. Within the BN gel, subunits of the large hetero-oligomeric pyruvate dehydrogenase complex (PDC) clustered together. This oligomeric protein complex is composed of three enzymes that act sequentially: pyruvate dehydrogenase (named E1), dihydrolipoamide S-acetyltransferase (E2), and dihydrolipoamide dehydrogenase (E3). It catalyzes the overall conversion of pyruvate to acetyl-CoA and CO_2. PDCs are known to form large complexes composed of a core complex of eight trimers (cube) or 20 trimers (pentagonal dodecahedron) of E2 with E1 and with E3, that promotes substrate channeling across the three enzyme components [73]. In our study an association between the core-subunit dihydrolipoamide S-acetyltransferase (E2) and pyruvate dehydrogenase (E1) was found at ~1100 kDa. The existence of such association was previously reported by Olinares et al. [30].

Two metabolic enzymes, glutamate synthase and carbamoyl phosphate transferase, were clustered together by all methods and functional annotation reveals they both are involved in the glutamate metabolism, more specific glutamine family amino acid biosynthetic process. Evidence of a link between these metabolic enzymes was also found within the STRING database. Other enzymes involved within the same metabolism are also found within the same clusters and are suggested to be putative interaction partners of this multi-enzyme complex, e.g., two dehydrogenases (ADL5F1 and ALD12A1).

Sequential enzymes of the Calvin cycle also clustered together. Glyceraldehyde-3-phosphate dehydrogenase (GADPH) and fructose-bisphosphate aldolase are co-migrating at 310 kDa and are functionally associated together within the glucose metabolic process.

At 290 kDa, the p-protein of the glycine decarboxylase complex (GDC) was clustered with serine hydromethyltransferase. In plants, the GDC cooperates with serine hydroxymethyltransferase (SHMT) to mediate photorespiratory glycine–serine interconversion [74].

Two isoforms of malate dehydrogenase (cytosolic NADP-malic enzyme 3 and plastidic NADP-malic enzyme 4) were grouped several times by the biclustering approach. They co-migrated together at different places between 623 and 587 kDa within the BN gel of the exponential BY-2 sample. These proteins have the highest catalytic efficiency for NADP and malate and can be co-expressed within the same subcellular compartment. They are involved in a variety of metabolic pathways. NADP-ME4 (~75 kDa) exists as an active dimer (~150 kDa) or tetramer (~300 kDa) while NADP-ME3 (65 kDa) is present as a hexamer (~390 kDa) or octamer (520 kDa) [75]. In the stationary BY-2 cells, both proteins were only found together between 500 and 605 kDa. Here, different associations are possible but the most likely is between a NADP-ME4 dimer and NADP-ME3 hexamer (540 kDa). An association between both isoforms is confirmed in the AtPIN database. Protein identification of NADP-ME3 by the Phenyx engine showed that this protein, while identified in several complexes in this study, only contained a phosphopeptide (serine phosphorylation) in a protein complex at 587 kDa (see supplemental data). This serine phosphorylation was also predicted with a high score (0.924) by NetPhos (http://www.cbs.dtu.dk/services/NetPhos/). The phosphorylated state of the mitochondrial isoform of malic enzyme (NAD-ME) in plants was previously reported [76].

4. Conclusion

This study employs a combination of bio-analytical and computational methods to screen for protein complexes of whole plant cell lysates in a discovery driven approach. By coupling LC–MS/MS to one-dimensional BN PAGE, multiple protein complexes were simultaneously detected in relatively complex samples, even within part of the BN gel slices that had no clear CBB bands. With the classical 2D BN/SDS PAGE approach these proteins would not have been selected for further analysis. Contrary to 2D BN/SDS PAGE, 1D BN PAGE/LC–MS/MS does not suffer from incomplete spot detection due to the limited dynamic range of the staining methods. Another advantage of the 1D BN PAGE/LC–MS/MS approach is that connectivity between compound protein interactions in a single complex is kept as long as possible and less manual intervention (and thereby technical variance) is needed prior to their identification.

In this report, we propose a workflow for the analysis of protein complexes by a data-mining technique (biclustering) that groups proteins by searching subsets of proteins that co-migrate frequently over different fractions of the native separation, even across multiple biological samples. As far as we are aware, this approach has never been applied before to protein complexes separated by BN PAGE/LC–MS/MS. The biclustering approach is a very powerful tool for the exploration of protein complexes in this data flood, since it allows for the participation of a protein within multiple biclusters, consistent with the promiscuous nature of proteins. The proteins shared between biclusters represent interesting links between larger protein complexes and occasional, e.g., condition-dependent, protein associations. In this context, biclustering complements the classical approaches. The complementarity of the different clustering approaches reveals interesting perspectives into the modularity of protein complexes. While cluster analysis based on the protein migration profiles is a powerful discovery method, co-clustering of unrelated proteins is hard to exclude. Therefore, this study employs a GO-driven association analysis of proteins within a bicluster to elucidate the biological relevance of the putative protein complexes herein, even if at present this may create false negatives due to incomplete annotation. We believe that the extension of combining BN PAGE/LC–MS/MS with a biclustering approach to whole plant cell lysates increases its application as an analytical semi-high throughput discovery tool

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

for functional proteomics and can be useful in large-scale mapping of protein–protein interactions within a cellular context. Its successful application to an unsequenced and recalcitrant heteropolyploid plant model like *N. tabacum* is illustrative for its discovery potential and the ability to study a wide variety of non-genomic biological models.

Supplementary materials related to this article can be found online at doi:10.1016/j.jprot.2011.03.023.

Acknowledgments

This work was supported by a PhD scholarship of the IWT (NR), a SBO grant [IWT-600450] of the IWT (KL), a Ph.D. Fellowship (KS) and post-doctoral Fellowships (JV and SC) of the Research Foundation—Flanders (FWO).

REFERENCES

[1] Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 1998;92:291–4.
[2] Nelson N, Ben-Shem A. The complex architecture of oxygenic photosynthesis. Nat Rev Mol Cell Biol 2005;6:818.
[3] Desprez T, Juraniec M, Crowell EF, Jouy H, Pochylova Z, Parcy F, et al. Organization of cellulose synthase complexes involved in primary cell wall synthesis in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 2007;104:15572–7.
[4] Pauwels L, Barbero GF, Geerinck J, Tilleman S, Grunewald W, Pérez AC, et al. NINJA connects the co-repressor TOPLESS to jasmonate signalling. Nature 2010;464:788–91.
[5] Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415:141–7.
[6] Van Leene J, Hollunder J, Eeckhout D, Persiau G, Van De Slijke E, Stals H, et al. Targeted interactomics reveals a complex core cell cycle machinery in *Arabidopsis thaliana*. Mol Syst Biol 2010;6:397.
[7] Fields S, Song O. A novel genetic system to detect protein protein interactions. Nature 1989;340:245–6.
[8] Williams NE. Immunoprecipitation procedures. Methods Cell Biol 2000;62:449–53.
[9] Kerppola TK. Bimolecular fluorescence complementation: visualization of molecular interactions in living cells. Methods Cell Biol 2008;85:431–70.
[10] Shoemaker BA, Panchenko AR. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. Plos Comput Biol 2007;3:e43.
[11] Hartman NT, Sicilia F, Lilley KS, Dupree P. Proteomic complex detection using sedimentation. Anal Chem 2007;79:2078–83.
[12] Kumaran S, Yi H, Krishnan HB, Jez JM. Assembly of the cysteine synthase complex and the regulatory role of protein–protein interactions. J Biol Chem 2009;284:10268–75.
[13] Remmerie N, Roef L, Van De Slijke E, Van Leene J, Persiau G, Eeckhout D, et al. A bioanalytical method for the proteome wide display and analysis of protein complexes from whole plant cell lysates. Proteomics 2009;9:598–609.
[14] Schägger H, Vonjagow G. Blue native electrophoresis for isolation of membrane–protein complexes in enzymatically active form. Anal Biochem 1991;199:223–31.
[15] Wittig I, Braun HP, Schägger H. Blue native PAGE. Nat Protoc 2006;1:418–28.
[16] Reisinger V, Eichacker LA. Solubilization of membrane protein complexes for blue native PAGE. J Proteomics 2008;71:277–83.
[17] Eubel H, Braun H, Millar AH. Blue-native PAGE in plants: a tool in analysis of protein–protein interactions. Plant Methods 2005;1:11.
[18] Wittig I, Schägger H. Native electrophoretic techniques to identify protein–protein interactions. Proteomics 2009;9:5214–23.
[19] Fandiño AS, Rais I, Vollmer M, Elgass H, Schägger H, Karas M. LC-nanospray-MS/MS analysis of hydrophobic proteins from membrane protein complexes isolated by blue-native electrophoresis. J Mass Spectrom 2005;40:1223–31.
[20] Wessels HJ, Vogel RO, van den Heuvel L, Smeitink JA, Rodenburg RJ, Nijtmans LG, et al. LC–MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. Proteomics 2009;9:4221–8.
[21] Helbig AO, de Groot MJ, van Gestel RA, Mohammed S, de Hulster EA, Luttik MA, et al. A three-way proteomics strategy allows differential analysis of yeast mitochondrial membrane protein complexes under anaerobic and aerobic conditions. Proteomics 2009;9:4787–98.
[22] Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. Proteomic characterization of the human centrosome by protein correlation profiling. Nature 2003;426:570–4.
[23] Freyhult E, Landfors M, Önskog J, Hvidsten TR, Rydén P. Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. BMC Bioinformatics 2010;11:503.
[24] Datta S, Datta S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics 2006;7:397.
[25] Yona G, Dirks W, Rahman S. Comparing algorithms for clustering of expression data: how to assess gene clusters. Methods Mol Biol 2009;541:479–509.
[26] Sardiu ME, Florens L, Washburn MP. Evaluation of clustering algorithms for protein complex and protein interaction network assembly. J Proteome Res 2009;8:2944–52.
[27] Sardiu ME, Gilmore JM, Carrozza MJ, Li B, Workman JL, Florens L, et al. Determining protein complex connectivity using a probabilistic deletion network derived from quantitative proteomics. PLoS One 2009;4:e7310.
[28] Choi H, Kim S, Gingras AC, Nesvizhskii AI. Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. Mol Syst Biol 2010;6:385.
[29] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. Bioinformatics 2008;24:i223–31.
[30] Olinares PD, Ponnala L, van Wijk KJ. Megadalton complexes in the chloroplast stroma of *Arabidopsis thaliana* characterized by size exclusion chromatography, mass spectrometry, and hierarchical clustering. Mol Cell Proteomics 2010;9:1594–615.
[31] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2005;25:25–9.
[32] Nagata T, Nemoto Y, Hasezawa S. Tobacco BY-2 cell line as the 'HeLa' cell in the cell biology of higher plants. Int Rev Cytol 1992;132:1–30.
[33] Carpentier SC, Panis B, Vertommen A, Swennen R, Sergeant K, Renaut J, et al. Proteome analysis of non-model plants: a challenging but powerful approach. Mass Spectrom Rev 2008;27:354–77.
[34] Searle BC, Turner M, Nesvizhskii AI. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. J Proteome Res 2008;7:245–53.
[35] Neuhoff V, Arold N, Taube D, Ehrhardt W. Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023

using Coomassie Brilliant Blue G-250 and R-250. Electrophoresis 1988;9:255–62.
[36] Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 2006;1:2856–60.
[37] Lauber WM, Carroll JA, Dufield DR, Kiesel JR, Radabaugh MR, Malone JP. Mass spectrometry compatibility of two-dimensional gel protein stains. Electrophoresis 2001;22: 906–18.
[38] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
[39] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003;75:4646–58.
[40] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[41] Spath H. Cluster dissection and analysis: theory, FORTRAN programs, examples. New York: Halsted Press; 1985.
[42] Geerts F, Goethals B, Mielikäinen T. Tiling databases. Proceedings of discovery science (DS). Lecture notes in computer scienceSpringer; 2004. p. 278–89.
[43] Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005;21:3448–9.
[44] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2007;2:2366–82.
[45] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009;37: D412–6.
[46] Brandão MM, Dantas LL, Silva-Filho MC. AtPIN: *Arabidopsis thaliana* protein interaction network. BMC Bioinformatics 2009;10:454.
[47] Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res 2010;38:D525–31.
[48] Wiles AM, Doderer M, Ruan J, Gu TT, Ravi D, Blackman B, et al. Building and analyzing protein interactome networks by cross-species comparisons. BMC Syst Biol 2010;4:36.
[49] Agrawal R, Srikant R. Fast algorithms for mining association rules. Proceedings of the Intl. Conf. on Very Large Data Bases (VLDB); 1994. p. 487–99.
[50] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006;440:631–6.
[51] Pang CN, Krycer JR, Lek A, Wilkins MR. Are protein complexes made of cores, modules and attachments? Proteomics 2008;8: 425–34.
[52] Leuendorf JE, Osorio S, Szewczyk A, Fernie AR, Hellmann H. Complex assembly and metabolic profiling of *Arabidopsis thaliana* plants overexpressing vitamin B6 biosynthesis proteins. Mol Plant 2010;3:890–903.
[53] Tanase K, Yamaki S. Purification and characterization of two sucrose synthase isoforms from Japanese pear fruit. Plant Cell Physiol 2000;41:408–14.
[54] Matic S, Akerlund HE, Everitt E, Widell S. Sucrose synthase isoforms in cultured tobacco cells. Plant Physiol Biochem 2004;42:299–306.
[55] Reiter WD, Vanzin GF. Molecular genetics of nucleotide sugar interconversion pathways in plants. Plant Mol Biol 2001;47: 95–113.
[56] Bush MS, Hutchins AP, Jones AM, Naldrett MJ, Jarmolowski A, Lloyd CW, et al. Selective recruitment of proteins to 5′ cap complexes during the growth cycle in *Arabidopsis*. Plant J 2009;59:400–12.
[57] Yang P, Fu H, Walker J, Papa CM, Smalle J, Ju YM, et al. Purification of the Arabidopsis 26S proteasome: biochemical and molecular analyses revealed the presence of multiple isoforms. J Biol Chem 2004;279:6401–13.
[58] Peltier JB, Cai Y, Sun Q, Zabrouskov V, Giacomelli L, Rudella A, et al. The oligomeric stromal proteome of *Arabidopsis thaliana* chloroplasts. Mol Cell Proteomics 2006;5:114–33.
[59] Jordan DB, Bacot KO, Carlson TJ, Kessel M, Viitanen PV. Plant riboflavin biosynthesis. Cloning, chloroplast localization, expression, purification, and partial characterization of spinach lumazine synthase. J Biol Chem 1999;274:22114–21.
[60] Xiao S, Dai L, Liu F, Wang Z, Peng W, Xie D. COS1: an *Arabidopsis* coronatine insensitive1 suppressor essential for regulation of jasmonate-mediated plant defense and senescence. Plant Cell 2004;16:1132–42.
[61] Book AJ, Yang P, Scalf M, Smith LM, Vierstra RD. Tripeptidyl peptidase II. An oligomeric protease complex from *Arabidopsis*. Plant Physiol 2005;138:1046–57.
[62] Nikolau BJ, Ohlrogge JB, Wurtele ES. Plant biotin-containing carboxylases. Arch Biochem Biophys 2003;414:211–22.
[63] Bates PW, Vierstra RD. UPL1 and 2, two 405 kDa ubiquitin-protein ligases from *Arabidopsis thaliana* related to the HECT-domain protein family. Plant J 1999;20:183–95.
[64] Kim DY, Kim KK. Structure and function of HtrA family proteins, the key players in protein quality control. J Biochem Mol Biol 2005;38:266–74.
[65] Sze H, Ward JM, Lai S. Vacuolar H(+)-translocating ATPases from plants: structure, function, and isoforms. J Bioenerg Biomembr 1992;24:371–81.
[66] Drobny M, Schnölzer M, Fiedler S, Lüttge U, Fischer-Schliebs E, Christian AL, et al. Phenotypic subunit composition of the tobacco (*Nicotiana tabacum* L.) vacuolar-type H(+)-translocating ATPase. Biochim Biophys Acta 2002;1564: 243–55.
[67] Domgall I, Venzke D, Lüttge U, Ratajczak R, Böttcher B. Three-dimensional map of a plant V-ATPase based on electron microscopy. J Biol Chem 2002;277:13115–21.
[68] Williams Jr RC, Shah C, Sackett D. Separation of tubulin isoforms by isoelectric focusing in immobilized pH gradient gels. Anal Biochem 1999;275:265–7.
[69] Gardiner J, Marc J. Putative microtubule-associated proteins from the *Arabidopsis* genome. Protoplasma 2003;222:61–74.
[70] Chuong SD, Good AG, Taylor GJ, Freeman MC, Moorhead GB, Muench DG. Large-scale identification of tubulin-binding proteins provides insight on subcellular trafficking, metabolic channeling, and signaling in plant cells. Mol Cell Proteomics 2004;3:970–83.
[71] Young JC, Barral JM, Ulrich Hartl F. More than folding: localized functions of cytosolic chaperones. Trends Biochem Sci 2003;28:541–7.
[72] Srere PA. Complexes of sequential metabolic enzymes. Annu Rev Biochem 1987;56:89–124.
[73] Mooney BP, Miernyk JA, Randall DD. The complex fate of alpha-ketoacids. Annu Rev Plant Biol 2002;53:357–75.
[74] Douce R, Bourguignon J, Neuburger M, Rébeillé F. The glycine decarboxylase system: a fascinating complex. Trends Plant Sci 2001;6:167–76.
[75] Wheeler MC, Tronconi MA, Drincovich MF, Andreo CS, Flügge UI, Maurino VG. A comprehensive analysis of the NADP-malic enzyme gene family of *Arabidopsis*. Plant Physiol 2005;139: 39–51.
[76] Bykova NV, Egsgaard H, Møller IM. Identification of 14 new phosphoproteins involved in important plant mitochondrial processes. FEBS Lett 2003;540:141–6.

Please cite this article as: Remmerie N, et al, Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods, J Prot (2011), doi:10.1016/j.jprot.2011.03.023