SYSTEMS BIOTECHNOLOGY

SIMB
Society for Industrial Microbiology and Biotechnology

Protein–protein interaction network of the marine microalga *Tetraselmis subcordiformis*: prediction and application for starch metabolism analysis

Chaofan Ji · Xupeng Cao · Changhong Yao · Song Xue · Zhilong Xiu

Received: 1 December 2013 / Accepted: 9 May 2014 / Published online: 31 May 2014

Abstract Under stressful conditions, the non-model marine microalga *Tetraselmis subcordiformis* can accumulate a substantial amount of starch, making it a potential feedstock for the production of fuel ethanol. Investigating the interactions of the enzymes and the regulatory factors involved in starch metabolism will provide potential genetic manipulation targets for optimising the starch productivity of *T. subcordiformis*. For this reason, the proteome of *T. subcordiformis* was utilised to predict the first protein–protein interaction (PPI) network for this marine alga based on orthologous interactions, mainly from the general PPI repositories. Different methods were introduced to evaluate the credibility of the predicted interactome, including the confidence value of each PPI pair and Pfam-based and subcellular location-based enrichment analysis. Functional subnetworks analysis suggested that the two enzymes involved in starch metabolism, starch phosphorylase and trehalose-phosphate synthase may be the potential ideal genetic engineering targets.

Keywords Chlorophyceae · Interolog · Ortholog · Post-translational modification

Introduction

The marine green microalga *Tetraselmis subcordiformis* has been shown to accumulate a substantial amount of starch under different stress conditions [33–35]; thus, this microalga is considered to be a potential sustainable feedstock for the large-scale production of fuel ethanol. Until now, however, optimization for both high biomass productivity and high starch content in algal cells has been proven to be difficult [24]. In order to improve algal starch productivity, it would be of great benefit to develop genetic strategies to increase starch accumulation without hampering algal growth [23]. The basis for the genetic strategies is the understanding of biological functions, especially on the protein and enzyme levels.

The study of protein–protein interaction (PPI) networks is currently one of the most active fields because it is a useful tool to investigate functional information in a wide range of biological processes, for instance, transcriptional activation/repression, signal transduction and metabolic regulation [21]. High-throughput screening methods for PPIs, including the yeast two-hybrid system and affinity capture mass spectrometry, are relatively time-consuming and costly. As alternatives, several computational methods based on gene-neighbourhoods, gene fusion, phylogeny, docking, co-expression or interologs were developed to predict PPIs [21]. Of these methods, the interolog approach has been widely used for PPI prediction in several organisms [11–13, 16]. This approach is based on the idea that orthologs are likely to share common functions, including PPIs, in a target organism and the model organisms for which experimental PPI information is available [31]. The

C. Ji and X. Cao have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10295-014-1462-z) contains supplementary material, which is available to authorized users.

C. Ji · Z. Xiu (✉)
School of Life Sciences and Biotechnology, Dalian University of Technology, Dalian 116023, China
e-mail: zhlxiu@dlut.edu.cn

X. Cao · C. Yao · S. Xue (✉)
Marine Bioproducts Engineering Group, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China
e-mail: xuesong@dicp.ac.cn

software InParanoid [22] can accurately identify orthologs in different species. There are several publicly accessible databases from which PPI information can be retrieved, including DIP [32], MINT [6] and BioGRID [5].

Although *T. subcordiformis* is a non-model organism whose entire genome has not been sequenced, its transcriptome and proteome at different growth stages have already been obtained. From the protein sequences, a reliable PPI network will be constructed for *T. subcordiformis* and key components in the regulation of starch metabolism in this alga will be identified, which will potentially provide targets for genetic manipulations on the strain through rational designs to improve starch accumulation.

Materials and methods

The marine green microalga *T. subcordiformis* FACHB-1751 was isolated from the Huanghai Sea near Dalian, Liaoning Province, P.R. China and maintained by the Freshwater Algae Culture Collection of the Institute of Hydrobiology (FACHB-collection), Chinese Academy of Sciences. The medium and photobioreactors for *T. subcordiformis* cultivation were the same as that in the Yao's publication [35]. Aeration was kept at 0.4 vvm with 3 % CO_2 enriched air and the temperature was maintained at 25 ± 2 °C; cool white fluorescent lamps provided an average irradiance of 200 μmol m^{-2} s^{-1}. First, *T. subcordiformis* cells cultivated in a medium full of nitrogen and sulphur were harvested from the late exponential phase. Then, the *T. subcordiformis* cells were resuspended in three media (full of *N* and *S*, depleted *S* or depleted *N*) with an initial cell density of 1.5×10^6 cells mL^{-1}. Algal samples were taken from the three media after 12, 24, and 48 h.

Constructing protein datasets of *T. subcordiformis* based on transcriptome

The transcriptome database of *T. subcordiformis* was obtained from GenBank (http://www.ncbi.nlm.nih.gov/genbank/, accession number: PRJNA203523/GANN01000000). Protein sequences were first aligned by BLAST to protein databases (e-value < 1e−5), including the nr nucleotide database of the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/), the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) and Swiss-Prot terms (http://www.uniprot.org/downloads); the proteins with the highest sequence similarity with the given sequences were retrieved along with the functional annotations of these proteins. Gene Ontology (GO, http://geneontology.org) functional annotation was obtained with nr annotation by the Blast2GO program [8].

Total protein extraction, LC–MS/MS analysis and data analysis were performed according to the published protocol with minor modifications [3, 20, 30]. A brief procedure is as follows: firstly, the algal cells from different cultural conditions were lysed ultrasonically in lysis buffer. Then the protein samples were digested by lysine C. The protein digests were labelled by light, medium, heavy dimethylation reagents on column, then the product was separated by SCX and RPLC in tandem. Eluted peptides were analysed with the LTQ-Orbitrap mass spectrometer equipped with a nano-spray source. Secondly, all acquired files were searched against the transcriptome database of *T. subcordiformis*. Protein quantification was performed using a dimethyl-adapted version of MSQuant (v2.0a81). The quantified proteins were normalised against the log2 of the median ratio of all peptides quantified. For detailed information, please see Electronic Supplementary Material 1.

Protein–protein interactome construction

The complete proteome sets of 12 model organisms (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Zea mays*, *Triticum aestivum* and *Synechocystis* sp. PCC 6803) were downloaded from UniProt (http://www.uniprot.org/downloads) and Ensembl (http://www.ensembl.org). Then, protein sequences from *T. subcordiformis* were aligned to the proteome data of each model organisms using InParanoid 4.1 to identify ortholog pairs. The block substitution matrix (BLOSSUM) was set to 62 for the perl script of InParanoid.

The interactome data of model organisms were collected from the BioGRID database (February 25th, 2013 release, http://thebiogrid.org/download.php) and the DIP database (August 18th, 2012 release http://dip.doe-mbi.ucla.edu/dip/). The PPIs for starch metabolism in maize and wheat were extracted from publications [14, 28, 29].

Orthologous proteins of *T. subcordiformis* that were identified by InParanoid were mapped onto the interactome data from the corresponding model organisms. The confidence value (CV) is the most direct parameter for evaluating the credibility of predicted interactions [11, 16], and CV can be calculated according to the following formula: $CV = N \times E \times S$, where *N* the total number of publications in which the same interaction appeared; *E* the total number of experimental methods by which the same interaction was predicted; and *S* the number of reference species from which the same interaction was recorded.

Protein family (PFAM) domains annotation and enrichment analysis

The domain annotation for the orthologs of *T. subcordiformis* was performed by multiple sequence alignments and the hidden Markov models based on the Pfam-A database (http://pfam.sanger.ac.uk/) [25]. Interactions between the domains in Pfam-A (iPfam) were also downloaded from the FTP site of PFAM (July 29th, 2013 release ver 27.0). We then counted the number of predicted PPIs associated with domain interaction pairs based on the iPfam database. For enrichment analysis, we also constructed the full set of pairwise PPIs using all the nodes and then calculated how many of the generated pairs could be associated with the iPfam database. Finally, a hypergeometric distribution was utilised to calculate the P value:

$$p = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-m}{N-i}}{\binom{N}{n}},$$

where N is the number of all protein pairs constructed by pairing all the nodes; n is the number of predicted PPIs; M is the number of all protein pairs that are associated with iPfam; and m is the number of predicted PPIs that are associated with iPfam.

Subcellular location prediction and enrichment analysis

The subcellular localisation of each protein was predicted by Plant-mPLoc (http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/) [7] based on its amino acid sequence. This predictor for plant protein samples covers the following 12 subcellular compartments: cell wall, chloroplast, cytoplasm, endoplasmic reticulum, extracellular, Golgi apparatus, mitochondrion, nucleus, peroxisome, cell membrane, plastid, and vacuole.

A randomisation algorithm, which is described in detail by Gandhi et al. [10], was used to determine the statistical significance of the enrichment of the interactions in different categories. The self-interacting proteins were deleted in this analysis to avoid spurious enrichment. The following equation was used to calculate the observed number of interologs where one protein is in compartment α ($c_{i\alpha}$) and its pair in β ($c_{j\beta}$). If these two proteins interact, $e_{ij} = 1$; otherwise, $e_{ij} = 0$:

$$n_{\alpha\beta}(\text{obs}) = \sum_{j}\sum_{i<j} \left(c_{i\alpha}c_{j\beta}\ \text{OR}\ c_{i\beta}c_{j\alpha}\right)e_{ij}$$

The probability distribution for the ensemble of random networks maintained the protein annotation, the degree (k) of each protein and the total number of interacting pairs (E). The $\bar{n}_{\alpha\beta}$ was given by:

$$n_{\alpha\beta}(\text{obs}) = \sum_{j}\sum_{i<j} \frac{(c_{i\alpha}c_{j\beta}\ \text{OR}\ c_{i\beta}c_{j\alpha})k_ik_j}{(2E + k_ik_j)}$$

The P value for the observed number of interologs $n_{\alpha\beta}$ was calculated by a Poisson distribution:

$$P(n_{\alpha\beta}) = \begin{cases} \sum_{j=0}^{n_{\alpha\beta}} \bar{n}_{\alpha\beta}^{i} \exp(-\bar{n}_{\alpha\beta})/j! & n_{\alpha\beta} < \bar{n}_{\alpha\beta}\ \text{(depletion)} \\ \sum_{j=n_{\alpha\beta}}^{\infty} \bar{n}_{\alpha\beta}^{i} \exp(-\bar{n}_{\alpha\beta})/j! & n_{\alpha\beta} \geqslant \bar{n}_{\alpha\beta}\ \text{(enrichment)} \end{cases}$$

Finally, the P values were applied to a multiple-testing correction P (multi) $= 1 - (1 - P)^m$, where P is the single-test P value. For enrichment, m equals the number of $\alpha\beta$ pairs with at least one edge in the observed network. For depletion, m equals the number of $\alpha\beta$ pairs possible in the ensemble of random networks.

Topology analysis of PPI network

The Cytoscape [26, 27] plugin NetworkAnalyzer was utilised to compute specific parameters that describe the network topology, including the distribution of node degrees, average clustering coefficients, and the shortest path lengths. The plugin Randomnetworks which creates random networks according to three main models (Erdos–Renyi, Watts-Strogatz, Barabasi–Albert) was used to randomise the predicted PPI network and generate 50 randomised network by the degree preserving random shuffle algorithm. For more information, please visit the following website: http://chianti.ucsd.edu/svn/csplugins/trunk/soc/pjmcswee/src/cytoscape/randomnetwork/.

Results and discussion

Predicted PPIs in *T. subcordiformis*

Based on the proteomic analysis, a total of 2,627 protein sequences were finally identified from the algal cells with the transcriptome of *T. subcordiformis* (26,428 protein coding sequences) as the protein dataset. To construct a map of the PPI networks, first, orthologous proteins were located from the 2,627 sequences. At the end, 1,182 sequences (45 % from the *T. subcordiformis* proteome and 3 % from the *T. subcordiformis* transcriptome) were determined to have orthologs that scored 100 % and matched at least one protein from the reference organisms. Then, the orthology results were used to replace the proteins that interacted with each other in the reference species with the corresponding *T. subcordiformis* proteins. Finally, 12,887 original interactions constructed by 938 algal proteins were identified (Table S1). Then, a total of 7,773 unduplicated PPIs (Table S2) were filtered out from the original 12,887 PPIs, because some of them may have been predicted several times by different methods, in different organisms and/or of different publications. The functional annotation of the 938 nodes is presented in Table S3. A total of 444 nodes received one or more hits from three ontologies (molecular function, cellular component and biological process) of the GO annotation; 767 nodes obtained K numbers and functional

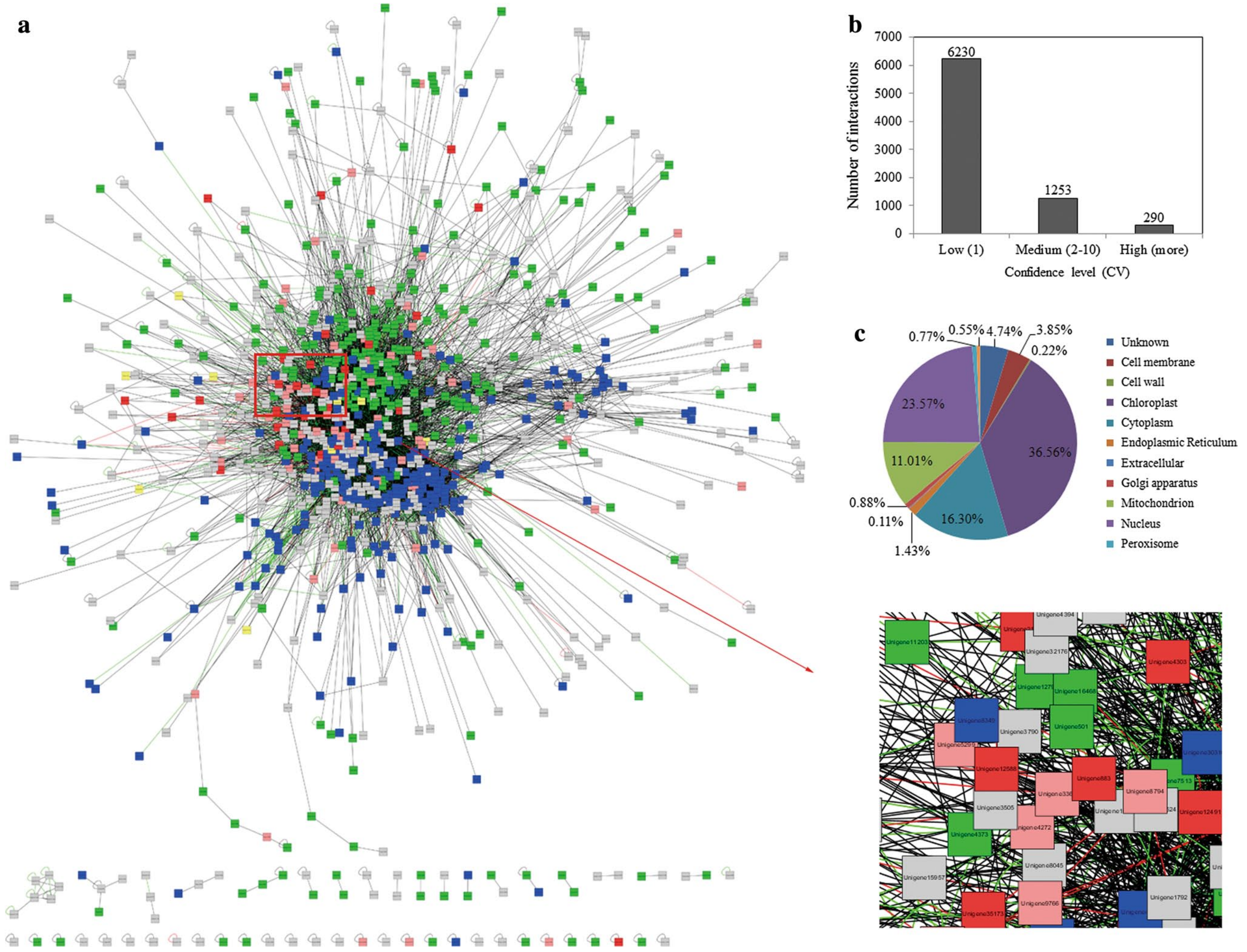


Fig. 1 The predicted protein–protein interaction network of *T. subcordiformis*. **a** The overall view of the PPI network. **b** Confidence values of the predicted PPIs. **c** Subcellular localisation prediction of the nodes. (The *node colour* is based on KEGG molecular functions: *red* cellular processes, *yellow* environmental information processing, *green* metabolism, *blue* genetic information processing, *pink* nodes related to multifunction, and *grey* function unknown. The *edge colour* is based on the CV: *black* low-confidence interactions, *green* medium-confidence interactions, and *red* high-confidence interactions)

annotation from KEGG. Then, based on the interactome of the source species, the interaction map was constructed with 938 proteins and 7,773 PPIs; the Cytoscape visualisation of the map is shown in Fig. 1a.

Validation of the predicted PPI interactions

The CV was calculated for all 7,773 PPIs, which were then divided into three different groups (Fig. 1b): 290 high-confidence interactions (CV > 10), 1,253 medium-confidence interactions (CV = 2–10) and 6,230 low-confidence interactions (CV = 1). Generally, PPIs with a medium- or high-CV are much more credible than low-CV PPIs as they are predicted using various experimental methods, species and/or publications. The PPIs with a low CV are likely to include more false positives, although PPIs with a low CV could be considered more reliable if additional experimental proof becomes available in the future. The 20 PPIs with the highest CV are presented in Table 1. Most of the reliable interactions are highly conserved among eukaryotes and involve proteins with critical functions, such as ubiquitin-related proteins, proteasome complexes, translation-related protein complexes.

Two different methods were introduced in this study to evaluate the global quality of the predicted interactome of *T. subcordiformis* indirectly. First it is known that protein interactions are mediated through the interaction domains [4]. The interaction information between domains can be utilised either to predict PPIs between proteins [4, 21] or to validate the constructed PPI network [13]. In this study, the annotations based on the Pfam-A database indicated that 892 nodes involved in the interactome of *T. subcordiformis*

Table 1 Twenty most conserved protein interactions

Interactor A	Annotation	Interactor B	Annotation	CV
Unigene471	Translation termination factor eRF3	Unigene471	Translation termination factor eRF3	240
Unigene17188	Ubiquitin-conjugating enzyme E2 variant	Unigene30310	Ubiquitin-conjugating enzyme E2 N	210
Unigene16115	26S proteasome regulatory subunit N11	Unigene5018	26S proteasome regulatory subunit N8	200
Unigene35173	Cofilin	Unigene5852	Actin beta/gamma 1	170
Unigene11774	26S proteasome regulatory subunit N1	Unigene15788	26S proteasome regulatory subunit T2	168
Unigene13200	Profilin	Unigene5852	Actin beta/gamma 1	168
Unigene11774	26S proteasome regulatory subunit N1	Unigene16115	26S proteasome regulatory subunit N11	150
Unigene4701	E3 ubiquitin-protein ligase HERC4	Unigene5193	GTP-binding nuclear protein Ran	144
Unigene16326	Translation initiation factor eIF-3 subunit 4	Unigene5144	Translation initiation factor eIF-3 subunit 2	130
Unigene8137	Glutathione S-transferase	Unigene8137	Glutathione S-transferase	128
Unigene11569	26S proteasome regulatory subunit T5	Unigene16115	26S proteasome regulatory subunit N11	120
Unigene12261	26S proteasome regulatory subunit T1	Unigene16115	26S proteasome regulatory subunit N11	120
Unigene11569	26S proteasome regulatory subunit T5	Unigene4533	26S proteasome regulatory subunit T3	110
Unigene16115	26S proteasome regulatory subunit N11	Unigene4296	26S proteasome regulatory subunit N5	108
Unigene11774	26S proteasome regulatory subunit N1	Unigene16359	UV excision repair protein RAD23	104
Unigene15442	Translation initiation factor eIF-3 subunit 9	Unigene5144	Translation initiation factor eIF-3 subunit 2	102
Unigene17201	Histone H2A	Unigene18499	Histone H2A	100
Unigene16115	26S proteasome regulatory subunit N11	Unigene750	26S proteasome regulatory subunit N3	96
Unigene16115	26S proteasome regulatory subunit N11	Unigene9395	26S proteasome regulatory subunit N6	90
Unigene10615	20S proteasome subunit alpha 1	Unigene7939	20S proteasome subunit alpha 7	88

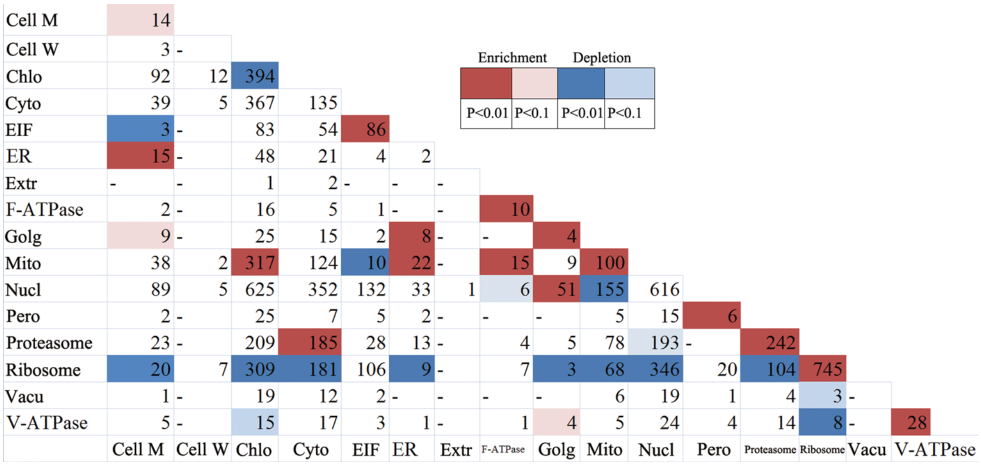
presented different domains (Table S4). These proteins formed 7,353 non-self PPI pairs in the predicted PPI network of *T. subcordiformis*; of these, 283 were associated with Pfam-A interacting domain pairs. In contrast, the full set of 397,386 non-self protein pairs were constructed by 892 nodes, and 4,512 PPIs of these were associated with Pfam-A interacting domain pairs. This result demonstrates that domain-associated interactions are significantly enriched in the predicted PPI network of *T. subcordiformis* (the hypergeometric P value <0.001).

Second, enrichment analysis of the subcellular locations was used to analyse the credibility of the PPIs [11]. With the help of the Plant-mPLoc server, the proteins in the PPI network were categorised into different subcellular locations, as presented in Fig. 1c. We therefore examined the PPIs for enrichment or depletion in the predicted subcellular compartment to validate this trend. In addition, several multi-located protein complexes, such as the proteasome and ribosome, were manually extracted as independent categories for the P value calculation. Most of the statistically significant enriched compartment pairs were those paired with the same categories, as shown in Fig. 2. Furthermore, interactions between two functionally related groups (e.g. mitochondrial and F-type ATPase) are also significantly enriched. Other interactions that are enriched across subcellular compartments (e.g. endoplasmic reticulum–mitochondrion, endoplasmic reticulum–Golgi, mitochondrion–chloroplast and Golgi–nucleus) indicate that some proteins, such as regulatory factors, were localised to more than one compartment.

Topology of the predicted PPI network of *T. subcordiformis*

As for complex biological networks, the topological and dynamic properties control the behaviour of the cell. It is important to analyse the topology of the network since the network construction disclose the intrinsic and/or potential capacity of the organisms and is a prerequisite for the study of dynamic network evolution [1, 9]. Basic network measures to characterise different complex networks include degree, degree distribution, clustering coefficient and shortest path. For a PPI network, the number of connections of a node is called its degree (k) [13]. For the PPI network of *T. subcordiformis*, the value of k_{av} is 16, which indicates that one protein interacts with other 16 proteins on average. In a scale-free network, the degree distribution of the nodes follows a power law, i.e. $P(k) \sim k^{-\gamma}$. The network of *T. subcordiformis* is characterised by a power law with $\gamma = 1.4$ ($R^2 = 0.82$), as shown in Fig. 3a.

The nodes of the PPI network were divided into free ends (111 proteins with single interactions), pipes (80 proteins with two interactions) and other hubs (747 proteins with three to more than 100 interactions), as described in Fig. 3b. The highly connected nodes (the so-called hubs) ensure the topological integrity of the network.



	Cell M	Cell W	Chlo	Cyto	EIF	ER	Extr	F-ATPase	Golg	Mito	Nucl	Pero	Proteasome	Ribosome	Vacu	V-ATPase
Cell M	14															
Cell W	3	-														
Chlo	92	12	394													
Cyto	39	5	367	135												
EIF	3	-	83	54	86											
ER	15	-	48	21	4	2										
Extr	-	-	1	2	-	-	-									
F-ATPase	2	-	16	5	1	-	-	10								
Golg	9	-	25	15	2	8	-	-	4							
Mito	38	2	317	124	10	22	-	15	9	100						
Nucl	89	5	625	352	132	33	1	6	51	155	616					
Pero	2	-	25	7	5	2	-	-	-	5	15	6				
Proteasome	23	-	209	185	28	13	-	4	5	78	193	-	242			
Ribosome	20	7	309	181	106	9	-	7	3	68	346	20	104	745		
Vacu	1	-	19	12	2	-	-	-	-	6	19	1	4	3	-	
V-ATPase	5	-	15	17	3	1	-	1	4	5	24	4	14	8	-	28

Fig. 2 Enrichment analyses of PPIs with subcellular localisation. The numbers in the matrix are compartment pairs, whose colour demonstrates the fold enrichment or depletion compared with an ensemble of random networks, as indicated in the figure. (*Cell W* cell wall, *Cell M* cell membrane, *Chlo* chloroplast, *Cyto* cytoplasm, *ER* endoplasmic reticulum, *Extr* extracellular, *Glog* Golgi apparatus, *Mito* mitochondrion, *Nucl* nucleus, *Pero* peroxisome, *Vacu* vacuole)

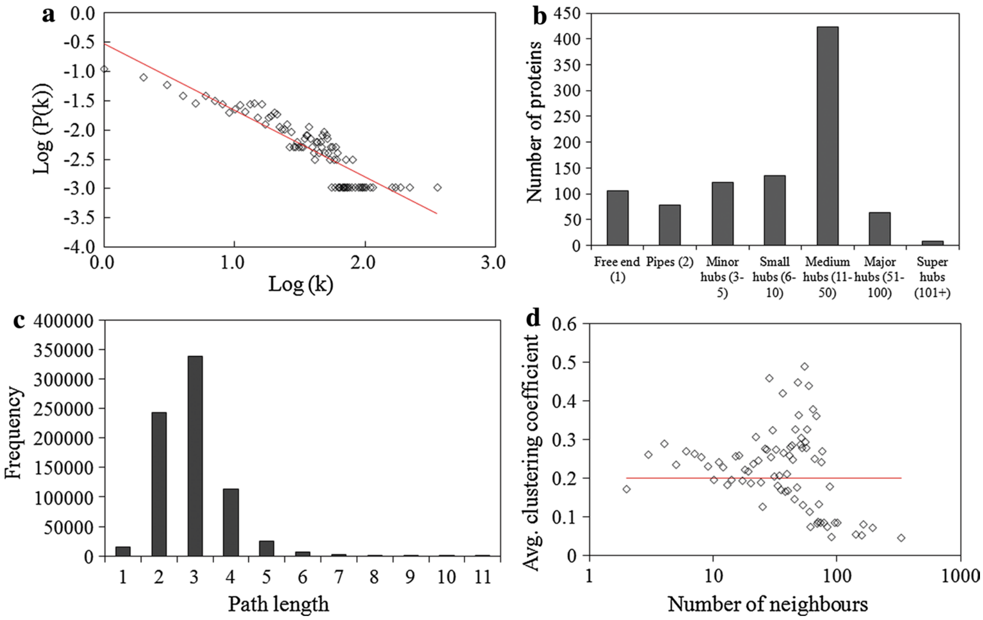


Fig. 3 Topological parameters of the PPI network of *T. subcordiformis*. **a** Distribution of the number of degrees of nodes in the PPI network with both axes plotted on logarithmic scales. **b** Different types of protein nodes classified based on their degree. **c** The shortest path length distribution. **d** Clustering coefficient of the nodes

Genome-wide studies demonstrated that knocking out the hub genes appears to confer a greater rate of lethality than knocking out other genes, known as the centrality–lethality rule [17]. Table 2 lists the 20 most highly connected protein interaction hubs, including subunits of the proteasome, ribosomal proteins, and heat shock protein. Mapping the biological function of these based on the dataset of KEGG revealed that most of the nodes are involved in important genetic information processing and cellular processes.

We calculated the shortest path and the diameter of the PPI network of *T. subcordiformis*; the results demonstrate that the average path length is 2.9 and that the lengths of 99 % of the paths are less than 6 (a 'small word' property), as presented in Fig. 3c. The clustering coefficient of a node in a network is a measure of the inter-connectivity between its neighbours. The mean clustering coefficient of the PPI network of *T. subcordiformis* is 0.20, as shown in Fig. 3d. The average clustering coefficient of 50 randomised PPI networks generated by the plugin Randomnetworks is 0.11 ± 0.005. Compared with randomised networks, the predicted PPI network shows a higher clustering coefficient among the nodes, indicating that the predicted PPI network is densely connected.

Starch metabolism related PPI subnetworks

For *T. subcordiformis*, nutrient stress has been the traditional method for increasing cellular starch accumulation [35]. However, the stress factors are harmful to the algae, decreasing the growth rate simultaneously. Compared with nutrient manipulation approaches, genetic engineering results in an increasingly reproducible and predictable system [2]. Identifying functional regulatory factors involved in starch metabolism through the analysis the PPI

Table 2 Twenty most highly connected protein interaction hubs

ID	Edges	Annotation	Biological function[a]
Unigene16115	330	26S proteasome regulatory subunit N11	G
Unigene11774	195	26S proteasome regulatory subunit N1	G
Unigene11569	165	26S proteasome regulatory subunit T5	G
Unigene17242	158	Small ubiquitin-related modifier	G
Unigene26714	143	Molecular chaperone HtpG	E, G
Unigene5852	102	Actin beta/gamma 1	C, E
Unigene4401	95	Histone H3	G
Unigene12160	91	Far upstream element-binding protein	G
Unigene8197	89	–	Unknown
Unigene12522	83	Rab family, other	G
Unigene3106	78	Protein phosphatase 1, catalytic subunit	C, E, G
Unigene15155	75	Small subunit ribosomal protein S3e	G
Unigene1711	75	Threonine synthase	M
Unigene30314	74	Small subunit ribosomal protein S8e	G
Unigene18499	73	Histone H2A	G
Unigene497	70	Cell division protease FtsH	G
Unigene11994	69	Histone acetyltransferase MYST1	G
Unigene8825	68	Heat shock 70 kDa protein 1/8	C, E, G
Unigene8952	68	Large subunit ribosomal protein L3e	G
Unigene2206	66	Small subunit ribosomal protein S5e	G

[a] Annotation for biological function of each node is based on the KEGG PATHWAY (http://www.genome.jp/kegg-bin/get_htext?br08901.keg) representing the knowledge on the molecular interaction and reaction networks for: cellular processes (C), genetic information processing (G), environmental information processing (E), and metabolism (M)

subnetwork would provide the potential genetic targets to improve starch accumulation in *T. subcordiformis*.

The starch metabolism related subnetwork was constructed from the PPIs for which at least one node was annotated as a starch and sucrose metabolism pathway (map00500 in KEGG, the partial map is presented in Fig. 4a) according to the *K* numbers of the nodes, as presented in Fig. 4b. Besides, evidence has been reported that higher plants contain multienzyme complexes comprising starch synthase (SS), 1,4-alpha-glucan branching enzyme (SBE) and other enzymes [14, 28, 29]. Starch biosynthesis is accomplished largely by the coordinated actions of these enzymes; however, the PPI datasets did not have such information. As a result, published studies about these complexes were mined to identify the PPIs; meanwhile, the protein sequences involved in the starch metabolism of these two organisms were downloaded from UniProt. Then, the proteome of *T. subcordiformis* was searched for orthologs. With these results, the different isoforms of SS, SBE and both subunits of ADP-glucose pyrophosphorylase (AGPase) were mapped into the subnetworks. As presented in Fig. 4b, UDP-glucose-6-dehydrogenase (UGDH), 4-alpha-glucanotransferase (malQ), glucokinase (GLK) and SS IV form homo- or heteropolymers separately. Other enzymes and their neighbours form a more complex subnetwork, with trehalose-phosphate synthase (TPS), starch phosphorylase (PYG), UTP-glucose-1-phosphate uridylyltransferase (UGP2), phosphoglucomutase (PGM), glucose-6-phosphate isomerase (GPI) and SBE as hubs. These predicted results were supplemented by the data in higher plants, where multienzyme complexes mainly comprise SS and SBE. TPS is a core hub in this subnetwork because it possesses the most neighbours and connects to other hubs to maintain the integrity of the network. Perhaps the reason for this partially lies in the product of TPS, trehalose, which serves as a signalling molecule to regulate carbohydrate metabolism in plants.

Since protein kinases and phosphatases are critical for regulating starch-synthesising enzymes [18], protein kinase- and phosphatase-related subnetworks were composed from the PPIs for which at least one node was functionally annotated as a kinase or protein phosphatase, as shown in Fig. 4c, d, respectively. Eleven types of protein kinases were predicted to be involved in 216 interactions, and eight types of protein phosphatases were predicted to interact with 176 proteins in the PPI subnetworks of *T. subcordiformis*. In our PPI network, PGM is demonstrated to interact with protein phosphatase 2 (PP, 2) (Fig. 4d). The most connected hub TPS in Fig. 4b is regulated by several types of kinases, including cGMP-dependent protein kinase (PKG), extracellular signal-regulated kinase (ERK) and cycling-dependent kinase 1 (CDK1) (Fig. 4c). UGP2 interacts with the catalytic subunit of protein phosphatase 1 (PP1, C) (Fig. 4d).

In maize and wheat, the stability of an enzyme complex involved in starch metabolism depends on the

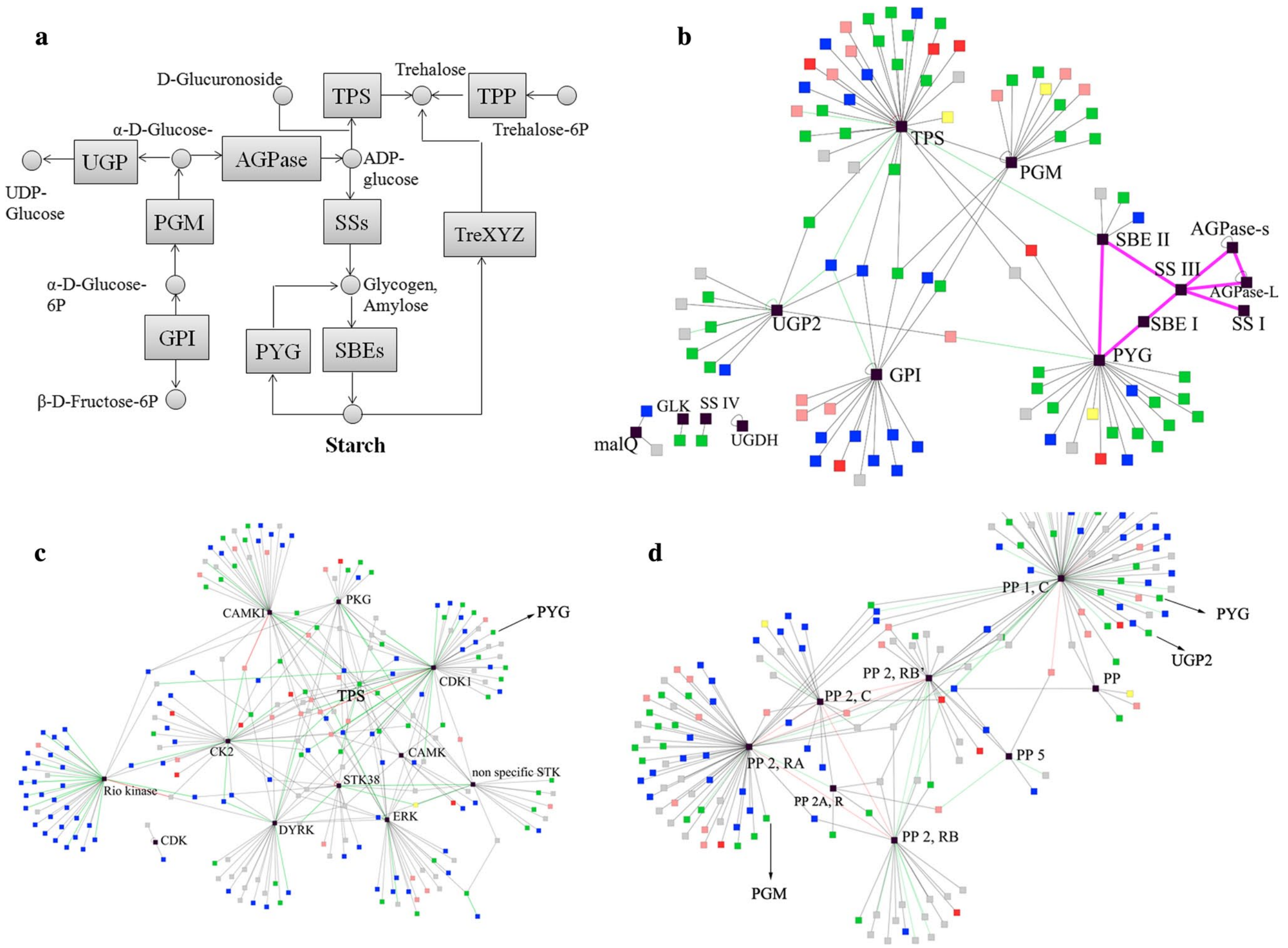


Fig. 4 PPI subnetworks involved in starch metabolism and phosphorylation modification. **a** Starch metabolism pathway based on KEGG PATHWAY (map00500) information. **b** The subnetworks were constructed using proteins involved in starch metabolism. **c**, **d** The subnetworks were constructed using interactions that involve proteins with kinases and protein phosphatases and their interacting neighbours, respectively. [The *black* nodes represent enzymes involved in starch metabolism in Fig. 4a, kinases in Fig. 4c, protein phosphatases in Fig. 4d. The nodes connected with lilac and bold edges represent the PPI pairs predicted from maize (*Z. mays*) and wheat (*T. aestivum*) though literature mining. The *other colours* of the nodes and edges have the same meaning as in Fig. 1]

phosphorylation status of the constituent proteins [15, 29]. However, the kinase and the corresponding target protein are still unknown. In starch metabolism, a dynamic mediatory role between starch synthesis and degradation has been ascribed to PYG, which is also an important hub in the predicted PPI subnetwork for starch metabolism. This enzyme interacts with the enzymes that catalyse starch anabolism and other enzymes involved in the relevant up- or down-stream reactions. The subnetwork of *T. subcordiformis* presented in Fig. 4c, d suggest that PP1, C and CDK 1 may be responsible for the dephosphorylation and phosphorylation of PYG, respectively. Meanwhile, PYG is also an important component of the protein complexes in starch metabolism, as shown in Fig. 4b. It is assumed that the phosphorylation status of PYG would influence the catalytic activity or the stability of the enzyme complexes. The subnetwork analysis suggested that two enzymes involving in starch metabolism, TPS and PYG may be the potential ideal genetic engineering targets for optimising starch accumulation in *T. subcordiformis*. There has been experimental evidence supporting this prediction. In *C. reinhardtii*, quantitative proteomics results indicated that PYG was the most significantly up-regulated enzyme in starch metabolism when the ammonium in the culture was generally exhausted [19]. For *T. subcordiformis*, three unigenes were annotated to different isoforms of PYG, among which Unigene12096 was predicted to have PPI partners as presented in Table S1. The expression level of PYG was 4- to 10-fold higher when placed in *N*-depleted medium: three isoforms of PYG were all significantly up-regulated at 24 h (data not shown), namely at the moment when *T. subcordiformis* cells possessed a maximum starch productivity [35].

As a result, further experimental investigations to prove this presumption are planned and are expected to provide a better understanding of the role of starch phosphorylase and trehalose-phosphate synthase in the regulation of starch biosynthesis.

Conclusions

Deep analysis of regulatory mechanism of starch metabolism in *T. subcordiformis* is an example to utilise the PPI network. In addition, this first PPI map for a marine microalga will be a powerful tool for predicting the biological functions of unknown genes and discovering the essential regulatory proteins in various metabolic reactions. This map could provide more information about the utilisation of eukaryotic algae for producing renewable biofuels.

Acknowledgments The authors gratefully acknowledge the support of the Hundred Talent Program of the Chinese Academy of Sciences (No. A1097) and thank Dr. Hongwei Liu and Mr. Keyue Wang for the proteomic data collection.

References

1. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113. doi:10.1038/nrg1272
2. Beer LL, Boyd ES, Peters JW, Posewitz MC (2009) Engineering algae for biohydrogen and biofuel production. Curr Opin Biotech 20:264–271. doi:10.1016/j.copbio.2009.06.002
3. Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat Protocols 4:484–494. doi:10.1038/nprot.2009.21
4. Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, Mana M, de Lichtervelde L, Mul JD, van de Peut D, Devos M, Simonis N, Yildirim MA, Cokol M, Kao H-L, de Smet A-S, Wang H, Schlaitz A-L, Hao T, Milstein S, Fan C, Tipsword M, Drew K, Galli M, Rhrissorrakrai K, Drechsel D, Koller D, Roth FP, Iakoucheva LM, Dunker AK, Bonneau R, Gunsalus KC, Hill DE, Piano F, Tavernier J, van den Heuvel S, Hyman AA, Vidal M (2008) A protein domain-based interactome network for *C. elegans* early embryogenesis. Cell 134:534–545. doi:10.1016/j.cell.2008.07.009
5. Chatr-aryamontri A, Breitkreutz B-J, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res 41:D816–D823. doi:10.1093/nar/gks1158
6. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the molecular interaction database. Nucleic Acids Res 35:D572–D574. doi:10.1093/nar/gkl950
7. Chou K-C, Shen H-B (2010) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS One 5:e11335. doi:10.1371/journal.pone.0011335
8. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676. doi:10.1093/bioinformatics/bti610
9. Florez A, Park D, Bhak J, Kim B-C, Kuchinsky A, Morris J, Espinosa J, Muskus C (2010) Protein network prediction and topological analysis in *Leishmania* major as a tool for drug target selection. BMC Bioinform 11:484. doi:10.1186/1471-2105-11-484
10. Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38:285–293. doi:10.1038/ng1747
11. Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M (2007) A predicted interactome for *Arabidopsis*. Plant Physiol 145:317–329. doi:10.1104/pp.107.103465
12. Guo J, Li H, Chang J-W, Lei Y, Li S, Chen L–L (2013) Prediction and characterization of protein–protein interaction network in *Xanthomonas oryzae* pv. *oryzae* PXO99A. Res Microbiol 164:1035–1044. doi:10.1016/j.resmic.2013.09.001
13. He F, Zhang Y, Chen H, Zhang Z, Peng Y-L (2008) The prediction of protein–protein interaction networks in rice blast fungus. BMC Genom 9:519. doi:10.1186/1471-2164-9-519
14. Hennen-Bierwagen TA, Lin Q, Grimaud F, Planchot V, Keeling PL, James MG, Myers AM (2009) Proteins from multiple metabolic pathways associate with starch biosynthetic enzymes in high molecular weight complexes: a model for regulation of carbon allocation in maize amyloplasts. Plant Physiol 149:1541–1559. doi:10.1104/pp.109.135293
15. Hennen-Bierwagen TA, Liu F, Marsh RS, Kim S, Gan Q, Tetlow IJ, Emes MJ, James MG, Myers AM (2008) Starch biosynthetic enzymes from developing maize endosperm associate in multisubunit complexes. Plant Physiol 146:1892–1908. doi:10.1104/pp.108.116285
16. Ho C-L, Wu Y, Shen H-b, Provart N, Geisler M (2012) A predicted protein interactome for rice. Rice 5:1–14. doi:10.1186/1939-8433-5-15
17. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42. doi:10.1038/35075138
18. Kötting O, Kossmann J, Zeeman SC, Lloyd JR (2010) Regulation of starch metabolism: the age of enlightenment? Curr Opin Plant Biol 13:320–328. doi:10.1016/j.pbi.2010.01.003
19. Lee DY, Park J–J, Barupal DK, Fiehn O (2012) System response of metabolic networks in *Chlamydomonas reinhardtii* to total available ammonium. Mol Cell Proteomics 11:973–988. doi:10.1074/mcp.M111.016733
20. Lemeer S, Jopling C, Gouw J, Mohammed S, Heck AJR, Slijper M, den Hertog J (2008) Comparative phosphoproteomics of zebrafish Fyn/Yes morpholino knockdown embryos. Mol Cell Proteomics 7:2176–2187. doi:10.1074/mcp.M800081-MCP200
21. Liu Z-P, Chen L (2012) Proteome-wide prediction of protein–protein interactions from high-throughput data. Protein Cell 3:508–520. doi:10.1007/s13238-012-2945-1
22. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38:D196–D203. doi:10.1093/nar/gkp931
23. Pignolet O, Jubeau S, Vaca-Garcia C, Michaud P (2013) Highly valuable microalgae: biochemical and topological aspects. J Ind Microbiol Biotechnol 40:781–796. doi:10.1007/s10295-013-1281-7
24. Procházková G, Brányiková I, Zachleder V, Brányik T (2013) Effect of nutrient supply status on biomass composition of eukaryotic green microalgae. J Appl Phycol pp 1–19. doi:10.1007/s10811-013-0154-922

25. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301. doi:10.1093/nar/gkr1065
26. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P, Lotia S, Pico AR, Bader GD, Ideker T (2012) A travel guide to cytoscape plugins. Nat Meth 9:1069–1076. doi:10.1038/nmeth.2212
27. Smoot ME, Ono K, Ruscheinski J, Wang P, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431–432. doi:10.1093/bioinformatics/btq675
28. Tetlow IJ, Beisel KG, Cameron S, Makhmoudova A, Liu F, Bresolin NS, Wait R, Morell MK, Emes MJ (2008) Analysis of protein complexes in wheat amyloplasts reveals functional interactions among starch biosynthetic enzymes. Plant Physiol 146:1878–1891. doi:10.1104/pp.108.116244
29. Tetlow IJ, Wait R, Lu Z, Akkasaeng R, Bowsher CG, Esposito S, Kosar-Hashemi B, Morell MK, Emes MJ (2004) Protein phosphorylation in amyloplasts regulates starch branching enzyme activity and protein–protein interactions. Plant Cell 16:694–708. doi:10.1105/tpc.017400
30. van Breukelen B, van den Toorn HWP, Drugan MM, Heck AJR (2009) StatQuant: a post-quantification analysis toolbox for improving quantitative mass spectrometry. Bioinformatics 25:1472–1473. doi:10.1093/bioinformatics/btp181
31. Walhout AJM, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science 287:116–122. doi:10.1126/science.287.5450.116
32. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim S-M, Eisenberg D (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30:303–305. doi:10.1093/nar/30.1.303
33. Y ao C, Ai J, Cao X, Xue S (2013) Characterization of cell growth and starch production in the marine green microalga *Tetraselmis subcordiformis* under extracellular phosphorus-deprived and sequentially phosphorus-replete conditions. Appl Microbiol Biot 97:6099–6110. doi:10.1007/s00253-013-4983-x
34. Yao C, Ai J, Cao X, Xue S (2013) Salinity manipulation as an effective method for enhanced starch production in the marine microalga *Tetraselmis subcordiformis*. Biores Technol 146:663–671. doi:10.1016/j.biortech.2013.07.134
35. Yao C, Ai J, Cao X, Xue S, Zhang W (2012) Enhancing starch production of a marine green microalga *Tetraselmis subcordiformis* through nutrient limitation. Bioresource Technol 118:438–444. doi:10.1016/j.biortech.2012.05.030