bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Molecular Mass and Isoelectric Point of Plant Proteomes

Tapan Kumar Mohanta*[1], Abdullatif Khan[1], Abeer Hashem[2], Elsayed Fathi Abd_Allah[3], Ahmed Al-Harrasi[1]

[1]Natural and Medical Science Research Centre, University of Nizwa, 616, Oman

[2]Botany and Microbiology Department, King Saud University, Riyadh, 11451, Saudi Arabia

[3]Plant Production Department, King Saud University, Riyadh, 11451, Saudi Arabia

Corresponding author: Tapan Kumar Mohanta, E-mail: nostoc.tapan@gmail.com

Abstract

A proteomic analysis of proteomes from 145 plant species revealed a *pI* range of 1.99 (epsin) to 13.96 (hypothetical protein). The molecular mass of the plant proteins ranged from 0.54 to 2236.8 kDa. A putative Type-I polyketide synthase (22244 amino acids) in *Volvox carteri* was found to be the largest protein in the plant kingdom and was not found in higher plant species. Titin (806.46 kDa) and misin/midasin (730.02 kDa) were the largest proteins identified in higher plant species. The *pI* and molecular weight of the plant proteome exhibited a trimodal distribution. An acidic *pI* (56.44% of proteins) was found to be predominant over a basic *pI* (43.34% of proteins) and the abundance of acidic *pI* proteins was higher in unicellular algae species relative to multicellular higher plants. In contrast, the seaweed, *Porphyra umbilicalis*, possesses a higher proportion of basic *pI* proteins (70.09%). Plant proteomes were also found to contain the amino acid, selenocysteine (Sec), which is the first report of the presence of this amino acid in plants. Additionally, plant proteomes also possess ambiguous amino acids Xaa (unknown), Asx (asparagine or aspartic acid), Glx (glutamine or glutamic acid), and Xle (leucine or isoleucine) as well.

Key words: Proteome, amino acids, Isoelectric point, Molecular weight, Selenocysteine, Pyrrolysine

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The isoelectric or isoionic point of a protein is the pH at which a protein carries no net electrical charge and hence is considered neutral [1–4]. The zwitterion form of a protein becomes dominant at neutral pH. The *pI* of polypeptides is largely dependent on the dissociation constant of the ionisable groups [5]. The major ionisable groups present in the amino acids are arginine, aspartate, cysteine, histidine, glutamate, lysine, and glutamate, where they play a major role in determining the *pI* of a protein [6–8]. Co-translational and post-translational modifications of a protein, however, can also play a significant role in determining the *pI* of a protein [9,10]. The exposure of charged residues to the solvents, hydrogen bonds (diploe interactions) and dehydration also impact the *pI* of a protein [11,12]. The inherent *pI* of protein, however, is primarily based on its native protein sequence. The *pI* of a protein is crucial to understanding its biochemical function and thus determining *pI* is an essential aspect of proteomic studies. During electrophoresis, the direction of movement of a protein in a gel or other matrix depends its'*pI*, hence numerous proteins can be separated based on their *pI* [13–16]. Given the impact of post-translational modifications and other biochemical alterations (phosphorylation, methylation, alkylation), however, the predicted *pI* of a protein will certainly be different than the predicted *pI*; the latter of which is based on the composition of amino acids in a protein [9,17,18]. Nonetheless, an estimated isoelectric point is highly important and a commonly identified parameter.

Several studies have been conducted to understand the *pI* of proteins/polypeptides [3,19–21]. These studies have been mainly based on animal, bacteria, and virus models and databases containing the *pI* of experimentally verified proteins. None of these databases, however, contain more than ten thousand proteins sequences which is very few relative to the availability of proteomic data. Therefore, an analysis was conducted of the *pI* and molecular weight of proteins from 144 plant

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

species which included 5.87 million protein sequences. This analysis provides an in-depth analysis of the *pI* and molecular mass of the proteins in the plant kingdom.

Results and Discussion

Plant proteins range from 0.54 kDa to 2236.8 kDa

A proteome-based analysis of plant proteins of 144 plant species that included more than 5.86 million protein sequences was conducted to determine the molecular mass, *pI*, and amino acid composition of proteins that exist in plant proteomes (Table 1). The analysis indicated that *Hordeum vulgare* possessed the highest number (248180) of protein sequences, while *Helicosporidium* sp. had the lowest number (6033). On average, plant proteomes possess 40988.66 protein sequences per species. The analysis also revealed that the molecular mass of plant proteomes ranged from 0.54 kDa to 2236.8 kDa. *Volvox carteri* was found to possess the largest plant protein (XP_002951836.1) of 2236.8 kDa, containing 22244 amino acids (*pI* 5.94), while *Citrus unshiu* possessed the smallest protein of 0.54 kDa, containing only four amino acids (*pI* 5.98) (id: GAY42954.1). This is the first analysis to document the largest (2236.8 kDa) and smallest (0.54 kDa) protein in the plant kingdom. These two proteins have not been functionally annotated and BLASTP analysis in the NCBI database did not identify suitable similarity with any other proteins. A few domains present in the largest protein, however, were found to be conserved with Type-I polyketide synthase. The molecular mass of some other high molecular mass proteins were: 2056.44 kDa (id: XP_001698501.1, type-1 polyketide synthase, *pI*: 6.00, aa: 21004, *Chlamydomonas reinhardtii*); 1994.71 kDa (id: XP_001416378.1, polyketide synthase, *pI*: 7.38, aa: 18193, *Ostreococcus lucimarinus*); 1932.21 kDa (id: Cz02g22160.t1, unknown protein, *pI*: 5.7, aa: 18533, *Chromochloris zofingiensis*); 1814.1 kDa (id: XP_007509537.1, unknown protein, *pI*: 4.46, aa: 16310, *Bathycoccus prasinos*); 1649.26 kDa (id: XP_011401890.1, polyketide synthase, *pI*: 5.53, aa: 16440, *Auxenochlorella protothecoides*); 1632.35 kDa (id: XP_005650993.1, ketoacyl-synt-

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

domain-containing protein, *pI*: 5.86, aa: 15797, *Coccomyxa subellipsoidea*); 1532.91 kDa (id: XP_002507643.1, polyketide synthase, *pI*: 7.07, aa: 14149, *Micromonas commoda*); 1370.23 kDa (id: GAX78753.1, hypothetical protein CEUSTIGMA, *pI*: 5.97, aa: 13200, *Chlamydomonas eustigma*); 1300.83 kDa (id: XP_022026115.1, unknown protein/filaggrin-like, *pI*: 11.75, aa: 12581, *Helianthus annuus*); 1269.42 kDa (id: XP_009350379.1, unknown protein, *pI*: 5.37, aa: 11880, *Pyrus bretschneideri*); 1237.34 kDa (id: XP_022840687.1, polyketide synthase, *pI*: 7.30, aa: 11265, *Ostreococcus tauri*); 1159.35 kDa (id: XP_005847912.1, polyketide synthase, *pI*: 5.91, aa: 11464, *Chlorella variabilis*); 1150.02 kDa (id: PKI66547.1, unknown protein, *pI*: 3.87, aa: 11234, *Punica granatum*); 1027.64 kDa (id: Sphfalx0133s0012.1, unknown protein, *pI*: 4.05, aa: 9126, *Sphagnum fallax*); 909.93 kDa (id: XP_002985373.1, unknown/titin-like protein, *pI*: 4.02, aa: 8462, *Selaginella moellendorffii*); 881.59 kDa (id: KXZ46216.1, hypothetical protein, *pI*: 5.80, aa: 8881, *Gonium pectorale*); 848.29 kDa (id: XP_003056330.1, *pI*: 6.12, aa: 7926, *Micromonas pusilla*); 813.31 kDa (id: GAQ82263.1, unknown protein, *pI*: 4.60, aa: 7617, *Klebsormidium nitens*), 806.46 kDa (id: XP_017639830.1, titin-like, *pI*: 4.21, aa: 7209, *Gossypium arboreum*); 806.12 kDa (id: OAE35580.1, *pI*: 4.83, hypothetical protein, aa: 7651, *Marchantia polymorpha*); and 802.74 kDa, (id: XP_012444755.1, titin-like, *pI*: 4.19, aa: 7181, *Gossypium raimondii*) (Table 1).

On average, approximately 7.38 % of the analysed proteins were found to contain ≥ 100 kDa proteins. The algal species, *V. carteri*, was found to encode largest plant protein (putative polyketide synthase); while other unicellular algae, and multi-cellular lower eukaryotic plants, including bryophytes and pteridophytes, were also found to encode some of the larger proteins (e.g. ketoacyl synthase) in the plant kingdom. The higher eukaryotic plants, including gymnosperms and angiosperms, were not found to encode a high molecular mass polyketide synthase protein. They did, however, possess the high molecular mass proteins; titin (806.46 kDa), misin/midasin (730.02 kDa), futsch (622.14 kDa), filaggrin (644.4 kDa), auxin transport

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

protein BIG (568.4 kDa), and von Willebrand factor (624.74 kDa) (Table 1). Titin is an extremely large protein that is greater than 1 µM in length and found in human striated muscle [22,23]. The largest titin protein found in plants, however, was only 806.46 kDa (*Gossypium arboreum*). The predicted formula of the 806.46 kDa titin protein was $C_{33863}H_{54610}N_{9232}O_{13061}S_{200}$ and its estimated half-life was 10-30 hours; whereas the predicted formula of the 2236.8 kDa protein of *V. carteri* was $C_{97783}H_{157401}N_{28489}O_{30265}S_{637}$. Almost all of the higher eukaryotic plants were found to possess titin, misin/midasin, and auxin transport protein BIG proteins. Species of unicellular algae were not found to possess titin or misin/midasin proteins. This suggests that titin and misin/midasin proteins originated and evolved in more complex, multicellular organisms rather than unicellular organisms. Thus, the evolution of titin, misin/midasin proteins may also be associated with the evolution of terrestrial plants from aquatic plants.

The presence of the smallest molecular mass protein, other than the tripeptide glutathione (Cys, Gly, and Glu), was also determined. A 0.54 kDa molecular mass protein, containing only four amino acids (MIMF) and starting with methionine and ending with phenylalanine, was identified in *Citrus unshiu* (id: GAY42954.1) (Table 1). Other low molecular mass plant proteins were 0.57 kDa (NP_001336532.1/ AT5G23115, *Arabidopsis thaliana*) and 0.63 kDa (AH003201-RA, *Amaranthus hypochondriacus*). Small proteins found in *A. thaliana* was MNPKS and that found in *A. hypochondriacus* was MLPYN, contained only five amino acids. These low molecular mass proteins were not present in all of the studied species and their cellular and molecular functions have not been reported yet. One of the universal small molecular weight plant proteins, however, was identified as cytochrome b6/f complex subunit VIII (chloroplast) (MDIVSLAWAALMVVFTFSLSLVVWGRSGL) that contains only 29 amino acids. Cytochrome b6/f is actively involved in the electron transfer system of

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

photosystem II and regulates photosynthesis [24–28]. It is commonly known that glutathione is the smallest functional polypeptide and that it plays diverse roles in cell signaling [29–31]. The tetra and penta peptides identified in the present analysis, however, were quite different from glutathione and none of them contained Cys, Gly, or Glu amino acids, as found in glutathione. Polypeptides with less than 100 amino acids are considered small proteins and studies indicate that many small proteins are involved in cell metabolism, cell signaling, cell growth, and DNA damage [32–35]. In the era of next-generation sequencing, small protein-coding genes are completely overlooked during genome annotation and get buried amongst an enormous number of open reading frames [36]. Therefore, it is difficult to identify more numbers of small proteins in plants.

A previously conducted comparative study revealed that plant proteins are comparatively smaller than animal proteins, as the former are encoded by fewer exons [37]. Longer proteins harbour more conserved domains and hence display a greater number of biological functions than short proteins. The average protein length of the studied plant species was 424.34 amino acids. A previous study reported the average length of eukaryotic proteins to be 472 amino acids and that the average length of plant proteins is approximately 81 amino acids shorter than animal proteins [37]. Our analysis indicates, however, that plant proteins are approximately 47.66 amino acid shorter than animal proteins. In addition, studies have also indicated that eukaryotic proteins are longer than bacterial proteins and that eukaryote genomes contain approximately 7 fold more proteins (48% larger) than bacterial genomes [38]. Although the average size of plant proteins was found to be 424.34 amino acids, the average protein size of lower, eukaryotic unicellular aquatic plant species; including *Chlamydomonas eustigma*, *Volvox carteri*, *Klebsordium nitens*, *Bathycoccus prasinos*, and *Durio zibethinus*, was found to be 576.56, 568.22, 538.73, 521.05, and 504.36 amino acids, respectively. This indicates that

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

unicellular plant species have an average protein size that is larger than terrestrial multicellular complex plant species, suggesting that the evolution of plant proteins involved a loss of protein size and hence gene size. The cause of the variability in protein length in the phylogenetic lineage of eukaryotic plants has yet to be elucidated. A multitude of evolutionary factors, including deletion (loss of exons) or fusion of multiple domains of proteins, may have played critical roles in shaping the size of higher plant proteins. Transposon insertion and splitting of genes increases the number of proteins but reduces the average size of the proteins [39–42]. Higher plants contain a very large number of transposable elements and therefore these elements are the most responsible factor to expect to have played a major role in increasing protein numbers and reducing the protein size in higher plants. The percentage of transposable elements in a genome is directly proportional to the genome size of the organism and varies from approximately 3% in small genomes to approximately 85% in large genomes [41]. Kirag et al (2007) reported a significant correlation between protein length and the *pI* of a protein [19]. In our analysis, however, no correlation was found between protein length and the *pI* of a protein. For example, titin and misin are two of the larger proteins in plants and they fall in the acidic *pI* range, but not the alkaline *pI* range.

Plant encode a higher number of proteins than animals and fungi

Our analysis identified an average of 40469.47 proteins per genome (Table 1). Previously the number of proteins in plant species was reported as 36795 per genome [37]. On average, animals and fungi encode 25189 and 9113 proteins per genome, respectively [37]. An average of 40469.47 proteins per plant genome is 62.24% higher than in animals and 444.08% higher than in fungi. Although, plant species encode a higher number of proteins, their size is smaller than the average size of animal proteins. Notably, green algae contains a smaller number of proteins than higher plants but their average protein size is 1.27 times larger. The average protein size (low to high) in the species of green algae ranged from 273.08 (*Helicosporidium* sp.) to 576.56

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(*Chlamydomonas eustigma*) amino acids, dicots ranged from 253.34 (*Trifolium pratense*) to 498.49 (*Vitis vinifera*), and monocots ranged from 111.54 (*Hordeum vulgare*) to 473.35 (*Brachypodium distachyon*) amino acids. The average protein size of monocot proteins (431.07 amino acids), however, is slightly larger than dicots (424.3 amino acids). In addition to transposons, previous studies have reported that endosymbiosis may have also played an important role in the reduction of protein size in plant genomes [37,43,44]. This would have been due to the post endosymbiosis acquisition of thousands of genes from the chloroplast, since cyanobacterial proteins are smaller than eukaryotic proteins and cyanobacteria are the ancestors of plastids [37,45]. In this hypothesis, the intermediate size of plant proteins would be the result of the migration of proteins from cyanobacteria (chloroplast) to the plant nucleus, thereby reducing the overall average size of the protein by a dilution effect [46,47].

The pI of plant proteins ranges from 1.99 to 13.96

Results indicated that the *pI* of analysed plant proteins ranged from 1.99 (id: PHT45033.1, *Capsicum baccatum*) to 13.96 (id: PKI59361.1, *Punica granatum*). The protein with the lowest *pI* (1.99) was epsin and the protein with the highest *pI* (13.96) was a hypothetical protein. This is the first study to report on the plant proteins with the lowest and highest *pI*. The *C. baccatum* protein with *pI* 1.99 contains 271 amino acids, whereas the *P. granatum* protein with *pI* 13.96 contains 986 amino acids. The epsin protein (*pI* 1.99) is composed of 16 amino acid repeats (GWIDGWIDGWIDGW), while the hypothetical protein (*pI* 13.06) is composed of 64 QKLKSGLT and 31 TRRGLTAV repeats. From among the 20 essential amino acids, the epsin protein only contained five amino acids, namely Asp (68), Gly (68), Ile (65), Met (3), and Trp (67). The amino acids were arranged in a repeating manner within the full-length epsin protein. This study is the first to report a full-length protein composed of such a minimum number of essential amino acids. Similarly, the hypothetical protein with the highest *pI* (13.96) was composed of only nine amino acids, namely Ala (62), Gly (132), Lys (127), Leu (197), Met

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(M), Pro (4), Gln (64), Arg (132), and Ser (66). Intriguingly, cysteine, which is one of the most important amino acids as it is responsible for the formation of disulphide bonds, was not found in either the smallest or largest protein. Disulphide bonds maintain the conformation and stability of a protein and are typically found in extracellular proteins and only rarely in intracellular proteins [48]. The absence of Cys amino acids in these proteins suggests that these proteins are localized to the intracellular compartments within the cell.

The plant proteome is primarily composed of acidic *pI* proteins rather than basic *pI* proteins (Table 1). Approximately, 56.44% of the analysed proteins had a *pI* within the acidic *pI* range. The average percentage of acidic *pI* proteins was comparatively higher in the lower eukaryotic plants, algae, and bryophytes, than in the higher land plants. A total of 64.18% of proteins in *Chlamydomonas eustigma* were found in the acidic *pI* region, followed by *Ostreococcus lucimarinus* (64.17%), *Micromonas commoda* (63.30%), *Helicosporium* sp. (62.97%), *Gonium pectoral* (62.76%), *Chromochloris zofingiensis* (62.41%), *Coccomyxa subellipsoidea* (62.12%), and *Sphagnum fallax* (61.83%). The algal species, *Porphyra umbilicalis,* had the lowest percentage (29.80%) of acidic *pI* proteins. The dicot plant, *Punica granatum,* and the algal species, *Botrycoccus braunii,* had a significantly lower percentage of acidic *pI* proteins (45.72% and 47.18%, respectively) relative to other plants. Principal component analysis (PCA) of acidic *pI* protein content revealed that the acidic proteins of bryophytes and monocots cluster closely to each other compared to algae and eudicot plants (Figure 1). Similarly, in the case of basic *pI* proteins, a great variation was observed for algae, eudicot and monocot plants (Figure 2). The basic *pI* proteins of bryophytes, however, were found to be consistent. A previous study reported that protein *pI* values are correlated with the sub-cellular localization of the proteins, and that the *pI* of cytosolic proteins fall below 7 [21]. Among cytosolic proteins are those involved in 26S proteasome degradation, oxidative pentose phosphate pathway, actin/tubulin, mevalonate pathway, sugar and nucleotide biosynthesis, glycolysis, RNA

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

processing, and several other cellular process. Our analysis indicated that the *pI* of all cytosolic proteins does not fall in the acidic *pI* range. Ribosomal proteins, pre-mRNA splicing factors, transcription factors, auxin induced protein, extensin, senescence associated protein, cyclin dependent protein kinase and other cytoplasmic proteins had a *pI* greater than 7.

In contrast to acidic *pI* proteins, plants possess a comparatively low number of basic *pI* proteins. On average, 43.34% of the analysed plant proteins possessed a *pI* in the basic range. The highest percentage of basic *pI* proteins was found in *Porphyra umbilicalis,* where 70.09% of the proteins had a basic *pI* (Table 1). *Punica granatum* also had a high percentage (54.11%) of basic *pI* proteins (Table 1). The lowest percentage of basic *pI* proteins was found in the algal species, *Chlamydomonas eustigma* (35.56%), followed by *Ostreococcus lucimarinus* (35.65%), *Micromonas commoda* (36.52%), *Helicosporidium* sp. (36.89%), and *Gonium pectorale* (37.04%). It is difficult to establish the reason that algal species contain more acidic *pI* and less basic *pI* proteins. *Porphyra umbilicalis* is a cold-water seaweed within the family, Bangiophyceae, and it is the most domesticated marine algae. The 87.7 Mbp haploid genome of *P. umbilicalis* has a 65.8% GC content and an evolutionary study reported that the genome of *Porphyra umbilicalis* had undergone a reduction in size [49]. Since this species is found in the intertidal region of the ocean, it has developed the ability to cope with mid-to-high levels of tidal stress. *Porphyra* is also tolerant to UV-A and UV-B radiation [49–51]. The high GC content in *Porphyra umbilicalis* is directly proportional to the high percentage of basic proteins. The GC content of algal species is higher relative to other plant species and algal species possess a lower percentage of basic *pI* proteins. This suggests that, in algae, percentage GC content is inversely proportional to percentage of proteins with a basic *pI*. However, this is not true in the case of higher plants.

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The pI of plant proteomes exhibits a trimodal distribution

The *pI* of the analysed plant proteins ranged from 1.99 to 13.96 and exhibited a trimodal distribution (Figure 3). Schwartz et al., previously reported a trimodal distribution of the *pI* of eukaryotic proteins[21], however, they did not provide information on the number of sequences/species considered in their study. Proteins are typically soluble near their isoelectric point and the cytoplasm possesses a pH that is close to neutral. This may be the reason for the trimodal distribution of *pI*. Although the *pI* values of proteins estimated *in silico* or experimentally might be different *in vivo*, they are typically in close agreement [52]. Kiraga et al., (2006) reported a bimodal distribution of the *pI* of proteins from all organisms, citing acidic and basic *pI* as the basis of the modality [19], where modality is defined as the set of data values that appears most often. They reported that taxonomy, ecological niche, proteome size, and sub-cellular localization are correlated with acidic and basic proteins. However, no correlation was observed in the current study between either acidic or basic *pI* of proteins with regard to taxonomy, ecological niche, or proteome size. For example, *Hordeum vulgare* and *Brassica napus* possess the largest proteomes among the studied plant species, possessing 248180 and 123465 proteins, respectively. In *H. vulgare*, 53.28% of the proteins fall in the acidic and 46.50% fall in the basic *pI* ranges; while in *B. napus*, 55.28% of the proteins have an acidic *pI* and 44.48% have a basic *pI*. Other species with smaller proteomes, however, possess a higher percentage of acidic or basic proteins (Table 1). Therefore, no correlation exists between the percentage of either acidic or basic proteins and proteome size, taxonomy, or the ecological niche of an organism. Knight et al. also reported a negative correlation between the *pI* of a protein with phylogeny of the organism [53]. The existence of a trimodal distribution of the *pI* of the plant proteome can be considered as a virtual 3D-gel of a plant's proteins where the *pI* of

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the protein is plotted against the molecular weight of the protein. On average, 0.21% of the analysed proteins were found to have a neutral *pI* (*pI* 7), while only 0.09% of the proteins in *O. lucimarinus* fall in neutral *pI*.

Leu is a high- and Trp is a low-abundant amino acid in the plant proteome

The plant-kingdom-wide proteome analysis revealed that Leu was the most (9.62%) while Trp was the least (1.28%) abundant amino acid (Figure 4, Supplementary File 1). Leu is a nonpolar amino acid, whereas Trp contains an aromatic ring. The distribution of amino acids indicates that the synthesis of nonpolar amino acids is more favoured in the plant proteomes than the polar amino acids or those containing an aromatic ring. The average abundance of other nonpolar amino acids Ala, Gly, Ile, Met, and Val was 6.68%, 6.80%, 4.94%, 2.40%, and 6.55%, respectively (Table 2, Supplementary File 1). Trp and Tyr amino acid contain an aromatic ring and the abundance of these two proteins is relatively low in the plant proteome compared to other amino acids. Results of the conducted analysis indicated that the abundance of Ala (17.58%), Gly (11.76%), Pro (9.2%), and Arg (9.81%) were the highest; whereas, Tyr (1.33%), Gln (2.04%), Asn (1.53%), Met (1.45%), Lys (7.07%), Lys (2.08%), Ile (1.77%), Phe (2.01%), and Glu (3.52%) were the lowest in *Porphyra umbilicalis*. In a few algae and seaweeds Ala, Asp, Glu, Gly, Pro, Gln, Arg, Thr, and Val were found in high percentage while Asp, Glu, Phe, His, Ile, Lys, Leu, Met, Asn, Gln, and Ser were found in low percentage (Table 2). These observations indicate that the composition of amino acids in unicellular algae, seaweeds, and gymnosperms are more dynamic and variable than in angiosperms and other terrestrial land plants. Principal component analysis revealed that the low abundant amino acids, Trp, Tyr, His, Met, Cys, and Xaa (unknown), cluster in one group while the high abundant amino acids, Leu, Glu, Ile, Lys, and Ser, cluster in another group (Figure 5). None of the terrestrial land plants were located in the high- and low-abundant amino acid clusters. This suggests that the

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

proteome and amino acid composition of the land plants are more conserved and stable relative to the algae and seaweeds. PCA analysis further revealed that the *pI* of algae, eudicots, and monocots are lineage specific. The *pI* of algae, monocots, and eudicots were strongly correlated and clustered together (Figure 5). The question arises, however, as to why the plant proteome contains the highest percentage of Leu and of the lowest percentage of Trp amino acids. Do the energy requirements of the different biosynthetic pathways play a pivotal role in deciding the abundance of amino acids in a proteome? To address this question, an attempt was made to understand the role of amino acid biosynthetic pathways in determining the abundance of specific amino acids in the proteome.

Various amino acids are produced in different biosynthetic pathways [54–58] (Figure 6). In some cases, a few amino acids act as the substrate for the biosynthesis of other amino acids; whereas in other cases, allosteric inhibition of the biosynthesis of amino acids occurs [59–61]. In all of these biosynthetic pathways, ATP or NADH/NADPH are used as a source of energy, along with substrate that play a vital role in the biosynthesis of amino acids. Overall, the biosynthesis of 20 essential amino acid families are grouped by metabolic precursors [62] (Table 3); namely α-ketoglutarate (Arg, Gln, Glu, Pro), pyruvate (Ala, Ile, Leu, Val), 3-phosphoglycerate (Cys, Gly, Ser), phosphoenolpyruvate and erythrose 4-phosphate (Phe, Trp, Tyr), oxaloacetate (Asn, Asp, Lys, Met, Thr), and ribose 5-phosphate (His) (Table 3) [62]. Ala, Ile, Leu, and Val are synthesized from pyruvate; Arg, Glu, Gln, and Pro are synthesized from α-ketoglutarate and Gly and Ser are synthesized from 3-phosphoglycerate[62]; all of which have a higher abundance in the plant proteome relative to the other amino acids (Figure 6, Table 3). 3-phosphoglycerate and pyruvate are intermediates of glycolysis and the amino acids synthesized from these intermediates maintain a high abundance in the plant proteome. The intermediate, 3-phosphoglycerate, is formed in an early step of glycolysis[62]. The amino acids Gly and Ser are

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

synthesized from 3-phosphoglycerate and are also found abundantly in the plant proteome (Figure 6, Table 3). The amino acid Cys, which is also synthesized from 3-phosphoglycerate[62], however, is present in low abundance (1.85%) in the plant proteome. The low abundance of Cys may be due to the allosteric inhibition. Phe (3.97), Trp (1.28%), and Try (2.67%) contain an aromatic ring and are synthesized via phosphoenolpyruvate and erythrose 4-phospahte. The aromatic amino acids are in low abundance in the plant proteome (Table). Since Phe also plays a role in the biosynthesis of Tyr, the abundance of Phe is relatively higher than Trp and Tyr. Glucose 6-phosphate gives rise to ribose 5-phosphate in a complex reaction of four steps[62] and His gets subsequently synthesized from ribose 5-phosphate. It is possible that the complexity of the biosynthetic pathways of amino acids containing ring compounds might be the reason for their low abundance in the plant proteome.

Plants possess selenocysteine (Sec) and other novel amino acids

A few of the plant proteomes that were analysed had proteins containing the amino acid, selenocysteine (Sec). *C. reinhardtii*, *M. pusilla*, and *V. carteri* contained 9, 16, and 11 Sec amino acids in their proteome, respectively. Selenium containing selenoproteins are commonly found in animals but have been reported to be present in plant species. Novoselov et al., (2002) reported the presence of a selenoprotein in *C. reinhardtii* [63]. In our analysis, nine selenoproteins (selenoprotein H, selenoprotein K1, selenoprotein M2, selenoprotein T, selenoprotein U, selenoprotein W1, selenoprotein W2, NADPH-dependent thioredoxin reductase 1, and glutathione peroxidase) were identified in *C. reinhardtii*. In addition, *M. pusilla* was found to possess 14 Sec-containing proteins [DSBA oxidoreductase (2 no.), selenoprotein T, glutathione peroxidase (4 no.), selenoprotein W, selenoprotein, selenoprotein M, selenoprotein H, selenoprotein O, methyltransferase, and peroxiredoxin). In addition, *V. carteri* was found to possess 10 Sec containing proteins (selenoprotein T, selenoprotein K1, selenoprotein H,

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

selenoprotein W1, selenoprotein M2, selenoprotein U, glutathione peroxidase, membrane selenoprotein, NADPH-dependent thioredoxin reductase, and peptide methionine-S-sulfoxide reductase). To the best of our knowledge, this is the first report of selenoproteins in *M. pusilla* and *V. carteri* and the first to report of H, K1, T, U, M, M2, O, W1, and W2 selenoprotein families collectively in *C. reinhardtii*, *M. pusilla*, and *V. carteri*. The I, N, P, R, S, and V selenoprotein family members are commonly found in the animal kingdom[64] but are absent in *C. reinhardtii*, *M. pusilla*, and *V. carteri*. This is also the first report of the Sec-containing proteins, DSBA oxidoreductase, methyltransferase, peroxiredoxin, peptide methionine-S-sulfoxide reductase, and membrane selenoprotein in plants lineage (algae). Notably, the selenoproteins DSBA oxidoreductase, methyltransferase, peroxiredoxin, peptide methionine-S-sulfoxide reductase, and membrane selenoprotein have not been reported in animal species. Outside of algal species, no other plant species; including bryophytes, pteridophytes gymnosperms and angiosperms, were found to possess a selenoprotein.

Some plant proteomes were also found to possess a few unknown or unspecified amino acids, commonly designated as Xaa (X). Among the analysed plant species, *Aegilops tauschii*, *Amaranthus hypochondriacus*, and *Amborella trichocarpa* encoded 149377, 55412, and 25843 X amino acids, respectively. *Solanum lycopersicum* was found to contain only one X amino acid, while at least 18 species (*Solanum pennellii*, *Solanum tuberosum*, *Sorghum bicolor*, *Sphagnum fallax*, *Spinacia oleracea*, *Spirodela polyrhiza*, *Tarenaya hassleriana*, *Theobroma cacao*, *Trifolium pratense*, *Trifolium subterraneum*, *Triticum aestivum*, *Triticum urartu*, *Vigna angularis*, *Vigna radiata*, *Vigna anguiculata*, *Vitis vinifera*, *Volvox carteri*, and *Zostera marina*) were found to lack any Xaa amino acids in their proteome. Xaa amino acids are known as non-protein amino acids as they have not been associated with any specific codons. Among the studied plant species, ten were found to contain amino acid B (Asx) that codes for the

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ambiguous amino acid Asn or Asp that is translated as Asp. Species that were found to possess an Asx amino acid included *Arachis duranensis* (1), *Brachypodium stacei* (40), *Dichanthelium oligosanthes* (20), *Dunaliella salina* (31), *Glycine max* (1), *Malus domestica* (4080), *Momordica charantia* (98), *Nelumbo nucifera* (64), *Prunus persica* (1), and *Trifolium pratense* (76). At least six species were found to possess a J (Xle) amino acid. Xle amino acid can encode either Leu or Ile but during translation produces Leu. Species that were found to possess Xle amino acids included *Arabidopsis thaliana* (10), *Dichanthelium oligosanthes* (11), *Malus domestica* (2175), *Momordica charantia* (39), *Nelumbo nucifera* (29), and *Trifolium pratense* (39). At least seven species were found to possess a Z (Glx) amino acid that codes for either Glu or Gln, which is subsequently translated as Glu. Species that were found to encode a Glx amino acid included *Brachypodium stacei* (20), *Dichanthelium oligosanthes* (16), *Dunaliella salina* (7), *Malus domestica* (1552), *Momordica charantia* (28), *Nelumbo nucifera* (14), and *Trifolium pratense* (25). Among the studied species, *Malus domestica* was found to contain highest number of ambiguous amino acids (Asx, Xle, and Glx). Bodley and Davie (1966) reported the incorporation of ambiguous amino acids in a peptide chain [65]. The presence of ethanol or streptomycin or a high magnesium ion concentration induces ambiguous coding in the peptide chain [65]. They reported that poly-U (uridylic acid) in the presence of a high concentration of magnesium ions or ethanol or streptomycin induces the incorporation of Leu/Ile amino acids in a peptide chain [65]. This explains how the specificity of the protein translation process can be altered by the presence of environmental factors. A high concentration of magnesium ions, organic solvents, antibiotics, pH, and low temperature have the ability to modify the coding specificity of a peptide chain [65]. Under some conditions poly-U triggers the incorporation of Leu and Ile or Phe [65]. *Malus domestica* is rich in magnesium ions (1%) and this might explain the presence of such a high number of ambiguous amino acids in its proteome.

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Conclusion

A proteomic analysis of the plant kingdom identified proteins with a great range of molecular mass and isoelectric points. Isoelectric points ranged from 1.99 to 13.96, covering almost the entire pH range. It is quite intriguing to think about the functions of protein at *pI* 1.99 or 13.96. Proteins with an acidic *pI* predominate over the proteins with an alkaline *pI*, and the presence of proteins with a *pI* that is near neutral is very negligible. The percentage of proteins with acidic or basic *pI*s is not related to the host cell, alkalinity or acidity of the environment. Additionally, the GC content of a genome or size distribution of the overall proteome are not directly proportional to the distribution (percent basic vs. percent acidic) of the *pI* in the proteome. The presence of Sec-containing proteins in some plant species needs to be further investigated to determine their functional role. Similarly, the presence of ambiguous amino acids in plant species should be further evaluated individually to understand whether these ambiguous amino acids are encoded by any specific codon in mRNA. A major question arises from this study is, whether the incorporation of ambiguous amino acids in the peptide chain of the protein brings any impact at the gene/genome level and incorporate the respective codon for the ambiguous amino acid and regulated genomic rearrangement through reverse central dogma approach. The presence of a pyrrolysine amino acid in the plant kingdom was not observed in the present study.

Methods

Protein sequences of the entire proteome of the analysed plant species were downloaded from the National Center for Biotechnology Information (NCBI) and Phytozome, DOE Joint Genome Institute (https://phytozome.jgi.doe.gov/pz/portal.html). All of the studied sequences were annotated nuclear-encoded proteins. The isoelectric point of each protein of each of the analysed plant species was calculated individually using the Python-based command line

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

"Protein isoelectric point calculator" (IPC Python) in a Linux platform [2]. The source code used was as written by Kozlowski (2016).

Once the molecular mass and isoelectric point of the proteins in each species was determined, they were separated into acidic and basic *pI* categories. Subsequently, the average *pI* and percentage of proteins in each category was calculated using a Microsoft excel worksheet. A graph comparison of isoelectric point versus molecular mass was prepared using a python-based platform. Pearson-correlation (r=0.19, p = 0) was used for the association analysis of molecular mass and isoelectric point. The X-axis data statistics were as follows: mean, 4.717365e+01; std, 3.662983e+01; min, 8.909000e-02; 25%, 2.279452e+01; 50%, 3.874486e+01; 75%, 5.999628e+01, and max, 2.236803e+03. The Y-axis data statistics were: mean, 6.840657e+00; std, 1.594912e+00; min, 1.990000e+00; 25%, 5.537000e+00; 50%, 6.605000e+00; 75%, 8.053000e+00, and max, 1.396300e+01.

Principal component analysis

Principal component analysis of the plant proteome parameters was carried out using a portable Unscrambler software version 9.7 using the excel file format. For acidic and basic *pI*, the plant proteome data were grouped according to the plant lineage algae, bryophyte, monocot, and eudicot plants. The average of acidic and basic *pI* was used to construct the PCA plot. Similarly, amino acid abundance was also analysed in relation to algae, bryophyte, eudicot, and monocot lineage.

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Kirkwood, J., Hargreaves, D., O'Keefe, S. & Wilson, J. Using isoelectric point to determine the pH for initial protein crystallization trials. *Bioinformatics* **31**, 1444–1451 (2015).
2. Kozlowski, L. P. IPC – Isoelectric Point Calculator. *Biol. Direct* **11**, 55 (2016).
3. Kozlowski, L. P. Proteome-pI: proteome isoelectric point database. *Nucleic Acids Res.* **45**, D1112–D1116 (2017).
4. Stekhoven, F., Gorissen, M. & Flik, G. The isoelectric point, a key to understanding a variety of biochemical problems: a minireview. *Fish Physiol. Biochem.* **34**, 1–8 (2008).
5. Kenneth, W., Kenneth, G. & Raymond, E. *General Chemistry*. (Saunders College Publishing, 1992).
6. Pace, C. N., Grimsley, G. R. & Scholtz, J. M. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *J. Biol. Chem.* **284**, 13285–13289 (2009).
7. Grimsley, G. R., Scholtz, J. M. & Pace, C. N. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci.* **18**, 247–251 (2009).
8. Masunov, A. & Lazaridis, T. Potentials of Mean Force between Ionizable Amino Acid Side Chains in Water. *J. Am. Chem. Soc.* **125**, 1722–1730 (2003).
9. Zhu, K., Zhao, J., Lubman, D. M., Miller, F. R. & Barder, T. J. Protein pI Shifts due to Posttranslational Modifications in the Separation and Characterization of Proteins. *Anal. Chem.* **77**, 2745–2755 (2005).
10. Locke, D., Koreen, I. V & Harris, A. L. Isoelectric points and post-translational modifications of connexin26 and connexin32. *FASEB J.* **20**, 1221–1223 (2006).
11. Youden, W. J. & Denny, F. E. Factors Influencing the pH Equilibrium Known as the Isoelectric Point of Plant Tissue. *Am. J. Bot.* **13**, 743–753 (1926).
12. Shaw, K. L., Grimsley, G. R., Yakovlev, G. I., Makarov, A. A. & Pace, C. N. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. *Protein Sci.* **10**, 1206–1215 (2001).
13. Saraswathy, N. & Ramalingam, P. Introduction to proteomics. in *Woodhead Publishing Series in Biomedicine* (eds. Saraswathy, N. & Ramalingam, P. B. T.-C. and T. in G. and P.) 147–158 (Woodhead Publishing, 2011). doi:https://doi.org/10.1533/9781908818058.147
14. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci.* **97**, 9390 LP-9395 (2000).
15. Rabilloud, T. & Lelong, C. Two-dimensional gel electrophoresis in proteomics: A tutorial. *J. Proteomics* **74**, 1829–1841 (2011).
16. Kumar, M. *et al.* An Improved 2-Dimensional Gel Electrophoresis Method for Resolving Human Erythrocyte Membrane Proteins. *Proteomics Insights* **8**, 1178641817700880 (2017).
17. Anderson, J. C. & Peck, S. C. A simple and rapid technique for detecting protein phosphorylation using one-dimensional isoelectric focusing gels and immunoblot analysis. *Plant J.* **55**, 881–885 (2008).
18. Jones, L. R., Simmerman, H. K., Wilson, W. W., Gurd, F. R. & Wegener, A. D. Purification and characterization of phospholamban from canine cardiac sarcoplasmic reticulum. *J. Biol. Chem.* **260**, 7721–7730 (1985).
19. Kiraga, J. *et al.* The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* **8**, 163 (2007).
20. Carugo, O. Isoelectric points of multi-domain proteins. *Bioinformation* **2**, 101–104 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

21. Schwartz, R., Ting, C. S. & King, J. Whole proteome pl values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res.* **11**, 703–709 (2001).
22. Labeit, S. *et al.* A regular pattern of two types of 100-residue motif in the sequence of titin. *Nature* **345**, 273 (1990).
23. Kurzban, G. P. & Wang, K. Giant polypeptides of skeletal muscle titin: Sedimentation equilibrium in guanidine hydrochloride. *Biochem. Biophys. Res. Commun.* **150**, 1155–1161 (1988).
24. Kurisu, G., Zhang, H., Smith, J. L. & Cramer, W. A. Structure of the Cytochrome b6f Complex of Oxygenic Photosynthesis: Tuning the Cavity. *Science (80-.).* **302**, 1009 LP-1014 (2003).
25. Leister, D. & Schneider, A. B. T.-. From Genes to Photosynthesis in Arabidopsis thaliana. *Int. Rev. Cytol.* **228**, 31–83 (2003).
26. Barbagallo, R. P., Finazzi, G. & Forti, G. Effects of Inhibitors on the Activity of the Cytochrome b6f Complex: Evidence for the Existence of Two Binding Pockets in the Lumenal Site. *Biochemistry* **38**, 12814–12821 (1999).
27. Cramer, W. A. *et al.* Some New Structural Aspects And Old Controversies Concerning The Cytochrome B6f Complex Of Oxygenic Photosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 477–508 (1996).
28. Kallas, T. The Cytochrome b6f Complex. in *The Molecular Biology of Cyanobacteria* (ed. Bryant, D. A.) 259–317 (Springer Netherlands, 1994). doi:10.1007/978-94-011-0227-8_9
29. Filomeni, G., Rotilio, G. & Ciriolo, M. R. Cell signalling and the glutathione redox system. *Biochem. Pharmacol.* **64**, 1057–1064 (2002).
30. Zhang, H. & Forman, H. J. Glutathione synthesis and its role in redox signaling. *Semin. Cell Dev. Biol.* **23**, 722–728 (2012).
31. Aquilano, K., Baldelli, S. & Ciriolo, M. R. Glutathione: new roles in redox signaling for an old antioxidant. *Front. Pharmacol.* **5**, 196 (2014).
32. Su, M., Ling, Y., Yu, J., Wu, J. & Xiao, J. Small proteins: untapped area of potential biological importance. *Front. Genet.* **4**, 286 (2013).
33. Setlow, P. I will survive: DNA protection in bacterial spores. *Trends Microbiol.* **15**, 172–180 (2007).
34. Schalk, C. *et al.* Small RNA-mediated repair of UV-induced DNA lesions by the DNA DAMAGE-BINDING PROTEIN 2 and ARGONAUTE 1. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E2965–E2974 (2017).
35. Xue, Y. *et al.* Tff3, as a Novel Peptide, Regulates Hepatic Glucose Metabolism. *PLoS One* **8**, e75240 (2013).
36. Basrai, M. a, Hieter, P. & Boeke, J. D. Small Open Reading Frames : Beautiful Needles in the Haystack Small Open Reading Frames : Beautiful Needles in the Haystack. *Genome Res.* **7**, 768–771 (1997).
37. Ramírez-Sánchez, O., Pérez-Rodríguez, P., Delaye, L. & Tiessen, A. Plant Proteins Are Smaller Because They Are Encoded by Fewer Exons than Animal Proteins. *Genomics. Proteomics Bioinformatics* **14**, 357–370 (2016).
38. Tiessen, A., Pérez-Rodríguez, P. & Delaye-Arredondo, L. J. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes* **5**, 85 (2012).
39. Purugganan, M. & Wessler, S. The splicing of transposable elements and its role in intron evolution. *Genetica* **86**, 295–303 (1992).

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

40. Huff, J. T., Zilberman, D. & Roy, S. W. Mechanism for DNA transposons to generate introns on genomic scales. *Nature* **538**, 533–536 (2016).
41. Lee, S.-I. & Kim, N.-S. Transposable elements and genome size variations in plants. *Genomics Inform.* **12**, 87–97 (2014).
42. Plasterk, R. H. A. *Transposable Elements*. *Encyclopedia of Genetics* (Academic Press, 2001). doi:https://doi.org/10.1006/rwgn.2001.1316
43. Martin, W. *et al.* Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12246–12251 (2002).
44. Sloan, D. B. & Moran, N. A. Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids. *Mol. Biol. Evol.* **29**, 3781–3792 (2012).
45. Mohanta, T. K., Pudake, R. N. & Bae, H. Genome-wide identification of major protein families of cyanobacteria and genomic insight into the circadian rhythm. *Eur. J. Phycol.* **52**, (2017).
46. Reyes-Prieto, A., Weber, A. P. M. & Bhattacharya, D. The Origin and Establishment of the Plastid in Algae and Plants. *Annu. Rev. Genet.* **41**, 147–168 (2007).
47. Martin, W. *et al.* Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**, 162 (1998).
48. Mallick, P., Boutz, D. R., Eisenberg, D. & Yeates, T. O. Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 9679–9684 (2002).
49. Brawley, S. H. *et al.* Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta). *Proc. Natl. Acad. Sci.* **114**, E6361 LP-E6370 (2017).
50. Dring, M. J., Wagner, A., Boeskov, J. & Lüning, K. Sensitivity of intertidal and subtidal red algae to UVA and UVB radiation, as monitored by Chlorophyll fluorescence measurements: Influence of collection depth and season, and length of irradiation. *Eur. J. Phycol.* **31**, 293–302 (1996).
51. Gröniger, A., Hallier, C. & Häder, D. P. Influence of UV radiation and visible light on Porphyra umbilicalis: Photoinhibition and MAA concentration. *J. Appl. Phycol.* **11**, 437 (1999).
52. Sillero, A. & Ribeiro, J. M. Isoelectric points of proteins: Theoretical determination. *Anal. Biochem.* **179**, 319–325 (1989).
53. Knight, C. G., Kassen, R., Hebestreit, H. & Rainey, P. B. Global analysis of predicted proteomes: functional adaptation of physical properties. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 8390–8395 (2004).
54. Gibson, F. & Pittard, J. Pathways of biosynthesis of aromatic amino acids and vitamins and their control in microorganisms. *Bacteriol. Rev.* **32**, 465–492 (1968).
55. Tzin, V. & Galili, G. New Insights into the Shikimate and Aromatic Amino Acids Biosynthesis Pathways in Plants. *Mol. Plant* **3**, 956–972 (2010).
56. Szabados, L. & Savouré, A. Proline: a multifunctional amino acid. *Trends Plant Sci.* **15**, 89–97 (2010).
57. Galili, G. & Höfgen, R. Metabolic Engineering of Amino Acids and Storage Proteins in Plants. *Metab. Eng.* **4**, 3–11 (2002).
58. Yoshida, S. Biosynthesis and Conversion of Aromatic Amino Acids in Plants. *Annu. Rev. Plant Physiol.* **20**, 41–62 (1969).
59. Light, S. H. & Anderson, W. F. The diversity of allosteric controls at the gateway to aromatic amino acid biosynthesis. *Protein Sci.* **22**, 395–404 (2013).
60. Berg, J., Tymoczko, J. & Stryer, L. Amino Acid Biosynthesis Is Regulated by

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Feedback Inhibition. in *Biochemistry* (WH Freeman, 2002).

61. Nelson, D. & Cox, M. Amino acid biosynthesis is under allosteric regulation. in *Lehlingers Principle of Biochemistry* (ed. Lehninger, A. L.) 833–880 (2005).
62. Nelson, D. & Cox, M. Biosynthesis of amino acids, nucleotides, and related molecules. in *Lehlinger Principles of Biochemistry* 842 (W.H. Freeman and Company, 2005).
63. Novoselov, S. V *et al.* Selenoproteins and selenocysteine insertion system in the model plant cell system, Chlamydomonas reinhardtii. *EMBO J.* **21**, 3681–3693 (2002).
64. Reeves, M. A. & Hoffmann, P. R. The human selenoproteome: recent insights into functions and regulation. *Cell. Mol. Life Sci.* **66**, 2457–2478 (2009).
65. Bodley, J. W. & Davie, E. W. A study of the mechanism of ambiguous amino acid coding by poly U: the nature of the products. *J. Mol. Biol.* **18**, 344–355 (1966).

Table 1

Details of plant proteome. Table shows acidic pI of proteins predominates the basic *pI*. However, in sea weed Porphyra umbilicalis, basic *pI* predominates over the acidic *pI*. Putative polyketide synthase type I found in lower eukaryote *Volvox carteri* was found to be the largest protein in the plant lineage. However, titin was found to be the largest protein in the higher eukaryotic land plants. Asterisks represents no specific data available for the said item.

Name of the species	**Total No. Of Protein Sequences Studied**	**Highest Mol. Wt. (kDa) of Protein**	**Name of the Protein with Highest Mol. Wt.**	**Lowest Mol. Wt. (kDa) of Protein**	**Name of the Protein with lowest Mol. Wt.**	**Highest *pI* of Protein**	**Name of the Protein with highest *pI***	**Lowest *pI* of Protein**	**Name of the Protein with lowest *pI***	**No. Of Proteins in Acidic *pI***	**No. Of Proteins in Basic *pI***	**No. Of Proteins in Neutral *pI***	**Average of Acidic *pI***	**Average of basic *pI***
Aegilops tauschii	55713	605.202	Misin	3.169	Cyt b6/f VIII	13.159	SARMP 2-like	2.409	Nucleolin-like	30488	25116	109	5.64	8.75
Amaranthus hypochondriacus	23879	596.11	Unknown	0.63	Unknown	12.67	Unknown	2.498	Unknown	13009	10833	37	5.58	8.36
Amborella trichopoda	27313	554.72	Unknown	4.32	Unknown	12.93	Unknown	2.587	Unknown	14583	12677	53	5.53	8.56
Anacardium occidentale	82170	457.50	Unknown	3.10	Unknown	12.61	Unknown	2.58	Unknown	46898	35100	172	5.65	8.27
Ananas comosus	35775	605.85	Midasin	3.16	Cyt b6/f VIII	12.89	Unknown	3.14	PPCD VHS3-like	20268	15399	108	5.68	8.35
Aquilegia coerulia	41063	620.28	Unknown	3.82	Unknown	12.74	Unknown	2.85	Unknown	23818	17169	76	5.64	8.26
Arabidopsis halleri	26911	609.56	Unknown	3.18	Unknown	12.50	Unknown	3.10	Unknown	14727	12123	61	5.59	8.32
Arabidopsis lyrata	39161	611.10	Midasin	2.51	Unknown	12.74	60S RP L41	2.85	RNA Pol. II Med 17	22213	16854	94	5.62	8.26
Arabidopsis thaliana	48350	611.88	Misin-like	0.57	Hypothetical	12.74	60S RP L41	2.75	Glycine-rich protein	27305	20926	119	5.61	8.31
Arabis alpina	23286	565.05	Unknown	***	****	12.79	Unknown	2.14	Unknown	13427	9810	49	5.52	8.37
Arachis duranensis	52826	617.77	Misin	4.11	DDHGT 4A	12.72	PERK2	2.96	Unknown	29514	23184	128	5.67	8.25
Arachis ipaensis	57621	617.63	Misin	4.08	DDHGT 4A	12.36	Unknown	3.13	Small acidic protein	31471	26007	143	5.67	8.26
Asparagus officinalis	36763	608.71	Misin	4.28	DDHGT 4A	12.50	protein TPRXL	2.89	FS CAYBR BP	20748	15934	81	5.64	8.25
Auxenochlorella protothecoides	7014	1649.26	PKS	***	***	13.21	Mucin-I	2.35	Hypothetical protein	4131	2874	9	5.52	8.91
Bathycoccus prasinos	7900	1814.10	Unnamed	3.32	Ycf12	12.74	Unknown	3.42	Unknown	4799	3093	8	5.47	8.38
Beta vulgaris	32874	617.35	Misin	3.56	Unknown	12.55	Proline-rich P	3.16	Shematrin-like 2	18770	14017	87	5.68	8.19
Botrycoccus braunii	23685	522.853	Unknown	3.068	Unknown	12.82	Unknown	2.21	Unknown	11176	12439	70	5.61	8.66

Brachipodium stacei	36357	605.43	Unknown	3.023	Unknown	12.88	Unknown	3.05	Unknown	19318	16954	85	5.67	8.59
Brachipodium sylvaticum	50263	608.40	Unknown	3.14	Unknown	12.88	Unknown	3.00	Unknown	26022	24130	111	5.66	8.82
Brachipodium distachyon	33944	605.24	Misin	3.16	Cyt b6/f VIII	12.39	Unknown	3.03	Prothymosin α-B-like	19814	14052	78	5.67	8.35
Brachypodium hybridum	80980	605.62	Unknown	3.02	Unknown	13.14	Unknown	3.05	Unknown	40744	40092	144	5.66	8.79
Brassica napus	123465	606.90	Misin-like	3.15	petN	12.57	IQ domain 31	2.75	Shematrin-like 2	68255	54929	281	5.62	8.29
Brassica oleracea	56687	606.74	Misin	3.48	Unknown	12.54	IQ domain 31	2.75	Shematrin-like 2	31109	25432	146	5.62	8.29
Brassica rapa	52553	607.93	Misin	3.33	Unknown	12.57	IQ domain 31	2.48	Dentin-sialophospho Protein-like	29429	23004	120	5.62	8.29
Cajanus cajan	38965	619.70	Misin	3.42	Cyt b6/f VIII	12.44	CLAVATA 3/ESR	2.82	RPB1-like	22092	16793	80	5.71	8.22
Camelina sativa	107481	610.42	Misin-like	2.61	Peptide POLARIS	12.72	SARMP 2-like	2.13	TsetseEP-like	60623	46584	274	5.61	8.26
Capsella grandiflora	26561	593.86	Unknown	3.41	Unknown	12.44	Unknown	2.80	Unknown	14831	11661	61	5.62	8.30
Capsella rubella	34126	610.40	Misin	3.57	Unknown	12.44	Lifeguard 1	3.13	Unknown	19477	14578	71	5.63	8.25
Capsicum annuum	45410	617.56	Misin	3.16	Cyt b6/f VIII	12.79	GRCW protein	2.75	GRCW protein 1	25339	19970	101	5.67	8.24
Capsicum baccatum	35853	553.75	Auxin TP BIG	3.16	Cyt b6/f VIII	12.64	Hypothetical	1.99	Hypothetical	20479	15300	74	5.55	8.42
Capsicum chinense	34973	550.87	Auxin TP BIG	3.16	Cyt b6/f VIII	13.21	Hypothetical	2.24	Hypothetical	20497	14398	79	5.56	8.25
Carica papaya	26103	446.01	Unknown	3.16	Cyt b6/f VIII	12.29	50S RP L34	3.10	Small acidic protein	14512	11526	65	5.67	8.25
Cephalotus follicularis	36667	611.81	AAA_5	***	****	13.04	Hypothetical	2.9	Hypothetical	18643	17959	65	5.66	8.33
Chenopodium quinoa	63173	621.14	Misin-like	3.16	petN	12.22	SR45-like	3.02	Unknown	36037	26973	163	5.66	8.18
Chlamydomonas eustigma	14161	1370.23	Unknown	2.28	Unknown	13.14	Unknown	2.18	Unknown	9089	5036	36	5.64	8.23
Chlamydomonas reinhardtii	14488	2056.44	PKS	1.46	Unknown	13.34	Unknown	2.40	Unknown	7427	7029	32	5.64	8.7
Chlorella variabilis	9780	1159.35	Unknown	4.91	Unknown	12.73	Unknown	2.88	Unknown	5883	3876	21	5.55	8.55
Chromochloris zofingiensis	15369	1932.21	Unknown	2.34	Unknown	12.5	Unknown	2.54	Unknown	9592	5740	37	5.61	8.35
Cicer arietinum	33107	616.07	Unknown	3.02	Unknown	12.25	Unknown	3.14	Unknown	19043	13979	85	5.69	8.20
Citrus clementina	34557	588.15	Unknown	****	****	12.44	Unknown	2.95	Unknown	19332	15151	74	5.67	8.28

Citrus sinensis	35648	617.00	Misin	3.16	Cyt b6/f VIII	12.25	Unknown	2.93	Circumsporozoite protein-like	20887	14680	81	5.69	8.18
Citrus unshiu	37970	684.122	Unknown	0.54	Unknown	12.83	Unknown	2.74	Unknown	20954	16926	90	5.67	8.34
Coccomyxa subellipsoidea	9839	1632.35	Ketoacyl-synt	5.06	Unknown	12.99	Unknown	3.05	Unknown	6112	3708	19	5.53	8.55
Corchorus capsularis	29356	606.73	Unknown	1.6	Unknown	12.74	Unknown	2.72	IMP	15540	13783	33	5.49	8.67
Corchorus olitorius	35704	498.84	ZRF	2.79	Unknown	12.79	Unknown	2.56	Unknown	19280	16359	65	5.49	8.62
Cucumis melo	29796	616.61	Misin	2.87	Unknown	13.23	Unknown	2.56	Loricin-like	17151	12562	83	5.7	8.24
Cucumis sativus	29796	617.61	Misin	2.87	Unknown	13.23	Unknown	2.56	Loricin-like	17151	12562	83	5.7	8.24
Cucurbita maxima	42777	615.95	Misin	3.58	Unknown	12.22	60S RP L39	2.54	CWP gp 1-like	24870	17817	90	5.68	8.25
Cucurbita moschata	43715	615.32	Misin	4.1	DDGT 4A	12.85	EPR1	2.31	PKDP	25399	18233	83	5.67	8.25
Daucus carota	44655	619.12	Misin	3.16	Cyt b6/f VIII	12.32	Ribo Protein L32	2.79	Loricin	26135	18423	97	5.68	8.16
Dendrobium officinale	34527	616.94	Misin	3.16	Cyt b6/f VIII	12.45	60S RP L39	3.23	Unknown	19029	15425	73	5.69	8.26
Dichanthelium oligosanthes	26468	538.41	Auxin TP BIG	1.09	Unknown	13.18	Unknown	3.00	Unknown	14146	12261	61	5.6	8.63
Dorcoceras hygrometricum	47778	563.43	Midasin	4.75	Unknown	12.86	Unknown	2.68	Unknown	23461	24237	80	5.44	8.92
Dunaliella salina	18801	603.54	Unknown	2.92	Unknown	12.69	Unknown	2.38	Unknown	9927	8833	41	5.66	8.53
Durio zibethinus	63007	620.96	Misin	3.57	Unknown	12.35	SR45-like	3.12	Acidic protein	37032	25800	175	5.68	8.21
Elaeis guineensis	41887	614.29	Misin	3.19	Cyt b6/f VIII	12.35	50S RP L34	3.28	Calsequestrin 1-like	24529	17266	92	5.69	8.29
Erythranthe guttata	31861	611.80	Misin	3.77	Unknown	12.14	SR45	2.6	CSF subunit 2	18284	13500	77	5.63	8.22
Eucalyptus grandis	52554	644.40	Futsch	3.4	Cyt b6/f VIII	12.83	Unknown	3.07	Fimbrin 1-like	31377	21034	143	5.69	8.24
Eutrema salsugineum	29485	609.35	Unknown	***	****	12.32	Unknown	3.05	Unknown	16105	13300	80	5.63	8.32
Fragaria vesca	31387	609.82	Misin	3.19	Cyt b6/f VIII	12.34	50S RP L34	3.13	Prostatic spermine BP	18862	12429	96	5.65	8.20
Genlisea aurea	17685	559.24	Unknown	4.94	Unknown	12.89	Unknown	2.93	Unknown	9842	7814	29	5.54	8.51
Glycine max	71523	619.18	Misin-like	1.34	AAPT	12.28	Unknown	2.67	HC1-like	41545	29785	193	5.68	8.22
Glycine soja	50399	603.79	Misin	1.59	Hypothetical	12.42	Dynein	2.21	RBP 12B	28254	22054	91	5.62	8.34
Gonium pectorale	16290	881.59	Unknown	5.02	Unknown	12.88	Unknown	2.48	Unknown	10225	6034	31	5.52	8.54

Gossypium arboreum	47568	806.46	Titin-like	3.16	Cyt b6/f VIII	12.23	SR45-like	2.91	Loricin-like	26844	20612	103	5.69	8.23
Gossypium hirsutum	90927	750.00	Titin-like	3.16	Cyt b6/f VIII	12.26	SARMP 2-like	2.67	ER TF TINY-like	51343	39336	248	5.67	8.25
Gossypium raimondii	59057	802.74	Titin-like	3.16	Cyt b6/f VIII	12.25	SR45	2.98	Arabinogalactan 11	33377	25561	119	5.67	8.25
Handroanthus impetiginosus	30271	475.38	Unknown	2.91	Unknown	12.66	Unknown	2.95	Unknown	16255	13959	57	5.64	8.36
Helianthus annuus	73839	1300.83	Unknown	2.76	Unknown	12.88	C1E8.05-like	2.62	Prostatic spermine BP	40355	33310	174	5.63	8.25
Helicosporidium sp.	6033	245.59	Unknown	***	***	12.79	Unknown	2.82	Unknown	3799	2226	8	5.32	8.92
Herrania umbratica	27748	620.76	Misin	4.23	DDGT 4A	12.66	Unknown	3.12	Small acidic protein 1	16209	11492	47	5.7	8.21
Hevea brasiliensis	58062	620.10	Misin	3.16	Cyt b6/f VIII	12.16	50S RP L35	3.12	Small acidic protein 1	33994	23952	116	5.69	8.20
Hordeum vulgare	248180	607.08	Unknown	1.75	Unknown	13.11	****	2.52	Unknown	132250	115421	509	5.60	8.62
Ipomoea nil	51054	636.34	Filaggrin-like	3.16	Cyt b6/f VIII	12.38	Formin 2-like	2.82	Nucleolin-like	29183	21727	144	5.69	8.18
Jatropha curcas	32547	628.22	Titin	3.16	petN	12.14	60S RP L39-3	3.13	Glycin-rich protein	18939	13531	77	5.69	8.18
Juglans regia	55627	624.15	Midasin	4.14	40S RP S29-like	12.98	Formin-like 3	2.46	Glycine-rich protein	32225	23292	110	5.69	8.22
Kalanchoe fedtschenkoi	45190	563.28	Unknown	3.16	Unknown	12.98	Unknown	2.88	Unknown	25024	20062	104	5.64	8.37
Kalanchoe laxiflora	69177	619.54	Unknown	3.01	Unknown	12.47	Unknown	3.09	Unknown	39148	29888	141	5.66	8.32
Klebsormidium nitens	16282	813.31	Unknown	***	***	12.83	Ser/Thr Prot Kin	3.09	Unknown	9751	6500	31	5.57	8.39
Lactuca sativa	45242	609.98	Misin	3.16	Cyt b6/f VIII	12.44	Glh-2-like	2.56	Ctenidin-3-like	25604	19492	146	5.67	8.20
Linum usitatissimum	43484	544.30	Unknown	4.95	Unknown	12.41	Unknown	2.49	Unknown	24926	18459	99	5.58	8.34
Lupinus angustifolius	52821	619.31	Misin	5.59	Arabinogalactan peptide 23	13.07	Collagen α-2(V) chain-like	2.76	Glutamic acid rich protein	31045	21650	126	5.67	8.21
Macleaya cordata	21911	624.74	Von Willbrand factor	3.87	Unknown	12.32	Unknown	2.94	Unknown	12657	9206	48	5.61	8.29
Malus domestica	60544	551.44	Auxin TP BIG	3.45	Unknown	12.91	CDPK	2.79	IFF6-like	34853	25548	143	5.65	8.25
Manihot acuminata	36528	516.85	Unknown	1.01	Unknown	12.88	Unknown	2.88	Unknown	18516	17936	76	5.64	8.58
Manihot esculenta	43286	621.75	Midasin	3.16	Cyt b6/f VIII	12.19	SR45-like	2.65	ASF1-like	25698	17491	97	5.68	8.18

Marchantia polymorpha	17956	806.12	Unknown	7.02	Unknown	12.10	Unknown	3.10	Unknown	10142	7793	21	5.54	8.48
Medicago truncatula	57661	611.94	Misin	1.90	NCR peptide	12.74	Unknown	2.57	LEA	30526	27027	108	5.61	8.40
Micromonas commoda	10137	1532.91	PKS	2.78	Antisense noncoding	12.61	Unknown	2.95	Unknown	6417	3703	17	5.4	8.61
Micromonas pusilla	10242	848.29	Unknown	5.07	Unknown	13.37	Unknown	2.80	Unknown	5985	4242	15	5.37	8.97
Miscanthus sinensis	89486	615.13	Unknown	2.90	Unknown	13.27	Unknown	2.88	Unknown	45710	43546	230	5.65	8.76
Momordica charantia	28666	616.95	Misin	4.21	DDG 4A	12.22	60S RP L39	3.16	Small acidic protein 1	16621	11997	48	5.69	8.20
Monoraphidium neglectum	16755	730.02	Misin	4.12	Unknown	13.01	Unknown	2.79	Unknown	8940	7783	32	5.44	8.91
Morus notabilis	26965	566.71	Auxin TP BIG	5.11	Unknown	12.32	Unknown	3.10	Unknown	13932	12984	49	5.60	8.55
Musa acuminata	47707	616.23	Misin	3.79	DGG 4A	12.44	Unknown	3.09	TUB8	27400	20184	123	5.69	8.28
Nelumbo nucifera	38191	797.88	Titin	3.16	Cyt b6/f VIII	12.22	60S RP L39	2.95	Prostatic spermine BP	22308	15782	101	5.70	8.21
Nicotiana attenuate	44491	616.08	Misin	4.15	DGG 4A	12.19	SR45	2.99	Unknown	23898	20492	101	5.67	8.24
Nicotiana sylvestris	48160	564.57	Auxin TP BIG	1.30	Unknown	12.44	Unknown	2.95	Unknown	26496	21565	99	5.66	8.25
Nicotiana tabacum	84255	539.85	Auxin TP BIG	3.16	Cyt b6/f VIII	12.44	Unknown	2.86	mRNA decay protein	46302	37768	185	5.65	8.24
Nicotiana tomentosiformis	48962	564.92	Aux TP BIG	3.18	Cyt b6/f VIII	12.25	Cell wall protein	3.09	Arabinogalactan peptide 14-like	27278	21567	117	5.67	8.22
Olea europaea	58334	567.49	Misin	3.16	Cyt b6/f VIII	12.36	Unknown	3.18	Acidic protein 2-like	32519	25690	125	5.65	8.23
Oryza brachyantha	26803	597.14	Misin	5.73	Unknown	12.44	NFD6	3.03	Prostatic spermine BP	16083	10659	61	5.64	8.30
Oryza sativa	37358	567.80	Unknown	3.22	Unknown	12.64	Unknown	2.65	Unknown	20560	16732	66	5.56	8.69
Ostreococcus lucimarinus	7603	1994.71	PKS	4.12	Cysteine protein	12.36	Unknown	2.85	Unknown	4879	2711	13	5.40	8.62
Ostreococcus tauri	7766	1237.34	Unknown	3.29	Ycf12	12.32	L39e	3.28	SVC	4732	3018	16	5.47	8.67
Panicum hallii	49825	608.88	Unknown	1.87	Unknown	13.07	Unknown	3.05	Unknown	24843	24878	104	5.68	8.79
Phalaenopsis equestris	29894	568.40	Auxin TP BIG	3.20	Cyt b6/f VIII	12.5	50S L34	2.75	Unknown	16701	13099	94	5.69	8.27
Phaseolus vulgaris	32720	617.41	Unknown	***	***	12.74	Unknown	3.09	Unknown	18577	14073	70	5.68	8.29
Phoenix dactylifera	38570	617.67	Misin	3.19	Cyt b6/f	12.45	60S RP l39	3.31	PPCD VHS3-like	22015	16450	105	5.67	8.31

Physcomitrella patens	35934	553.40	Unknown	3.21	Unknown	12.72	Unknown	2.77	Unknown	19655	16205	74	5.60	8.47
Populus deltoides	57249	567.25	Unknown	3.14	Unknown	12.54	Unknown	2.85	Unknown	31040	26084	125	5.64	8.37
Populus euphratica	49760	619.99	Misin	3.16	Cyt b6/f	12.22	60S RP L39	3.02	Calsequestrin -1-like	29358	20278	124	5.68	8.19
Populust richocarpa	45942	603.68	Misin	***	***	12.74	Unknown	2.54	Unknown	25431	20413	98	5.62	8.35
Porphyra umbilicalis	13360	480.60	Unknown	3.18	Cyt b6/f	13.33	Unknown	2.6	Unknown	3982	9365	13	5.47	10.40
Prunus avium	35009	607.05	Misin	3.84	Unknown	12.22	SR45	2.86	Unknown	20962	13972	75	5.68	8.18
Prunus mume	29705	678.34	Unknown	3.19	Cyt b6/f VIII	12.22	60S RP L39	2.54	Cell wall protein gp1	17588	12054	63	5.68	8.21
Prunus persica	32595	607.17	Misin	3.19	Cyt b6/f VIII	12.22	SR45	2.98	SCP SP60-like	19315	13201	79	5.70	8.18
Punica granatum	50476	1150.02	Unknown	1.21	Unknown	13.96	Unknown	2.07	Unknown	23078	27314	84	5.52	8.93
Pyrus bretschneideri	47086	1269.42	Unknown	4.13	Unknown	12.23	50S RP L34	3.10	Unknown	27610	19365	111	5.67	8.25
Raphanus sativus	61216	607.96	Misin	2.89	Unknown	12.58	IQ domain 31	2.79	Shematrin-like 2	34204	26871	141	5.61	8.30
Ricinus communis	27998	619.84	Misin	3.16	Cyt b6/f VIII	12.47	Unknown	3.12	Small acidic protein	16138	11789	71	5.68	8.21
Salix purpurea	61520	621.82	Unknown	3.26	Unknown	12.61	Unknown	2.77	Unknown	35045	26358	117	5.64	8.29
Selaginella moellendorffii	34746	909.93	Unknown	5.18	Unknown	12.45	Unknown	3.28	Unknown	20404	14243	99	5.66	8.27
Sesamum indicum	35410	614.62	Misin	3.16	Cyt b6/f VIII	12.64	L32	2.88	BCP1-like	20353	14974	83	5.68	8.21
Setaria italica	35844	608.65	Misin	3.16	Cyt b6/f VIII	12.44	NFD6	3.13	Prostatic spermine BP	20727	15027	90	5.7	8.36
Solanum lycopersicum	36008	620.33	Misin	3.16	Cyt b6/f VIII	12.29	SR45	2.46	Cell wall protein	21001	14925	82	5.67	8.20
Solanum pennellii	35068	620.19	Misin	3.66	Unknown	12.16	SR45	2.20	Myb-like	19944	15043	81	5.67	8.20
Solanum tuberosum	37960	618.76	Misin	3.16	Cyt b6/f VIII	13.24	Extension-like	2.95	Tripartite motif 44	22261	15614	85	5.67	8.19
Sorghum bicolor	39248	615.07	Misin	3.16	Cyt b6/f VIII	12.57	Unknown	3.00	Unknown	21947	17200	101	5.68	8.43
Sphagnum fallax	32298	1027.64	Unknown	3.19	Unknown	12.19	Unknown	2.62	Unknown	19972	12256	70	5.61	8.32
Spinacia oleracea	32794	615.44	Misin	2.50	SpolCp151	12.42	EPR1	2.56	Unknown	18350	14357	87	5.66	8.20
Spirodela polyrhiza	19623	596.16	Unknown	3.16	Cyt b6/f VIII	12.69	Unknown	2.74	Unknown	10680	8903	40	5.60	8.52
Tarenaya hassleriana	41094	614.85	Midasin	4.03	Unknown	12.41	IQ-domain 14	3.04	Unknown	23230	17762	102	5.65	8.28

Theobroma cacao	30854	621.12	Midasin	3.16	Cyt b6/f VIII	12.21	SR45	3.12	Small acidic protein	17897	12869	88	5.70	8.20
Trifolium pratense	63799	566.60	Auxin TP BIG	***	***	12.88	Unknown	2.34	Unknown	37487	26211	101	5.36	8.45
Trifolium subterraneum	42059	571.83	Unknown	***	***	12.33	Unknown	2.6	Unknown	24271	17701	87	5.28	8.28
Triticum aestivum	250	230.08	Unknown	6.54	Unknown	12.15	Unknown	3.75	Unknown	147	103	0	5.59	8.22
Triticum urartu	24169	559.02	UBR4	4.5	Unknown	13.27	Unknown	2.96	Unknown	13783	10339	47	5.56	8.51
Vigna angularis	37769	621.26	Midasin	2.96	Unknown	12.22	50S RP L34	2.81	Clumping factor A	21862	15798	109	5.7	8.21
Vigna radiata	42284	624.61	Midasin	3.16	Cyt b6/f VIII	12.41	Formin-like 6	2.89	ATF7IP	24545	17654	83	5.69	8.21
Vinga unguiculata	42287	616.90	Unknown	3.17	Cyt b6/f VIII	12.91	Unknown	2.95	Unknown	24082	18103	102	5.69	8.35
Vitis vinifera	41208	622.14	Midasin	3.16	Cyt b6/f VIII	12.19	Orf19	3.23	Circumsporozite	25295	15837	76	5.70	8.19
Volvox carteri	14436	2236.80	Putative PKS I	4.89	Unknown	13.79	Unknown	2.42	Unknown	7560	6849	27	5.64	8.53
Zostera marina	20450	606.86	*******	1.68	********	12.75	*****	2.42	*****	10988	9417	45	5.61	8.36
Average	40469.47	707.23				12.62				22820	17794.26	91.36	5.62	8.37

Abbreviations: PPCD VHS3-like: phosphopantothenoylcysteine decarboxylase subunit VHS3-like isoform X2, 60S RP L41: 60S ribosomal protein subunit L41, DDHGT 4A: Dolichyl-diphosphooligosaccharideprotein glycosyltransferase subunit 4A, PERK2: proline-rich receptor-like protein kinase PERK2, RNA Pol. II Med 17: mediator of RNA polymerase II subunit 17, FS CAYBR BP: fibrous sheath CABYR-binding protein-like, PKS: polyketide synthase, PS-I RC N: photosystem I reaction center subunit N, chloroplastic, partial; RPB1: DNA-directed RNA polymerase II subunit RPB1-like isoform X3, GRCW protein: glycine-rich cell wall structural protein, SR45: serine/arginine-rich splicing factor SR45-like, IMP:inosine-5'-monophosphate cyclohydrolase, ZRF: zinc ring finger-type, CWP:cell wall protein, DDGT: dolichyl-diphosphooligosaccharideprotein glycosyltransferase subunit 4A, PKDP: polycystic kidney disease protein 1-like 3, CSF: cleavage stimulation factor subunit 2 tau variant-like, AAPT: aminoalcoholphosphotransferase, RBP: RNA binding protein, SARMP: serine/arginine repetitive matrix protein 2-like, ER TF: ethylene responsive transcription factor, CDPK: cyclin-dependent serine/threonine-protein kinase, LEA: late embryogenesis associated protein, DDG: dolichyl-diphosphooligosaccharideprotein glycosyltransferase subunit 4A, NFD: nuclear fusion defective, SVC: satellite virus coat protein, PPCD: phosphopantothenoylcysteine decarboxylase, SCP: spore coat protein, ATF7IP: activating transcription factor 7-interacting protein 1.

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2

Abundance of various amino acids in different species. The second column represents the average abundance whereas the third and fourth column represent variation (high and low) in amino acid composition in different species from the average.

Amino Acids	Average abundance (%)	High Abundance (%)	Low Abundance (%)
Ala	7.68	*Porphyra umbilicalis* (17.58), *Monoraphidium neglectum* (16.58), *Gonium pectorale* (15.45), *Chlorella variabilis* (15.15), *Chlamydomonas reinhardtii* (14.57), *Micromonas pusilla* (14.20), *Auxenochlorella protothecoides* (13.69), *Volvox carteri* (13.29), *Helicosporidium* sp. (12.75), *Micromonas commoda* (12.61), *Coccomyxa subellipsoidea* (12.29), *Ostreococcus lucimarinus* (11.89), *Chromochloris zofingiensis* (11.57), *Dunaliella salina* (11.53), *Ostreococcus tauri* (11.47), *Klebsormidium nitens* (10.62), *Botrycoccus braunii* (9.86), *Dichanthelium oligosanthes* (9.71), *Chlamydomonas eustigma* (9.53)	*Picea glauca* (5.12), *Medicago truncatula* (5.8), *Lactua sativa* (5.81), *Zostera marina* (5.86)
Asp	5.32	*Micromonas pusilla* (6.68), *Micromonas commode* (6.46), *Ostreococcus lucimarinus* (6.38), *Ostreococcus tauri* (6.33), *Bathycoccus prasinos* (6.05)	*Picea glauca* (3.46), *Dunaliella salina* (4.28), *Chlorella variabilis* (4.41), *Chlamydomonas reinhardtii* (4.61), *Porphyra umbilicalis* (4.63), *Volvox carteri* (4.71), *Monoraphidium neglectum* (4.71), *Botrycoccus braunii* (4.78), *Gonium pectorale* (4.78)
Glu	6.43	*Bathycoccus prasinos* (8.44), *Klebsormidium nitens* (7.2),	*Porphyra umbilicalis* (3.52), *Picea glauca*

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		Ostreococcus tauri (7.18), *Ostreococcus lucimarinus* (7.03)	(4.02), *Chromochloris zofingiensis* (4.66), *Chlamydomonas reinhardtii* (5.15), *Volvox carteri* (5.21), *Monoraphidium neglectum* (5.35)
Phe	3.97	*Arachis duranensis* (6.18)	*Porphyra umbilicalis* (2.01), *Gonium pectorale* (2.4), *Monoraphidium neglectum* (2.44), *Volvox carteri* (2.49), *Dunaliella salina* (2.56), *Chromochloris zofingiensis* (2.58), *Chlamydomonas reinhardtii* (2.59), *Chlorella variabilis* (2.68), *Chlamydomonas eustigma* (2.89), *Auxenochlorella protothecoides* (2.92)
Gly	6.80	*Porphyra umbilicalis* (11.76), *Monoraphidium neglectum* (10.51), *Gonium pectorale* (10.33), *Chlamydomonas reinhardtii* (9.58), *Volvox carteri* (9.54), *Chlorella variabilis* (9.23), *Auxenochlorella protothecoides* (8.82), *Micromonas commoda* (8.81), *Micromonas pusilla* (8.59), *Klebsormidium nitens* (8.44), *Helicosporidium* sp. (8.29), *Dunaliella salina* (8.11), *Botrycoccus braunii* (8.03)	*Arachis duranensis* (4.19)
His	2.4	*Dunaliella salina* (3.13)	*Bathycoccus prasinos* (1.87), *Ostreococcus tauri* (1.93), *Monoraphidium neglectum* (1.93), *Micromonas pusilla* (1.94), *Ostreococcus lucimarinus* (1.96)

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ile	4.94	*Zostera marina* (6.09)	*Porphyra umbilicalis* (1.77), *Monoraphidium neglectum* (2.34), *Gonium pectorale* (2.37), *Chlorella variabilis* (2.52), *Chlamydomonas reinhardtii* (2.60), *Volvox carteri* (2.78), *Helicosporidium* sp. (2.80), *Dunaliella salina* (2.82), *Auxenochlorella protothecoides* (2.96), *Micromonas pusilla* (2.98)
Lys	5.73	*Bathycoccus prasinos* (7.33)	*Porphyra umbilicalis* (2.08), *Gonium pectorale* (2.79), *Monoraphidium neglectum* (2.86), *Volvox carteri* (3.03), *Auxenochlorella protothecoides* (3.06), *Chlorella variabilis* (3.06), *Chlamydomonas reinhardtii* (3.07), *Helicosporidium* sp. (3.17), *Dunaliella salina* (3.39)
Leu	9.62	*Picea glauca* (13.12)	*Porphyra umbilicalis* (7.07), *Micromonas pusilla* (7.70), *Bathycoccus prasinos* (7.85), *Micromonas commoda* (7.98), *Ostreococcus tauri* (8.18), *Ostreococcus lucimarinus* (8.41)
Met	2.40	*Picea glauca* (3.75)	*Porphyra umbilicalis* (1.45), *Monoraphidium neglectum* (1.78),

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

			Klebsormidium nitens (1.84), *Auxenochlorella protothecoides* (1.95), *Gonium pectorale* (1.97), *Chlorella variabilis* (1.98), *Micromonas pusilla* (1.98), *Helicosporidium* sp. (1.99)
Asn	4.13	*Medicago truncatula* (5.09)	*Porphyra umbilicalis* (1.53), *Monoraphidium neglectum* (1.88), *Chlorella variabilis* (1.93), *Auxenochlorella protothecoides* (2.02), *Gonium pectorale* (2.11), *Helicosporidium* sp. (2.15), *Chlamydomonas reinhardtii* (2.38), *Micromonas pusilla* (2.49), *Volvox carteri* (2.49), *Dunaliella salina* (2.61), *Micromonas commoda* (2.71), *Coccomyxa subellipsoidea* (2.75), *Klebsormidium nitens* (2.82), *Botrycoccus braunii* (2.89), *Ostreococcus tauri* (2.93), *Ostreococcus lucimarinus* (2.99)
Pro	5.10	*Porphyra umbilicalis* (9.2), *Botrycoccus braunii* (7.10), *Dunaliella salina* (7.08), *Volvox carteri* (7.00), *Gonium pectorale* (6.85), *Chlamydomonas reinhardtii* (6.49), *Picea glauca* (6.45), *Chlorella variabilis* (6.41),	*Bathycoccus prasinos* (3.81)

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		Monoraphidium neglectum (6.40), *Auxenochlorella protothecoides* (6.25), *Klebsormidium nitens* (6.11)	
Gln	3.74	*Dunaliella salina* (7.00), *Chromochloris zofingiensis* (6.20), *Monoraphidium neglectum* (5.60), *Chlorella variabilis* (5.56), *Chlamydomonas eustigma* (4.83), *Sphagnum fallax* (4.67), *Coccomyxa subellipsoidea* (4.59), *Volvox carteri* (4.46), *Botrycoccus braunii* (4.35), *Chlamydomonas reinhardtii* (4.28), *Physcomitrella patens* (4.08), *Dorcoceras hygrometricum* (4.04), *Klebsormidium nitens* (4.03)	*Porphyra umbilicalis* (2.04), *Micromonas pusilla* (2.06), *Micromonas commoda* (2.58), *Ostreococcus tauri* (2.58), *Ostreococcus lucimarinus* (2.64)
Arg	5.68	*Porphyra umbilicalis* (9.81), *Micromonas pusilla* (8.12), *Ostreococcus tauri* (7.96), *Helicosporidium* sp. (7.74), *Micromonas commoda* (7.65), *Ostreococcus lucimarinus* (7.46), *Auxenochlorella protothecoides* (7.13)	*Medicago truncatula* (4.84), *Trifolium pratense* (4.95), *Cicer arietinum* (4.97), *Lactuca sativa* (4.99)
Ser	8.71	*Chlamydomonas eustigma* (9.72)	*Monoraphidium neglectum* (6.33), *Chlorella variabilis* (6.44), *Auxenochlorella protothecoides* (6.49), *Porphyra umbilicalis* (6.52), *Micromonas commoda* (6.57), *Ostreococcus lucimarinus* (6.72), *Micromonas pusilla* (6.74), *Gonium pectorale* (7.06), *Chlamydomonas reinhardtii* (7.11)
Thr	4.90	*Chromochloris zofingiensis* (5.80), *Ostreococcus tauri* (5.77), *Bathycoccus prasinos* (5.72), *Ostreococcus lucimarinus* (5.70), *Chlamydomonas eustigma* (5.44),	

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		Porphyra umbilicalis (5.39), *Micromonas pusilla* (5.33), *Volvox carteri* (5.30)	
Val	6.55	*Ostreococcus tauri* (7.42), *Ostreococcus lucimarinus* (7.35), *Porphyra umbilicalis* (7.32), *Micromonas commoda* (7.22), *Helicosporidium* sp. (7.20), *Micromonas pusilla* (7.11)	*Picea glauca* (5.26), *Dunaliella salina* (5.78)

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3

Average abundance of different amino acids in plant proteome. Leu was high abundant whereas Trp was the low abundant amino acid in the plant kingdom. The average amino acid composition includes 5.8 million protein sequences from 145 plant species.

Biosynthetic pathways/Substrate	Amino acids	Average abundance (%) in Proteome
α-Ketoglutarate	Arginine	6.68
	Glutamate	6.43
	Glutamine	3.74
	Proline	5.10
Pyruvate	Alanine	7.68
	Isoleucine	4.94
	Leucine	9.62
	Valine	6.55
3-Phosphoglycerate	Glycine	6.80
	Cysteine	1.85
	Serine	8.71
Oxaloacetate	Asparagine	4.13
	Aspartate	5.32
	Lysine	5.73
	Methionine	2.40
	Threonine	4.90
Phosphoenolpyruvate & Erythrose 4-phosphate	Phenylalanine	3.97
	Tryptophan	1.28
	Tyrosine	2.67
Ribose 5-phosphate	Histidine	2.4

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure legends

Figure 1

Principal component analysis (PCA) of acidic *pI* proteins. The PCA plot illustrates the relationship between the acidic *pI* of bryophytes and monocot plants which exhibit a linear correlation relative to algae and eudicots.

Figure 2

Principal component analysis (PCA) of basic *pI* proteins. The PCA plot illustrates that the basic *pI* of algae, bryophytes, eudicots, and monocot plants cluster distinctly from each other and that there is no lineage-specific correlation with basic *pI* proteins.

Figure 3

Trimodal distribution of isoelectric points (*pI*) and the molecular mass (kDa) of plant proteins. The *pI* of plant proteins ranged from 1.99 (epsin) to 13.96 (hypothetical protein), while the molecular mass ranged from 0.54 (unknown) to 2236.8 (type I polyketide synthase) kDa. The X-axis represents the *pI* and the Y- axis represents the molecular mass of the proteins.

Figure 4

Average amino acid composition of proteins in the plant kingdom. Leu is the most abundant while Trp is the least abundant. The amino acid, Sec, was only found in a few species of algae and was absent from all other species. Ambiguous amino acids were found in a few species as well.

Figure 5

Principal component analysis (PCA) of amino acid abundance in plant proteomes. The PCA plot shows that Tyr, Trp, Cys, His, Met, and Xaa (unknown) amino acids are low-abundance and cluster together. The abundance of Leu, Ser, Ile, Lys, and Gln was higher and grouped together. The plot shows that the abundance of amino acids is lineage specific. Algae, eudicots, and monocot plants exhibit a lineage specific correlation.

bioRxiv preprint doi: https://doi.org/10.1101/546077; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 6

Schematic illustration of the biosynthetic pathway of amino acids. The abundance of aromatic ring containing amino acids is lower relative to other amino acids. The average abundance of the aromatic ring containing amino acid, Trp, is the lowest amongst others that are biosynthesized via phosphoenolpyruvate and erythrose 4-phosphate. Similarly, the abundance of Cys is relatively low compared to other amino acids. Ser is biosynthesized from 3-phosphoglycerate and Ser is subsequently used to produce Gly and Cys amino acids. The abundance of Cys is lower relative to Gly, suggesting the existence of allosteric feed-back inhibition of the biosynthesis of Cys by Ser.

Supplementary Data

Supplementary File 1

Supplementary file containing average amino acid composition of plant proteomes. The file is present in excel sheet and individual sheet represents details of different amino acid.

Figure 1

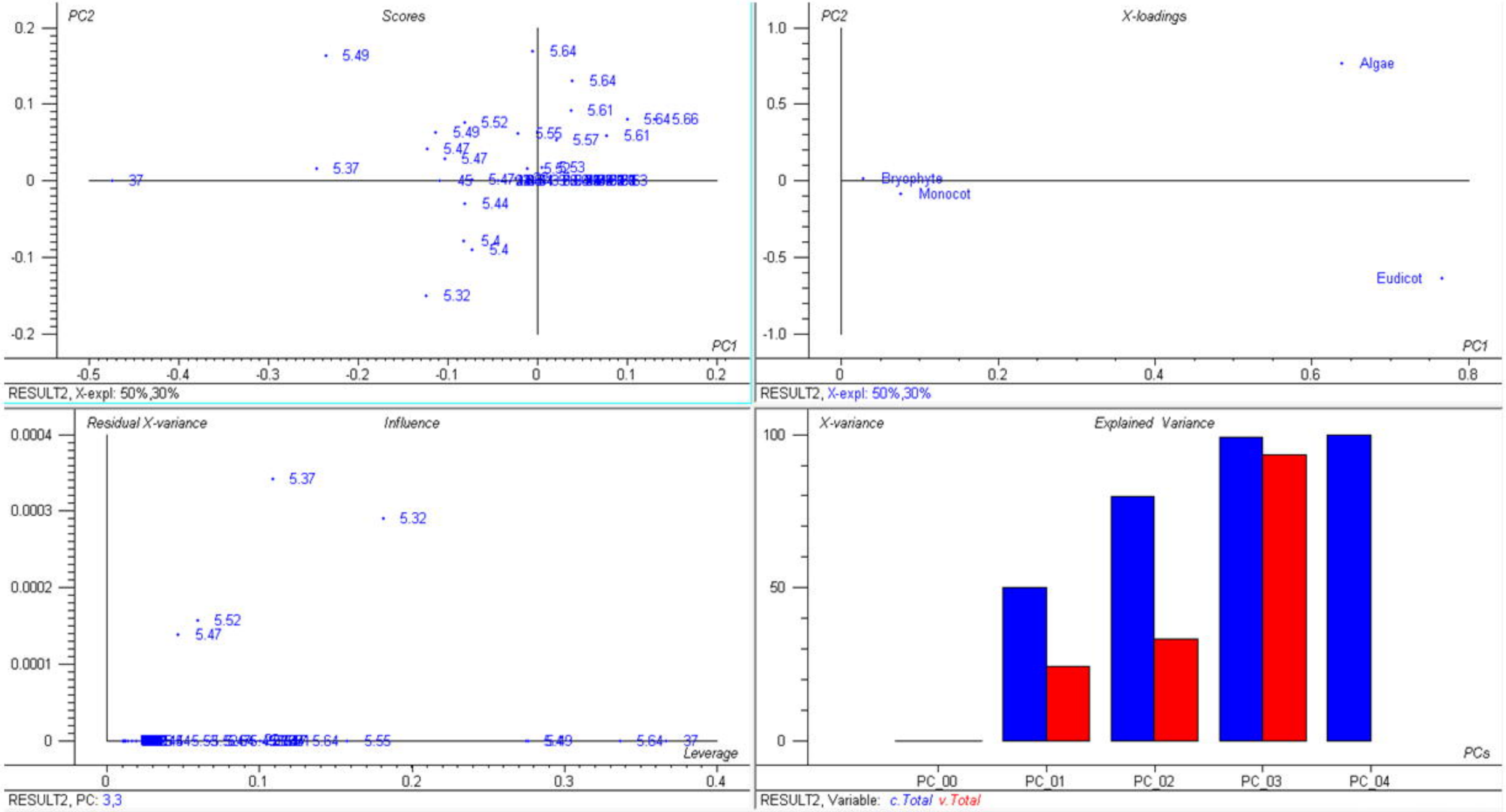

Scores
PC2
PC1
RESULT2, X-expl: 50%,30%
X-loadings
Algae
Bryophyte
Monocot
Eudicot
RESULT2, X-expl: 50%,30%
Residual X-variance
Influence
Leverage
RESULT2, PC: 3,3
X-variance
Explained Variance
PCs
PC_00
PC_01
PC_02
PC_03
PC_04
RESULT2, Variable: c.Total v.Total

Figure 2

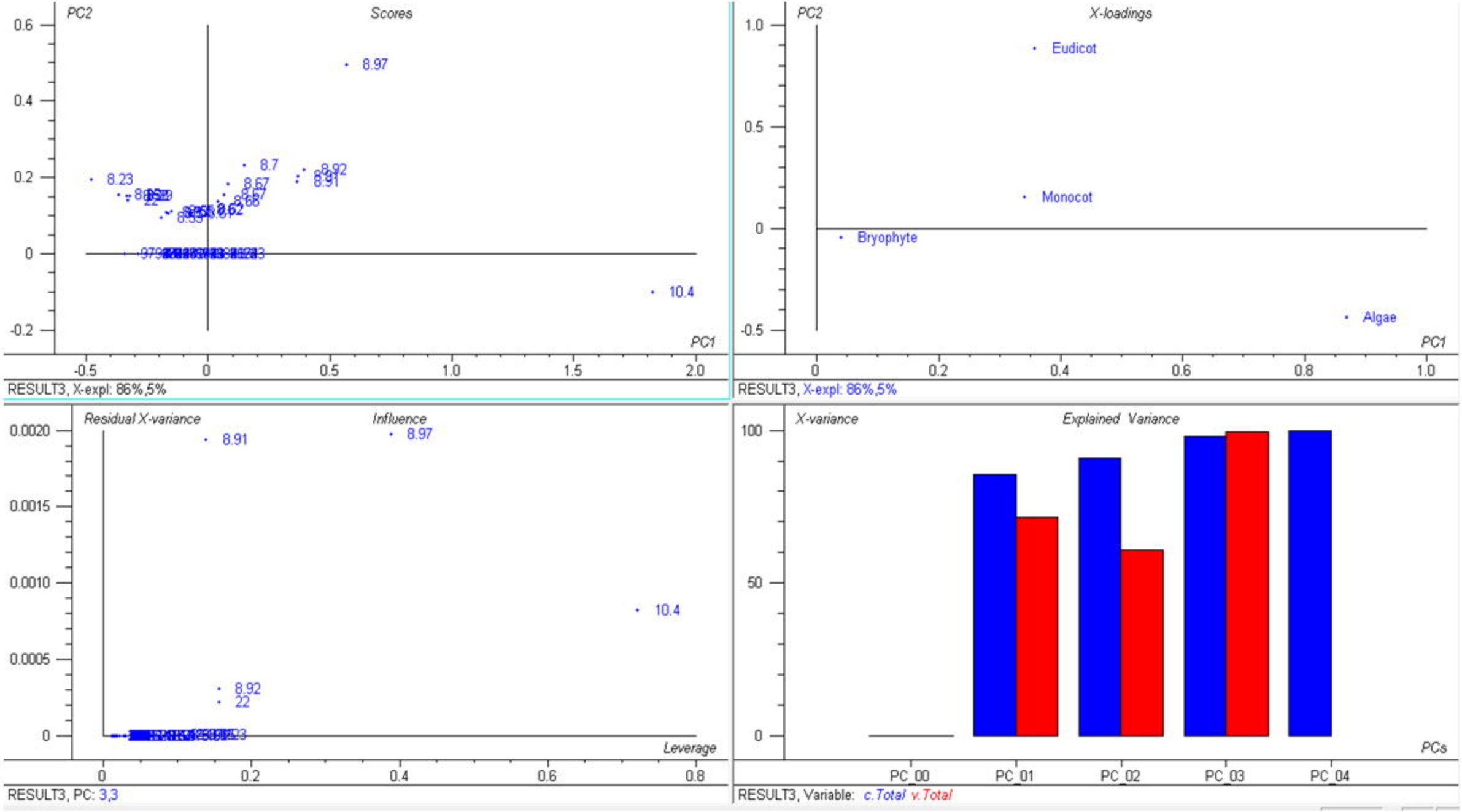

Scores
PC2
PC1
8.97
8.7
8.92
8.91
8.23
10.4
RESULT3, X-expl: 86%,5%
X-loadings
Eudicot
Monocot
Bryophyte
Algae
RESULT3, X-expl: 86%,5%
Residual X-variance
Influence
Leverage
8.91
8.97
10.4
8.92
22
RESULT3, PC: 3,3
X-variance
Explained Variance
PCs
PC_00
PC_01
PC_02
PC_03
PC_04
RESULT3, Variable: c.Total v.Total

Figure 3

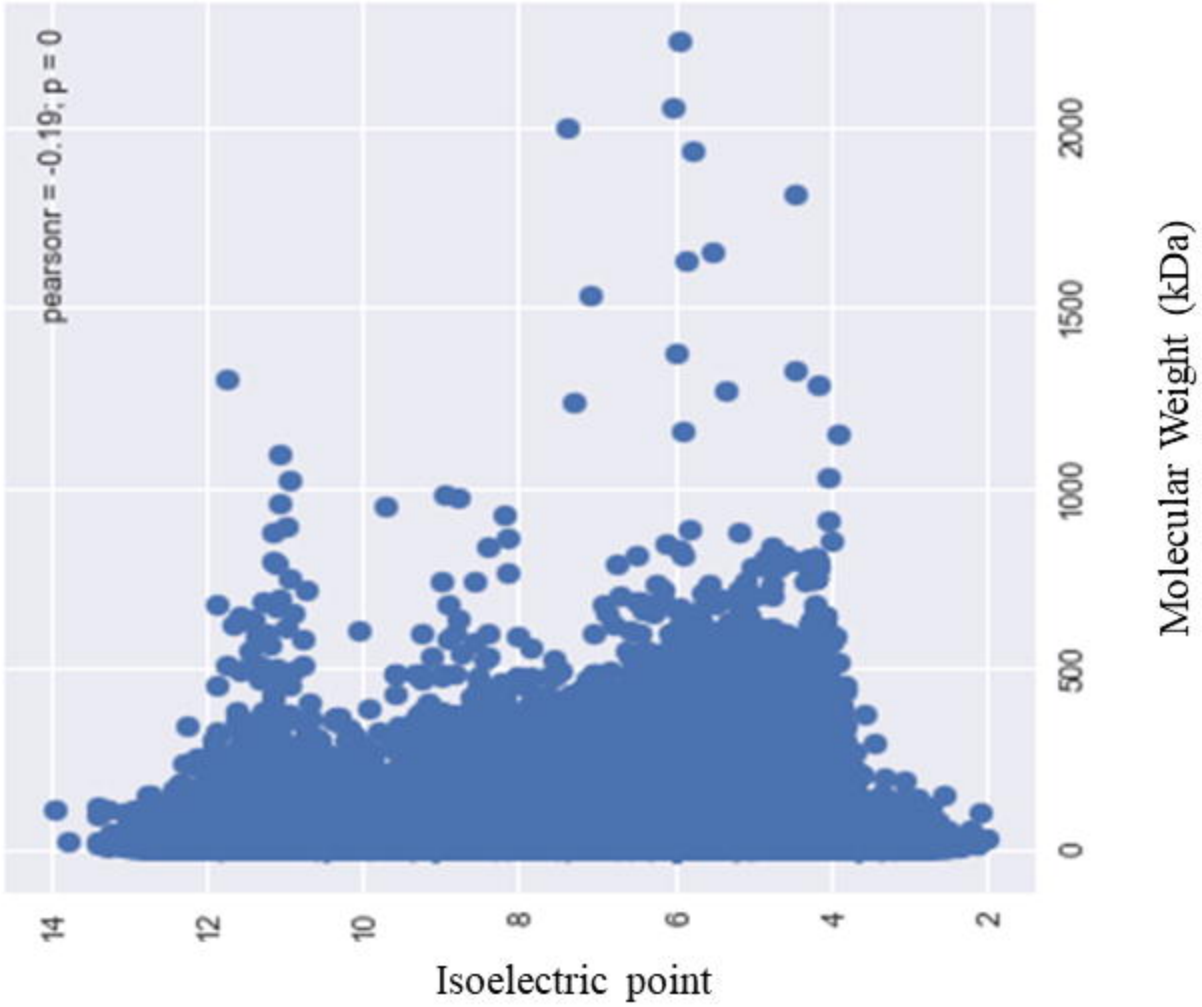

pearsonr = -0.19; p = 0
14
12
10
8
6
4
2
Isoelectric point
0
500
1000
1500
2000
Molecular Weight (kDa)

Figure 4

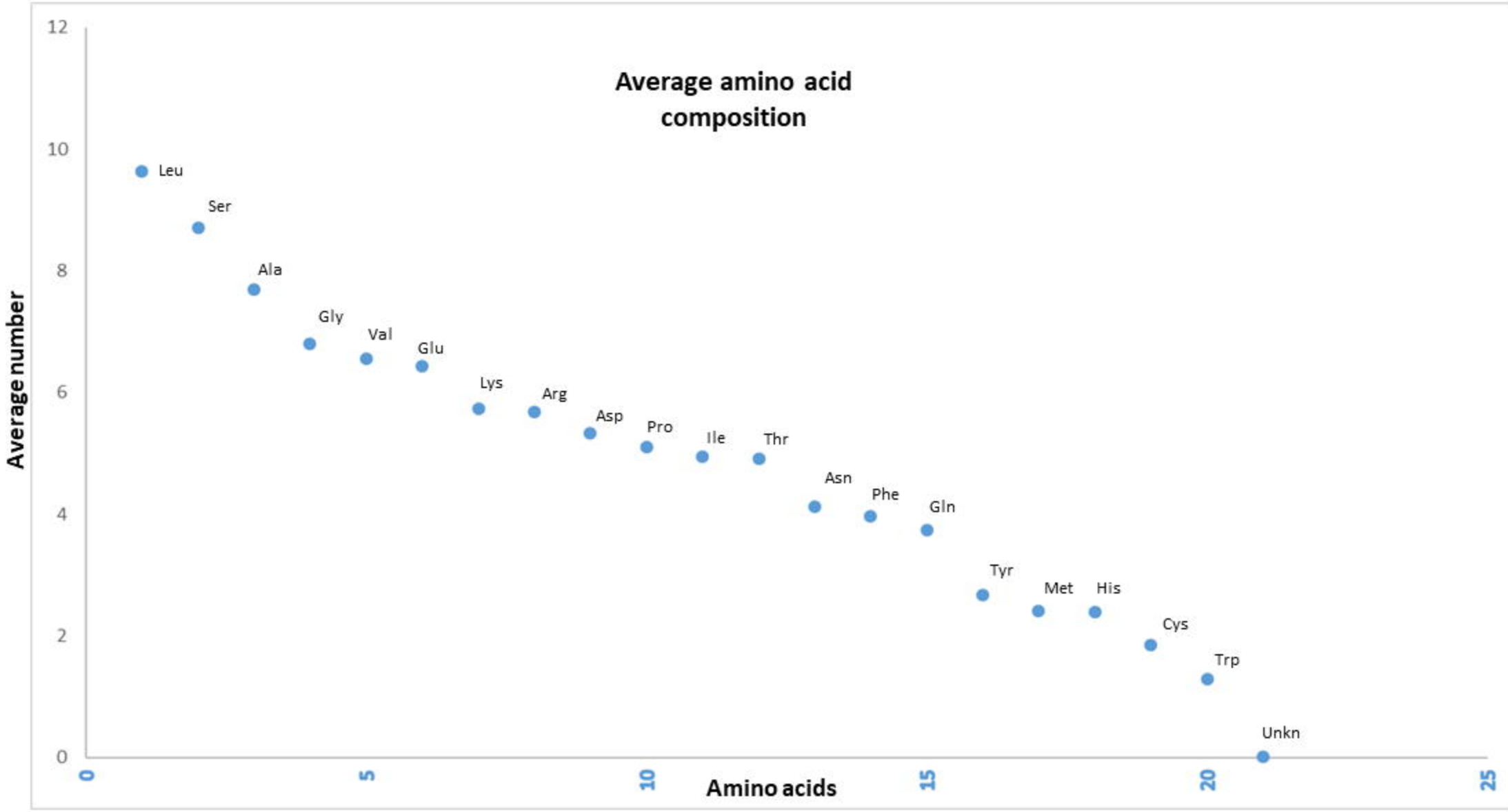

Average amino acid composition
Average number
Amino acids
12
10
8
6
4
2
0
0
5
10
15
20
25
Leu
Ser
Ala
Gly
Val
Glu
Lys
Arg
Asp
Pro
Ile
Thr
Asn
Phe
Gln
Tyr
Met
His
Cys
Trp
Unkn

Figure 5

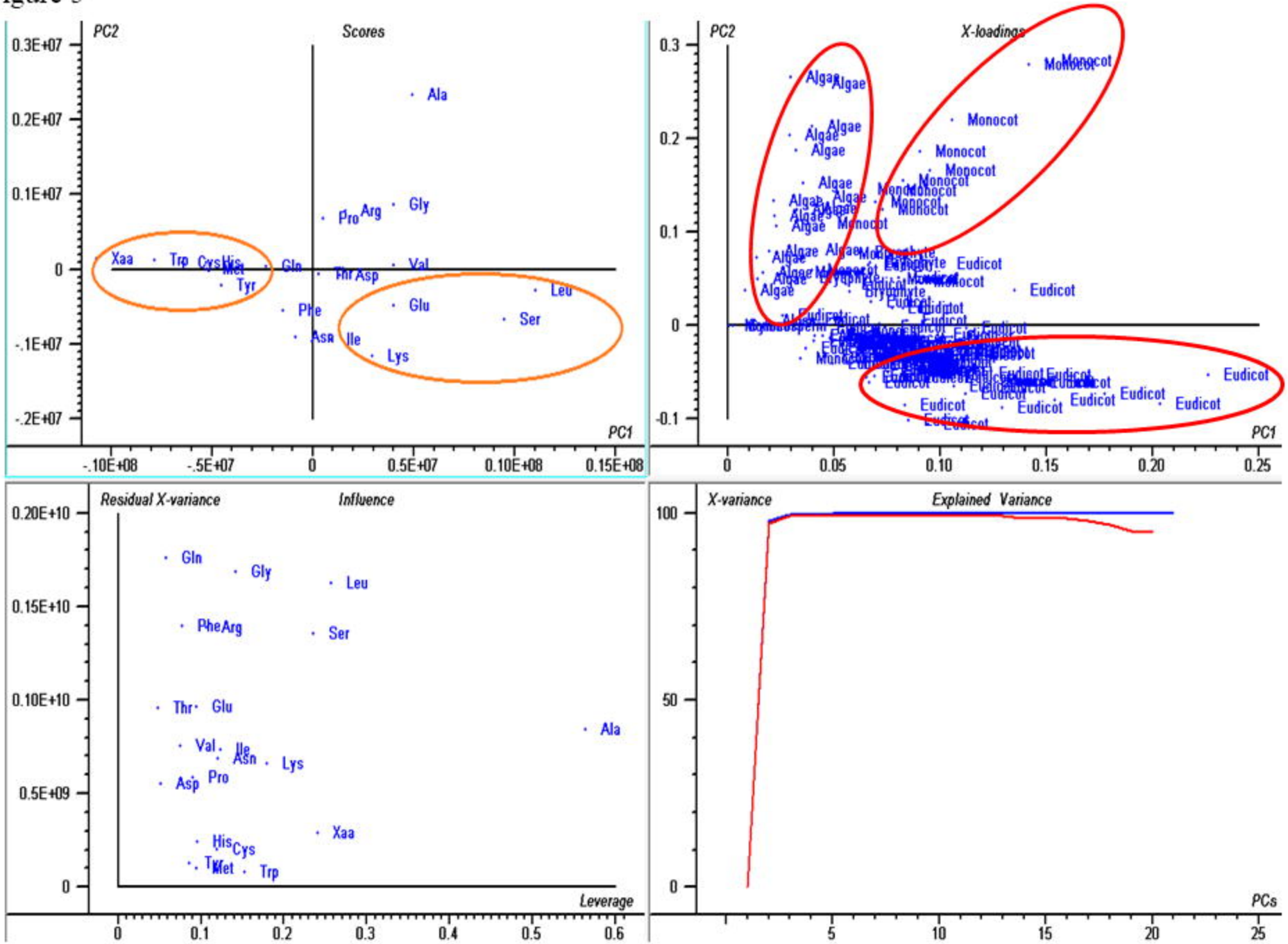
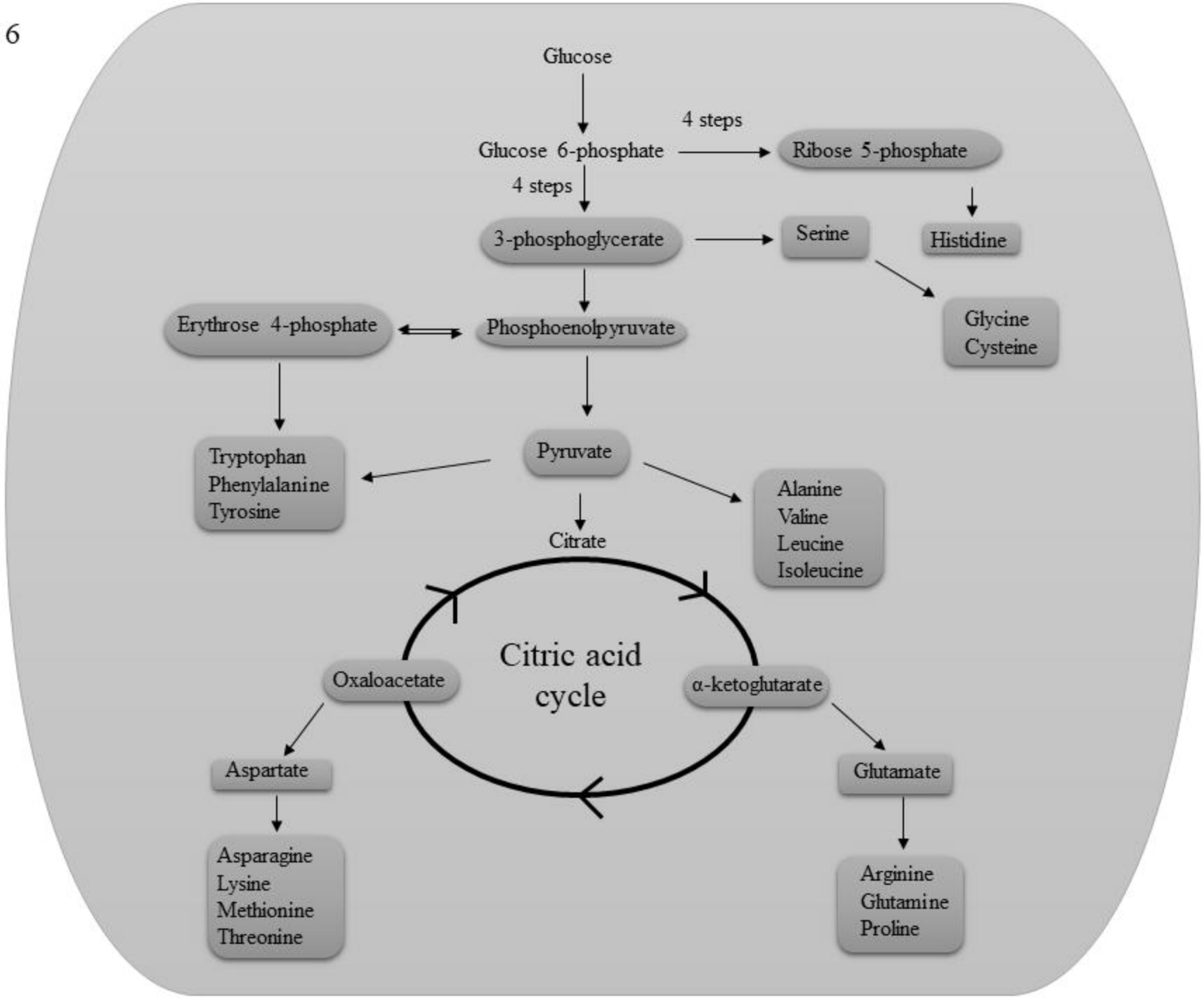

PC2
Scores
PC1
Ala
Gly
Pro
Arg
Xaa
Trp
Cys
His
Met
Gln
Thr
Asp
Val
Tyr
Leu
Phe
Glu
Ser
Asn
Ile
Lys
X-loadings
Monocot
Algae
Eudicot
Residual X-variance
Influence
Leverage
X-variance
Explained Variance
PCs

Figure 6


Glucose
4 steps
Glucose 6-phosphate
Ribose 5-phosphate
4 steps
3-phosphoglycerate
Serine
Histidine
Glycine
Cysteine
Erythrose 4-phosphate
Phosphoenolpyruvate
Tryptophan
Phenylalanine
Tyrosine
Pyruvate
Alanine
Valine
Leucine
Isoleucine
Citrate
Citric acid cycle
Oxaloacetate
α-ketoglutarate
Aspartate
Glutamate
Asparagine
Lysine
Methionine
Threonine
Arginine
Glutamine
Proline