Journal of Bioinformatics and Computational Biology
Vol. 4, No. 4 (2006) 783–792

Imperial College Press
www.icpress.co.uk

TRANSLATION REGULATION OF INTRON-CONTAINING GENES IN CHLOROPLASTS

ALEXANDER SELIVERSTOV* and VASSILY LYUBETSKY†

Institute for Information Transmission Problems
Russian Academy of Sciences, B. Karetny
19, Moscow 127994, Russia
**slvstv@iitp.ru*
†*lyubetsk@iitp.ru*

Received 18 October 2005
Revised 14 February 2006
Accepted 16 February 2006

The paper provides a short description of the originally developed algorithm for searching of the conservative protein–RNA binding sites. The algorithm is applied to analyze chloroplast genes. The candidate protein–RNA binding sites were detected upstream of *atpF*, *petB*, *clpP*, *psaA*, *psbA*, and *psbB* genes in many chloroplasts of algae and plants. We suggest that some of these sites are involved in suppressing translation until splicing is completed.

Keywords: Translation; chloroplasts; multiple alignment.

1. Introduction

The gene expression in chloroplasts of algae and plants is regulated by binding of nuclear-encoded proteins to the chloroplast mRNA.[1] These proteins are involved in editing, translation and maintaining stability of the chloroplast mRNA. The detailed analysis of regulatory sites was published for alga *Chlamydomonas reinhardtii* and some plants.[1–3] For example, protein binding to the *psbA* 5′-untranslated region in *C. reinhardtii* leads to activation of translation.[3]

Here, we briefly describe an algorithm for identification of protein–RNA binding sites. The algorithm was used to identify regulatory sites in 5′-untranslated regions of protein-coding chloroplast genes.

Many chloroplast protein-coding genes contain introns. Thus, their translation should not start immediately after transcription. However, the translation machinery of chloroplasts closely resembles that of bacteria, particularly, in the ribosome being able to follow the RNA-polymerase on the mRNA strand. If the ribosome arrives at the end of an exon before splicing starts, it will preclude splicing. To avoid this, in some rare cases the AUG start codon is derived from ACG by editing of mRNA, which prevents translation from starting immediately.[2] RNA editing is

known for chloroplasts of higher plants and is absent, e.g. in liverwort *Marchantia polymorpha.*

Our algorithm detects candidate protein–RNA binding sites upstream of *atpF*, *petB*, *clpP*, *psaA*, *psbA*, and *psbB* genes in many chloroplasts.

We suggest that some of these sites are involved in suppressing translation until splicing is completed. This conjecture is also consistent with the presence of conserved sites upstream of these genes in multiple alignments (given below) and is supported by experiment in published evidence.[3]

2. Materials

Chloroplast genomes were obtained from GenBank (NCBI). The dataset contained 5′-untranslated intergenic regions from chloroplast genomes of algae and plants: *Euglena gracilis*,*Odontella sinensis*, *Guillardia theta*, *Cyanidioschyzon merolae*, *Cyanidium caldarium*,*Gracilaria tenuistipitata*, *Porphyra purpurea*, *Chlamydomonas reinhardtii*, *Nephroselmis olivacea*, *Chaetosphaeridium globosum*, *Mesostigma viride*, *Anthoceros formosae*, *Marchantia polymorpha*, *Adiantum capillus-veneris*, *Huperzia lucidula*, *Psilotum nudum*, *Pinus thunbergii*, *Amborella trichopoda*, *Arabidopsis thaliana*, *Atropa belladonna*, *Calycanthus floridus*, *Cucumis sativus*, *Epifagus virginiana*, *Lotus corniculatus*, *Nicotiana tabacum*, *Nymphaea alba*, *Oryza nivara*, *Oryza sativa*, *Panax ginseng*, *Spinacia oleracea*, *Triticum aestivum*, *Zea mays.*

The only nonphotosynthetic species in this group is *E. virginiana.* In *P. nudum*, the *psbB* gene is annotated *psbT.*

3. Methods

Consider a dataset of leader regions upstream of orthologous genes and a corresponding species tree. A set of shallow phylogenetic subtrees (groups) is selected in the species tree. For each of these groups, the algorithm searches for conserved regions of fixed length n (that can be varied) *via* finding cliques in a suitable multipartite graph. The basic idea is as follows. The algorithm finds clusters of very similar sites, called signals or motifs, of a fixed length n for each of these phylogenetic groups. A cluster of sites generates a weight matrix $4 \times n$, where the kth column of the matrix, $1 \leq k \leq n$, contains letter frequencies in the kth site position from the cluster. Further, the algorithm generates clusters of weight matrices for different suitable n. The clusters of matrices are generated accounting for distances between the ancestors of the initial phylogenetic groups in the species tree. The algorithm of finding cliques also constructs these clusters of matrices. In each matrix cluster, the matrices are replaced by corresponding site clusters and thus obtained the clusters of clusters are then combined. The described procedure can be iterated. The algorithm is described in detail in Ref. 4.

4. Notation

In the following sections, R denotes A or G, Y denotes C or U, W denotes A or U, S denotes C or G, D denotes A, G or U, N denotes any nucleotide and * denotes any nucleotide or gap.

5. Results

In many chloroplasts, the algorithm found long conserved binding sites containing conserved helices upstream of the genes *atpF*, *petB*, *clpP*, *psaA*, *psbA*, and *psbB.* A summary of the results is given in Table 1.

Table 1. Occurrence of introns and predicted sites upstream of chloroplast genes *atpF*, *petB*, *clpP*, *psaA*, *psbA*, and *psbB*. Notation: "+" — candidate protein binding site present; "−" — no candidate binding site; "s" — introns present; "n" — no gene homolog in the species; "&" — helices in the site; "E" — editing of the start codon.

Filum	Species	*atpF*	*clpP*	*petB*	*psaA*	*psbA*	*psbB*
Euglenozoa	*Euglena gracilis*	−s	−	−s	−s	−s	−s
Bacillariophyta	*Odontella sinensis*	−	−	−	+&	+&	−
Cryptophyta	*Guillardia theta*	−	−	−	+&	+&	−
Rhodophyta	*Cyanidioschyzon merolae*	−	−	−	−	+&	−
	Cyanidium caldarium	−	−	−	−	−	−
	Porphyra purpurea	−	−	−	+&	+&	+
	Gracilaria tenuistipitata	−	−	−	−	+&	−
Chlorophyta	*Chlamydomonas reinhardtii*	−	−	−	−s	+&s	−
	Nephroselmis olivacea	−	−	−	+&	+&	+
Streptophyta	*Chaetosphaeridium globosum*	−	+&s	−s	+&	+&	+
	Mesostigma viride	−	−	−	+&	−	−
Anthocerophyta	*Anthoceros formosae*	+s	+&s	+&s	+&	+&	+
Hepatophyta	*Marchantia polymorpha*	+s	+&s	+&s	+&	+&	+
Lycopodiophyta	*Huperzia lucidula*	+s	+&s	+&s	+&	+&	+
Pteridophyta	*Adiantum capillus-veneris*	+sE	+&s	−sE	+&	+&	+
Psilophyta	*Psilotum nudum*	+s	+&s	+&s	+&	+&	+
Pinophyta	*Pinus thunbergii*	+s	+&	+&s	+&	+&s	+
Magnoliophyta	*Amborella trichopoda*	+s	+&s	+&s	+&	−	+
(eudicotyledons)	*Arabidopsis thaliana*	+s	+&s	+&s	+&	+&	+
	Atropa belladonna	+s	+&s	+&s	+&	+&	+
	Calycanthus floridus	+s	+&s	+&s	+&	+&	+
	Cucumis sativus	+s	+&s	+&s	+&	+&	+
	Epifagus virginiana	n	+&s	n	n	n	n
	Lotus corniculatus	+s	+&s	+&s	+&	+&	+
	Nicotiana tabacum	+s	+&s	+&s	+&	+&	+
	Nymphaea alba	+s	+&s	+&s	+&	+&	+
	Panax ginseng	+s	+&s	+&s	+&	+&	+
	Spinacia oleracea	+s	+&s	+&s	+&	+&	+
Magnoliophyta	*Oryza nivara*	+s	+&s	+&s	+&	+&	+
(Liliopsida)	*Oryza sativa*	+s	+&s	+&s	+&	+&	+
	Triticum aestivum	+s	+&s	+&s	+&	+&	+
	Zea mays	+s	+&s	+&s	+&	+&	+

For the genes *atpF*, *clpP*, and *petB*, there is a strong correlation between the occurrence of splicing and the existence of the predicted protein-binding sites. On the other hand, for *psaA*, *psbA*, and *psbB* no such correlation is found. For *clpP*, *petB, psaA*, *psbA* the sites always contain helices, but for *atpF* and *psbB* they do not.

5.1. *ATP synthetase*

The gene *atpF* (ATP-synthetase subunit) contains introns in chloroplasts of studied plants, except for *E. virginiana*. In *A. capillus-veneris*, the start codon AUG is derived from ACG by mRNA editing. In other species, AUG or GUG are used as start codons.

The 5′-untranslated regions in the plant *atpF* genes contain two conserved motifs. One of them is A-rich, and the other has the consensus UAUCUADAA-GRGGAGADNR and overlaps the putative ribosome-binding sites. These two motifs are separated by a UA-rich region of length 19–61 residues, which is divergent even within Magnoliophyta. Figure 1 shows an alignment of 5′-untranslated regions adjacent to the start codon of *atpF*.

5.2. *Cytochrome b6*

The 5′-untranslated regions of the gene *petB* (cytochrome b6) in chloroplast genomes of all plants (except *A. capillus-veneris* and *E. virginiana*) were found to possess a conserved motif located upstream of the AUG start codon. Its consensus is GGUAGUUCGAYCGYGGAAUU*YUUU***GUUUNNGUAUUUYYGGAAU. The region also contains a conserved helix. *E. virginiana* does not contain a homolog of this gene. The alignment is shown in Fig. 2. In *A. capillus-veneris*, the corresponding 5'-untranslated region could not be aligned with the same profile.

Anthoceros formosae	ggauaaggaagagacauacuaagacuuaaagaaccuaugaugggagagagagu
Marchantia polymorpha	aaaaacuccauaauuuucaauaauauaauaacgaaaaaaagaggacagc****
Adiantum capillus-veneris	ggcuggaucagaaauugcccaaaacguaaaaccuucgaggagggaaaagaau*
Huperzia lucidula	auaaggaaaaaacuauguaaacuuggauaauaaccuguaaugggagaaaagu*
Psilotum nudum	uuaauaaaagaaaaacuuugucaaaauuagauagucauuaugggagagguauu
Pinus thunbergii	gaaaauaacaaaauucuguagaacauaucccuuaucuaugaggggagagcgu**
Amborella trichopoda	caaaaagaacuccguuugguuuuguuagucccuaucuagaagaggagaguau**
Arabidopsis thaliana	aaaaaaaggacagaguuccuuuuuuauaguuuagcuagaagaggagauuau**
Atropa belladonna	uacaaaaagaacucuguucgauuuuuuagucuaucuauaagaggagaucau**
Calycanthus floridus	agaacucugcgcaauuuuguuagcccuaucuauucuauaagaggaaagcau**
Cucumis sativus	aaaaaaauagaaagaaauagauaauuagguuuuaucuauaaaaggagaucau**
Lotus corniculatus	aaaaauaggaaucauauaaagagaauucguuuaucccauaagaggagaucau**
Nicotiana tabacum	uacaaaaagaacucuguucgauuuuuuagucuaucuauaagaggagaucau**
Nymphaea alba	uuuuuguauuuuguuagucccuauuccuauccauaagauaagaggagagcau**
Panax ginseng	aaagaacucuguucuuuuuuuuuuuuuagucuaucuauaagaggagaucau**
Spinacia oleracea	acaauugaaauaauacaacgauuuuuuuguuuaucuauaagaggagaucau**
Oryza nivara	gugaucgaaaaacuuuguucuuuguucguccuaucuauaagaggagagcau**
Oryza sativa	gugaucgaaaaacuuuguucuuuguucguccuaucuauaagaggagagcau**
Triticum aestivum	gugaucgaaaaacuuuguucuuuguucguccuaucuauaagaggagagcau**
Zea mays	guaaucgaaaaacuuucuucuuuguucguccuaucuauaagaggagagcau**
Consensus	**UAUCUADAAGRGGAGADNR**

Fig. 1. Alignment of 5′-untranslated regions upstream of the *atpF* start codon.

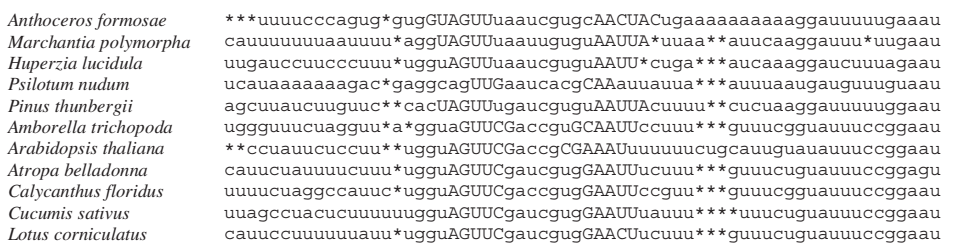

Fig. 2. Alignment of 5′-untranslated regions upstream of the *petB* start codon. Helices are shown in capitals.

5.3. *ATP-dependent clp protease proteolytic subunit*

The 5′-untranslated regions of the *clpP* gene from chloroplasts of plants and alga *C. globosum* contain a conserved motif with the consensus UUACGYUUYCAYAU-YARAGNRNARU. The corresponding alignment is shown in Fig. 3. The motif contains a conserved helix.

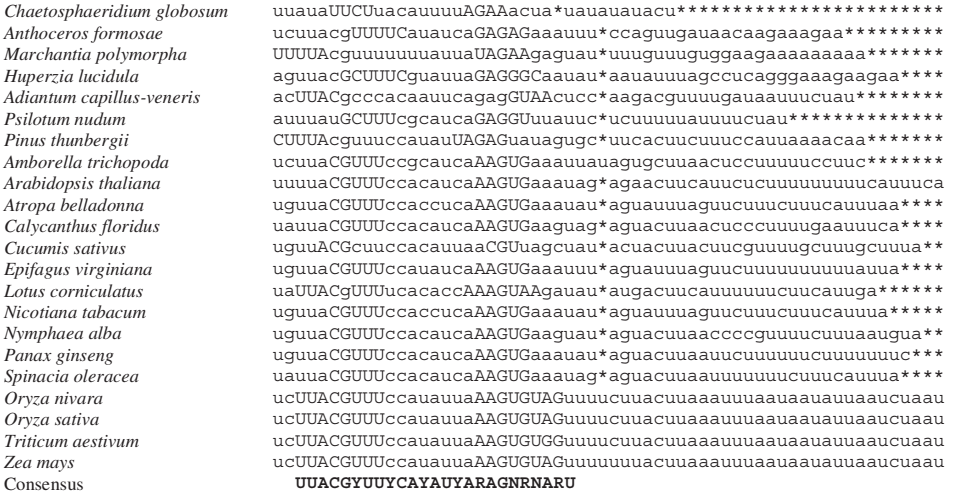

Fig. 3. Alignment of 5′-untranslated regions upstream of the *clpP* start codon. Helices are shown in capitals.

5.4. *Photosystem I*

In many chloroplasts, the 5′-untranslated regions of *psaA* gene (photosystem I P700 apoprotein A1) are found to possess a conservative motif adjacent upstream of the start codon AUG with the consensus

WGUURGYRRGUYUYUUY*UAUN*****
NUYGUCYGRARAGAGGAGRA*CUCR.

The motif contains a conservative helix near the putative ribosome-binding site. The corresponding alignment is shown in Fig. 4.

5.5. *Photosystem II*

In many chloroplasts, the 5′-untranslated regions of the *psbA* gene (photosystem II protein D1) and the *psbB* (photosystem II P680 chlorophyll A apoprotein) gene contain conserved motifs adjacent to the AUG start codon with the consensus

YUUGGGARYYYY**********NAAACYAAG

and

AAAGUNACRUAGU*GUCUAYUUNN**********NNNAAGGGGURUUU,

respectively. The first motif, upstream of *psbA* contains a conserved helix. These motifs were not found in 5′-untranslated regions of *M. viride* genes, nor in other algae.

Odontella sinensis	cuuaugagaguuucau*aaauu*****UUCgucUCCcaaaaGGAGAAaguca
Guillardia theta	auaaaguaagaguuuuuagauu*****gcugUCUCaaaagagGAGAaccuca
Porphyra purpurea	uagaaauaagcguuuu**gauu****ccuugUCUCaagagagGAGAaucuca
Nephroselmis olivacea	agccaggaagacuauuu*cauu****CCUCgugugaagaGAGGagaaucucg
Chaetosphaeridium globosum	uguuguuaaguauuuucuuagc*****CUCgUCUgaaaAGAgGAGaauuucg
Mesostigma viride	uagaggugaguuuuuuu*ugug****cCUCaUCUaaaaAGAgGAGaaucucc
Anthoceros formosae	uuguuggcggucuuuuc*caug****CCUCgucugaaagGAGGauaauaucg
Marchantia polymorpha	uguugguagguuuuucc*uaug****CCUCgucugaagaGAGGagaaccucg
Huperzia lucidula	ucuuggcggguuuuuuc*uaug****CCUCgucuggaaaGAGGagaaccucg
Adiantum capillus-veneris	uguugguagguuguugc*uauc****cCCUgCUCgaaGAGAGGagaguccca
Psilotum nudum	ugcuggcagguuguugc*uauu****CCUCgucucgagaGAGGagaaucuca
Pinus thunbergii	uauuggcagguuucuuauuauuuaagucccgUCCgaaaagaGGAgaa*uuca
Amborella trichopoda	ucuuggcgggucucuucguaug****uguugUCCggaaagaGGAgga*cuca
Arabidopsis thaliana	uguuggcggguuucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Atropa belladonna	uguuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Calycanthus floridus	uguuggcggguuucuuuguaug****uguugUCCggaaauaGGAgga*cuca
Cucumis sativus	uauuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Lotus corniculatus	uauuggcaggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Nicotiana tabacum	uguuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Nymphaea alba	uguuggcgggucucuucguaug****uguugUCCggaaagaGGAgga*cuca
Panax ginseng	uguuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Spinacia oleracea	uguuggcaggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuca
Oryza nivara	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Oryza sativa	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Triticum aestivum	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Zea mays	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Consensus	**WGUURGYRRGUYUYUUY*UAUN*****NUYGUCYGRARAGAGGAGRA*CUCR**

Fig. 4. Alignment of 5′-untranslated regions upstream of the *psaA* start codon. Helices are shown in capitals.

The GenBank annotation of the *psbA* gene of *A. trichopoda* probably misses a short N-terminal sequence, which might explain why the algorithm failed to find the corresponding motif. The *psbA* and *psbB* alignments are shown in Figs. 5 and 6.


```
Odontella sinensis          aGUAAAaacuugggaagaau*****************uaaaUUUAUaaaaauguuauu*
Guillardia theta            auagAUUAcuuggggaaguauauuacuauaaaauu****uaaacaUAAUauaaaaca****
Cyanidioschyzon merolae     aguacuuacUUGggaagagaaaaa******aucuuucuuaaaCGAcu**************
Porphyra purpurea           agauauuguUUGGgaagcaucuua**guuaaauccuaaaaaaUUGAuaaaaaauu*****
Gracilaria tenuistipitata   aaAUAUUUcuuGGGAaguauccuc******aauuaUCCUaaacuAAAUAUauaauauu**
Chlamydomonas reinhardtii   auauuuuacGGAGaaa*******************uuaaaaCUUUaaaaaaauuaacau
Nephroselmis olivacea       caaaaaagcUUGGgagaucacuu*************acaaaaUCAAuaaauucaccuacu
Chaetosphaeridium globosum  uuauaguUUGGguaaauuauuauaa********cuuuauaaaCCAAuauauacc******
Anthoceros formosae         uuauauGCUUGGgaacuuauuauua********uuucacaaaCCAAGUuuuacc******
Marchantia polymorpha       uuaaauAUUUGGgaaacucuuaauua*******uuuuaaaaaCCAAGUuuuacu******
Huperzia lucidula           uuauauGCUUGGggaacuuauuau****cacuuuauaaaaaaCUAAGU**uuuacc****
Adiantum capillus-veneris   uuacuagccUGGggaaucacuaau*****uccuuuucgaaaaCCAa**aaauuccc****
Psilotum nudum              uuuauugacUUGGgaacuuuagaa******uugcuuuagaaaCCAA**auauuauc****
Pinus thunbergii            gccuuauGCUUgggagccucugaugauu*******uuauaaaacgAAGU*ucugacc****
Arabidopsis thaliana        uuugugcgCUUGGgagucccugauuauu*******aaauaaaCCAAGgauuuuacc****
Atropa belladonna           cucgugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Calycanthus floridus        ucugugugCUUGGgagucccugaugauu*******caauaaaCCAAG*aucuuacc****
Cucumis sativus             uucaugugCUUGGgagucccugaugauuaauuauuaaauaaaCCAAG*auuuuacc****
Lotus corniculatus          uucgugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Nicotiana tabacum           cuagugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Nymphaea alba               ucugugugCUUGGgagcuccugaugauu*******gaauaaaCCAAG*aucuuacc****
Panax ginseng               uucgugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Spinacia oleracea           uuugugugCUUGGgagucccugaugauuaa**auuaaauaaaCCAAG*auuuuacc****
Oryza nivara                uacgugugCUUGGgaguccuugcaauuu*******gaauaaaCCAAG*aucuuacc****
Oryza sativa                uacgugugCUUGGgaguccuugcaauuu*******gaauaaaCCAAG*aucuuacc****
Triticum aestivum           uacguguGCUUGGggguccuugcaauuu*******gaauaaaCCAAGU*ucuuacc****
Zea mays                    uacgugugCUUGGgaguccuugcaauuu*******gaauaaaCCAAG*aucuuacc****
Consensus                          YUUGGGARYYYY******************NAAACYAAG
```

Fig. 5. Alignment of 5′-untranslated regions upstream of the *psbA* start codon. Helices are shown in capitals.

```
Porphyra purpurea           uuuguaagaaagucaacaaaguauguucuuauu******cauaggaggcauguaguca*
Nephroselmis olivacea       auagcaagaaagucauaccguuauaaaggucucgaau*agaccuagaaggaguauag**
Chaetosphaeridium globosum  acugcaagaaagucacaaaua*guuuguuuuuuuc*****uuaacaaagagguauuuac
Anthoceros formosae         aaugcaagaaauuuacguagu*gucuauucuucu*******ggauaaagggguaucuuc
Marchantia polymorpha       aaugcaaaaaaguuacauagc*gucuaauucucuu******ugagaaagggguauuuuu
Huperzia lucidula           aacguaagaaagucauaugau*gucuaccuaucuuuggguaaggggaaaggggacucaa
Adiantum capillus-veneris   auugcaagaaaguuacgcagugaucaguugucuccaauauucaagaaaggggguuuuuc*
Psilotum nudum              aacgcaagaaaguuacguagu*auugacuaaa**********aaaaagagguauuuaa
Pinus thunbergii            aaugugagaaaguuacauagu*gucuacuuuuuc*******cgauaaagggguguuugc
Amborella trichopoda        aaugcgauaaaguuacauagu*gucuauuuug**********augaagggguauuucc
Arabidopsis thaliana        aaugcaauaaaguuacauagu*gucuauuuuucgu******ugauaaagggguauuucc
Atropa belladonna           aaugcaauaaaguuacguaga*uuuaucuuuga**********uauaagggguauuucc
Calycanthus floridus        aaugcgauaaaguuacauagu*gucuauuuuuucu*****uugauaaagggguauuuuc
Cucumis sativus             agugcaauaaaguuacauagu*gucuauuuuuccu******ugauaaagggguauuucc
Lotus corniculatus          agugcaauaaagucacauagu*gucuauuuucugu*****ugauagaagaggguauucuc
Nicotiana tabacum           aaugcaauaaaguuacguagu*gucuauuuaucuu*****ugauauaagggguauuucc
Nymphaea alba               aaugugacaaaguuacauagu*gucuauuuuucuu******ugauagaagggguauuucc
Panax ginseng               aaugcaauaaaguuacguagu*gucuuuuuuucuuug*auauaaaauaagggguauuucc
Spinacia oleracea           aaugcaauaaaguuacauagu*gucauuuuucuu*******ugauaaagggguauuucc
Oryza nivara                gauaaaauaaagugacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Oryza sativa                gauaaaauaaaagugacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Triticum aestivum           gauaaaauaaagcgacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Zea mays                    gauaaaauaaagcgacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Consensus                        AAAGUNACRUAGU*GUCUAYUUNN**********NNNAAGGGGURUUU
```

Fig. 6. Alignment of 5′-untranslated regions upstream of the *psbB* start codon.

6. Discussion

Conserved motifs detected upstream of the *atpF*, *petB*, *clpP*, *psaA*, *psbA*, and *psbB* genes are likely to be involved in the regulation of translation.

The conserved region upstream of *atpF* contains an AG-rich motif typical for the ribosome-binding sites. This conservative region is considerably longer than typical binding sites. The presence of introns in the genes suggests that translation starts only after completion of splicing.

These *petB* genes do not have a typical ribosome-binding site but contain a conserved helix that might suggest posttranscriptional modification of the 5′-untranslated regions or binding of a transcription activator. In all plants, the *petB* gene contains introns.

The conserved regions in the 5′-untranslated regions of *clpP* and *psbA* genes were observed upstream of almost all their orthologs, even those lacking introns. The translational regulation of the *psbA* gene has been experimentally observed in *C. reinhardtii*, where transcription is constitutive, but translation is activated at light by a 47 kDa protein, forming a complex with other proteins and mRNA, but not interacting with mRNA directly.[3] The complex is inactivated in the dark.

The conserved nature of this region in plants and algae might suggest that the translation regulation machinery for the gene *psbA* preceded the emergence of introns.

Notably, the conserved RNA motifs in the transcripts of *clpP*, *petB*, *psaA* and *psbA* contain helices with conserved flanks likely interacting with protein mediator, which is typical for most regulatory systems.[5,6]

The long conserved motifs were found upstream of the *psaA* and *psbB* genes, which lack introns in all species containing the motifs. On the other hand, in chloroplasts of *A. capillus-veneris*, all studied 5′-untranslaled regions are considerably diverged. Thus, the motif was not found upstream of *petB*, while it was upstream of the other five cases. In the latter situation, however, site trees and species trees differed considerably in the node containing the name of the corresponding species.

Other intron-containing genes in the studied chloroplast genomes were not found to have conserved 5′-motifs, or their 5′-untranslated regions were too short or absent. Two such examples are discussed below.

In studied plants, the upstream regions of the *rbcL* genes encoding a ribulose 1,5-bisphosphate carboxylase/oxygenase subunit contain only a short conserved motif with the consensus ARGGAGGGACYT whose core is the ribosome-binding site. And we have no reason to assign a regulatory role to this motif, as the plant *rbcL* genes lack introns. On the other hand, the *rbcL* genes contain introns in chloroplasts of both algae *E. gracilis* and *C. reinhardtii*, and in the latter case, *rbcL* is regulated by mRNA-binding proteins.[3] This seeming discrepancy is not surprising, since in both algae the structure of 5'-untranslated region is completely different from that in studied plants.

A different situation is observed in the case of *ycf3* genes (photosystem I assembly protein Ycf3). It contains introns and a long 5'-untranslated region neither overlapping with other genes in plant chloroplasts, nor was it found to possess conserved motifs.

Acknowledgments

The authors are grateful to M.S. Gelfand for valuable discussion and to L.Y. Rusin for discussion and help. This study was partially supported by ISTC 2766.

References

1. Nickelsen J, Chloroplast RNA binding proteins, *Current Genet* **43**:392–399, 2003.
2. Zerges W, Translation in chloroplasts, *Biochimie* **82**:583–601, 2000.
3. Hauser CR, Gillham NW, Boynton JE, Translation regulation of chloroplast genes, *J Biol Chem* **271**:1486–1497, 1996.
4. Lyubetsky VA, Seliverstov AV, Note on cliques and alignments, *Inf Processes* **4**:241–246, 2004.
5. Vitreschak AG, *et al.*, Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet* **20**:44–50, 2004.
6. Seliverstov AV, *et al.*, Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria, *BMC Microbiol* **5**:54, 2005.

Vassily Lyubetsky obtained his M.Sc., Ph.D. and D.Sc. habilitation degrees in Math from Moscow State University and Russian Academy of Sciences, Russian Federation, in 1968, 1971, and 1991, respectively. His fields of academic expertise are mathematic logic, algebra, and theoretical computer science. Since 1995 he is a faculty professor of Moscow State University at the chair of mathematical logic and algorithmic theory. For over 10 years he is occupying permanent positions in Russian Academy of Sciences as a head of the Laboratory for mathematic methods and models in bioinformatics of the Institute for Information Transmission Problems, and as a principal researcher at the Federal State Enterprise, Institute of Strategic Stability, of the Ministry for Industry and Science of Russian Federation.

He has authored and co-authored over 150 scientific publications, four monographs and over 50 international conference publications.

Alexander Seliverstov received his Diploma in Applied Mathematics from State Classic Academy, Moscow, Russian Federation, in 1999. He is currently in the Laboratory for mathematic methods and models in bioinformatics of the Institute for Information Transmission Problems, Russian Academy of Science.