Laboratory of mathematical methods and models in bioinformatics

Institute for Information Transmission
Problems, Russian Academy of Sciences

lyubetsk@iitp.ru,
http://lab6.iitp.ru/en/

Directions 2-4 (thirteen tasks and corresponding results):

Regulation (itself)

and

evolution of cellular processes

1) Transcription regulation based on protein – DNA interaction. Evolution =Ev.

2) Allied to 1: Searching for conserved and highly labile promoters. Ev.

3) Translation regulation of gene expression based on RBS blocking by a secondary structure. Ev.

4) Transcription regulation of gene expression based on terminator and antiterminator secondary structures competition. Different regulation mechanisms based on it. Ev.

5) Transcription and translation regulation with triplexes and pseudoknots. Ev.

6) On the whole evolution of regulations based on the secondary structure dynamics (original modifications made to the parsimony functional)

7) **The role of RNA sites in pathogen invasion**: *Toxoplasma gondii* (Apicomplexa) switches on plastid genes in the host cell

8) **Transcription regulation in nucleus** in the Piroplasmida

9) **The role of RNA secondary structures in pathogen invasion**: *Brucella* (alpha-proteo) competes for the macrophage host cell resources (metal cations) using the RNA secondary structure

10) Secondary structures are often used in **bacteria in defense against their phages**

11) **Secondary structures in DNA (crest-hairpins)**

12-13) **RNA polymerase competition as an important transcription-related process**:

the competition drives **physiological responses** (e.g., to heat shock); and/or **regulation responses** (physiological or tissue-specific through the interaction of nuclear and plastid genomes)

Below there are some our results about all
13 tasks listed above

1) **Transcription regulation based on protein – DNA interaction**

Original results on the evolution of this regulation for genes *proA* (gamma-glutamyl phosphate reductase) and *proB* (gamma-glutamyl kinase) were already shown (the first presentation = Direction 1).

In such studies the regulation itself is to be found first.

Now is another example: original results on regulation of nitrogen metabolism

NtcA- or NtcB-regulated genes in cyanobacteria are listed according to their product function (genes with regulation in all species are in blue):

Nitrogen metabolism:

- nitrate/nitrite: ***narB, narK/nrtP***, ***nirA***, *nirB*
- glutamate/glutamine: ***glnA, glnN, glnB***, ***gifA, gifB***, *gltS, hisH, cobA, cobB*
- NAD-dependent isocitrate dehydrogenase: ***icd***
- arginase/agmatinase: *speB*
- urease accessory proteins: *ureE, ureG*.

Heterocyst differentiation protein: *hetC*.

Transcription factors: ***ntcA, ntcB***.

Carbon fixation: *ccmK, rbcL, rpe*

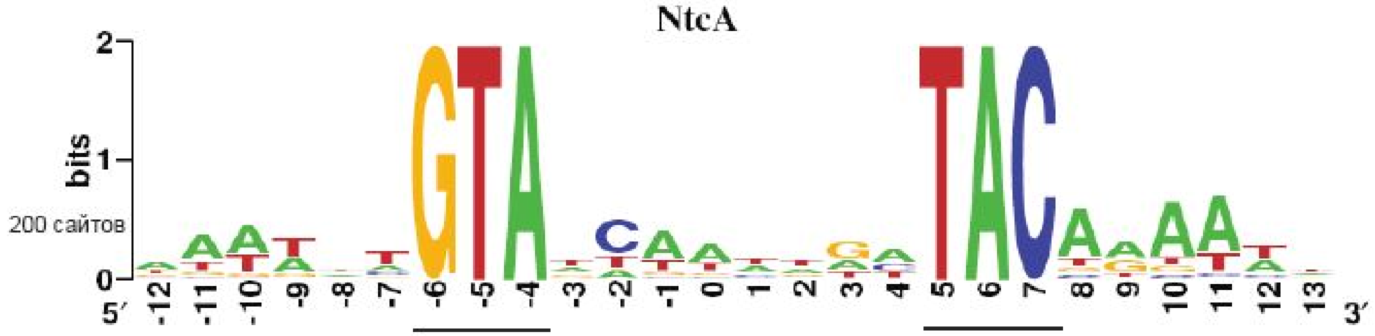
Transporters: *urtA, urtB, urtC, urtD, urtE*, ***amtB/amt1***, *tauA, tauB, tauC*, ***nrtA,*** *cmpA, cynA, cynB, cynD, cynS, devB, futC,*

Porin: ***som***; **Pigments and photosynthesis:** *psaI, psaB, psaL, psaF, psbA3, psbZ, psbB, psbO, psbW, psbE, psb27, isiB, isiA, pcbD, pcbA, ndhB, petH*, ***petF/fdx, apcF***, *apcE, apcA, cpcB, trxA, trxM*

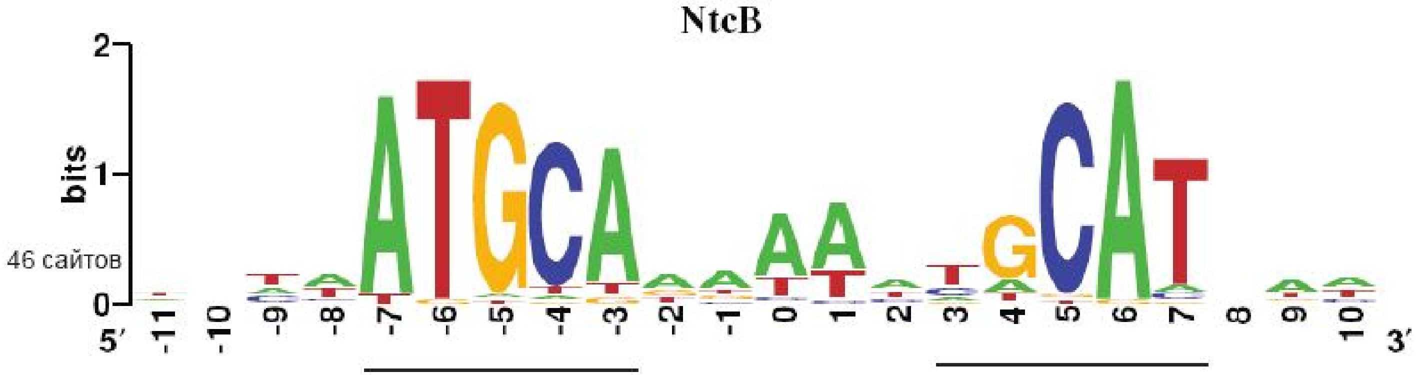
Metalloproteins: *hypA2, hypB, moaA, moaC, moeA*

Others: *metG, thrC, mutS, rnc, xisA, gor, aarF, rpoD*

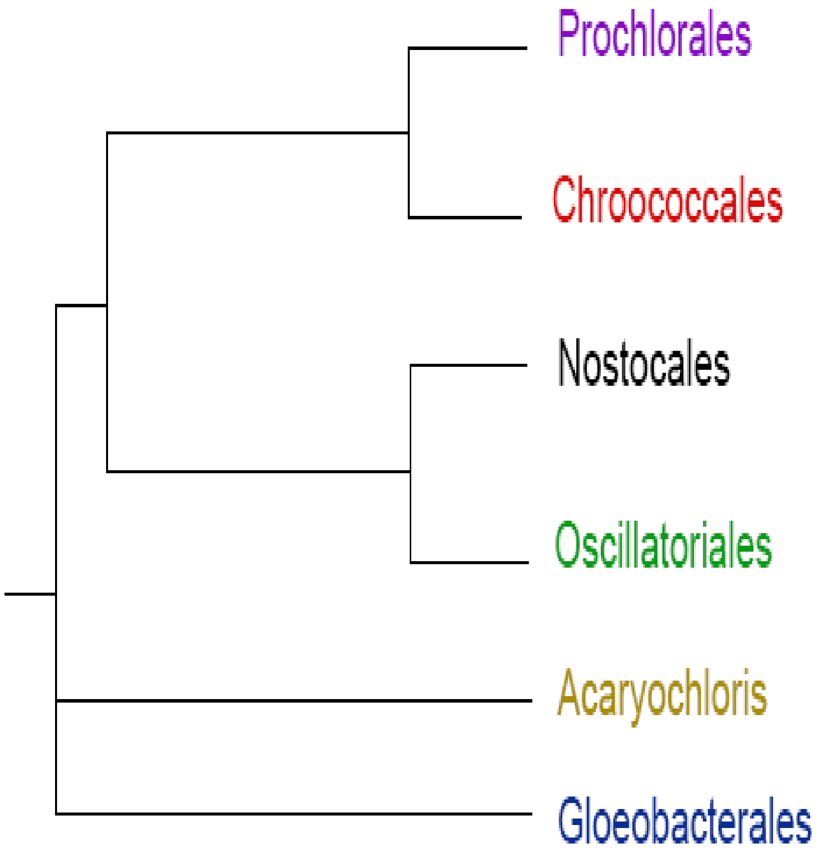
Nucleotide frequency profile of the NtcA binding motifs predicted in cyanobacteria. Underlined are conserved positions in the obtained consensus:



Nucleotide frequency profile of the NtcB binding motifs:



High taxonomy of Cyanobacteria



Site and gene are together				“there is gene and no site”	
–	+*glnA*	–	–	+*icd*	+*narB*
+*glnN*	+*glnA*	+*gifA*	+*gifB*	+*icd*	*narB*
–	+*glnA*	+*gifA*	+*gifB*	+*icd*	*narB*
–	+*glnA*	+*gifA*	+*gifB*	+*icd*	*narB*
+*glnN*	+*glnA*	+*gifA*	+*gifB*	*icd*	+*narB*
+*glnN*	+*glnA*	–	–	*icd*	*narB*

Site and gene present: «**+*gene***»; only gene present «***gene***», both absent «–»

The above results are obtained with two original programs:

1. “Twobox”: finding sites with complex structure
2. “Treealign”: tree-guided alignment

2) Technically related task is **searching for widely conserved and labile promoters**.

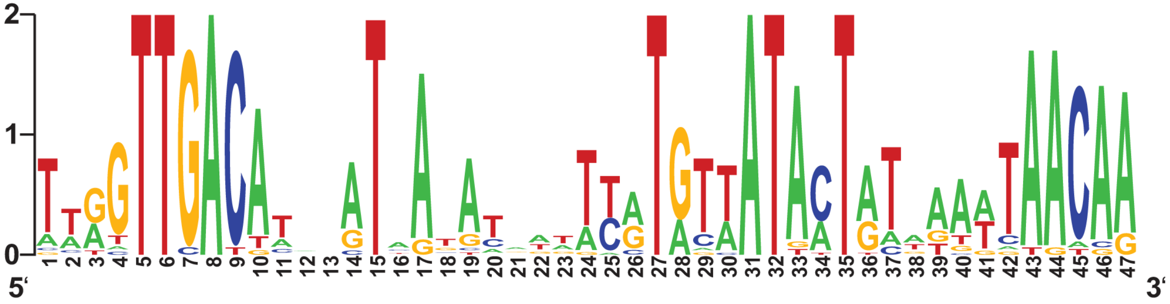
E.g., in plastids of plants and algae including secondary and tertiary endosymbionts

While the amino acid sequences of plastid-encoded proteins are highly conserved, the noncoding gene regions substantially vary even in closely related species, suggesting an important role in the regulation of gene expression

Multiple alignment keeping one of the candidate promoters in each region:

Arabidopsis thaliana	TTGGTTGACATGGCT-ATATAAGTCATGTTATACTGTTTCATAACAA	-74
Spinacia oleracea	TTGGTTGACACGGG-CATATAAGGCATGTTATACTGTTGAATAACAA	-79
Cycas taitungensis	TCGATTCACGATA--TATATAAGTCATACTATACTGTTAAATAACAA	-57
Cryptomeria japonica	TTGGTTGACATACA-GATATGTCTCATATTATACTGTTGAATAACAA	-55
Pinus koraiensis	TTGGTTGACATTGAT-ACATGGATCATATTATACTGTAAAATAACAA	-49
Pinus thunbergii	TTGGTTGACATTGAT-ACATGGATCATATTATACTGTAAAATAACAA	-49
Welwitschia mirabilis	ATAGTTGACTTTAAT-AAACCATTTCTGTTATACTGTTAAAATAACA	-48
Adiantum capillus-veneris	TTGGTTGACACGGAT-AGGTTTTT-GTGATATGCTACATAGTAACAG	-52
Angiopteris evecta	TAAGTTGACATCAAT-AGATAAGTTGTGTTATACTATGAAGTAACAA	-66
Psilotum nudum	TAAGTTGACATATAT-GGAAAGATCATGTTATACTTCAAATCAACAG	-50
Huperzia lucidula	TGGGTTGACACAAA-AAGAAAGATTGTGTAATATTATGGAATAACAA	-52
Aneura mirabilis	GATGTTGACATAC-TAATGGGATATGTGTAATAATATGGGTTAACAG	-51
Marchantia polymorpha	TTAGTTGACATAA-TCATATGTTATGTGTAATACTATAAGTTAACAA	-50
Physcomitrella patens	TCAGTTGACATAA-TAATACATTTTGTGTAATACTATAAATTAACAA	-50
Chara vulgaris	CTAGTTGACATTT-TTATACTTTACATACTATAATATCTAATAACAA	-118
Chaetosphaeridium globosum	TAGGTTGACATTAGTTATACGT-TTGTGCAATACTAAATATTAACAA	-54
Staurastrum punctulatum	AAGGTTGACAGCT-TAAGGTTAAT-ATGTAATAATATAATTTAACAA	-56
Zygnema circumcarinatum	TTAGTTGACAACAG-CATTAACTATCTGTAATAATATAAATTAACAA	-55
Mesostigma viride	TTATTTGACAAATA-AACATCATTT-TGGCATAATAATAATCAACAA	-50
Bigelowiella natans	TTTTTTGATTAATATAA-ATTAATTA-GTTATAATATTATAGAGTAA	-133
Cyanophora paradoxa	AAGCTTGACAAAT-TAGACCATTAA-TATTATTATAAGATTTAACGA	-58

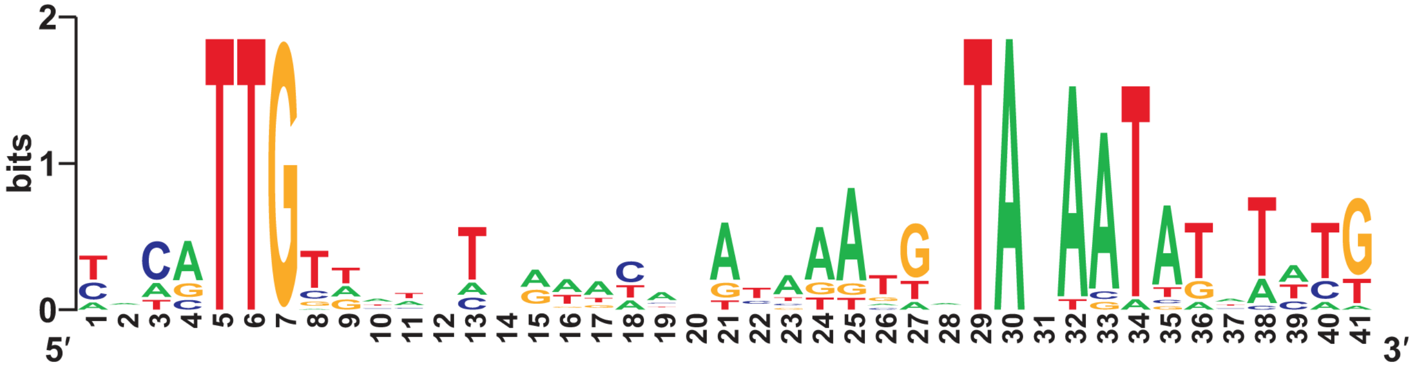
psbA



Multiple alignment keeping one of the candidate promoters in each region:

Arabidopsis thaliana	CCCATTGCATATTGGTACTTATCGGATATAGAATAGATCCG	-171
Spinacia oleracea	CCCATTGCGTATTGCTACTTATCGAGTATAGAATAGATTTGT	-176
Cycas taitungensis	CACATTGTGCATTGGTACACATAAATGATAAAATATTTACG	-171
Cryptomeria japonica	CACATTGTATATTGATACATATAAATGATAAAATATATCCG	-143
Pinus koraiensis	TACATTGTGTATTGGTACATACAAACGATAAAATATCTTTG	-194
Pinus thunbergii	TACATTGTGTATTGGTACATACAAACGATAAAATATCTTTG	-181
Welwitschia mirabilis	TCACTTGGACCCAAGCCTCC-CTTTTTCTACTATATATAAT	-272
Adiantum capillus-veneris	TACGTTGTTACATGGGGAATGAAAATGCTAAAATATTCACG	-292
Angiopteris evecta	CACATTGTTATGCAAAATCTGTGAATGCTAGAATATCTATG	-182
Psilotum nudum	CACATTGTTGCACAAATTGTGCAAATGTTAAAATATCTCTG	-179
Huperzia lucidula	TCCATTGCGATGTTAAACGCATGGATGTTAAACTATTTCTG	-188
Chara vulgaris	ATTCTTGGACGGTCAAGTTATAAAATGGTATAATATATAAA	-180
Chaetosphaeridium globosum	AATATTGATATATAAGACAAATTAATGTTAAAATAATAATT	-162
Staurastrum punctulatum	TGTGTTGTTCTGAT-AGAAAAGAAATGATACAATCAAAATG	-191
Zygnema circumcarinatum	TTAGTTGTAATCTC-ATAAGAGATAGAGTACAATGGAATTG	-160
Chlorokybus atmophyticus	AGACTTGTTATCCTAATTAG-TTTGGTATATAGTTTGTTTT	-267
Mesostigma viride	TTAGTTGTTATAATTATACGTTAATAATTATAAATGTATTT	-90

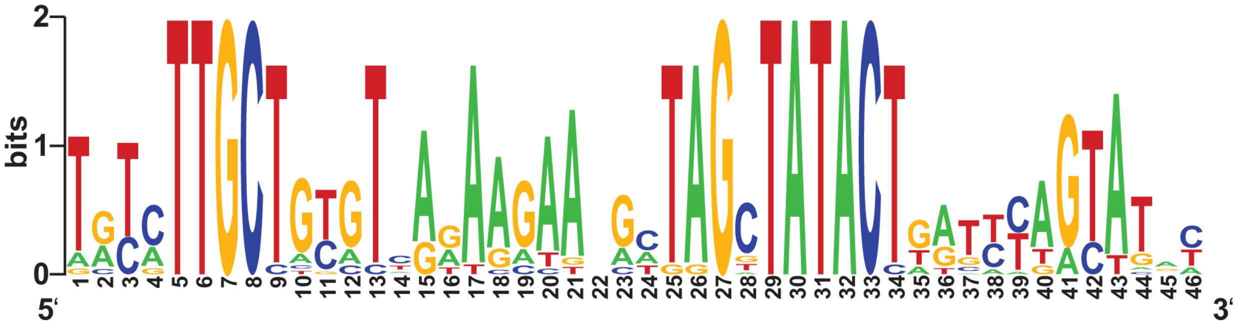
psbB



Multiple alignment keeping one of the candidate promoters in each region:

Arabidopsis thaliana	TGCGTTGCTGTGTCAGAAGAAGGATAGCTATACTGATTCGGTAGAC	-120
Spinacia oleracea	TGCCTTGCTGTGTCAGAAGAAGGATAGCTATACTGATTCGGTATAC	-145
Cycas taitungensis	TGTATTGCTGTGTCAGAGGAAGGCTAGCTATACCGGTCCAATATAC	-136
Cryptomeria japonica	TATATTGCTATGTTAGAAGCAGGCTAGCTATACTTAGTATACTTCA	-132
Pinus koraiensis	TGTATTGCTGTGTCAGAAGAAAGCTAGCTATACTGGTCCAGTTATA	-143
Pinus thunbergii	TGTATTGCTGTGTCAGAAGAAAGCTAGCTATACTGGTCCAGTAGAC	-140
Welwitschia mirabilis	TATATTGCTGTGTCATAAAAAAGTTGGTTATACTGGTCCAGTATTA	-26
Adiantum capillus-veneris	AACCTTGCCGCATTGTACGTGAAATAGCTATACTGACCCAGCATAT	-186
Angiopteris evecta	TATCTTGCTGCGTCAAAAGAAGGCTAGCTATACTGTTCTAGTATAT	-137
Psilotum nudum	TCTCTTGCTGTATAGGAAAAAAGATAGCTATACTGATACTATATAT	-122
Huperzia lucidula	TGTCTTGCTGCGTCAGAGGAACACTAGCTATACTAGTCTAGTATAC	-129
Anthoceros formosae	TACCTTGCTTCGTTGAAAGAACGCTAGCTATACTTATTTAGTATGC	-138
Marchantia polymorpha	TATCTTGCTGCGTAAAAAGAACATTAGCTATACTAAGTTAGTATGC	-127
Physcomitrella patens	TGTCTTGCTACGCTAAAACAACCCTAGATATACTTATTTAGTATGC	-140
Chaetosphaeridium globosum	TCTCTTGCTGGCTGGTTAGTTAAATAGGTATACTATAATTGTACGT	-114
Staurastrum punctulatum	GGCCTTGCTGTCTTAAAGAAATCTTAGTTATACTTACTTAGCATGT	-149
Zygnema circumcarinatum	AGTGTTGCTCTATAAAAACAATGTGAGGTATACTTAGTTAGCAGCT	-117

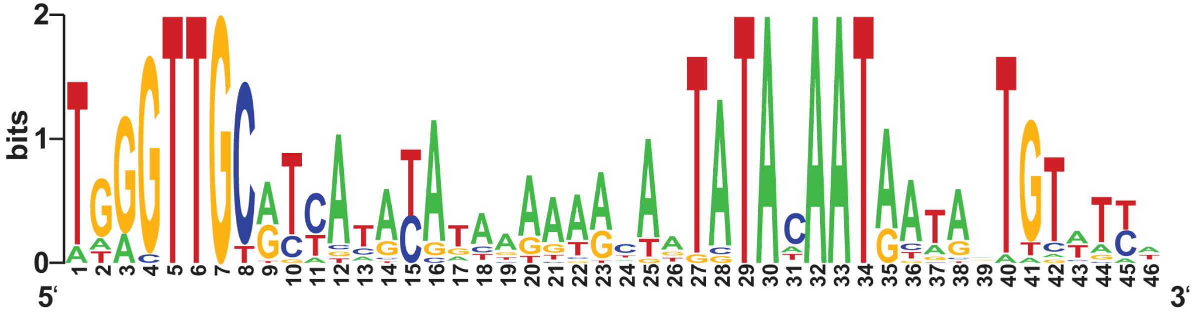
psbE



Multiple alignment keeping one of the candidate promoters in each region:

Arabidopsis thaliana	TAGGTTGCGCTATACATATGAAAGAATATACAATAATGATGTATTT	-172
Spinacia oleracea	TGGGTTGCGCCATATATATGAAAGAGTATACAATAATGATGTATTT	-171
Cycas taitungensis	AGGGTTGCGCCATACATAAAGAACATTATACAATAATAGTGTATTT	-151
Cryptomeria japonica	TGGGTTGCGTCATACATACATAACATGATACAATATCACTTGAAAG	-157
Pinus koraiensis	TGGGTTGCGTCATACATAAAGAACATTATACAATGAGAGTGTATCT	-131
Pinus thunbergii	TGGGTTGCGTCATACATAAAGAACACTATACAATGAGAGTGTATCT	-122
Welwitschia mirabilis	TGGGTTGCATTATATGGAAAAAACAATCTAAAATGATAGTGTATTT	-131
Adiantum capillus-veneris	TTAGTTGCACCCCGCATCGGACGCGGTATAAAATAATAATGTTCCA	-152
Angiopteris evecta	TGGGTTGCATTATACAGAAAATAATTTATAGAATACTAGTGTCTCA	-143
Psilotum nudum	TGGGTTGCATCATATAGCAACTGCAATATAAAATAATAGTGTTTCC	-135
Huperzia lucidula	TGGGTTGCATCACGTATCAAAAGCAATATACAATGATAATGTTTTA	-145
Anthoceros formosae	TAGGTTGCATCATATACTAGAAATAATATACAATAGTAATGTTTTA	-160
Aneura mirabilis	TGGGTTGCATTACGTCGGATAAGCAATATACAATAATGATGTTTCA	-143
Marchantia polymorpha	TAGGTTGCATTACATATAAAAAACAATATACAATAATAATGTTTTA	-119
Physcomitrella patens	TGAGTTGCATCAAATGTAGAAAATAATATACAATAATACTGTTTTG	-138
Chara vulgaris	TGGCTTGTGTAGAGTAAATATTTATATATATAATATACGTACCGCC	-97
Chaetosphaeridium globosum	TTAGTTGCGTCATCTATTCAAGAATGTGTATAATACAATATAGAAA	-149
Staurastrum punctulatum	TTAGTTGTTTTAATCAATGTATGTAGT-TACAATAAATTTGTAATA	-214
Zygnema circumcarinatum	AGGGTTGCAGATGATAAAAAA-GTAATATATAATGAAGTTGCTGCT	-163

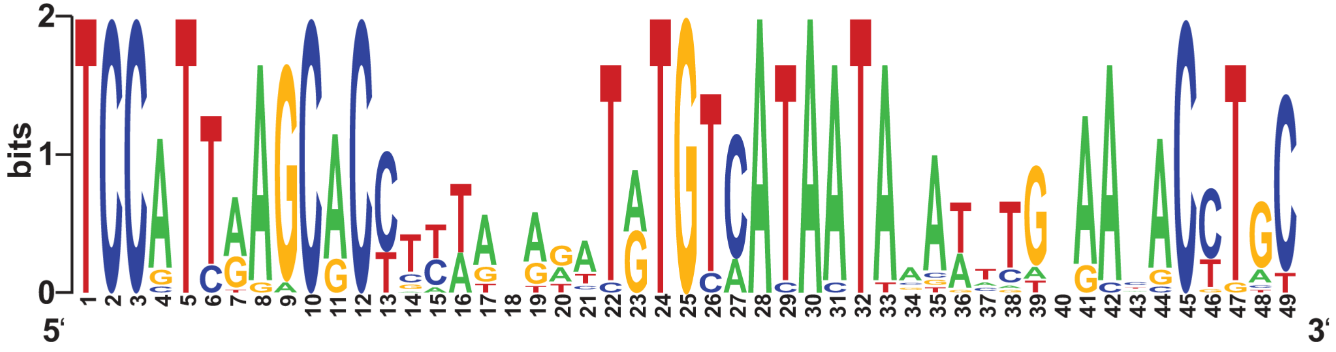
rbcL



Multiple alignment keeping one of the candidate promoters in each region:

Arabidopsis thaliana	TCCGTTGAGCACCCT-ATGGATATGTCATAATAGATCCG-AACACTTGC	-179
Spinacia oleracea	TCCGTTGAGCGCCAC-ACGTCTATGTCATAATAGATCCG-AACACTTGC	-171
Cycas taitungensis	TCCATTGAGCACCTC-AGGGATATGTCATAATAAATTTG-AACACCTGC	-147
Cryptomeria japonica	TCCATTAAGCACCTA-TCAGATATGTCATAATAAATATGAACACCTGTC	-133
Pinus koraiensis	TCCATTGAGCACCTC-GAAGATATGTCATAATAAAACTG-AACACCTGC	-149
Pinus thunbergii	TCCATTGAGCACCTCAAAAGATATGTCATAATAGAATTG-AACACCTGC	-149
Welwitschia mirabilis	TCCATTGAGCGCCTCTTGTATTATGTCATAATAAAAAGGAACACCTGC	-146
Adiantum capillus-veneris	TCCATCAGGCGCCGCT-AAGCCGTGTAATAATACCACCG-AAAGCCTAT	-154
Angiopteris evecta	TCCATTAAGCACTTTT-TGATTGTGTAATAATAAAATTG-AATGCCTGC	-143
Psilotum nudum	TCCATTAAGCACTTC-GATATTGTGTAATAATAAGTTTT-AATACCTGC	-138
Huperzia lucidula	TCCATTAAGCACCTTT-GATATGTGTAACAATAATTTTG-AATACCTGC	-144
Anthoceros formosae	TCCATTAAGCACCTTT-GAGATGTGTCATAATAAAAATG-AATACTTGC	-146
Marchantia polymorpha	TCCATTAAGCACCTT-AAAATTGTGTCATAATAAATTTG-AAGACCTGC	-140
Physcomitrella patens	TCCATTAAGCACCTT-AAAGATGTGTCATAATAAATTTG-AATACCTGC	-152
Chara vulgaris	TCCATTAAGCGCTCT-ATATATATGCCATACTACAGGTATGAAA-GTCT	-190
Chaetosphaeridium globosum	TCCATCAAGCAC-CTAAAAAATGTGTCATAATTTATTAG-AACACTTAC	-145
Staurastrum punctulatum	TCCCTTTAGCACT-AAAAAAATATGCCATAATATAAATA-GAAACCTAC	-226
Zygnema circumcarinatum	TCCATCAAACACTGT-GTGTGTGTGTCATAATACATTTTAGA-ACCTGC	-148

psaA



The first result here:

Lack of conservation of bacterial type promoters in plastids of Streptophyta.

Namely,

we found widely conserved PEP-promoters

ONLY for plastid genes ***psaA, psbA, psbB, psbE, rbcL***

The opposite case is evolutionary **labile promoters**

What is known about lability? Little:

there is ample published research on the promoter comparisons within small lineages, largely the studies of the promoters and their transcription factors in gamma- and alpha-proteobacteria [Collado-Vides, *J Bacteriol.* 2009].

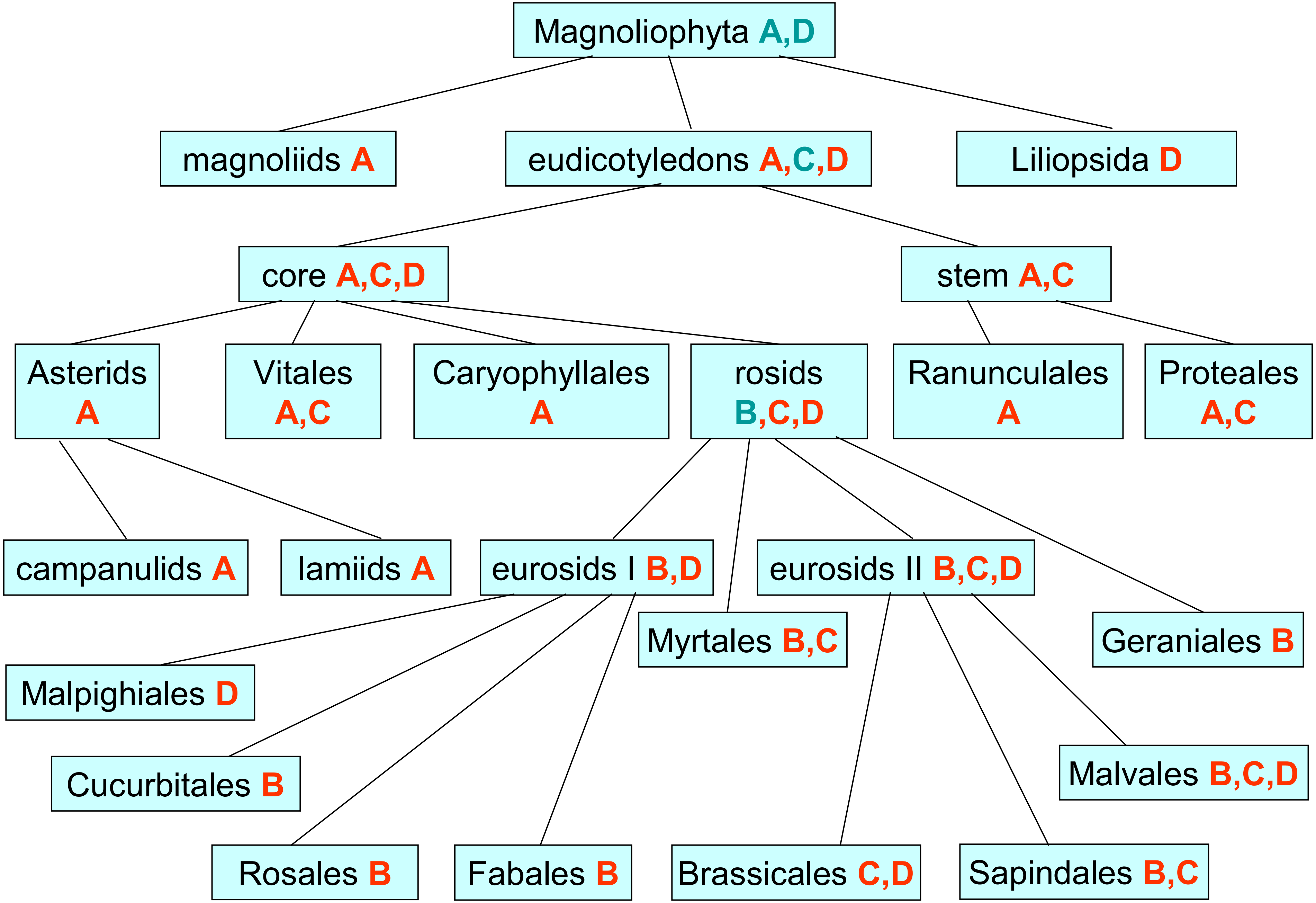
Further, some pairs of closely related species were shown to possess largely diverged promoters [Swiatecka-Hagenbruch, *Mol Genet Genomics* 2007; Hoffer, *Plant Physiol,* 1997]

We report on evolutionary **labile** PEP-promoters for some genes in **narrow lineages**, e.g. for the *ndhF* gene in **dicotyledonous angiosperm plants**.

The second result:

for *ndhF* we described **four different promoter types**, which are likely to have replaced each other during evolution

Suggested evolution of *ndhF* promoters in flowering plants



Magnoliophyta **A,D**

magnoliids **A**

eudicotyledons **A,C,D**

Liliopsida **D**

core **A,C,D**

stem **A,C**

Asterids **A**

Vitales **A,C**

Caryophyllales **A**

rosids **B,C,D**

Ranunculales **A**

Proteales **A,C**

campanulids **A**

lamiids **A**

eurosids I **B,D**

eurosids II **B,C,D**

Myrtales **B,C**

Geraniales **B**

Malpighiales **D**

Cucurbitales **B**

Malvales **B,C,D**

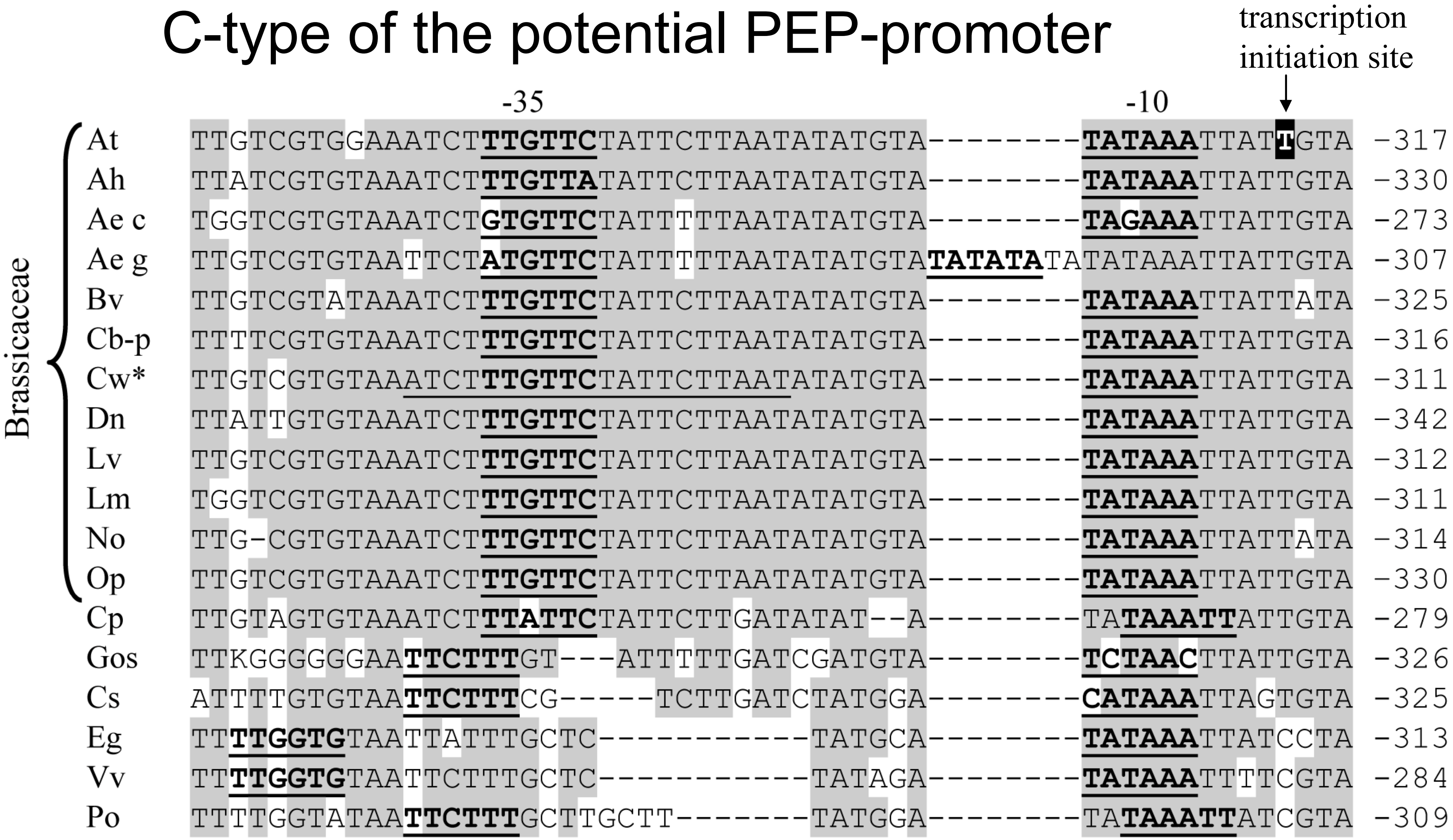
Rosales **B**

Fabales **B**

Brassicales **C,D**

Sapindales **B,C**

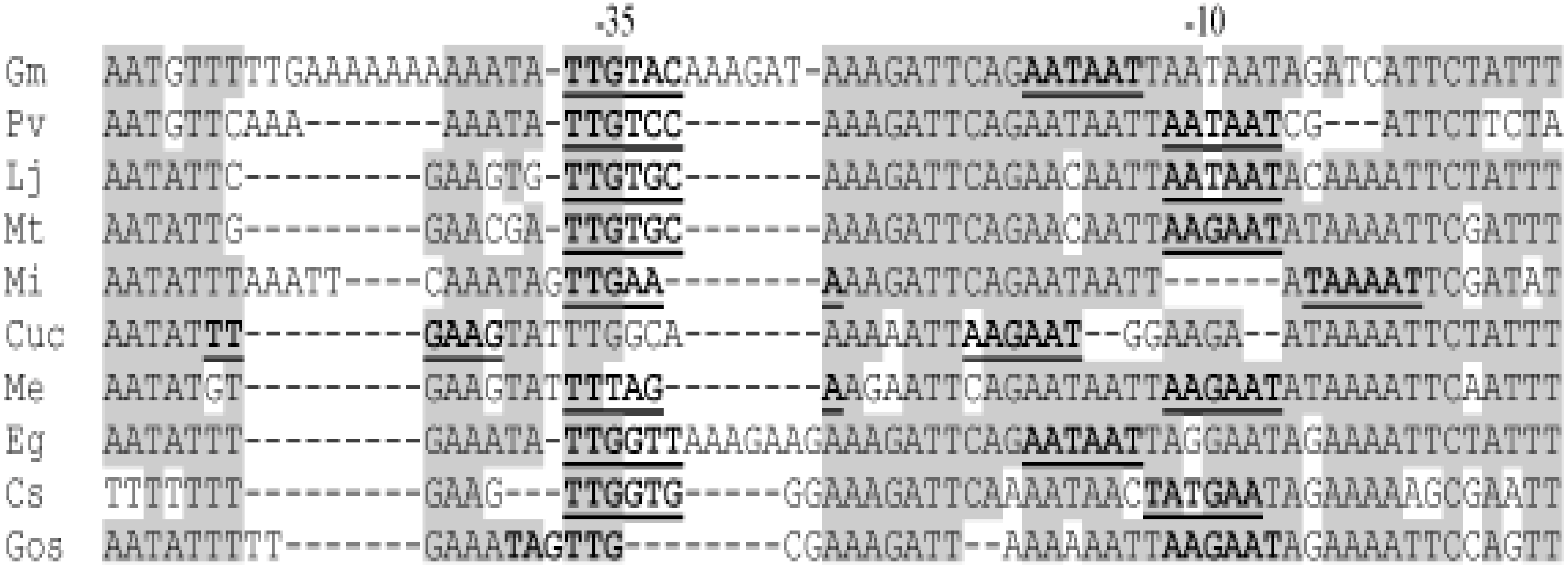
C-type of the potential PEP-promoter



```
At     TTGTCGTGGAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTGTA  -317
Ah     TTATCGTGTAAATCTTTGTTATATTCTTAATATATGTA--------TATAAATTATTGTA  -330
Ae c   TGGTCGTGTAAATCTGTGTTCTATTTTTAATATATGTA--------TAGAAATTATTGTA  -273
Ae g   TTGTCGTGTAATTCTATGTTCTATTTTTAATATATGTATATATATATATAAATTATTGTA  -307
Bv     TTGTCGTATAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTATA  -325
Cb-p   TTTTCGTGTAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTGTA  -316
Cw*    TTGTCGTGTAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTGTA  -311
Dn     TTATTGTGTAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTGTA  -342
Lv     TTGTCGTGTAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTGTA  -312
Lm     TGGTCGTGTAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTGTA  -311
No     TTG-CGTGTAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTATA  -314
Op     TTGTCGTGTAAATCTTTGTTCTATTCTTAATATATGTA--------TATAAATTATTGTA  -330
Cp     TTGTAGTGTAAATCTTTATTCTATTCTTGATATAT--A--------TATAAATTATTGTA  -279
Gos    TTKGGGGGGAATTCTTTGT---ATTTTTGATCGATGTA--------TCTAACTTATTGTA  -326
Cs     ATTTTGTGTAATTCTTTCG-----TCTTGATCTATGGA--------CATAAATTAGTGTA  -325
Eg     TTTTGGTGTAATTATTTGCTC-----------TATGCA--------TATAAATTATCCTA  -313
Vv     TTTTGGTGTAATTCTTTGCTC-----------TATAGA--------TATAAATTTTCGTA  -284
Po     TTTTGGTATAATTCTTTGCTTGCTT-------TATGGA--------TATAAATTATCGTA  -309
```

Brassicaceae and related groups: At = *Arabidopsis thaliana*, Ah = *Arabis hirsute*, Ae c = *Aethionema cordifolium*, Ae g = *Aethionema grandiflorum*, Bv = *Barbarea verna*, Cb-p = *Capsella bursa-pastoris*, Cw* = *Crucihimalaya wallichii* (20bp tandem insertion of the underlined region is omitted), Dn = *Draba nemorosa*, Lv = *Lepidium virginicum*, Lm = *Lobularia maritime*, No = *Nasturtium officinale*, Op = *Olimarabidopsis pumila;* Cp = *Carica papaya* (Brassicales), Gos = *Gossypium spp.* (*G. barbadense*, K=T; *G. hirsutum,* K=G) (eurosids II), Cs = *Citrus sinensis* (eurosids II), Eg = *Eucalyptus globulus* (rosids), Vv = *Vitis vinifera* (core eudicotyledons), Po = *Platanus occidentalis* (eudicotyledons).

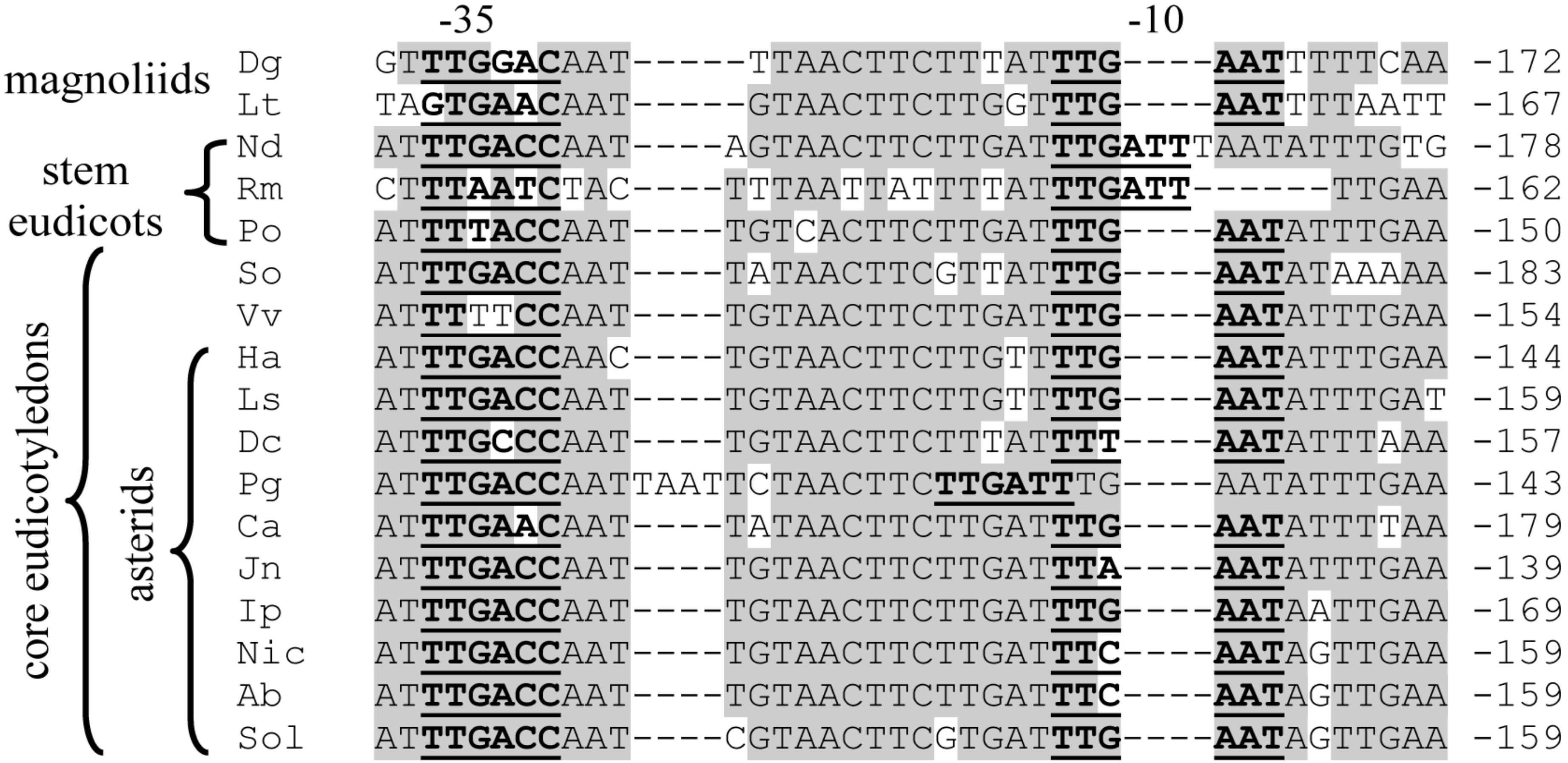
B-type of the potential PEP-promoter



```
                                 -35                        -10
Gm     AATGTTTTTGAAAAAAAAAAATA-TTGTACAAAGAT-AAAGATTCAGAATAATTAATAATAGATCATTCTATTT
Pv     AATGTTCAAA-------AAATA-TTGTCC-------AAAGATTCAGAATAATTAATAATCG---ATTCTTCTA
Lj     AATATTC---------GAAGTG-TTGTGC-------AAAGATTCAGAACAATTAATAATACAAAATTCTATTT
Mt     AATATTG---------GAACGA-TTGTGC-------AAAGATTCAGAACAATTAAGAATATAAAATTCGATTT
Mi     AATATTTAAATT----CAAATAGTTGAA--------AAAGATTCAGAATAATT------ATAAAATTCGATAT
Cuc    AATATTT---------GAAGTATTTGGCA-------AAAAATTAAGAAT--GGAAGA--ATAAAATTCTATTT
Me     AATATGT---------GAAGTATTTTAG--------AAGAATTCAGAATAATTAAGAATATAAAATTCAATTT
Eg     AATATTT---------GAAATA-TTGGTTAAAGAAGAAAGATTCAGAATAATTAGGAATAGAAAATTCTATTT
Cs     TTTTTTT---------GAAG---TTGGTG-----GGAAAGATTCAAAATAACTATGAATAGAAAAAGCGAATT
Gos    AATATTTTT-------GAAATAGTTG--------CGAAAGATT--AAAAAATTAAGAATAGAAAATTCCAGTT
```

eurosids I: Fabaceae/Papilionoideae: Mt = *Medicago truncatula*, Gm = *Glycine max*, Lj = *Lotus japonicus*, Pv = *Phaseolus vulgaris*;
other **eurosids I**: Mi = *Morus indica*, Cuc = *Cucumis sativus*, Me = *Manihot esculenta*;
eurosids II: Cit = *Citrus sinensis*, Gos = *Gossypium* spp.;
other **rosids**: Eg = *Eucalyptus globulus* (Myrtales).

A-type of the potential PEP-promoter



```
            -35                                   -10
Dg   GTTTGGACAAT-----TTAACTTCTTTATTTG----AATTTTTCAA -172
Lt   TAGTGAACAAT-----GTAACTTCTTGGTTTG----AATTTTAATT -167
Nd   ATTTGACCAAT----AGTAACTTCTTGATTTGATTTAATATTTGTG -178
Rm   CTTTAATCTAC----TTTAATTATTTTATTTGATT------TTGAA -162
Po   ATTTTACCAAT----TGTCACTTCTTGATTTG----AATATTTGAA -150
So   ATTTGACCAAT----TATAACTTCGTTATTTG----AATATAAAAA -183
Vv   ATTTTTCCAAT----TGTAACTTCTTGATTTG----AATATTTGAA -154
Ha   ATTTGACCAAC----TGTAACTTCTTGTTTTG----AATATTTGAA -144
Ls   ATTTGACCAAT----TGTAACTTCTTGTTTTG----AATATTTGAT -159
Dc   ATTTGCCCAAT----TGTAACTTCTTTATTTT----AATATTTAAA -157
Pg   ATTTGACCAATTAATTCTAACTTCTTGATTTG----AATATTTGAA -143
Ca   ATTTGAACAAT----TATAACTTCTTGATTTG----AATATTTTAA -179
Jn   ATTTGACCAAT----TGTAACTTCTTGATTTA----AATATTTGAA -139
Ip   ATTTGACCAAT----TGTAACTTCTTGATTTG----AATAATTGAA -169
Nic  ATTTGACCAAT----TGTAACTTCTTGATTTC----AATAGTTGAA -159
Ab   ATTTGACCAAT----TGTAACTTCTTGATTTC----AATAGTTGAA -159
Sol  ATTTGACCAAT----CGTAACTTCGTGATTTG----AATAGTTGAA -159
```

magnoliids: Dg = *Drimys granadensis,* Lt = *Liriodendron tulipifera*;
eudicotyledons: Nd = *Nandina domestica*, Rm = *Ranunculus macranthus*, Po = *Platanus occidentalis,* So = *Spinacia oleracea*, Vv = *Vitis vinifera*, Ha = *Helianthus annuus*, Ls = *Lactuca sativa*, Dc = *Daucus carota*, Pg = *Panax ginseng,* Ca = *Coffea arabica*, Jn = *Jasminum nudiflorum*, Ip = *Ipomoea purpurea*, Ab = *Atropa belladonna*, Nic = *Nicotiana spp.* (*N. tabacum, N. tomentosiformis*, *N. sylvestris*),
Sol = *Solanum spp.* (*S. bulbocastanum, S. lycopersicum, S. tuberosum*).

D-type of the potential PEP-promoter

Lm	CGCAA**TTGGAA**CTTTTTGAATTATTTAA**AATAAT**GGATG	-159
De	TTAGAT**TTGAAG**CCAATTAATAAAATTC**CATAAT**GAGTC	-386
Aco	ATAAAT**TTGATA**ATAACTTAACTCCAAA**TAAACT**AACTA	-267
Pop*	CGAGC**TTGAAG**GAGTTAATTCAATGTAA**TATTTT**TTTAG	-493/-552
Gos	ATAA**TTGATA**TTTCAGAAAAATATTAC**TATAAT**GAAAA	-254
At**	ATCTT**TTGACA**GTAACTTAGTAATATTT**TTTAAT**CATTT	-193

Liliopsida: Lm = *Lemna minor*, De = *Dioscorea elephantipes*, Aco = *Acorus* sp. (*A. americanus, A. calamus*)
rosids: Pop = *Populus* sp. (*P. alba*, *P. trichocarpa*), Gos = *Gossypium* sp., At = *Arabidopsis thaliana*.

* Distance from the coding region: -493 (*P. trichocarpa*) / -552 (*P. alba*).

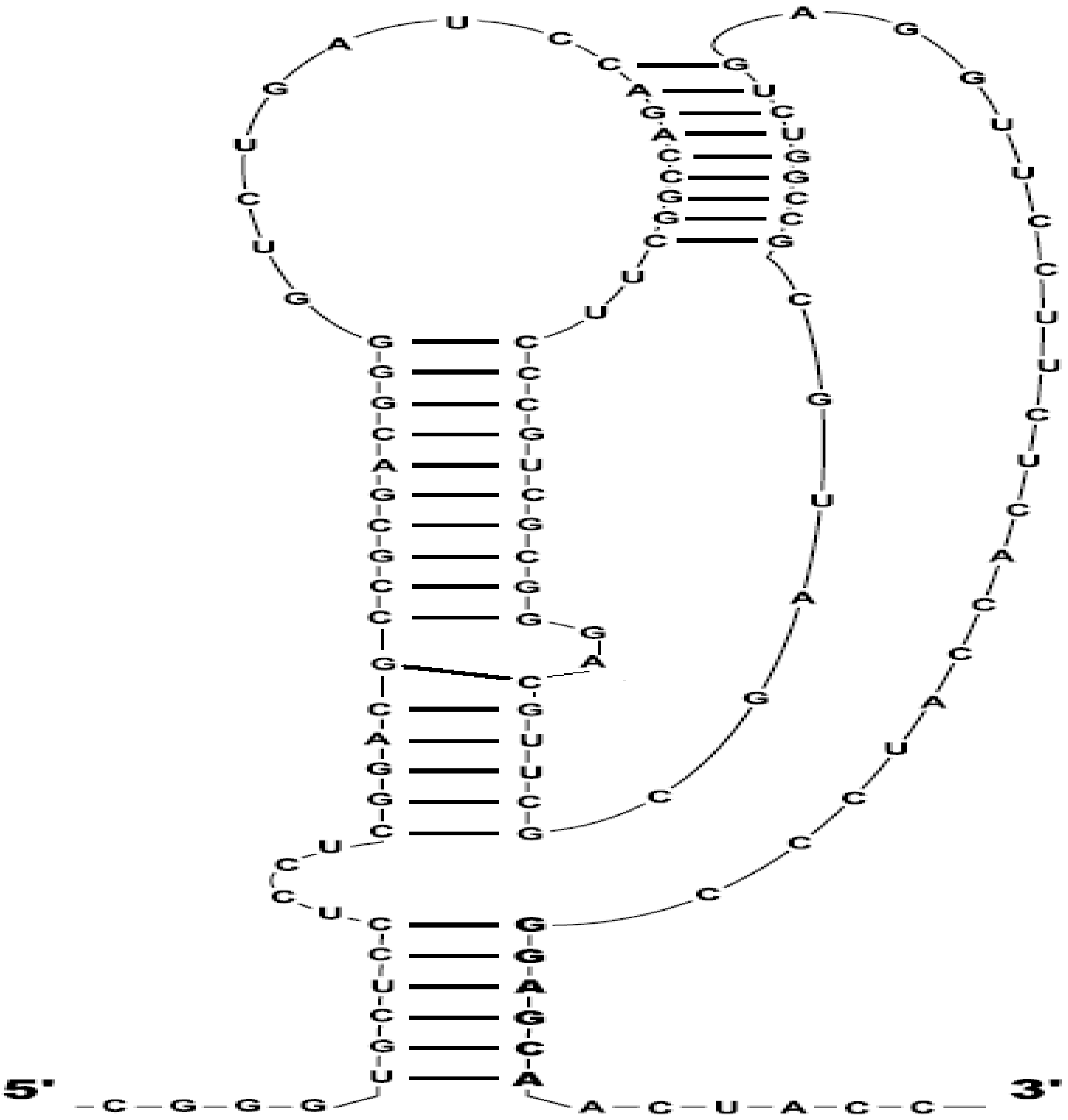
** Sequence is conserved in all Brassicales.

A next level task: searching for **transcription and translation regulations based on dynamics of RNA secondary structure** and inferring their evolution.

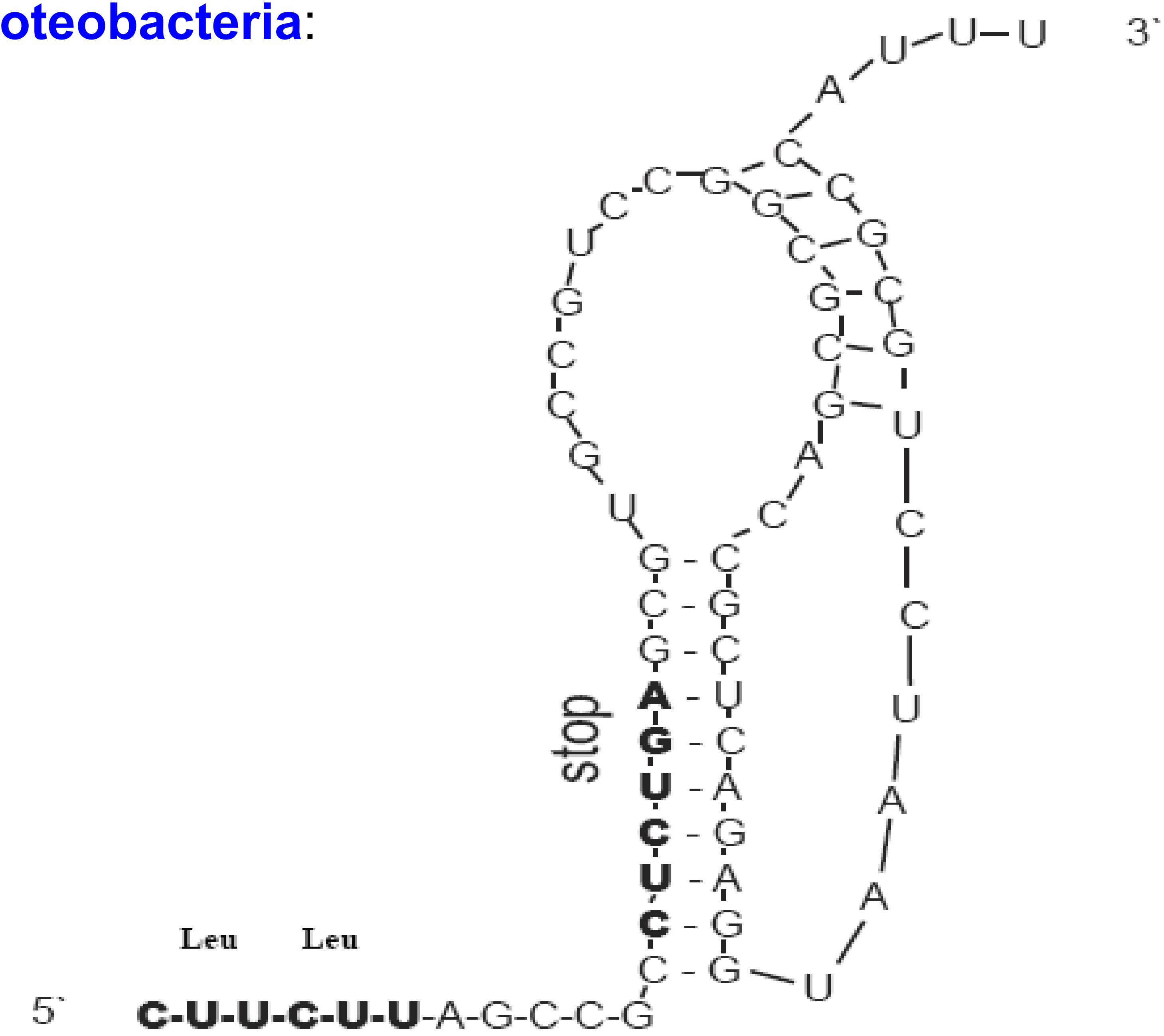
Here radically new approaches are required to both find the regulation and infer its evolution.

3) **Translation regulation** of gene expression through **blocking RBS by a secondary structure**

Translation regulation: LEU-pseudoknot in *Mycobacterium bovis*. It is conserved in **almost all Actinobateria**

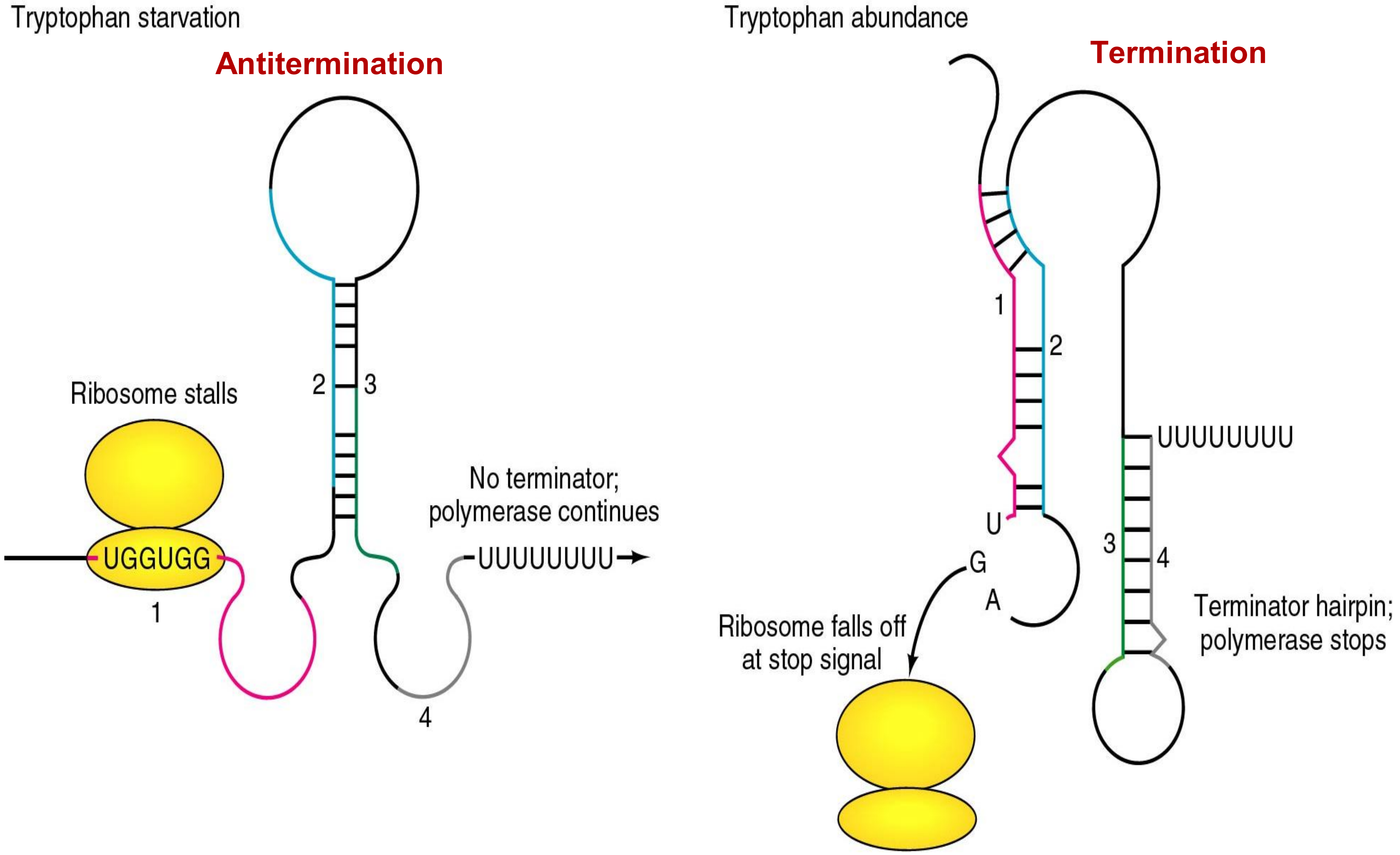


LEU1-pseudknot in *Dinoroseobacter shibae* and **many alphaproteobacteria**:



4) **Transcription regulation** of gene expression through competition of two structures: terminator and antiterminator

Classic attenuation by Yanofsky:
definitions of "**antiterminator**" and "**terminator**"



Types of classic attenuation regulation:

«**by Yanofsky**»: terminator and antiterminator have mutually exclusive structure; the 3’-end of terminator has an adjacent poly-U run;

«**succession of hairpins**»: **terminator and antiterminator are not mutually exclusive**, but there exists a succession of usually four hairpins, from which the first one is the antiterminator prohibiting the formation of the next hairpin (co-terminator) thus leading to the formation of the third hairpin (co-antiterminator) prohibiting the fourth one (terminator); 3’- poly-U run is present; in this case hairpins are usually stabilized with the formation of RNA triplexes;

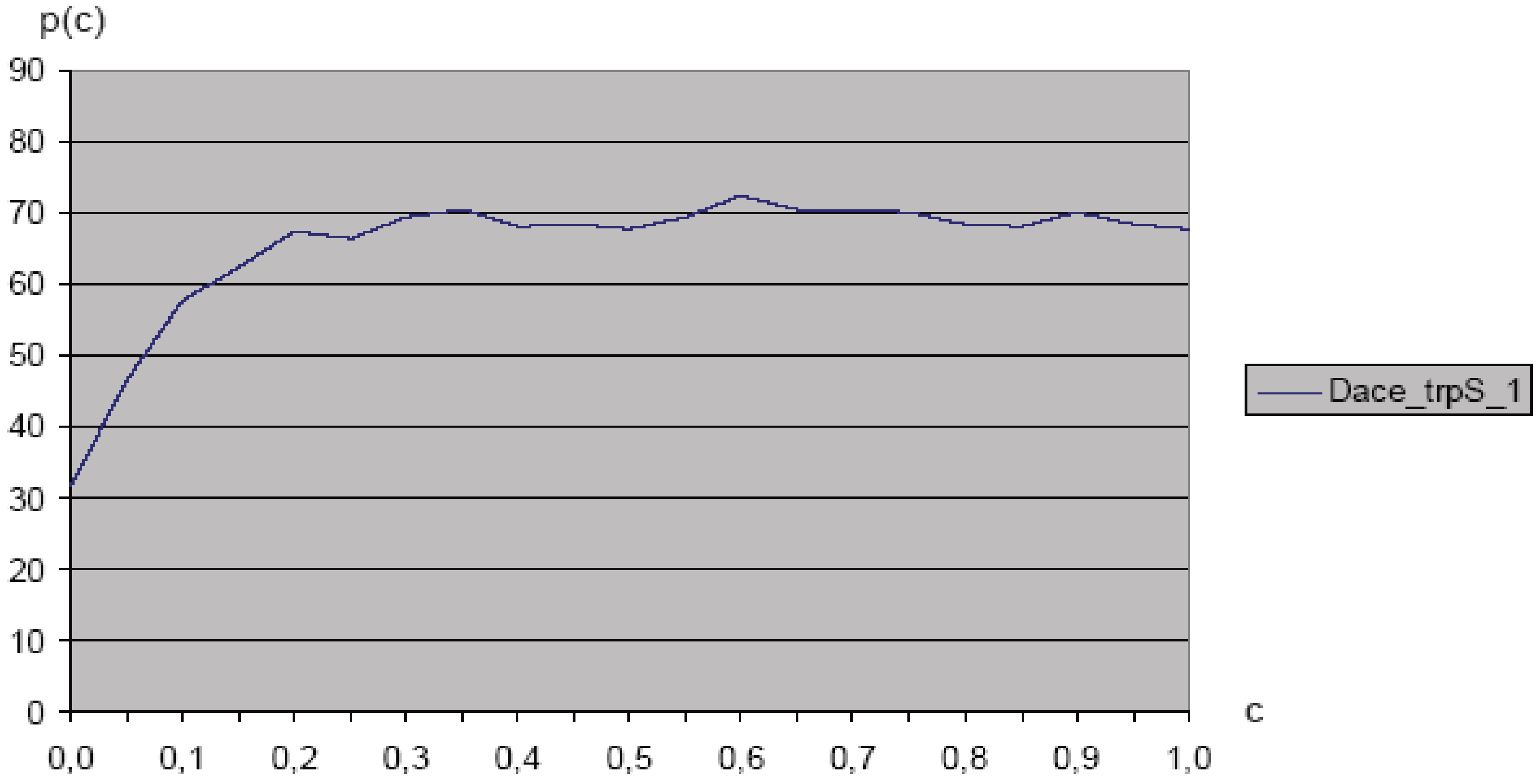
«**assembly of hairpins**»: conserved **antiterminator is replaced by a group of hairpins**, with each being exclusive to a conserved terminator; poly-U run is present, **hairpins may be stabilized by RNA triplexes**;

«**sequester-attenuation**» regulation follows below

Occurrence of **classical attenuation regulation** in major bacterial taxa (**non-classic attenuation** LEU and LEU1 in the last column)

Bacteria	**Gene: either with CAR**						**orNCA**
α-proteobacteria	***ilvB,I***	***trpE***	***hisS***	***pheST***	***thrA***	***leuA***	***leuA***
β-proteobacteria	***ilvB***	***trpE***		***pheA***	***thrS***	***leuA***	***leuA***
γ-proteobacteria	***ilvB,G***	***trpE***	***hisG***	***pheA, S***	***thrA***	***leuA***	
δ-proteobacteria	***ilvB***	***trpS***			***thrA,S***	***leuA***	
Actinobacteria	***ilvB,I, D***	***trpE,S,BE, BA***					***leuA***
Bacteroides/Chlorobi	***ilvD***	***trpE***	***hisG***				
Firmicutes	***ilvD, lysQ***	***trpB***	***hisZ***				
Thermotogae		***trpE***	***hisS***				
Chloroflexi	***ilvD***						

Assembly of antiterminators in *Desulfuromonas acetoxidans*

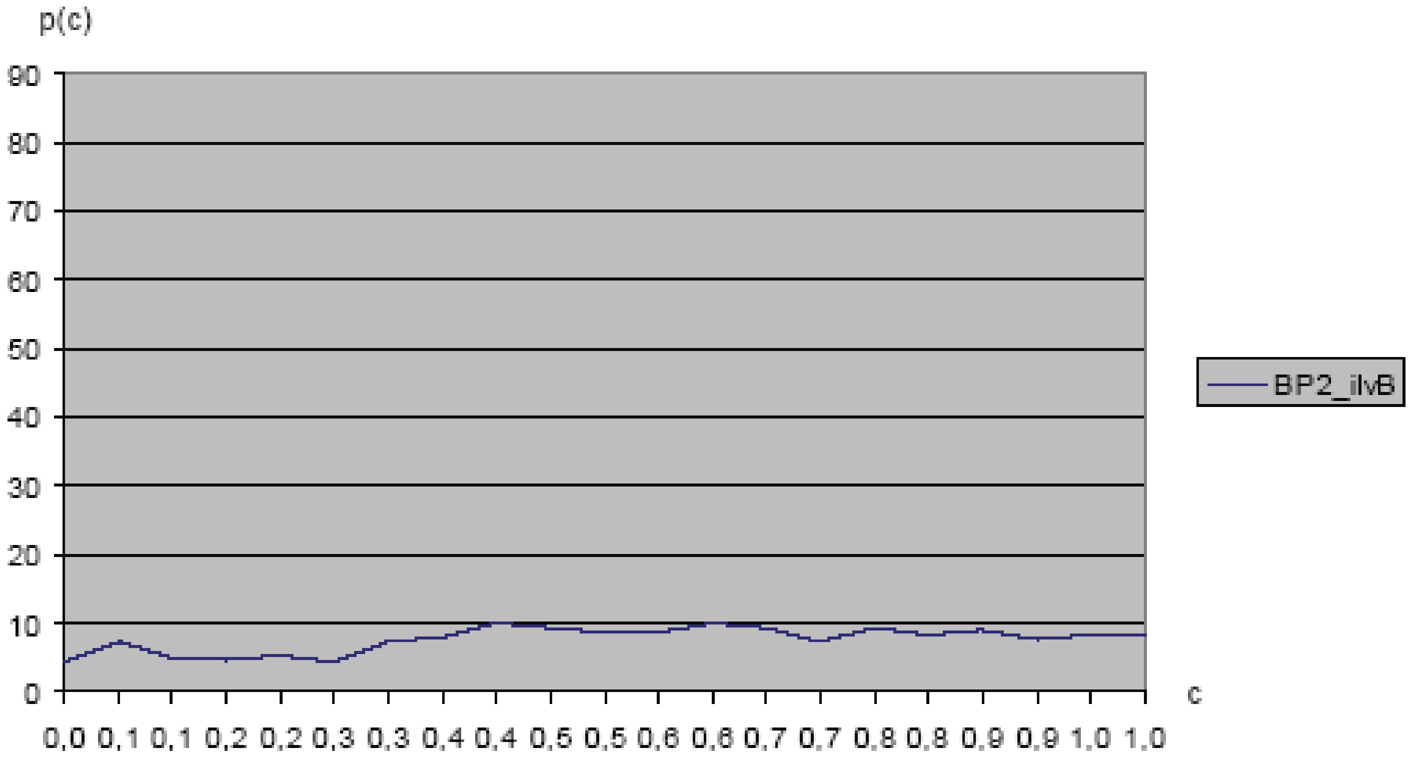


=========>Terminator<=========

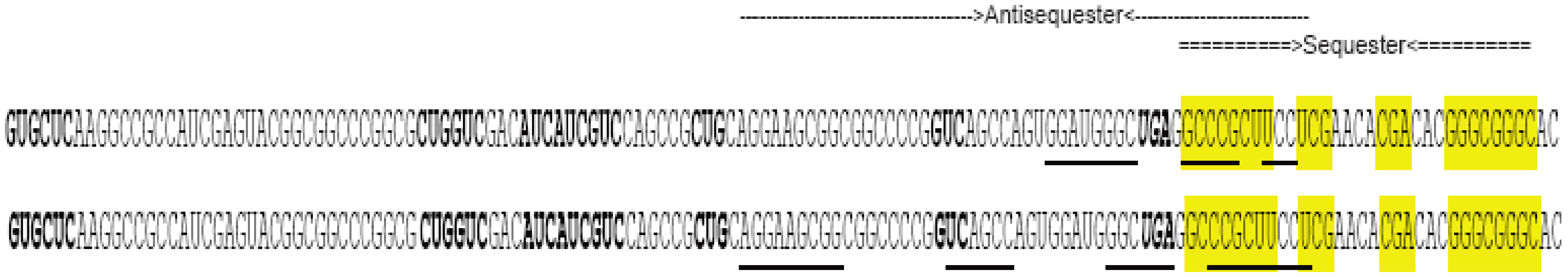
AUGACAUUACAGCGAUGGUGGUGGCGCUCAGCCUUUUAUGGACUGAUCGCGUGCCUUUCGCUUUUUUGACAUCACAGACUGCGCCGGGCACGGCGCGGCAACAAUGAAAUCACACAAGCCGCGAUGAUCGUAUCGCGGCUUGUUU

Sequester-attenuation regulation in *Bordetella pertussis*

No poly-U
and CAR,
but RBS
present



Here are two variants of antisequester but not many:


-------------------------------->Antisequester<--------------------------

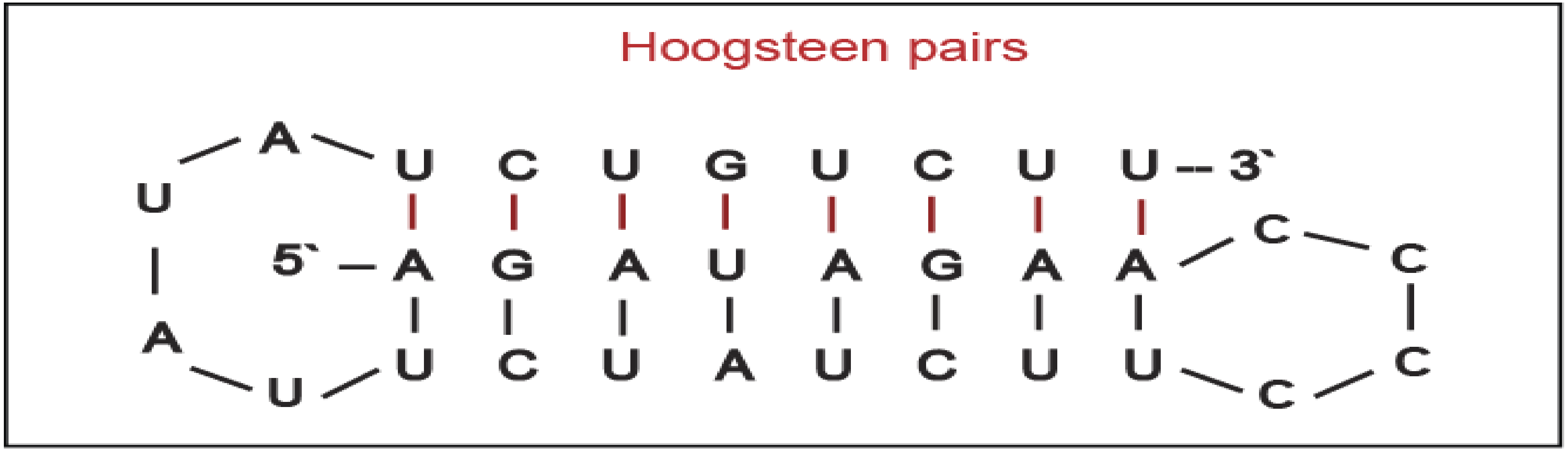
==========>Sequester<==========

GUGCUCAAGGCCGCCAUCGAGUACGGCGGCCCGGCG**CUGGUC**GAC**AUCAUCGUC**CAGCCG**CUG**CAGGAAGCGGCGGCCCCG**GUC**AGCCAGUGGAUGGGC**UGA**GGCCCGCUUCCUCGAACACGACACGGGCGGGCAC

GUGCUCAAGGCCGCCAUCGAGUACGGCGGCCCGGCG **CUGGUC**GAC**AUCAUCGUC**CAGCCG**CUG**CAGGAAGCGGCGGCCCCG**GUC**AGCCAGUGGAUGGGC**UGA**GGCCCGCUUCCUCGAACACGACACGGGCGGGCAC

5) **Transcription and translation** regulations with **triplexes and pseudoknots**

The role of RNA triplex



The RNA triplex is a structure formed with Hoogsteen hydrogen bonds between a region of mRNA («third shoulder») and the stem of a Pu-Py helix. The third shoulder may be located in front of the left shoulder. A triplex is formed with N--Pu-Py triads (usually U--A-U), where the Pu-Py pair comes from the helix. The third shoulder has the same orientation as the purine shoulder of the helix.

RNA triplexes in *his* regulation

Many γ-proteobacteria possess a **triplex** upstream gene *hisG* **stabilizing the co-terminator**. Its third shoulder contains many poly-U runs, which make it stable regardless of the cytoplasm acidity. Usually it is formed with Py--Pu-Py triads. In *Alteromonadales bacterium* and *Pseudoalteromonas haloplanktis*, however, the triplex CUGU--GAGG-CCUC is composite (Py--Pu-Py and Pu--Pu-Py triads).

The permease-coding gene *lysQ* in *Lactococcus lactis* (Firmicutes) has CAR with **histidine regulatory codons** and a **hairpin succession where the co-terminator is stabilized** with the Pu--Pu-Py **triplex** AGA--AGA-UCU

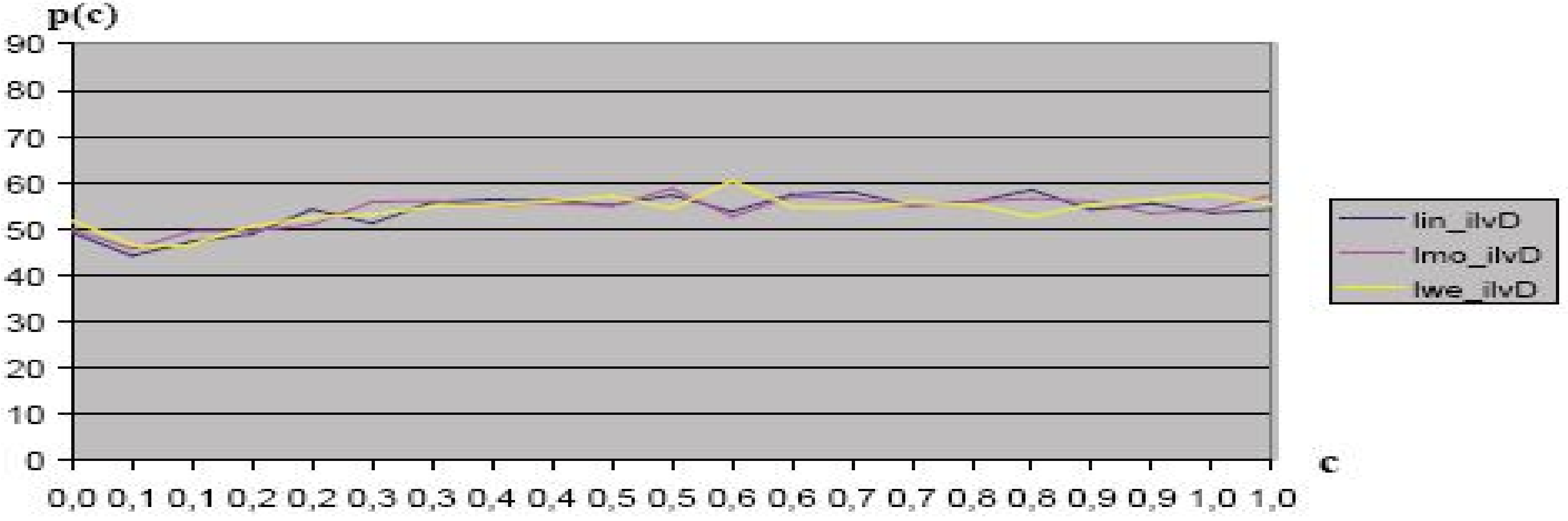
RNA triplexes in *ilvD* regulation

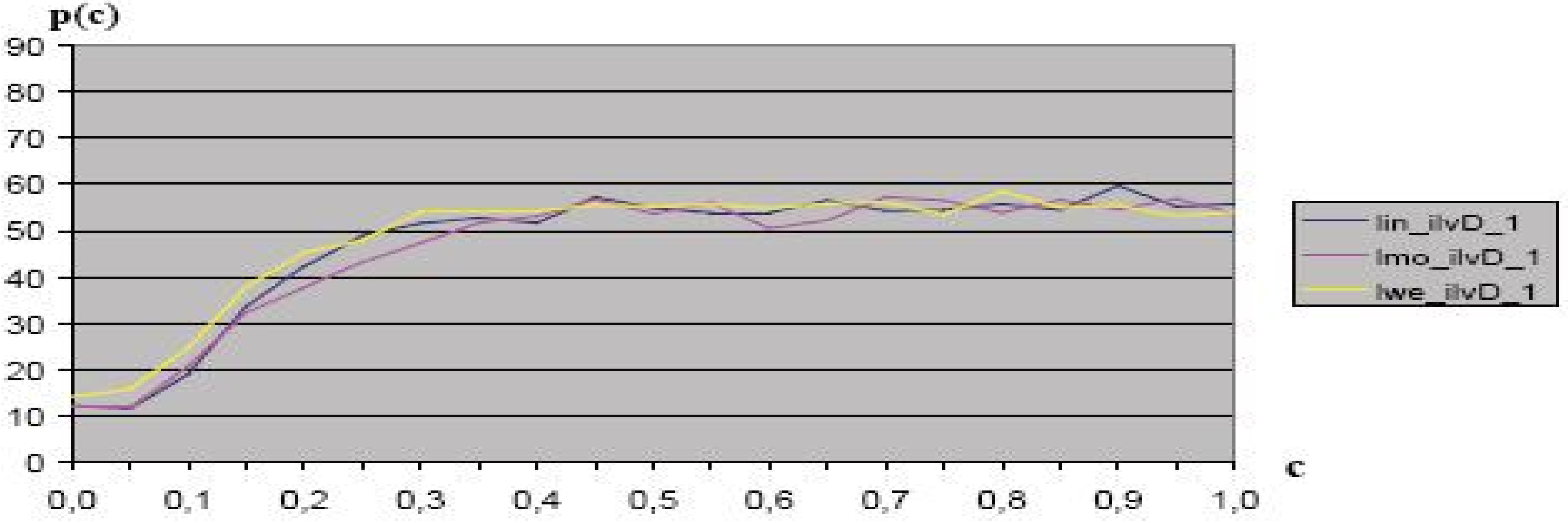
Regulatory regions upstream *ilvD* in Staphylococcus and Listeria (Firmicutes) **possess a succession of four conserved hairpins** forming a Py--Pu-Py **triplex** in the co-terminator.

Modeling this and the *hisG* cases suggests the **importance of taking the energy of the co-terminator triplex** into account.

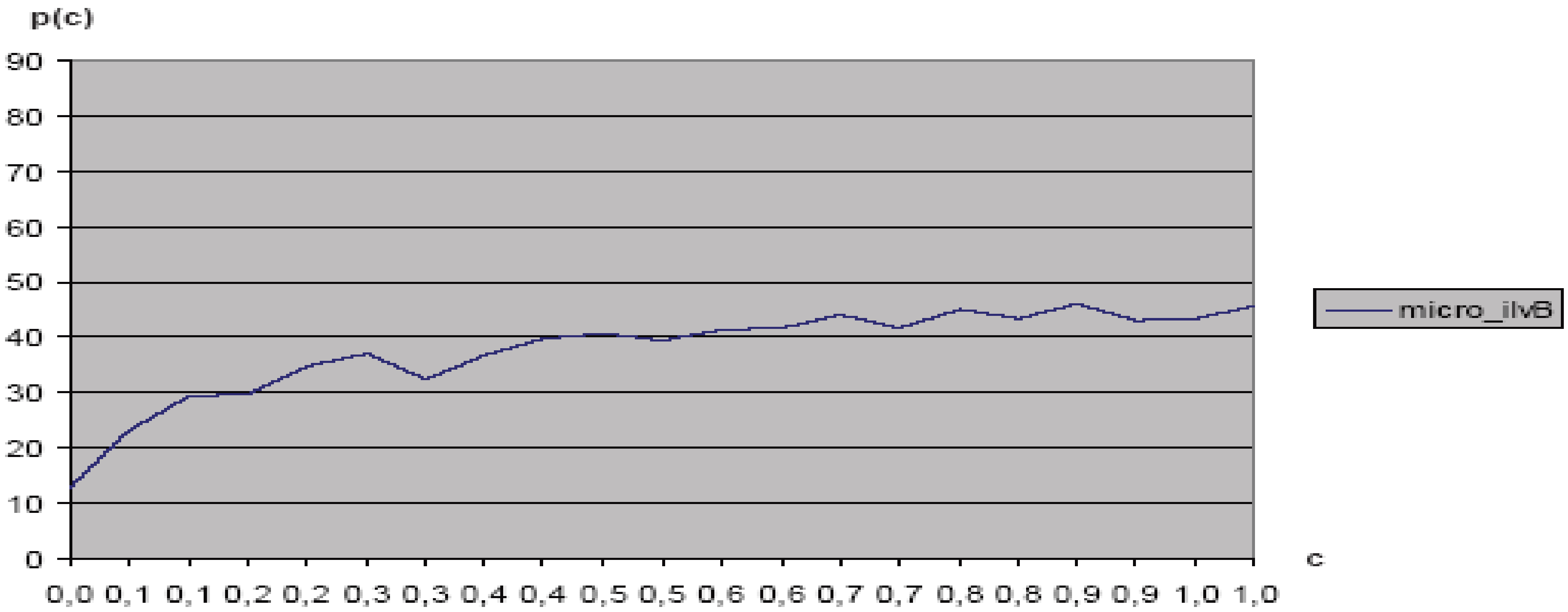
This result and the observed **conservativity of triplexes** support their importance in calculating RNA secondary structures

Probability of the *ilvD* operon termination in *Listeria* **without** and **with** RNA **triplex energies**

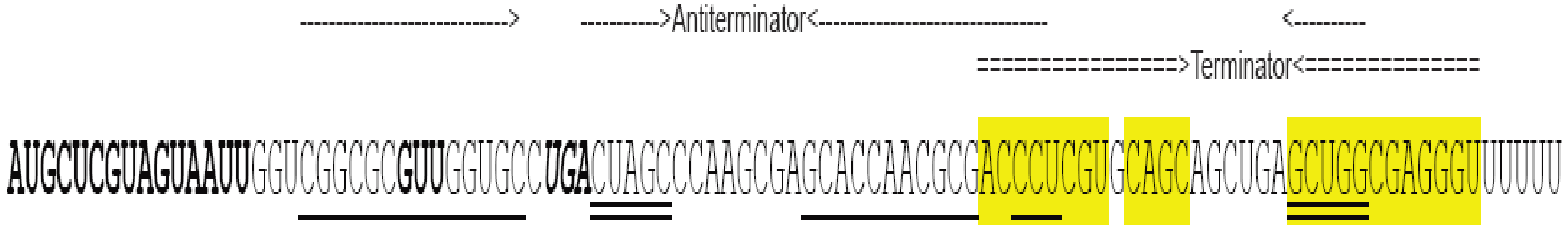




Antitarminator and terminator exist together under low concentration but terminator is blocked by pseudoknot



Mycobacterium microti (Actinobacteria, gene *ilvB*) ???



6) Evolution of regulations based on the secondary structure dynamics

7) Plastids of **Apicomplexa** resemble those of **red algae**.

- **Importance of plastids**: the **Apicomplexa** (secondary endosymbionts) are **protozoan infection agents, and many illnesses like that**;

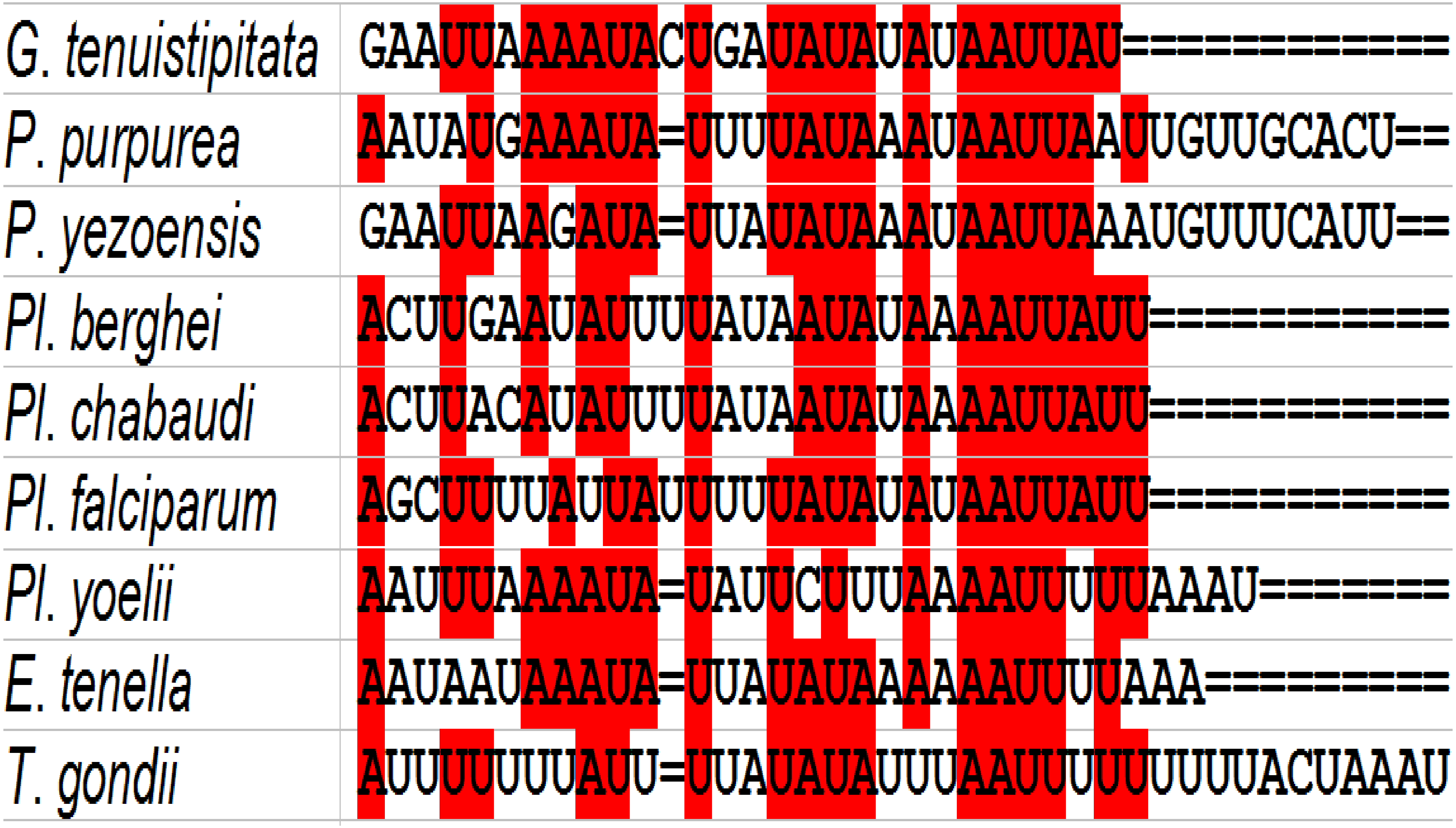
- **All plastids** are good **targets for drugs** affecting bacterial RNA polymerases or ribosomes, thus being safe for the eukaryotic host cell

The role of plastid RNA sites in pathogen invasion:

Toxoplasma gondii (Apicomplexa) switches on plastid genes in the host cell.

Also other **regulations**

Bacterial type regulation of gene ***ycf24*** in **rhodophytes, plasmodia, coccidia**: a hypothetic protein factor binding site overlapping the **RBS** in mRNA:



G. tenuistipitata	GAAUUAAAAUACUGAUAUAUAUAAUUAU============
P. purpurea	AAUAUGAAAUA=UUUUAUAAAUAAUUAAUUGUUGCACU==
P. yezoensis	GAAUUAAGAUA=UUAUAUAAAUAAUUAAAUGUUUCAUU==
Pl. berghei	ACUUGAAUAUUUUAUAAUAUAAAAUUAUU===========
Pl. chabaudi	ACUUACAUAUUUUAUAAUAUAAAAUUAUU===========
Pl. falciparum	AGCUUUUAUUAUUUUUAUAUAUAAUUAUU===========
Pl. yoelii	AAUUUAAAAUA=UAUUCUUUAAAAUUUUUAAAU=======
E. tenella	AAUAAUAAAUA=UUAUAUAAAAAAUUUUAAA=========
T. gondii	AUUUUUUUAUU=UUAUAUAUUUAAUUUUUUUUUACUAAAU

Here conserved regions in 5'-UTR adjoin the start codon of ***ycf24***.

This signal is **not detected in other orthologous groups**.

This signal existing in the ancestor of these species diversifies within one descendant:

In *Toxoplasma gondii* this regulation extends onto genes *rps4* and *rpoB*:

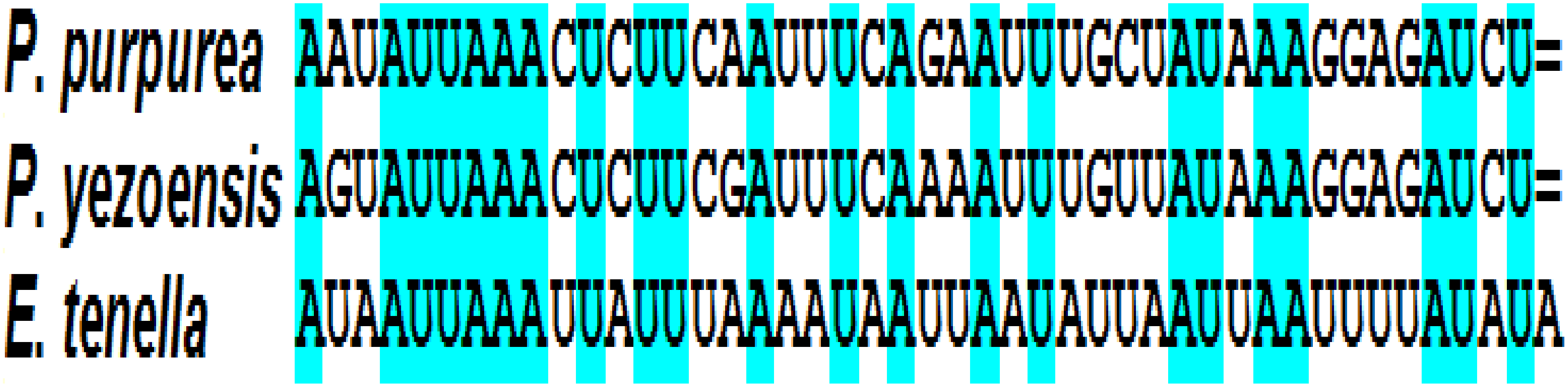
nearly identical regions upstream *ycf24, rps4* and *rpoB*; **signal absent upstream other genes**;

regulation affects all plastome genes through the regulation of **ribosome protein S4 and *β*-subunit** of RNA polymerase;

experimentally proved: in *Toxoplasma gondii* **plastids** are **essential for virulence but not critical for** *in vitro* **survival**, [Wilson et al. 2003. *Phil. Trans. R. Soc. Lond*.].

Hypothesis: this regulation sustains the pathogenicity of *T. gondii*

Translation regulation (excess of subunits) of RNA polymerase β-subunits (=*rpoB*) in plastids



```
P. purpurea   AAUAUUAAACUCUUCAAUUUCAGAAUUUGCUAUAAAGGAGAUCU=
P. yezoensis  AGUAUUAAACUCUUCGAUUUCAAAAUUUGUUAUAAAGGAGAUCU=
E. tenella    AUAAUUAAAUUAUUUAAAAUAAUUAAUAUUAAUUAAUUUUAUAUA
```

rhodophytes and coccidia; conserved sites (in blue) constitute a putative mRNA-protein binding site that overlaps the RBS (according to RpoB protein alignment);

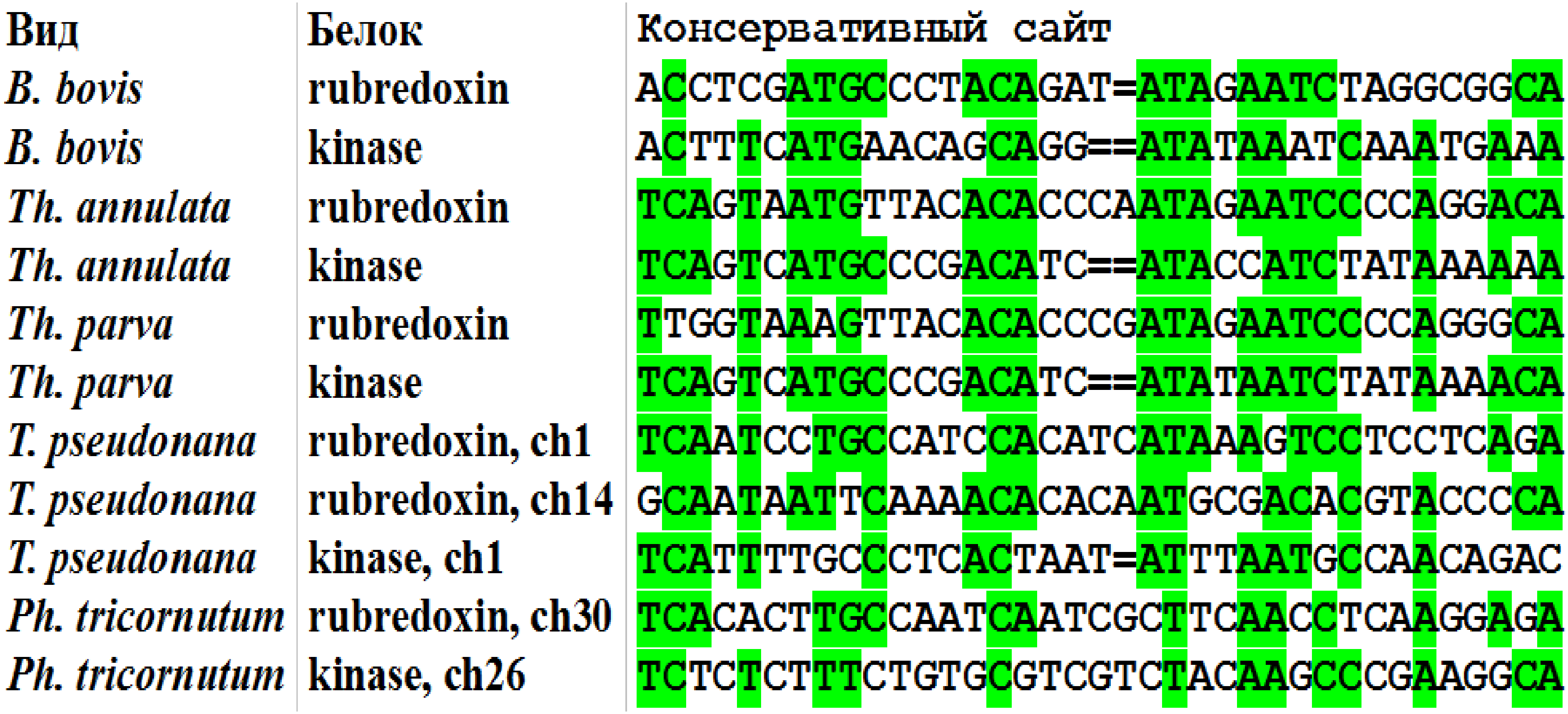
signal **not detected in other orthologous groups**;

in *E. coli* similar regulation is found at the translation level [Passador *et al*, 1992, *J. Bacteriology*]: a β-subunit binds to a specific mRNA site and **interrupts translation**

8) **Transcription regulation in nucleus**

in the Piroplasmida

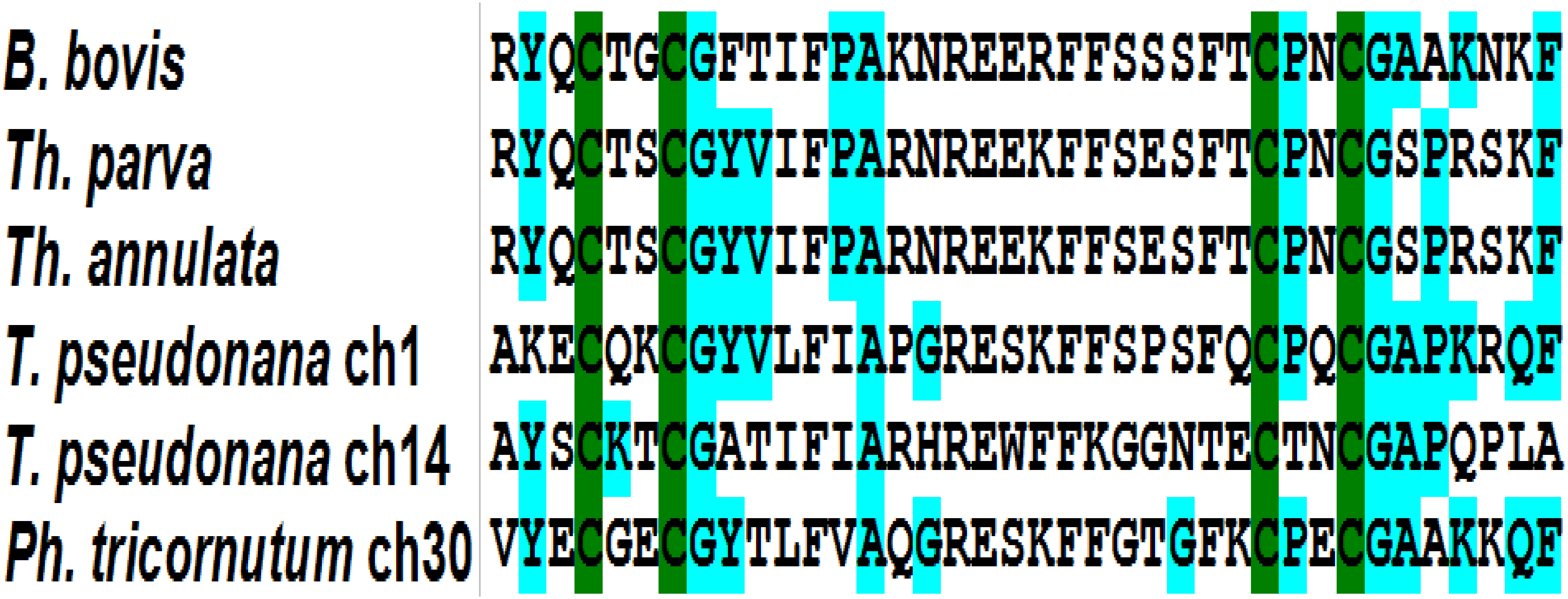
Alignment of the **rubredoxin and kinase** regulatory regions (transcription in nucleus)



Вид	Белок	Консервативный сайт
B. bovis	rubredoxin	ACCTCGATGCCCTACAGAT=ATAGAATCTAGGCGGCA
B. bovis	kinase	ACTTTCATGAACAGCAGG==ATATAAATCAAATGAAA
Th. annulata	rubredoxin	TCAGTAATGTTACACACCCAATAGAATCCCCAGGACA
Th. annulata	kinase	TCAGTCATGCCCGACATC==ATACCATCTATAAAAAA
Th. parva	rubredoxin	TTGGTAAAGTTACACACCCGATAGAATCCCCAGGGCA
Th. parva	kinase	TCAGTCATGCCCGACATC==ATATAATCTATAAAACA
T. pseudonana	rubredoxin, ch1	TCAATCCTGCCATCCACATCATAAAGTCCTCCTCAGA
T. pseudonana	rubredoxin, ch14	GCAATAATTCAAAACACACAATGCGACACGTACCCCA
T. pseudonana	kinase, ch1	TCATTTTGCCCTCACTAAT=ATTTAATGCCAACAGAC
Ph. tricornutum	rubredoxin, ch30	TCACACTTGCCAATCAATCGCTTCAACCTCAAGGAGA
Ph. tricornutum	kinase, ch26	TCTCTCTTTCTGTGCGTCGTCTACAAGCCCGAAGGCA

Species are piroplasmids and diatoms. Kinases phosphorilate tyrosine in proteins. **The signal is not predicted in other orthologous groups**.
In *Th. parva* the region begins upstream the kinase transcription start codon at position -46. We assume: **this is a protein-DNA regulation affecting a promoter**

Rubredoxins under regulation contain a very similar domain:

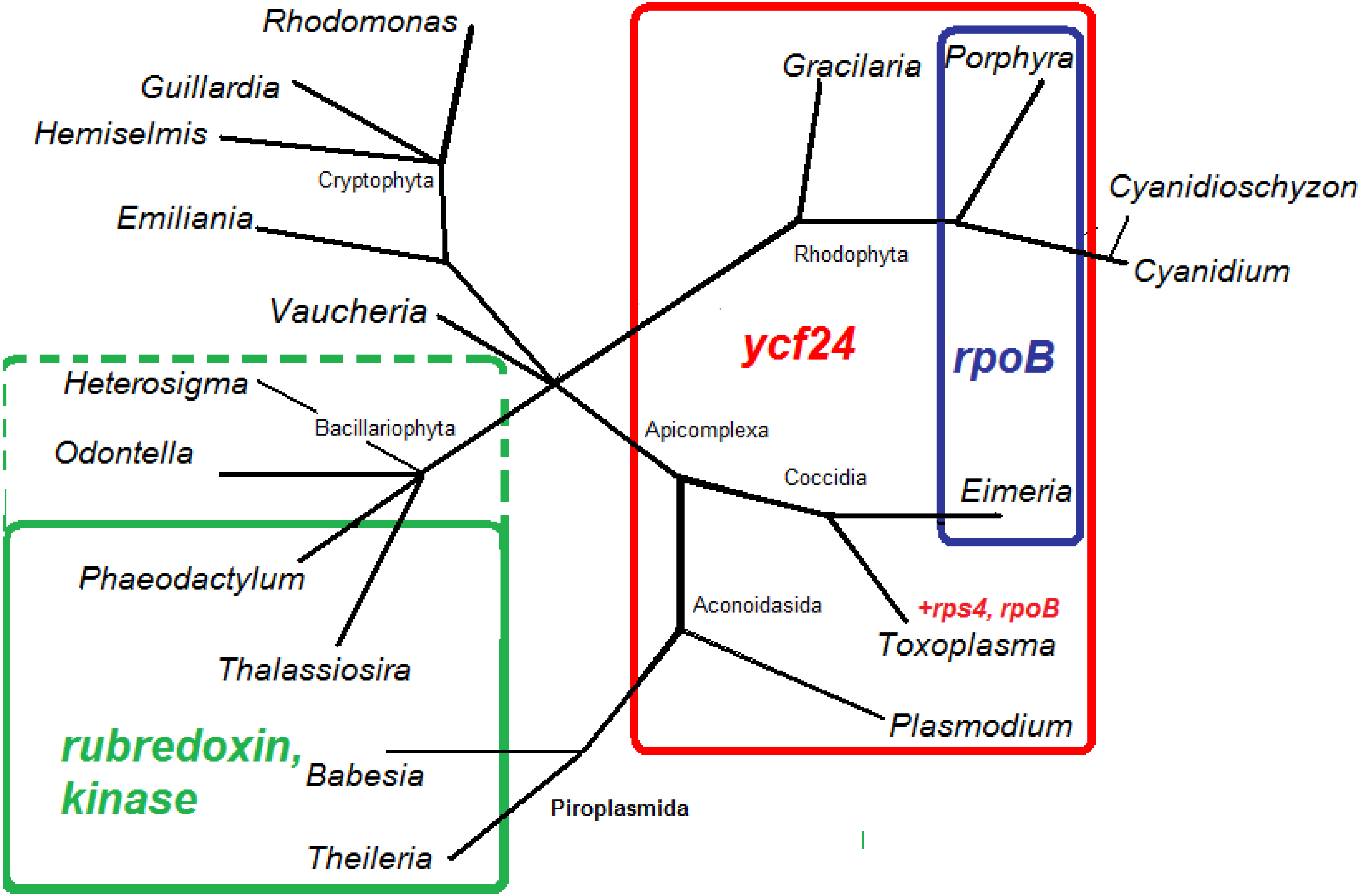


B. bovis	RYQCTGCGFTIFPAKNREERFFSSSFTCPNCGAAKNKF
Th. parva	RYQCTSCGYVIFPARNREEKFFSESFTCPNCGSPRSKF
Th. annulata	RYQCTSCGYVIFPARNREEKFFSESFTCPNCGSPRSKF
T. pseudonana ch1	AKECQKCGYVLFIAPGRESKFFSPSFQCPQCGAPKRQF
T. pseudonana ch14	AYSCKTCGATIFIARHREWFFKGGNTECTNCGAPQPLA
Ph. tricornutum ch30	VYECGECGYTLFVAQGRESKFFGTGFKCPECGAAKKQF

These (and two paralogs) are found in nuclei of diatom algae and parasitic Piroplasmida (*Theileria, Babesia*).

In the conserved (blue) active center of rubredoxins four cysteine residues (green) bind a Fe ion (the ferro-sulfuric center); **it is a subfamily of rubredoxins**

Regulation in Apicomplexa and algae (very similar plastids but not their regulations):

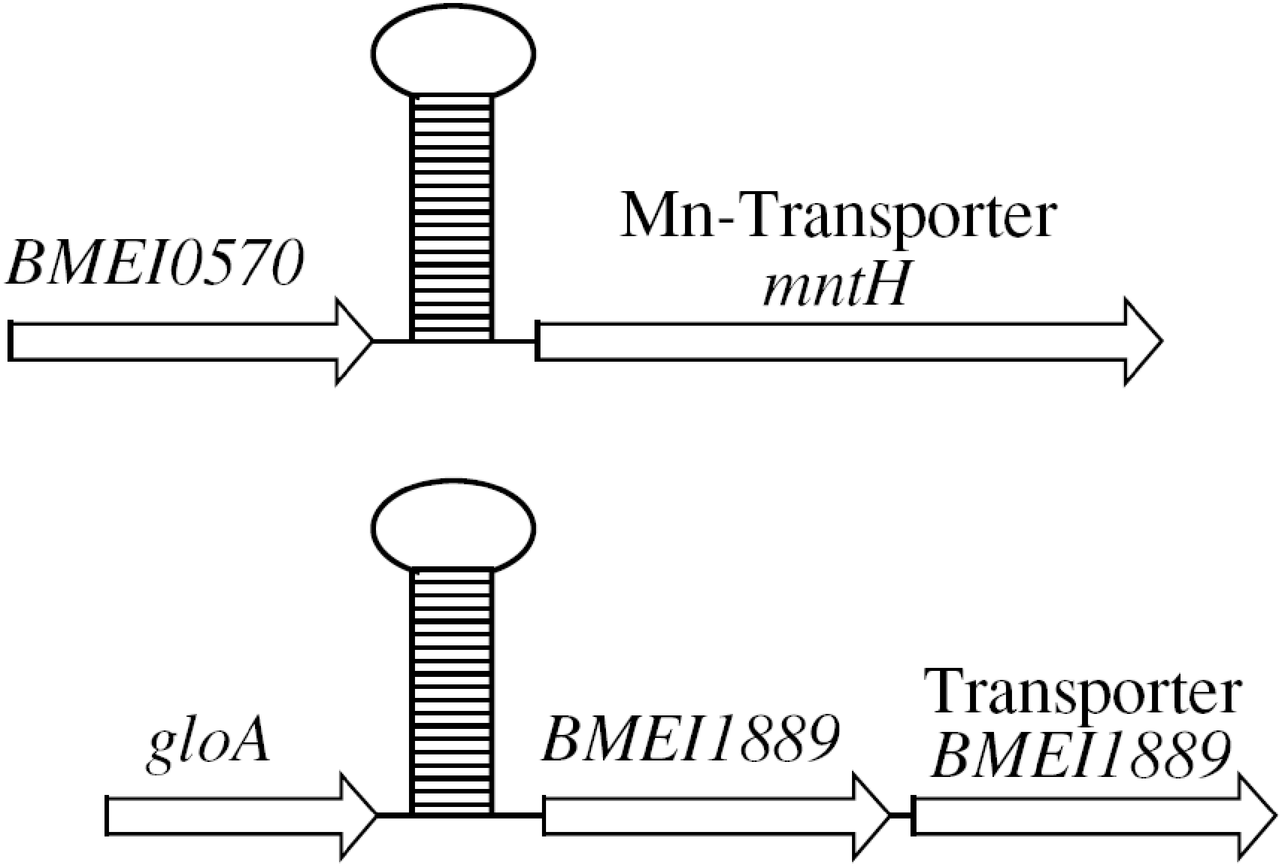


The role of RNA secondary structures in pathogen invasion:

9) *Brucella* (alpha-proteo) competes for the host macrophage cell resources (metal cations) using RNA secondary structures

It was found that divalent cation transporters of the Nramp family in eukaryotic cell phagosomes and bacteria that parasitize these cells compete for metals that are vital for bacterial survival. Long helices were determined in the 5'-untranslated region for each mRNA in *Brucella.*
Long helices of quite similar nucleotide composition were found in mRNAs that encode manganese transporters and Ni-dependent glyoxalase I. **We suggest** that long helices in these regions are involved in the regulation of RNA stability

The helices were found between close (up to 300 nt apart) genes on the same strand that are not separated by a usual terminator. Therefore, the genes might belong to the same operon



Brucella helices

GGAGUAAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUGAAGAA
UAUGGUCGCUGCGGCUAUGUACAACCAAAACAUACUCCCCUACUCCCUUAUUCCCCUAUU
CCCUUAACAUACUGCCUUACUGCCCUAUUGCCUUACUGCCUUAUUCC

UAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUGAAGAAUAUGG
UCGCUGCGGCCAUGCGCAACCAAAAACAUACUCCCCUACUCCCUUAUUCCCCUAUUCCCU
UAACAUACUGCCCUAUUGCCCUAUUGCCUUACUGCCUUA

GGAGUAAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUGAAGAAUAUGGUCGCUGCGGCU
AUGUACAACCAAAACAUACUCCCCUACUCCCUUAUUCCCCUAUUCCCUUAACAUACUGCC
UUACUGCCCUAUUGCCUUACUGCCUUAUUCC

UUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUAGGGCAGUGA
AGAAUAUGGUCGCUGCGGCCAUGCGCAACCAAAAACAUACUCCCCUACUCCCUUAUUCCC
CUAUUCCCUUAACAUACUGCCUUAUUGCCCUAUUGCCUUACUGCCUUAA

GGAGUAAGGGCAGUAGGGCAGUGAAGAAUAUGGUCGCUGCGGCUAUGUACAACCAAAACA
UACUCCCCCUACUCCCUUAUUCCCCUAUUCCCUUAACAUACUGCCUUACUGCCUUAUUCC

UUAGGGCAGUAGGGCAGUGAAGAAUAUGGUCGCUGCGGCCAUGCGCAACCAAAAACAUAC
UCCCCUACUCCCUUAUUCCCCUAUUCCCUUAACAUACUGCCCUAUUGCCCUAUUGCCUUA
CUGCCUUACUGCCUUAA

10) Secondary structures are often used in bacteria in defense against their phages

11) **Secondary DNA structures (crest-hairpins)**

T1 and T2 from BD-10.

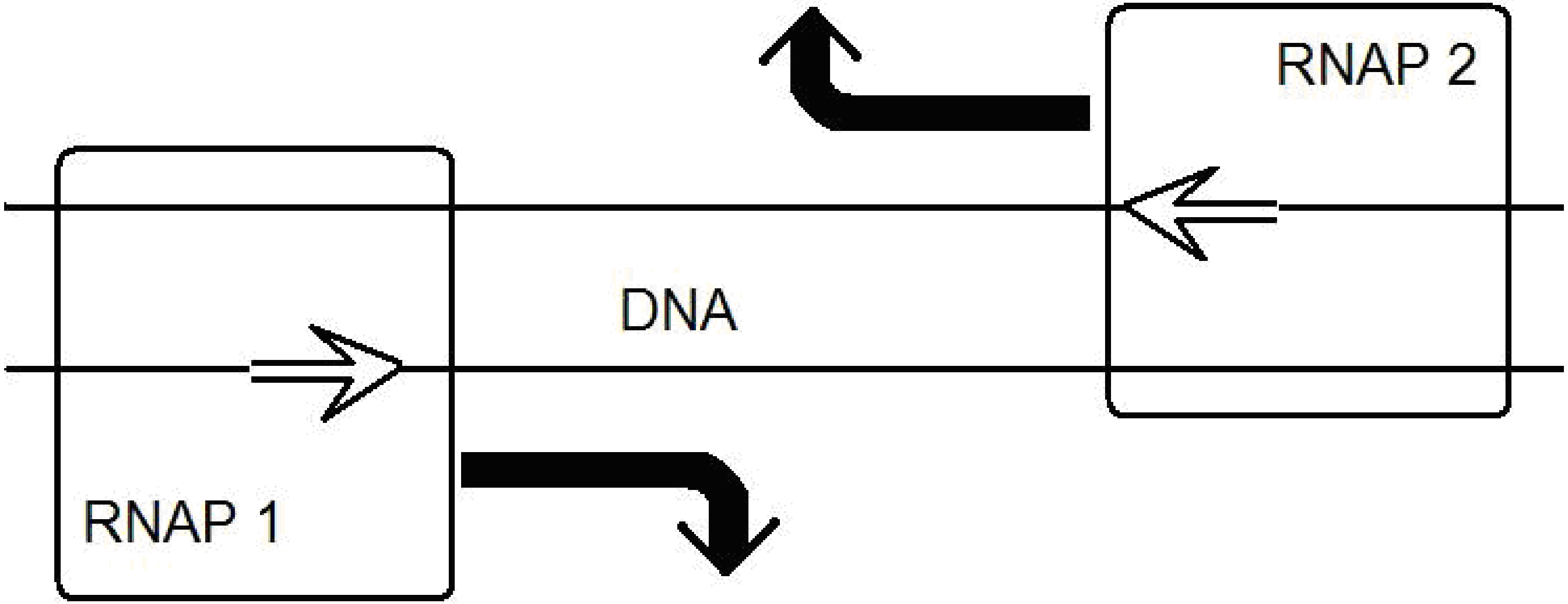
These hairpins are common also in actinobacteria in the trailer regions of highly transcribed genes

12-13) **RNA polymerase competition as an important transcription-related process**:

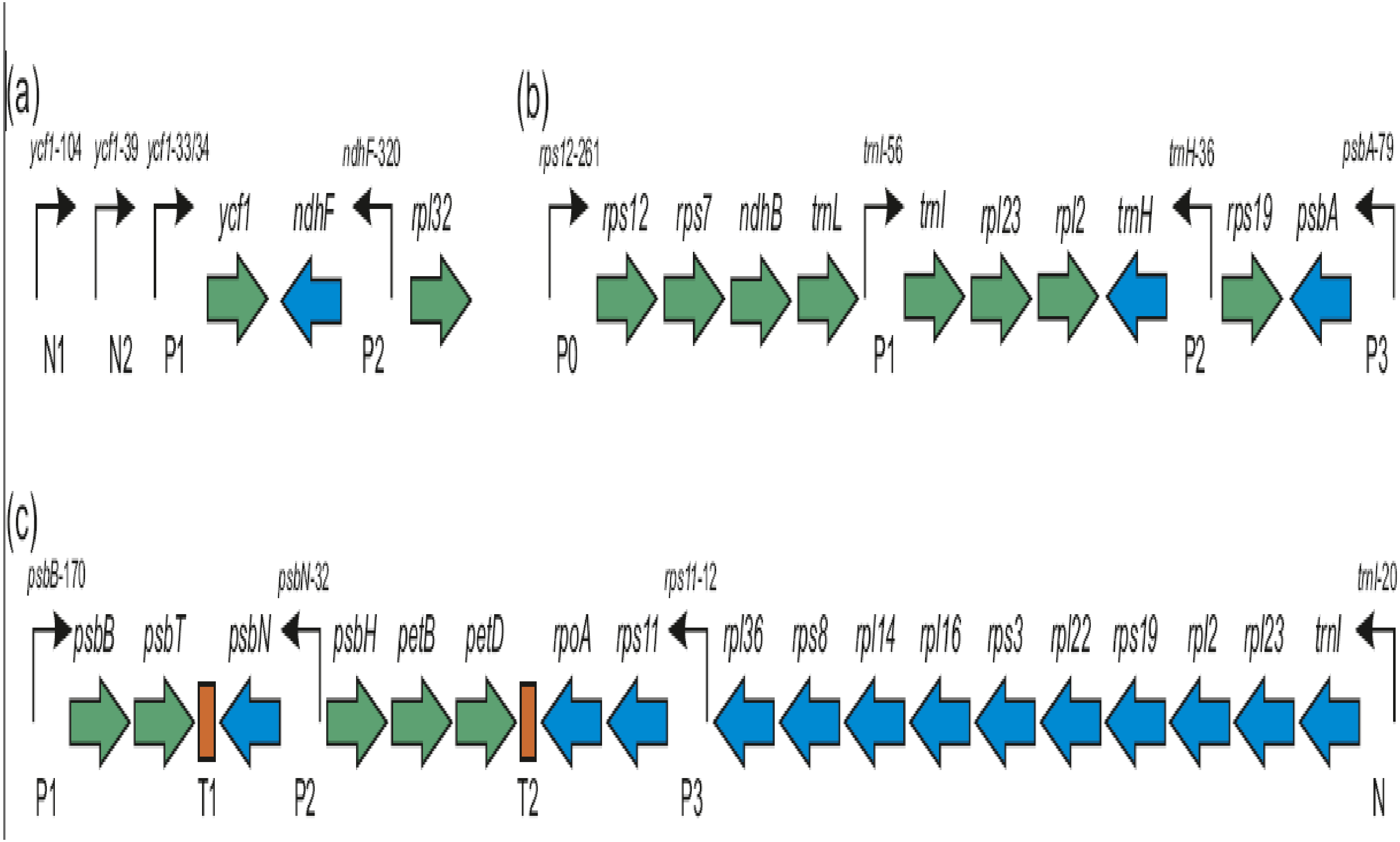
the competition drives physiological responses (e.g., to heat shock);

and regulation responses (physiological or tussue-specific through the interaction of nuclear and plastid genomes)

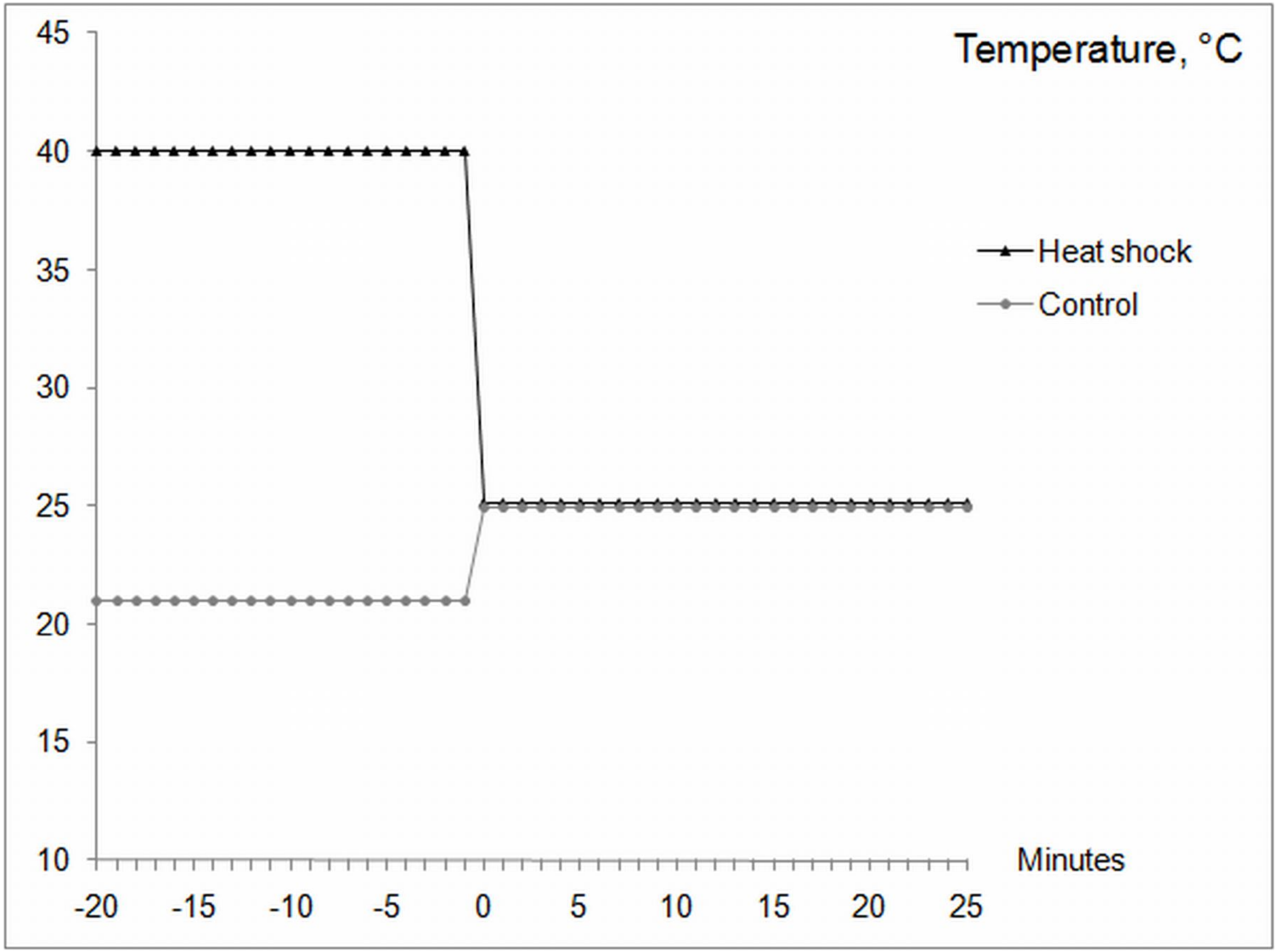
Competition:
RNA polymerases on complementary DNA strands collide and detach



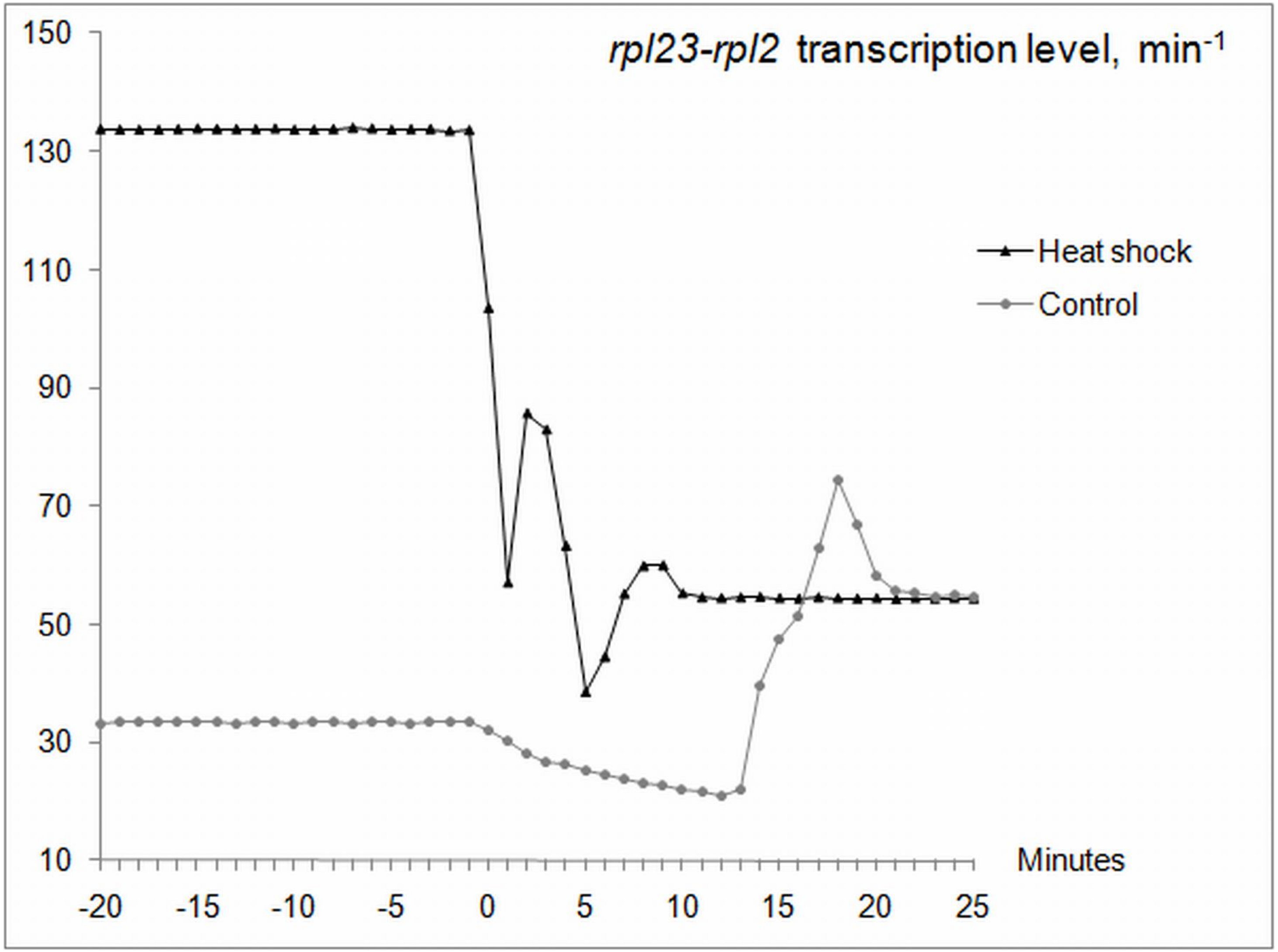
Examples of loci:



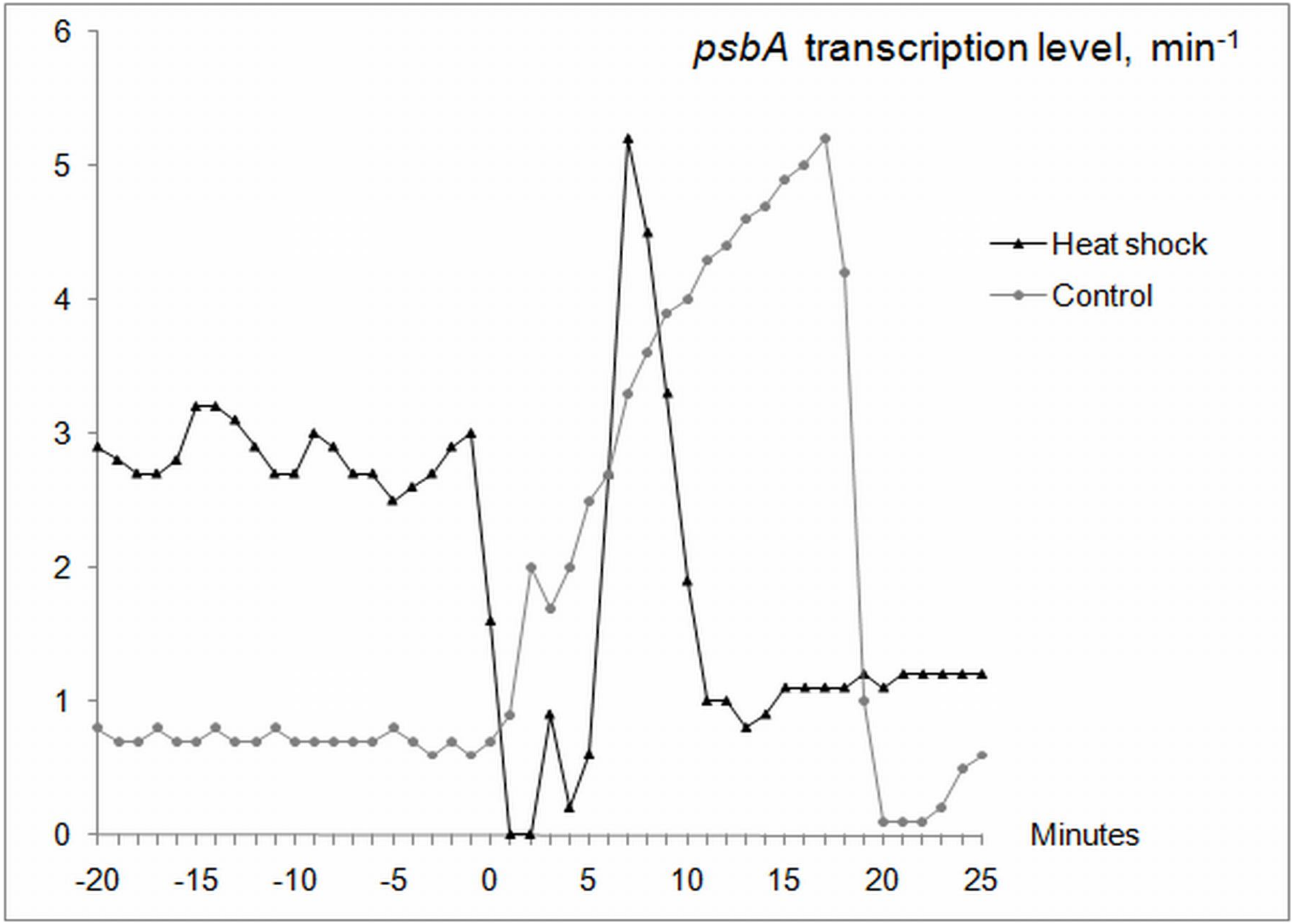
Experimental conditions. The zero on the horizontal axis corresponds to the heat shock termination



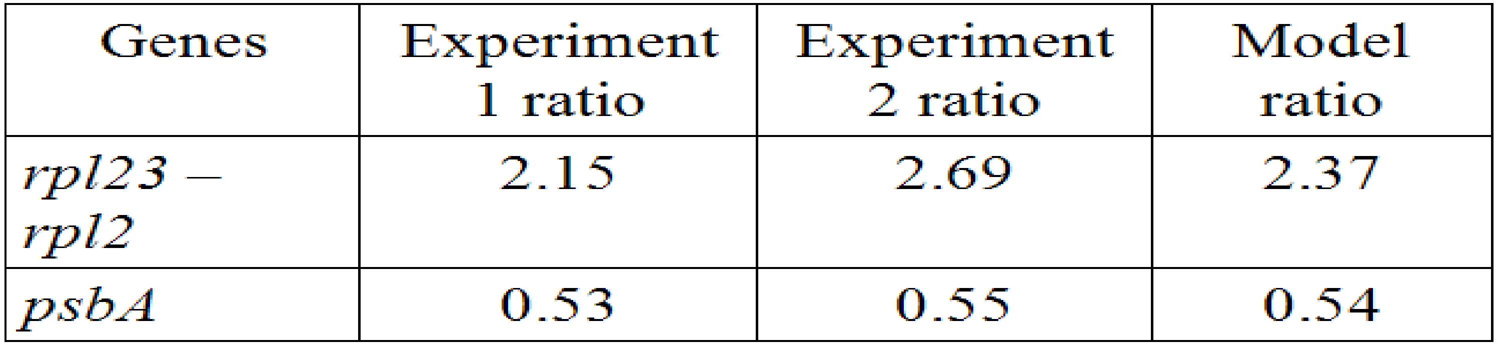
Transcription level of genes *rpl23-rpl2 vs.* time measured in the experiment. The zero on the horizontal axis corresponds to the heat shock termination



Transcription level of gene *psbA vs.* time measured in the experiment. The zero on the horizontal axis corresponds to the heat shock termination

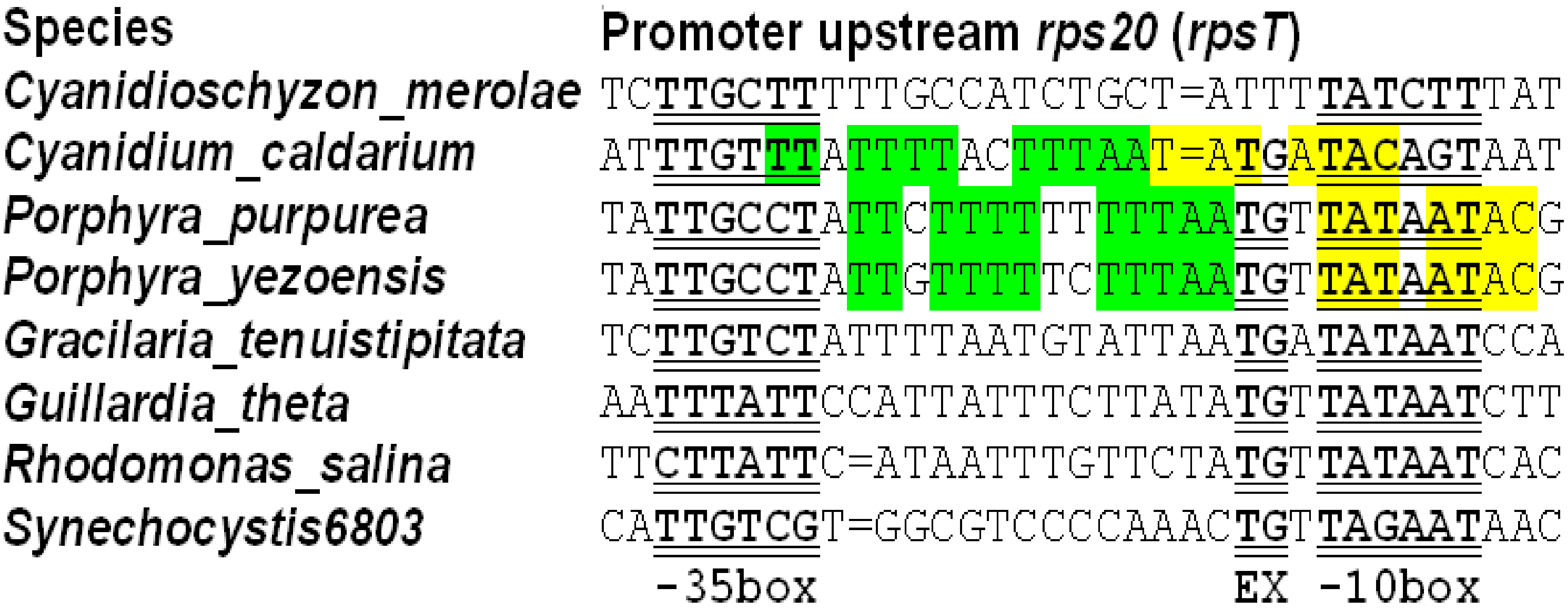


The *Hordeum vulgare* chloroplast contains two copies of the following set of genes: *rps12–rps7–ndhB–trnL–trnI–rpl*23–*rpl*2–(*trnH*)–*rps*19. One set competes with neighboring gene *psbA*: P0-*rps12-rps7-ndhB-trnL*-P1-*trnI-rpl23-rpl2*-(*trnH*-P2)-*rps19*-(*psbA*-P3), and the other set adjoins the next operon on the same strand: P0-*rps12-rps7-ndhB-trnL*-P1-*trnI-rpl23-rpl2*-(*trnH*-P2)-*rps19-rpl22-rps3-rpl16-rpl14-rps8-infA-rpl36-rps11-rpoA*. The transcription level ratios were measured experimentally for these sets at the temperatures of 21°C and 40°C. Our model predictions conform within experimental error with *in vitro* measurements (the table below) for values of the promoter binding efficiency P0=0.2, P1=0.9, P2=0.3, P3=0.1s^{-1}, and the RNA polymerase elongation rates R_{21}=9.2 and R_{40}=36.8bp/s at lower and higher temperatures, respectively.



Genes	Experiment 1 ratio	Experiment 2 ratio	Model ratio
rpl23 – rpl2	2.15	2.69	2.37
psbA	0.53	0.55	0.54

Multiple alignment of the *rps20* promoter regions
(rhodophytes, cryptophytes, cyanobacteria):

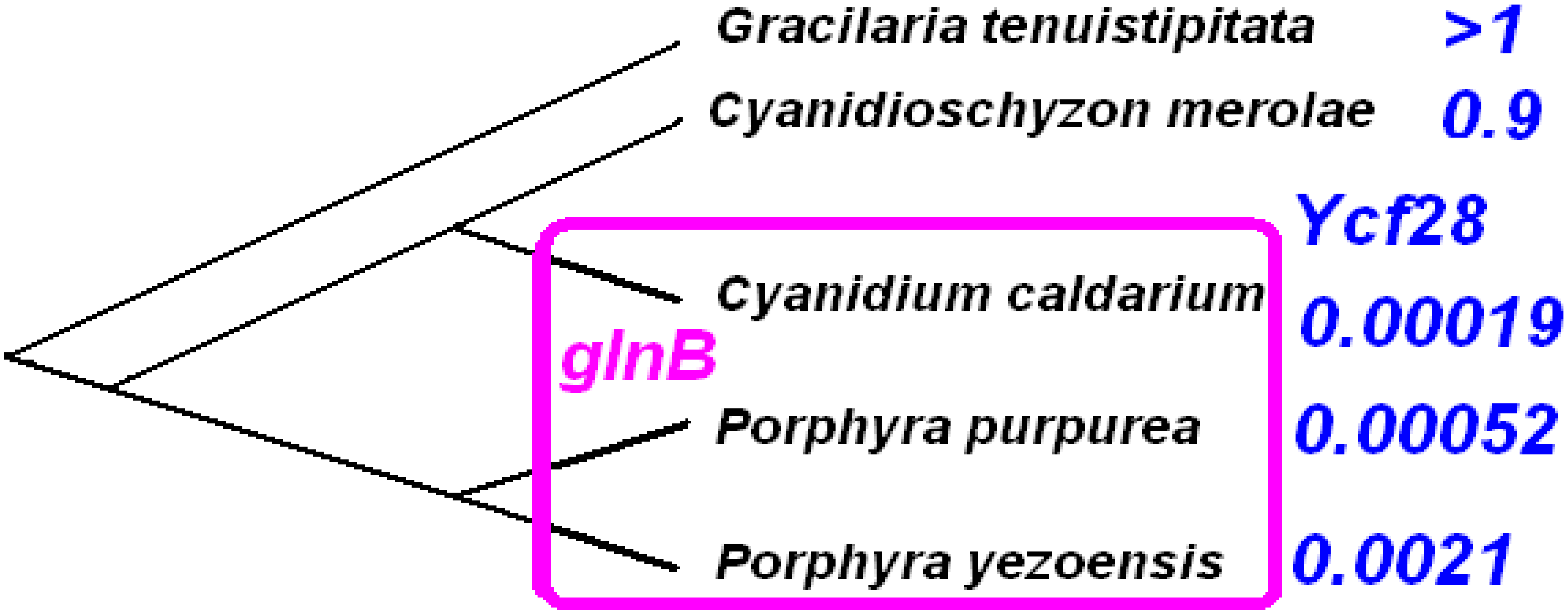


Species	**Promoter upstream *rps20* (*rpsT*)**
Cyanidioschyzon_merolae	`TCTTGCTTTTTGCCATCTGCT=ATTTTATCTTTAT`
Cyanidium_caldarium	`ATTTGTTTATTTTACTTTAAT=ATGATACAGTAAT`
Porphyra_purpurea	`TATTGCCTATTCTTTTTTTTTAATGTTATAATACG`
Porphyra_yezoensis	`TATTGCCTATTGTTTTTCTTTAATGTTATAATACG`
Gracilaria_tenuistipitata	`TCTTGTCTATTTTAATGTATTAATGATATAATCCA`
Guillardia_theta	`AATTTATTCCATTATTTCTTATATGTTATAATCTT`
Rhodomonas_salina	`TTCTTATTC=ATAATTTGTTCTATGTTATAATCAC`
Synechocystis6803	`CATTGTCGT=GGCGTCCCCAAACTGTTAGAATAAC`
	-35box EX -10box

the *rps20* promoter is conserved in cyanobacteria and plastids of red and cryptophyte algae;

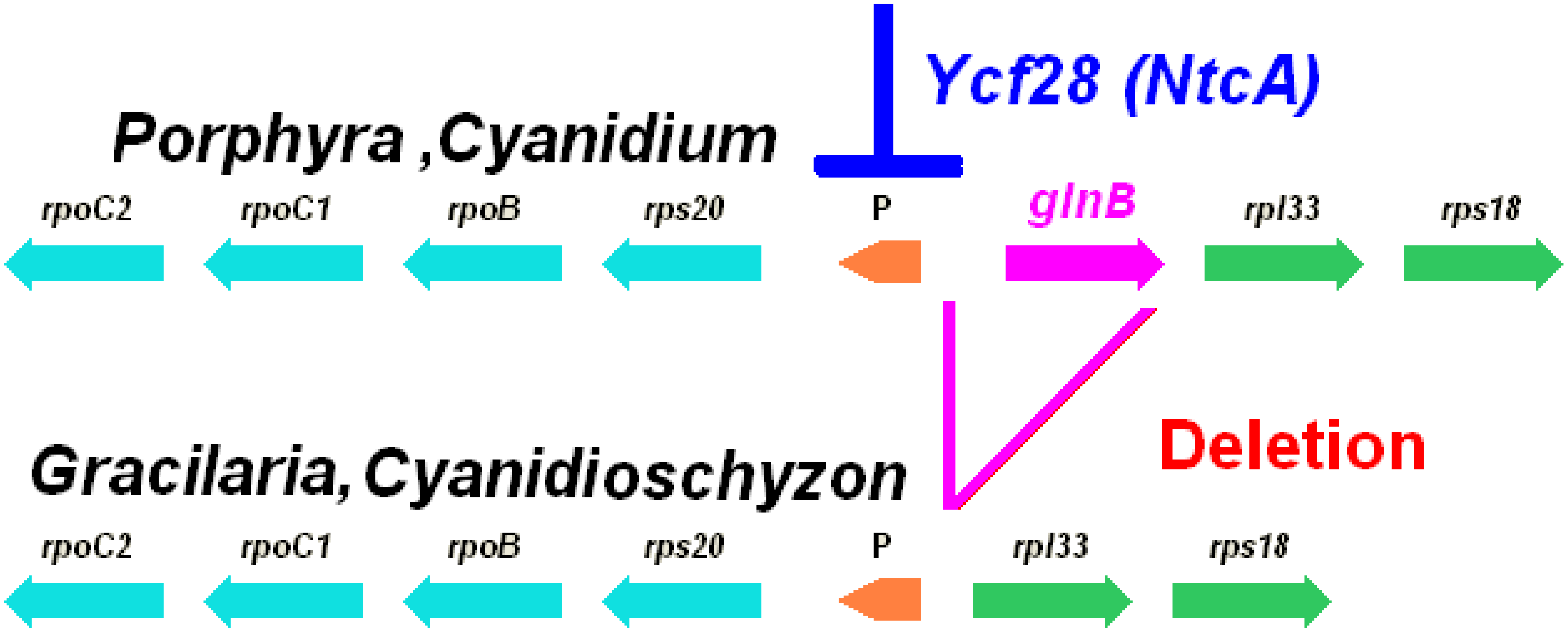
a two-boxed site binds **repressor Ycf28** (promoter close to the consensus); in 3 out of 8 species; in cyanobacteria **Ycf28** activates ***glnB*** transcription

1) when *rps20* is antisense to *glnB* ?,
2) under the presence of transcription factor Ycf28 ?:



GlnB is a factor from the PII family involved in protein-protein interaction. Ycf28 (=NtcA) is a transcription factor in plastids of red algae, i.e. it possesses a **crp-domain**;

e values are the similarity between the cpr-consensus from the Pfam database and Ycf28 domain



the RNA polymerase and ribosome proteins-coding locus; it is conserved among Rhodophyta, **while** *Gracilaria* and *Cyanidioschyzon* do not possess gene *glnB* due to a large chromosome deletion;

we found a repressor binding site near the *rps20* promoter **specifically in** *Porphyra* и *Cyanidium*

The role of the Ycf28 factor (=NtcA):

in *Porphyra* and *Cyanidium* the binding site overlaps the *rps20-rpoBC1C2* promoter, **Ycf28** silences ***rps20*** transcription and thus enhances the transcription of antisense ***glnB*** by relaxing the polymerase competition.

Hypothesis: repression of the *rps20-rpoBC1C2* operon relaxes RNA polymerase competition toward transcription of *glnB* → **transcription activation** of *glnB*.

In cyanobacteria, factor NtcA **activates** the *glnB* transcription through protein-DNA interaction [Muro-Pastor et al, 2003, *Plant Physiol Biochem*.]