

Russian Academy of Sciences,
Institute for Information Transmission Problems of RAS,
Moscow

Laboratory of mathematical methods and models in
bioinformatics

Head D.Sc. Prof. ***V.A. Lyubetsky***

***Gene expression regulation in
actinobacteria and chloroplasts***

In our institute and particularly in the lab we are
interested to study the principal succession:

Psychics, Intellect,, eukaryotic cell,
prokaryotic cell, biomolecules.

Our scope is applications of novel algorithms and models to **real biological** data and comparison of biological results generated using our implementations with **real experimental** evidence.

Our aim is to find novel **biological facts** with computer analyses.

Two research directions:

1) Searching for gene expression regulatory **signals** of various (also **novel**) **types** in the bacterial cell. Those are signals based on protein-DNA interaction, formation of RNA secondary structures and of **other types** acting at the level of both **transcription** and **translation**.

2) Inferring phylogeny (or the same, **evolution**) and gene evolution events in bacteria at the level of species, protein families and regulatory signals.

Method: developing **algorithms and models** of regulatory systems (and of their mechanisms) and of metabolism in bacteria.

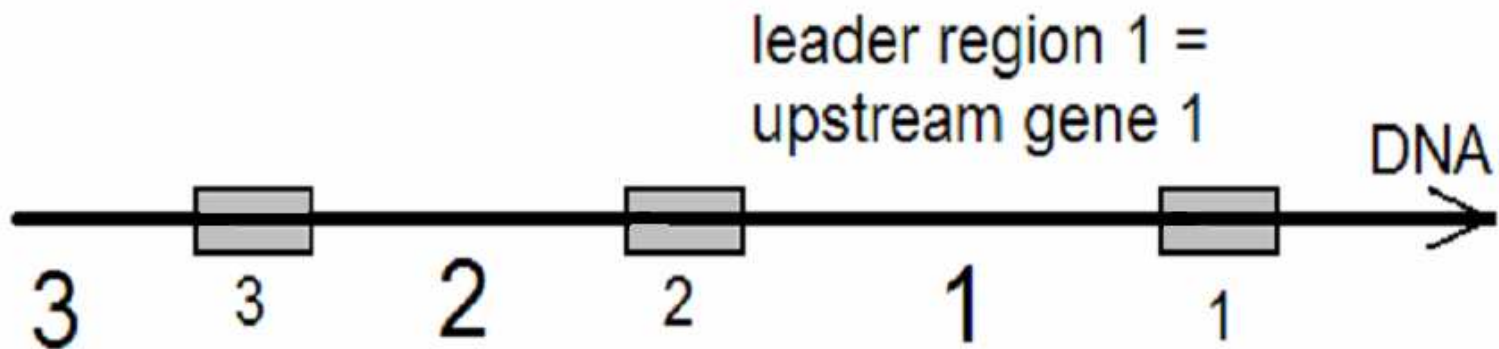
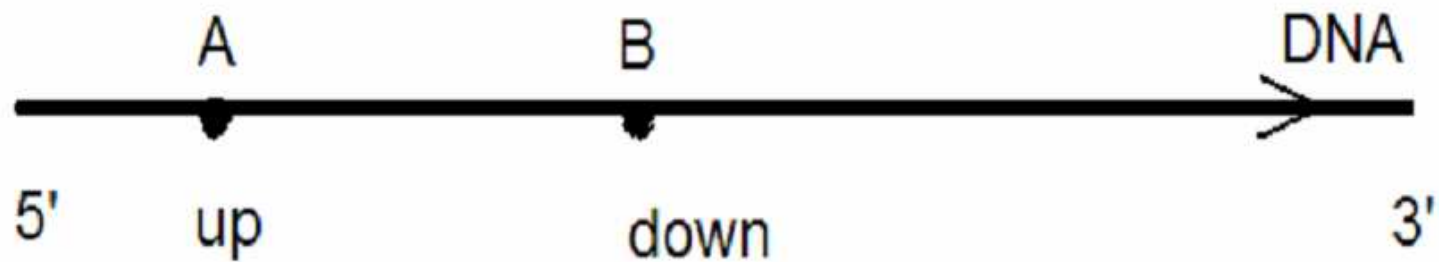
An example is algorithms for and a model of **classic attenuation regulation** of gene expression in bacteria in response to amino acid concentration.

The other two directions:

3) Computer analysis of stochastic processes (“games”) under various types and amount of **information available** to “players” for making a decision.

4) Effective set theory and non-standard set-theory analysis for **effective description** of mathematic **objects** and **structures**.

DNA = deoxyribonucleic acid = a sequence in {G,C,A,U} alphabet. Each letter is called a *nucleotide*. Here we do not discriminate between DNA and RNA.



trpE gene in some Actinobacteria: namely in *Corinebacterium* and *Streptomyces* spp. Gene length is about 1600-1900 nt. **Species** is a group of very closely related bacteria. **Operon** is a group of closely related genes.

Species	Operon structure
<i>C. diphtheriae</i>	<i>trpB</i> ₁ <i>E</i> <i>GDC</i> ₁
<i>C. efficiens</i>	<i>trpE</i> <i>GDCBA</i>
<i>C. glutamicum</i>	<i>trpE</i> <i>GDCBA</i>
<i>S. avermitilis</i>	<i>trpE</i> ₁
<i>S. coelicolor</i>	<i>trpE</i>
<i>S. venezuelae</i>	<i>trpE</i>

**Some characteristic parts of the corresponding
leader regions of operon (= gene) *trpE***

<i>C</i> <i>d</i>	CCGCGGGCCGUUUUCACGCAUUCAUUUCAACAGGCUCGCCUU GUCCAACAAGCAGCGGGCC
<i>C</i> <i>e</i>	GCGGGCCCACGGAUCACCAAGUUGUUUUCACACUGAAGAUUU CAAGGCUCGUGUACUUCGUUCGACGAAGCAGCGGGCCUUUUG UGGUU
<i>C</i> <i>g</i>	GCGAGCCUGACACCUCAAGUUGUUUUCACUUUGAUGAAUUUU UUAAGGCUCGUACUUCGUUCGACGAAGAAGCGGGCCUUUUGU GGUU
<i>S</i> <i>a</i>	CUGCGCGUACGCAAGACUUCGCGAAGGCCGCCCGAGGGGGCGG CCUUUCGUGUUUCC
<i>S</i> <i>c</i>	CUGCGCGCGACUCAAGACUCGCGAAGGCCGCCCGAGGGGGCGG CCUUCGGUGUUUUC
<i>S</i> <i>v</i>	UCGCGCGUACACGGAUCACACGCACAGGCCGCCCGAGGGGGCG GCCUUUCUCG

A triple of nucleotides (= «codon») translates into a new letter («amino acid») according to the Table. Some degenerate designations will also be used **R** = {A; G}, **Y** = {U; C}, **S** = {G, C}, **W** = {A, U}, **N** – wildcard (any letter)

Ala	A	GCN		Leu	L	UUR; CUN
Arg	R	AGR; CGN		Lys	K	AAR
Asp	D	GAY		Met	M	AUG
Asn	N	AAY		Phe	F	UUY
Cys	C	UGY		Pro	P	CCN
Glu	E	GAR		Ser	S	AGY; UCN
Gln	Q	CAR		Thr	T	ACN
Gly	G	GGN		Trp	W	UGG
His	H	CAY		Tyr	Y	UAY
Ile	I	AUA; AUU		Val	V	GUN

Each gene either “works” or “doesn’t work”. In fact, all processes in the cell are stochastic. A gene works stronger, weaker or in the “middle”, *etc.* A **gene** usually has one or more **signals** in its leader region that make it “switched on” or “switched off” (which means to **express** the gene).

Our **goal is to detect these signals** for all genes. By far not all signal types are currently discovered. And as they can vary greatly, we will find not only **signals** but novel **types** of signals as well.

Signal in leader region

“**terminate**”
(= “not working”)

“**antiterminate**”
(= “working”)

Gene

implies **not expressed**

implies **expressed**

The **simplest type** of signal is a **continuous part** (called “**site**”) of the leader region. More precisely, **signal is a set of sites**. This set is constructed by minimizing the pairwise **similarity** in nucleotide content between all sites from the signal. **ONE-box signal** means **each site in it is a ONE-box site**. This type is called *protein-RNA*:



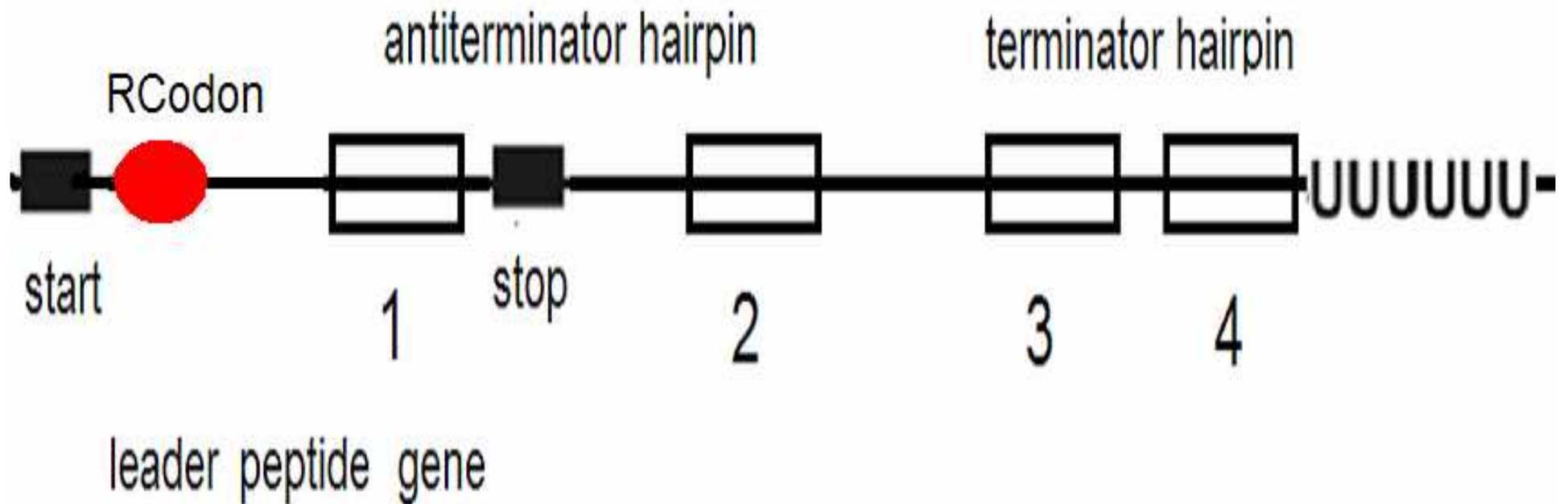
Two-box signal upstream *atpF* gene in plant chloroplasts

<i>Anthoceros formosae</i>	AAUGAAUAAU---AACCUAUGAUGGGAGAGAGAGU
<i>Marchantia polymorpha</i>	AAUGAAAAAA---CGAAAAAAGAGGACAGC***
<i>Adiantum capillus-veneris</i>	AAUAAUJAAU---CCUUCGAGGAGGGAAAAGAAU*
<i>Huperzia lucidula</i>	AAGGAAAAAA---AACCUGUAAUGGGAGAAAAGU*
<i>Psilotum nudum</i>	AAUAAAAGAA---UAGUCAUUAUGGGAGAGGUAAU
<i>Pinus thunbergii</i>	AAUAAGAAAA---UAUCUAUGAGGGGAGAGCGU**
<i>Amborella trichopoda</i>	AAUAAAAAUA---UAUCUAGAAGAGGAGAGUAU**
<i>Arabidopsis thaliana</i>	AAUAAAAAAA---UAGCUAGAAGAGGAGAUUAU**
<i>Atropa belladonna</i>	AAUAAAUAAA---UAUCUAUAAGAGGAGAUCAU**
<i>Calycanthus floridus</i>	AAUCAAAAAA---AUUCUAUAAGAGGAAAGCAU**
<i>Cucumis sativus</i>	AAUAAAAAAA---UAUCUAUAAAAGGAGAUCAU**
<i>Lotus corniculatus</i>	AAAAAAAAAA---UAUCCAUAAGAGGAGAUCAU**
<i>Nicotiana tabacum</i>	AAUAAAUAAA---UAUCUAUAAGAGGAGAUCAU**
<i>Nymphaea alba</i>	GAUACAAAAA---AUAAGAUAAAGAGGAGAGCAU**
<i>Panax ginseng</i>	UAUAAAAAAA---UAUCUAUAAGAGGAGAUCAU**
<i>Spinacia oleracea</i>	AAUAAAAAAA---UAUCUAUAAGAGGAGAUCAU**
<i>Oryza nivara, sativa</i>	AUUAAAAAAA---UAUCUAUAAGAGGAGAGCAU**
<i>Triticum aestivum, Zea m.</i>	GAUCAAAAAA---UAUCUAUAAGAGGAGAGCAU**

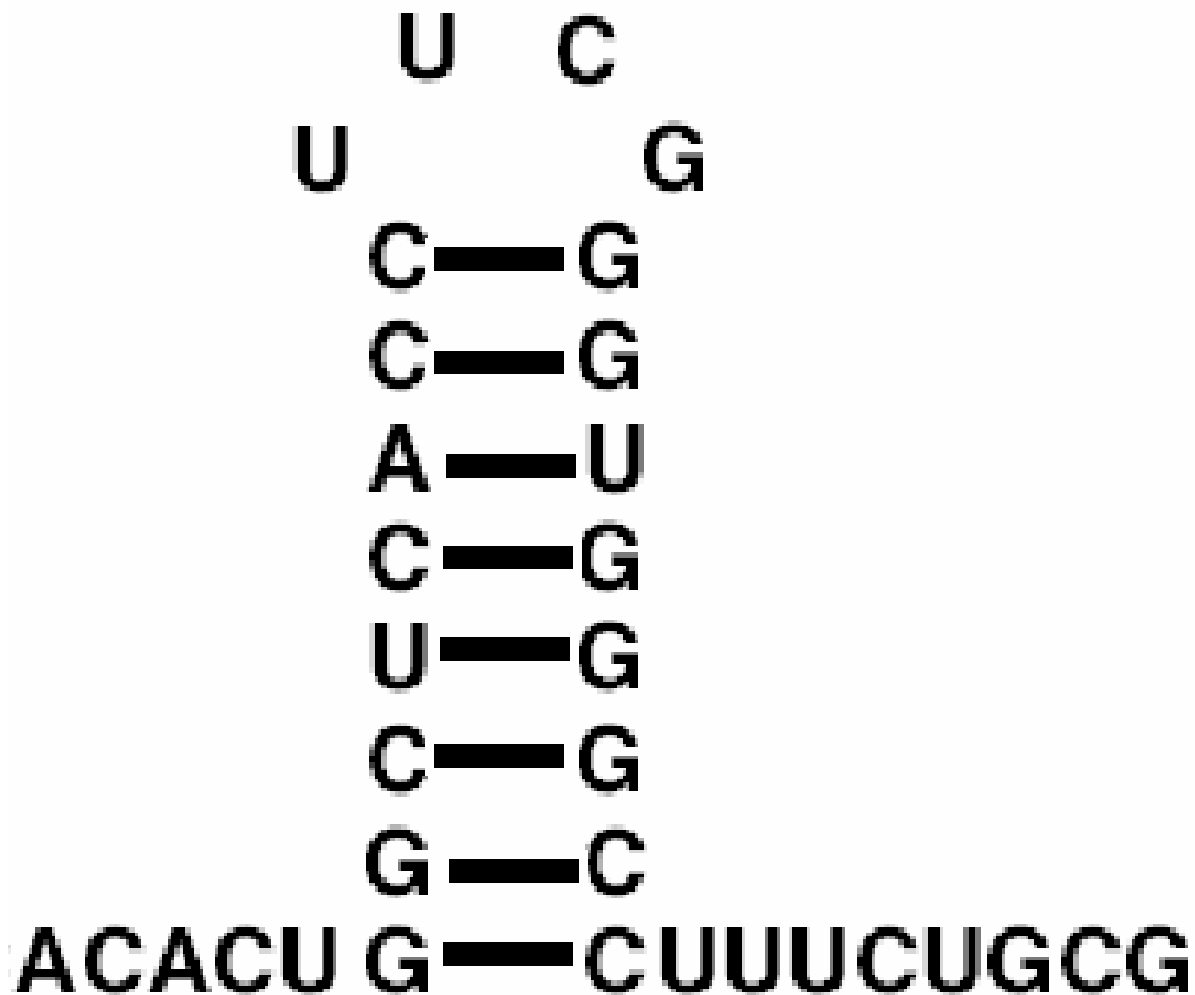
Multi-box signal:

“classical attenuation regulation”.

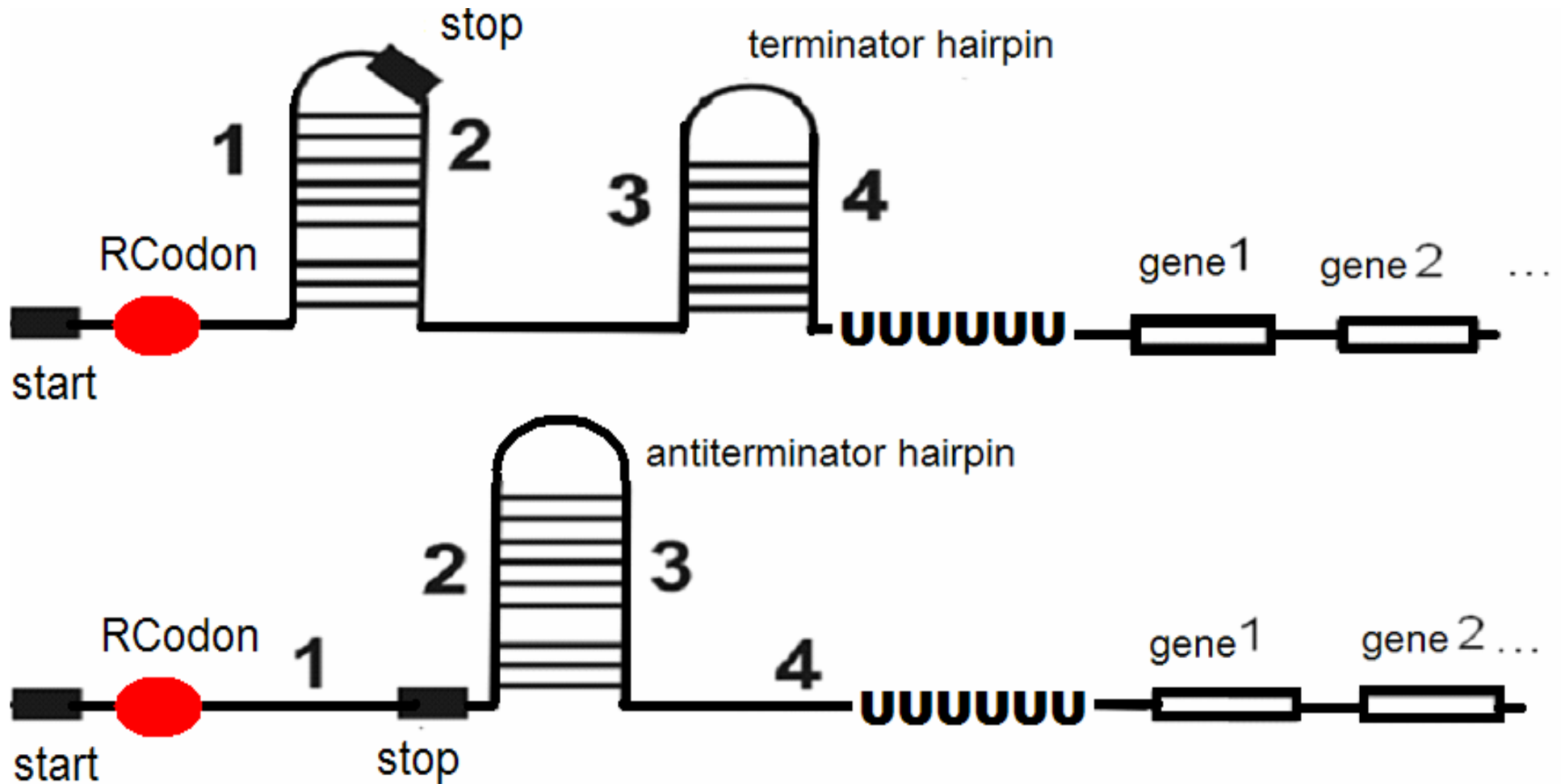
One **site** of this signal is shown;
here each site is highly multi-boxed:



Hairpin is pairing of nucleotides according to the rule: G-C and A-U (G-U allowed). A set of hairpins comprises the **secondary structure** of a region within RNA strand (= string of letters).



The above set of strings and hairpins is sought for **every site** in **CAR**. For this type it's important that combinations of hairpins (1-2, 3-4) and (2-3) are mutually exclusive.



The above Table for *trpE* gene is supplemented with two **related genes** from the same actinobacteria

Species	Operon structure
<i>C. diphtheriae</i>	<i>trpB</i> ₁ <i>E</i> <i>GDC</i> ₁
<i>C. diphtheriae</i>	<i>trpB</i> ₂ <i>A</i> (trp syntase)
<i>C. efficiens</i>	<i>trpE</i> <i>GDCBA</i>
<i>C. glutamicum</i>	<i>trpE</i> <i>GDCBA</i>
<i>S. avermitilis</i>	<i>trpE</i> ₁
<i>S. avermitilis</i>	<i>trpS</i> ₂ (trp-tRNA synt)
<i>S. coelicolor</i>	<i>trpE</i>
<i>S. venezuelae</i>	<i>trpE</i>

Classical attenuation regulation = **CAR**:
W=UGG is trp-codon

Species	Leader peptide & Rcodons
<i>C. diphtheriae</i>	* * * * * MNAHN WW RA * * * * *
<i>C. diphtheriae, trpB2A</i>	* * * * * MNAAF KFW RA * * * * *
<i>C. efficiens</i>	VNNFCQSQGTQ WWW RAR * * * * *
<i>C. glutamicum</i>	VNNSCLSQSTQ WWW RAN * * * * *
<i>S. avermitilis, tprS2</i>	* * * MTRTCTQ QW AA * * * * *
<i>S. avermitilis</i>	* * * MFAHSIQ NWWW TAHPAAH
<i>S. coelicolor</i>	* * * MFAHSTR NWWW TAHPAAH
<i>S. venezuelae</i>	* * * MFAHS?? NWWW TAHPAAH

Classical attenuation regulation = CAR:

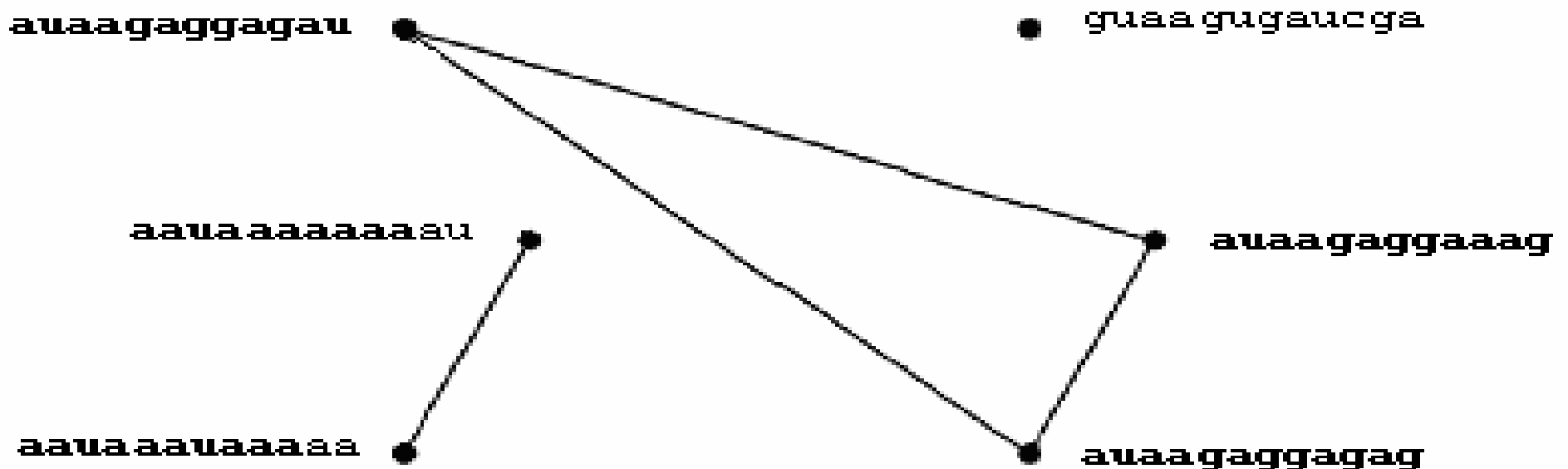
Species	Terminator hairpin is shown in capitals
<i>C. diphtheriae</i>	<u>aac</u> ** <u>AGGCUCGCCUUGU</u> cca***AC*AAGcaGCGGGCCUuuuuguuagc
<i>C. diphtheriae</i>	<u>aacacAAGCCCGC</u> Guau*****C*GCGGGCCUuuucguauau
<i>C. efficiens</i>	***c <u>AAGGCUCG</u> uguaCUUCGUucgACGAAGcagCGGGCCUuu*gugguu
<i>C. glutamicum</i>	uuuu <u>AAGGCUCG</u> u**aCUUCGUucgACGAAGaagCGGGCCUuu*gugguu
<i>S. avermitilis</i>	*** <u>AACGGC</u> * <u>CGCCG</u> ccu*****CGGCGGCCGUUcucguuucu
<i>S. avermitilis</i>	<u>CGCGAAGGC</u> * <u>CGCCC</u> *****gagGGGCGGCCUuCGUGuuucc
<i>S. coelicolor</i>	<u>cgCGAAGGC</u> * <u>CGCCC</u> *****gagGGGCGGCCUUCGguguuuuc
<i>S. venezuelae</i>	cgcac <u>AGGC</u> * <u>CGCCC</u> *****gagGGGCGGCCUuucucg

Classical attenuation regulation = **CAR**:

	Antiterminator hairpin is underlined
Cd	c* <u>gcgggcc</u> * <u>guuuu</u> ***cacgcauuc <u>uuuc</u> ***** <u>aac</u> ** <u>AGGCUCGCCU</u>
Cd	agg <u>cgggccc</u> <u>uuuug</u> ugugagcauucaccacaca <u>uuuug</u> gaa <u>acac</u> <u>AAGCCGCG</u> u
Ce	aag <u>cgggccc</u> <u>caggau</u> caccaagu <u>uuuuc</u> acacuga <u>aguuu</u> ***c <u>AAGGCUCG</u> ugu
Cg	aag <u>cgagccu</u> gacacc <u>uca</u> agu <u>uuuuc</u> acuu**uga <u>uga</u> uuuuuu <u>AAGGCUCG</u> u**
Sa	<u>cggcg</u> * <u>gccgu</u> acacacg <u>uaugu</u> acuc***** <u>AACGGC</u> * <u>CGCCG</u>
Sa	<u>cggcg</u> * <u>gccca</u> <u>cugac</u> ugcgcg <u>u</u> *****acgcaagacuu <u>CGCGAAGGC</u> * <u>CGCCC</u>
Sc	<u>cggcg</u> * <u>gccca</u> <u>cugac</u> ugcgcgcg*****acucaagac <u>ucg</u> <u>CGAAGGC</u> * <u>CGCCC</u>
Sv	<u>cggcg</u> * <u>gccca</u> <u>cugau</u> cgcgcg <u>u</u> *****acac <u>ggauc</u> acacgcac <u>AGGC</u> * <u>CGCCC</u>

One-box signal else: Initial set of nucleotide sequences is used to construct multipartite graph G . In it:

vertices are all strings of fixed length contained in at least one of the sequences. One **part** consists of strings originating from one sequence. **Edges** connect **similar** strings from different parts. **Signal (=clique)** is a subgraph with all its vertices connected pairwise.



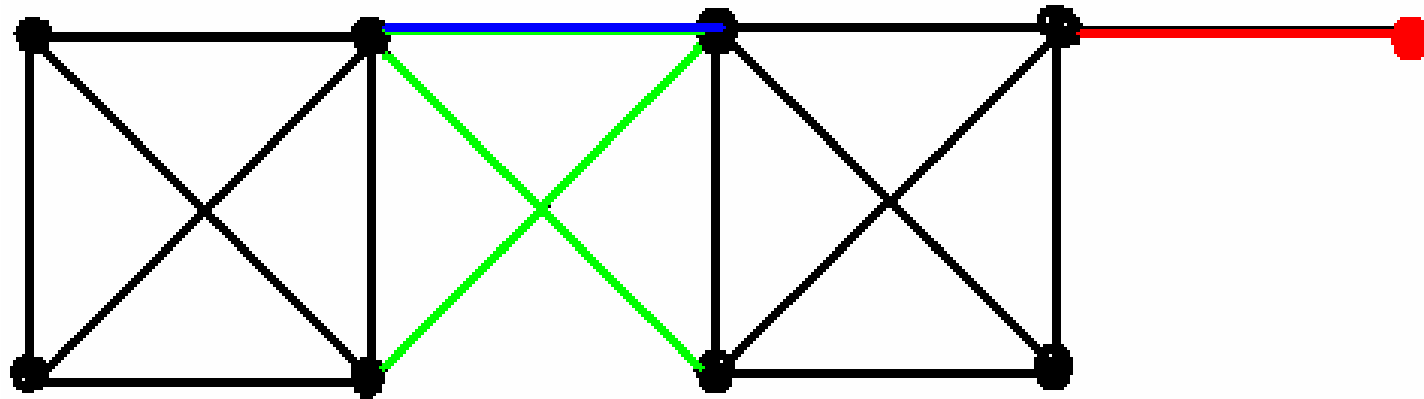
Example: algorithm finds two 4-cliques in G .

Here each part consists of one vertex.

Red vertex is excluded at first step.

Green edges are excluded at second step.

Blue edge is excluded at third step.



Idea of the alg: graph G is pruned to exclude all vertices and edges which can't be in a q -clique.

Algorithm for q -clique finding in more details

1. **Current graph** G' is *initial* graph G and **current list** of q -cliques CL is empty list.
2. Graph G' is pruned to **exclude** all vertices (and all incident edges) that have at least one edge connected with parts of the graph to provide for the sum of parts be strictly less than $q-1$. Edges are further **excluded** if they contained in strictly less than $q-2$ number of 3-cliques or strictly less than $(q-2)(q-3)/2$ number of 4-cliques. Such exclusion is repeated over all vertices and then all edges in G' until possible.
3. When not possible, and *all edges* are **excluded** from G' , the algorithm **halts** and outputs current list CL . If some edges are **still present** in G' , the algorithm searches for vertex R with **degree exactly** $q-1$. If such R **exists**, it is **excluded** from G' together with its incident edges and is verified to form a q -clique with all its adjacent vertices. If so, the q -clique is included in current list CL .
4. If such vertex R is **not found** in G' , an edge in G' shared by minimum number of 3-cliques is excluded from G' .
5. Steps 2-4 are applied to such new G' again until possible.

The originally found **CAR** of branched amino acid synthesis in actinobacteria:

Species	Leader peptide
<i>C. diphtheriae</i>	***MNIIRLVVITTRRLP
<i>C. efficiens</i>	**MTSIRPVVIVAARRLP*
<i>C. glutamicum</i>	***MTIIRLVVVTARRLP
<i>M. avium</i>	*****MLVVI*RRVGA
<i>M. tuberculosis</i>	MDKAGKPGMLVVIGRRVGA
<i>M. bovis</i>	MDKAGKPGMLVVIGRRVGA
<i>M. leprae</i>	*****MLVVICQRVGG
<i>M. marinum</i>	MDTAGTPGKLVVLGRRVVA
<i>S. avermitilis</i>	*****MRTRILVLGKRVG
<i>S. coelicolor</i>	*****MRTRILVLGKRVG

Continued: CAR terminators for branched amino acid synthesis

Species	TERMINATOR is in capitals
<i>C. diphtheriae</i>	*GCCCUCGaCAG***CAccacacaUGCUGAGCGGGGGCuuuccu
<i>C. efficiens</i>	*GCCCUCGACAGUACccaccacaGUGCUGuuUCGAGGGCuuuguu
<i>C.glutamicum</i>	*GCCCUCGaCAACACUcaccacAGUGUUGgaaCGAGGGCuuucuu
<i>M. avium</i>	AACCCUCGugCAGCaca*****aGCUGuCG*GGGGUUuuuu
<i>M.tuberculosis</i>	*ACCCUCGugCAGCagc*****ugaGCUGgCGA*GGGUuuuuu
<i>M. bovis</i>	*ACCCUCGugCAGCagc*****ugaGCUGgCGA*GGGUuuuuu
<i>M. leprae</i>	AACCCUCGugCAGCUag*****ucAGCUGuCGA*GGGUUUuuuu
<i>M. marinum</i>	AACCCUCGUgCAGCagc*****ugaGCUGACG*GGGGUUuuuu
<i>S. avermitilis</i>	uCCCCUCGcuUGCC*****ucacGGCACGAGGGGuuuuuu
<i>S. coelicolor</i>	uCCCCUCGcuUGCC*****uuacGGCACGAGGGGuuuuuu

The role of actinobacteria:

Industrial producers of amino acids (*Corynebacterium glutamicum*, *Corynebacterium efficiens*) and antibiotics (*Streptomyces* spp.), Human microflora components (*Bifidobacterium longum*, *Propionibacterium acnes*), Dangerous human pathogens (*Corynebacterium diphtheriae*, *Mycobacterium* spp.).

Chloroplasts of plants and algae and **Cyanobacteria** are very important from practical view as well .

Below we will demonstrate some **novel regulation types** as well as **signals** (neither protein-RNA signals, nor CARs), particularly, for the following six chloroplast genes *atpF*, *clpP*, *petB*, *psaA*, *psbA*, *psbB* **in many plants and algae**.

The signals belong to **several regulation types**, which will be specified below.



*Marchantia
polymorpha*



Adiantum capillus-veneris



*Pinus
thunbergii*

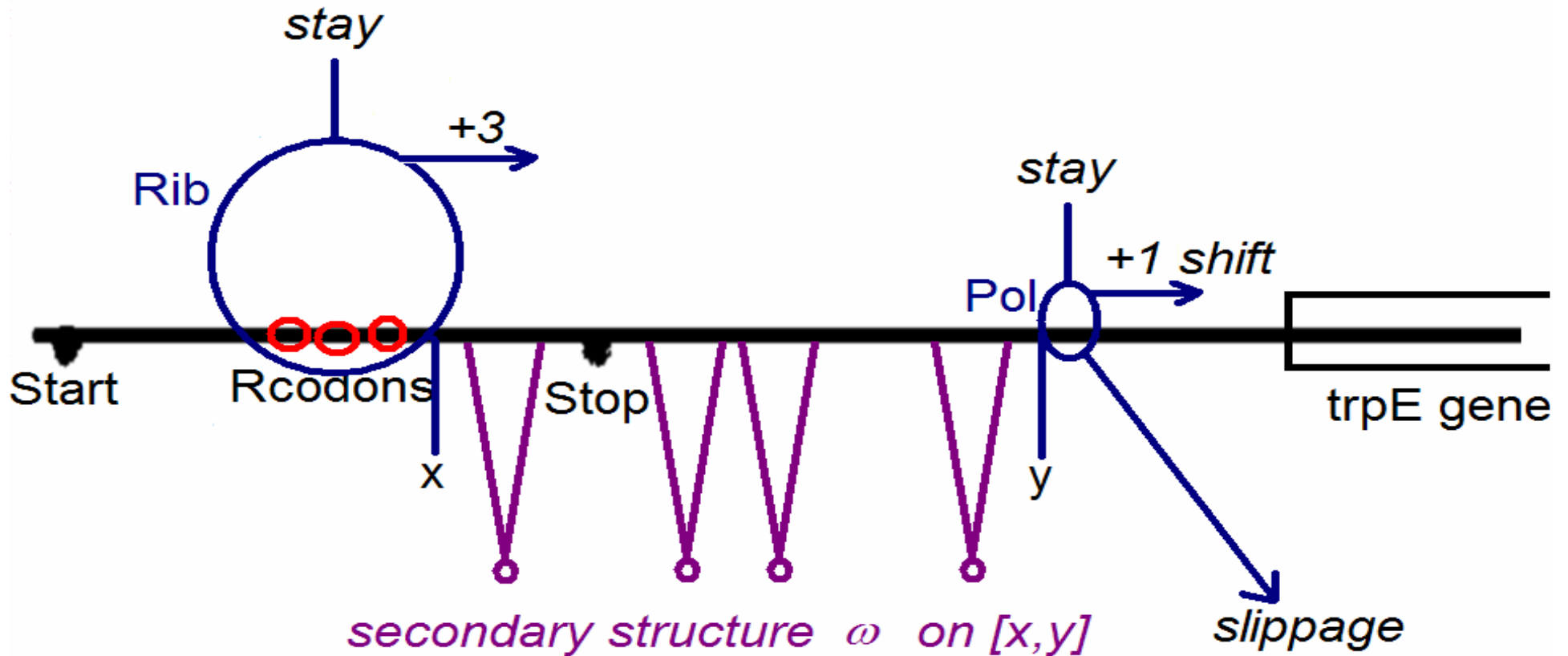
We developed a **mathematic model** of classical attenuation regulation (= **CAR**) of gene expression in bacteria.

The model **predicted the same CAR** initially found by searching for signals and, particularly, with the clique search algorithm.

The model **utilizes one initial sequence**, thus eliminating the need to construct multiple alignments, signals and analyze sequence profiles!

Mechanism of CAR:

In the cell:



The speed of *Rib* on Rcodons depends on the level of the **concentration c** of tryptophan = the acid of trpE gene.

The speed of *Rib* on Rcodons depends on the level of the **concentration c** of tryptophan = the acid of trpE gene.

Polymerase slippage means the signal of termination for the corresponding gene.

Hence the probability $p(c)$ of slippage is equal to the level of expression of the gene.

We want to determine the value $p(c)$, for each **argument c .**

Therefore $p(c)$ is the probability of **polymerase slippage** under fixed value of **concentration c** .

Now description of the model:

Transitions allowed in our model of **CAR**:

(1) **Polymerase *shifts*** by 1 nucleotide or ***slips*** from RNA, otherwise ***stays*** on same position;

(2) **Ribosome *shifts*** by 3 nucleotides, otherwise ***stays*** on same position;

(3) **Secondary structure rearranges** within the region between ribosome and polymerase, i.e., current secondary structure ω alters into new structure ω' .

Each of the three transitions is described as a **Poisson flow**:

$$P_t(u) = (kt)^u \cdot \exp(-kt) / u!$$

- probability of u events occurrence in time t , where k – flow **intensity** (=«**rate constant**»), i.e., average number of random events per time unit.

Event probability is then given by

$$1 - P_t(0) = 1 - \exp(-kt)$$

Only **rate constants** are now to be defined for each event! Those are given in **square brackets** on next two slides.

Transition probabilities in our model:

(1) **Polymerase shift** $1 - \exp\left(-\left[40 - F(\omega)\right] \cdot t\right)$

Polymerase slippage $1 - \exp\left(-\left[\frac{F(\omega)}{4}\right] \cdot t\right)$

Here $F(\omega) = \frac{\delta}{L_1^2 \cdot (p(\omega) - p_0)^2 + 1} \cdot \exp\left(-\frac{r}{r_0}\right)$

(2) **Ribosome shift** $1 - \exp\left(-\left[\frac{45 \cdot c}{c_0 + c}\right] \cdot t\right)$

Let us assume $c_0 = 1$

Definition of p :

For hairpin Ω consisting of the handle with the length h , i.e. the number of its pairs, and the loop with l as its length, p is defined as follows:

$$\operatorname{tg}(p \cdot h) = \frac{2}{p \cdot l}, \quad 0 < p \cdot h < \frac{\pi}{2}$$

For hairpin \mathcal{Q} with a set of paired segments separated by bulges and terminal loop let us define: if \mathcal{Q} contains s segments of length h_1, \dots, h_s ,

$s-1$ bulges of length l_1, \dots, l_{s-1}

and a loop of length l , then

$$p = \bar{p} \cdot \left(1 - \frac{1}{2h + l \cdot \sin^2(\bar{p} \cdot h)} \cdot \sum_{i=1}^{s-1} l_i \cdot \sin^2(\bar{p} \cdot h(i)) \right)$$

Here $h(i) = h_1 + \dots + h_i$ and $h = h(n) = h_1 + \dots + h_n$

and \bar{p} is found with equation

$$\text{tg}(\bar{p} \cdot h) = \frac{2}{\bar{p} \cdot l}, \quad 0 < \bar{p} \cdot h < \frac{\pi}{2}$$

Transition probabilities (continued):

(3) **Secondary structure *rearrangement***

from *state* ω into state ω' within the region between ribosome and polymerase:

$$1 - \exp\left(-\left[\kappa \cdot \exp\left(\frac{1}{2}\left((G_{loop}(\omega) + G_{hel}(\omega)) - (G_{loop}(\omega') + G_{hel}(\omega'))\right)\right)\right] \cdot t\right)$$

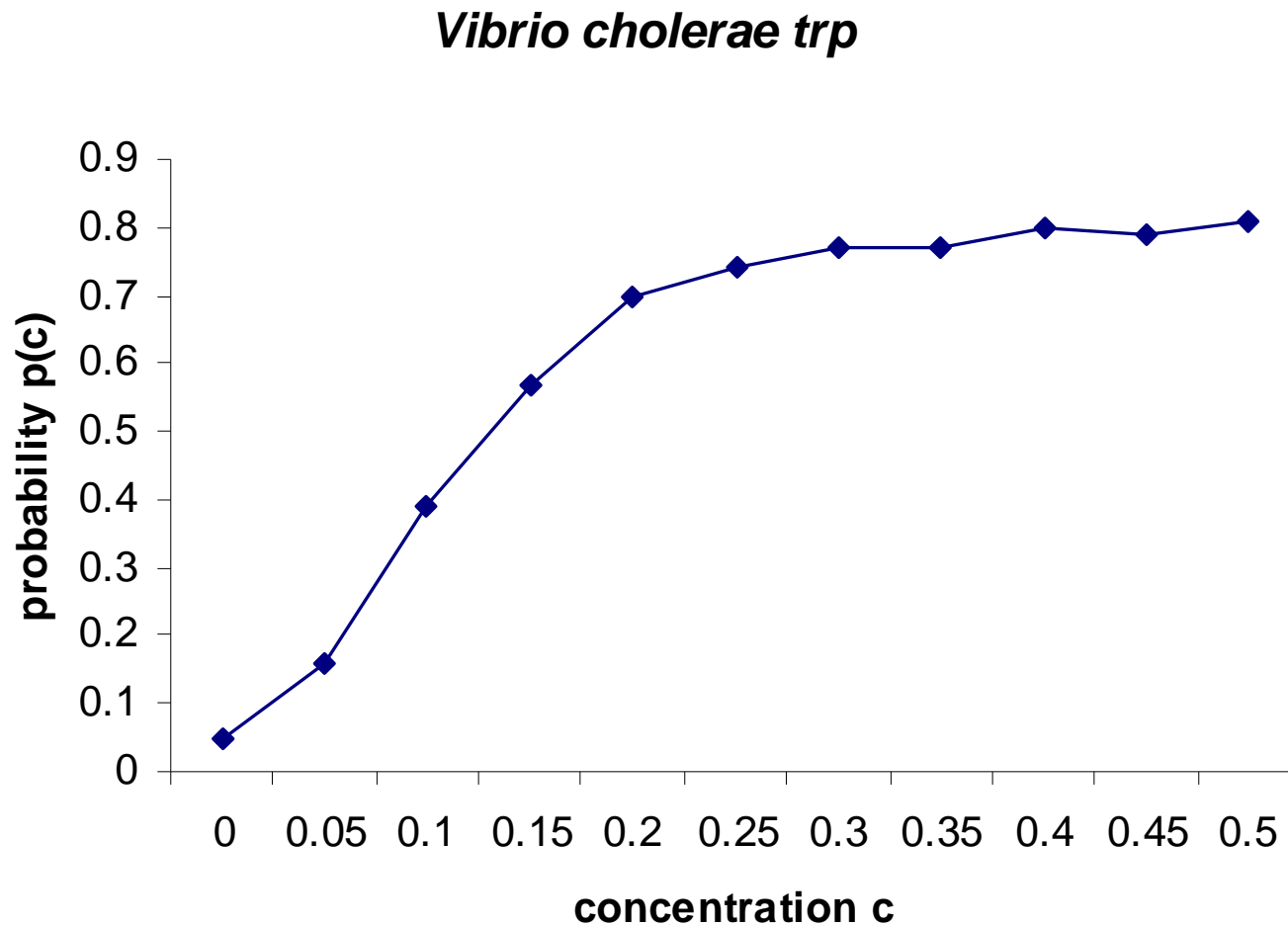
where $G_{loop}(\omega), G_{loop}(\omega') > 0$

- total loop free energy of ω and ω' ,

$$G_{hel}(\omega), G_{hel}(\omega') < 0$$

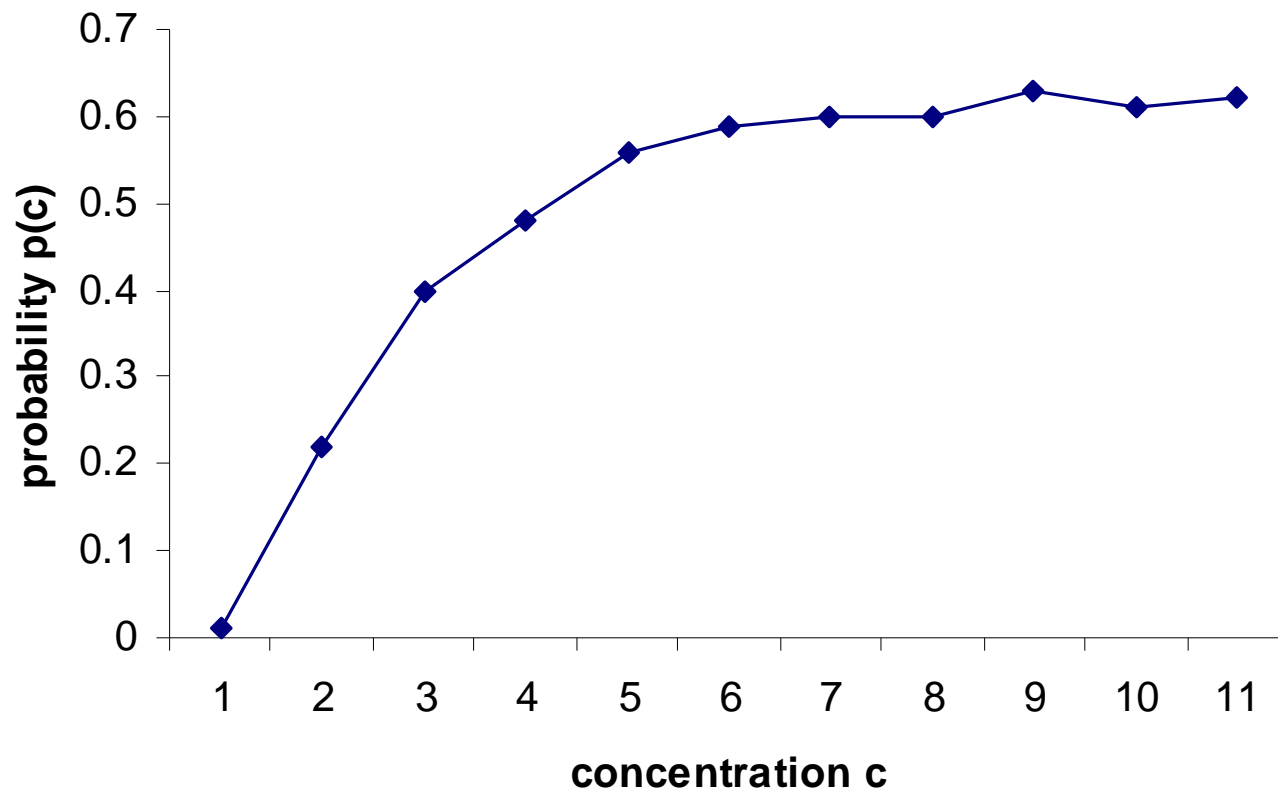
- total loop free energy (stacking) of neighboring base pairs in ω and ω' .

Model output for tryptophan biosynthesis regulation in *Vibrio cholerae* (gamma subdivision):



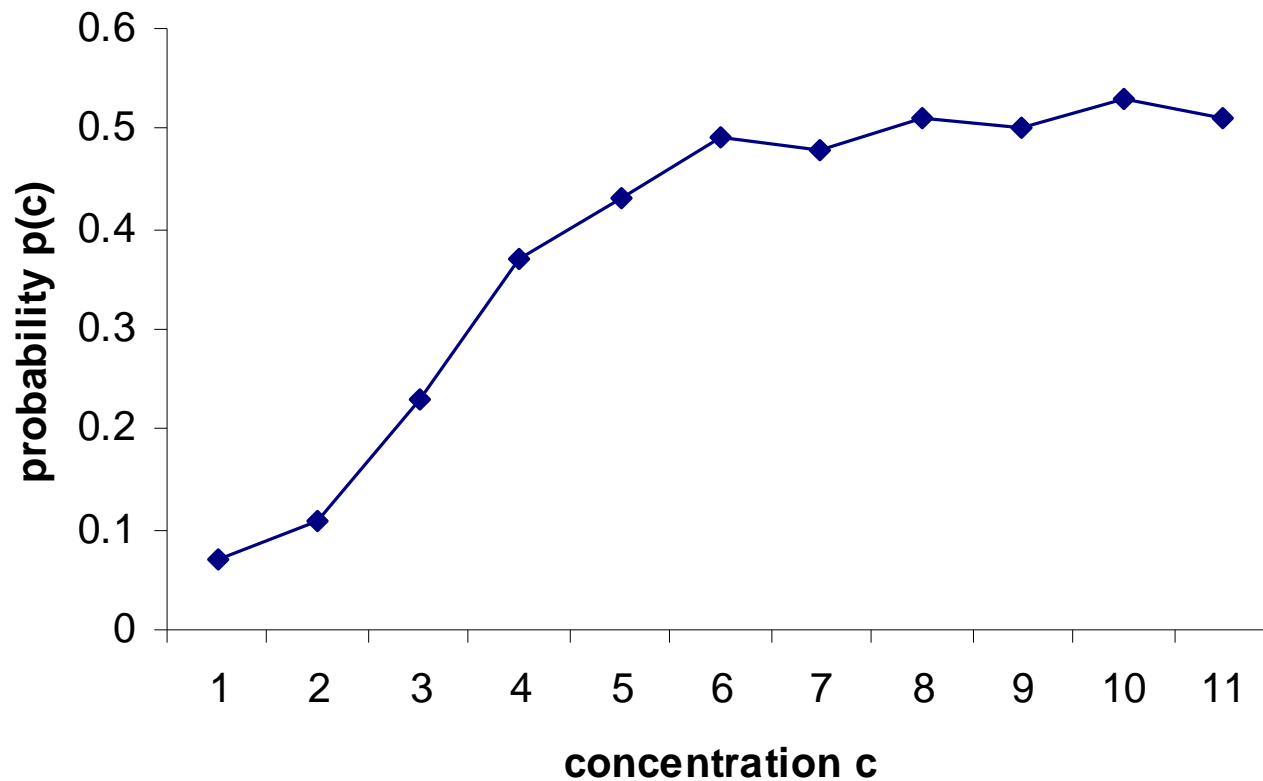
Model output for tryptophan biosynthesis regulation in *Rhodopseudomonas palustris* (alpha subdivision):

Rhodopseudomonas palustris, trp



Model output for tryptophan biosynthesis regulation in *Sinorhizobium meliloti* (alpha subdivision):

Sinorhizobium meliloti, trp



Termination probability $p(c)$ under all the same parameters for leader region upstream *trpE* and its “mutation” (in **only left box antiterminator hairpin and the end of leader peptide gene**, correspondingly):

Gene	Species	Concentration c										
		.00	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
<i>trpE</i>	<i>E. coli</i>	.34	.46	.54	.68	.70	.70	.71	.73	.75	.75	.74
<i>Mutant</i>		.71	.73	.83	.80	.77	.78	.69	.76	.74	.79	.82
<i>trpE</i>	<i>V. cholerae</i>	.05	.16	.39	.57	.70	.74	.77	.77	.80	.79	.81
<i>Mutant</i>		.80	.83	.77	.76	.79	.76	.79	.87	.82	.77	.76

Leader region upstream *trpE* and its “mutation”
in antiterminator hairpin region in *E. coli*

trpE___ATGAAAGCAATTTTCGTACTGAAAGGTTGG

Mutant_ATGAAAGCAATTTTCGTACTGAAAGGTTGG

trpE___TGGCGCACTTCCTGAAACGGGCAGT*GTAT

Mutant_TGGCGCACTTCCTGAAAATCG*A*TCG*A*

trpE___TC*A*CCAT*GCGTAAAGCAATCAGATACC

Mutant_TCGATCGATCGCGTAAAGCAATCAGATACC

trpE___CAGCCCGCCTAATGAGCGGGCTTTTTTTTGG

Mutant_CAGCCCGCCTAATGAGCGGGCTTTTTTTTGG

Leader region upstream *trpE* and its “mutation”
in *Vibrio cholerae*

trpE___ ATGTTACAAGAATTTAACCCCAAACCATAAA

Mutant_ ATGTTACAAGAATTTAACCCCAAACCATAAA

trpE___ CCCAATTTTCAGTCCAGCGGATGCTGAACTG

Mutant_ CCCAATTTTCAGTCCAGCGGATGCTGAACTG

trpE___ GCTTGGTGGCGCACTTGGACAAGTTCTTGGT

Mutant_ GCTTGGTGGCGCACTTGGACAAGTTCTTGGT

trpE___ GGGCTC*A*CGTGTATTTCTAAGTTTAGA

Mutant_ *GGATCGATCG*****ATCGTTCTAA******ATCGA****

trpE___ TACTC*ACACACCTAGCCCGCCAACCTTGA**

Mutant_ **TCGA***TCG**ACACACCTAGCCCGCCAACCTTGA**

trpE___ GCGGGCTTTTTATTGGTTTTT

Mutant_ GCGGGCTTTTTATTGGTTTTT

Remember: a **set** of strings and hairpins determines the gene **regulation type**.

In this **set**:

alternative combinations determine either **presence** or **absence** of gene expression.

An example of regulation type is the above described **CAR**.

SECOND PART of my report is about novel regulator signals in: *actinobacteria*, *chloroplasts* of plants and algae and *cyanobacteria*.

Novel Rho-mediated regulation type
is originally proposed for cysteine
biosynthesis:

No terminator. Regulation requires dedicated Rho-protein that hampers the ribosome attachment: when it binds gene is not expressed, otherwise it is.

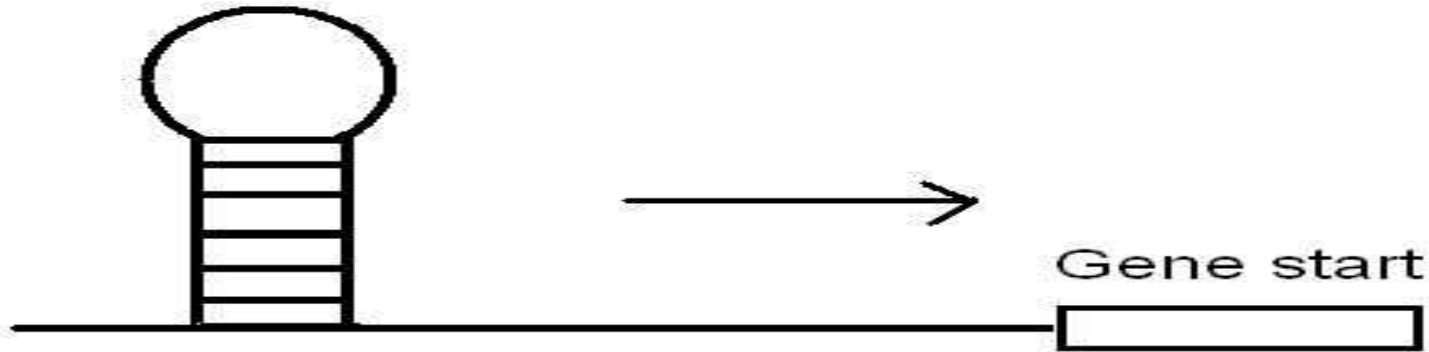
The mechanism: ribosome stays on the leader peptide and overlaps with the Rho protein binding site. Otherwise, it moves and opens the site allowing for binding of Rho-protein.

Cases of **this** Rho-mediated regulation of cysteine biosynthesis in actinobacteria

Species	Leader peptide
<i>M. avium</i>	MQHRLQPRFAPSRCLVVACCCCCR
<i>M. bovis</i>	MQQAIQLRFILPRRLAVGCCCC***
<i>M. tuberculosis</i>	MQQAIQLRFILPRRLAVGCCCC***
<i>M. leprae</i>	MHQSTQPRFVFTRRFTVDCYCRCC*
<i>M. marinum</i>	MQQAAQLSFVLTRCPAVDCCCC***
<i>P. acnes</i>	*****MTSAMMVCICRCCC*
<i>B. longum</i>	*****MQIISCCCR*

Other two **novel types** (= mechanisms) of regulation were originally proposed:

1st) Ribosome binding site is overlapped by a **hairpin** (gene not expressed) **or such hairpin is absent** (gene expressed).



2nd) The hairpin is always present in the binding site of the regulator protein that prevents the ribosome from binding. This **protein is either present or not**.

CHLOROPLASTS (**protein-mediated type = 2nd**)

Division	<i>atpF</i>	<i>clpP</i>	<i>petB</i>	<i>psaA</i>	<i>psbA</i>	<i>psbB</i>
Euglenozoa	-S	-	-S	-S	-S	-S
Bacillariophyta	-	-	-	+	+	-
Cryptophyta	-	-	-	+	+	-
Rhodophyta	-	-	-	+ -	+ -	+ -
Chlorophyta	-	-	-	+ -S	+S	-
Streptophyta	-	+ -S	-S	+	+ -	+ -
Anthoceroophyta	+s	+S	+S	+	+	+
Hepatophyta	+s	+S	+S	+	+	+
Lycopodiophyta	+s	+S	+S	+	+	+
Pteridophyta	+s	+S	-S	+	+	+
Psilophyta	+s	+S	+S	+	+	+
Pinophyta	+s	+	+S	+	+S	+
Magnoliophyta	+s	+S	+S	+	+	+

Table legend:

- Signal found in neither species from the division,
- + Signal found in all species from the division,
- + - Signal found in some species from the division,
- S** Gene expression requires splicing (i.e., excision of introns and ligation of exons or ligation of mRNAs from several genes [trans-splicing])

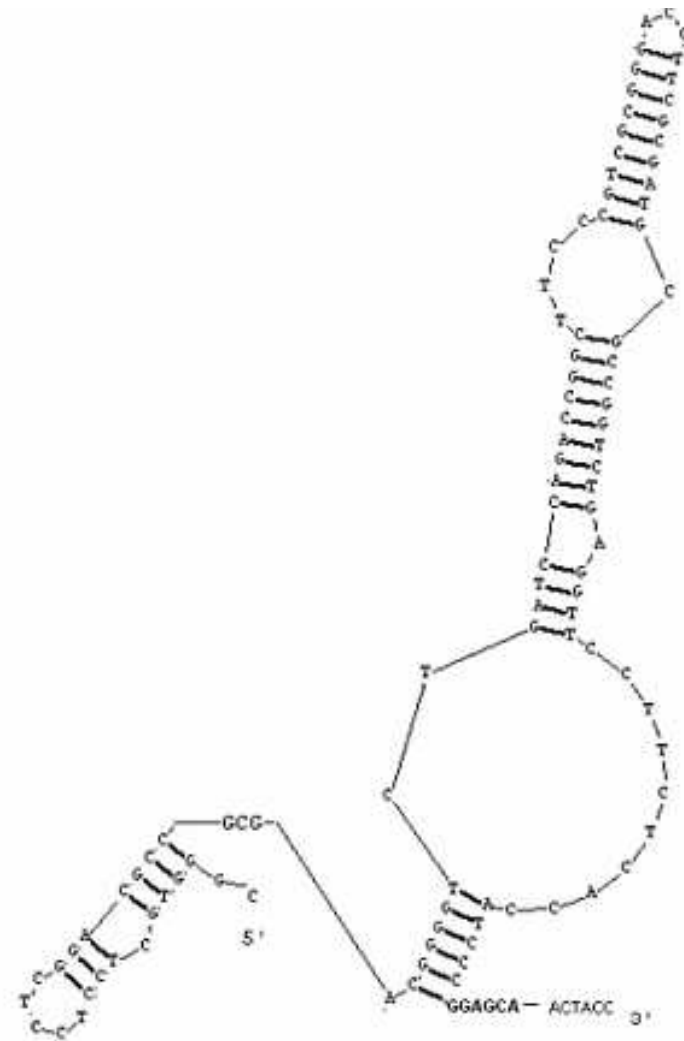
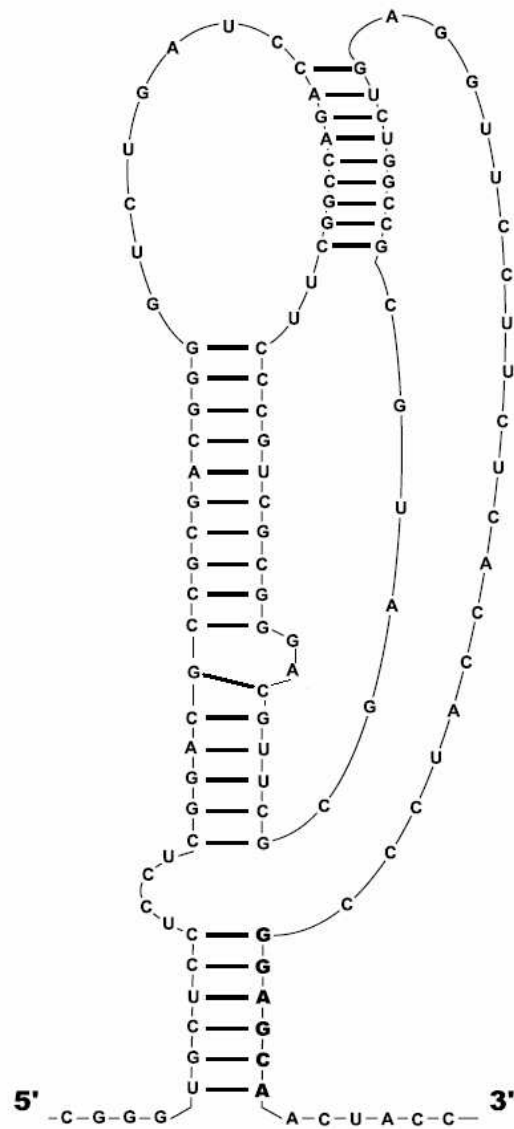
Examples of original biological hypotheses based on original results:

- **Delay** in ribosome binding facilitates **intron splicing** in chloroplast genes *atpF*, *clpP*, *petB*.
- Photosystem gene *psbA* regulation appeared in evolution before the acquisition of introns.
- Photosystem genes, *psaA*, *psbA*, *psbB*, regulation appeared **early** in evolution of chloroplasts.

Another **novel regulation type** (= mechanism) was originally proposed that is based on hairpin called **LEU-element** (next slide).
So called **LEU-regulation**.

LEU-elements were first discovered upstream of *leuA* genes of leucine biosynthesis in various actinobacteria.

LEU-element in *Mycobacterium bovis*



The regulation mechanism is based on overlapping the ribosome binding site (RBS) by LEU-element in **one** of its **two alternative** conformations.

And **the first one is conservative!**

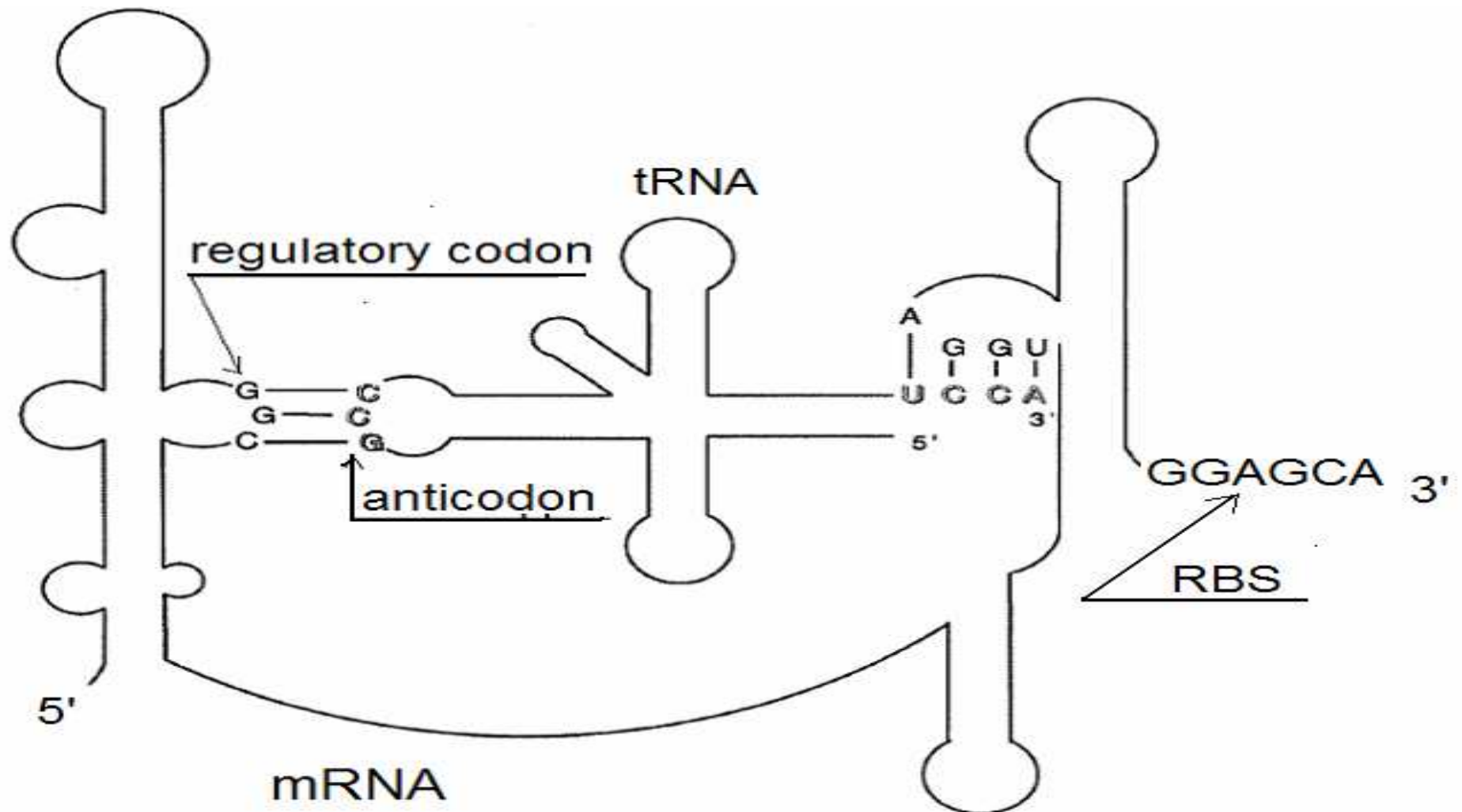
T-box as a **novel regulation type** (= mechanism):
also based on RBS overlap and
proposed by H. Putzer and our group:

T-box stabilized by unloaded tRNA that facilitates
ribosome binding and therefore **gene
expression**.

Otherwise, a part of T-box forms a hairpin that
now prevents ribosome from binding, and
precludes gene expression.

This type is found upstream of gene *ileS* in many
actinobacteria and upstream of gene *alr3806* in
cyanobacterium *Nostoc*.

A combination of strings and hairpins for a conventional T-box:



Distribution of LEU-elements and T-boxes in *actinobacteria* (**hairpin-based regulation types = 1st**)

Genus	Leucine biosynthesis, <i>leuA</i>	Isoleucyl-tRNA synthetase, <i>ileS</i>
<i>Actinomyces</i>	LEU	T
<i>Bifidobacterium</i>		T
<i>Corynebacterium</i>	LEU	T
<i>Kineococcus</i>	LEU	T
<i>Leifsonia</i>	LEU	
<i>Mycobacterium</i>	LEU	T
<i>Nocardia</i>	LEU	T
<i>Propionibacterium</i>		T
<i>Rubrobacter</i>		T
<i>Streptomyces</i>	LEU	T
<i>Thermobifida</i>	LEU	

We discovered novel thiamine riboswitches (**hairpin-mediated regulation type**), e.g. upstream ABC transport protein YkoE gene in actinobacteria:

Brevibacterium linens – new case

Kineococcus radiotolerans – new case

Leifsonia xyli

Propionibacterium acnes – new case

Thermobifida fusca

Corynebacterium diphtheriae

Corynebacterium glutamicum

For **CAR**, the originally found leader peptides upstream leucyl-tRNA synthetase genes in all *Streptomyces* spp. are identical. This is the following string in amino acid alphabet:

M R A V R L L L S E P R