Laboratory of mathematical methods and models in bioinformatics
Institute for Information Transmission Problems
Russian Academy of Sciences

**V.A. Lyubetsky, A.V. Seliverstov**

# Mathematical problems in biological evolution and molecular regulation

lyubetsk@iitp.ru, slvstv@iitp.ru

http://lab6.iitp.ru/

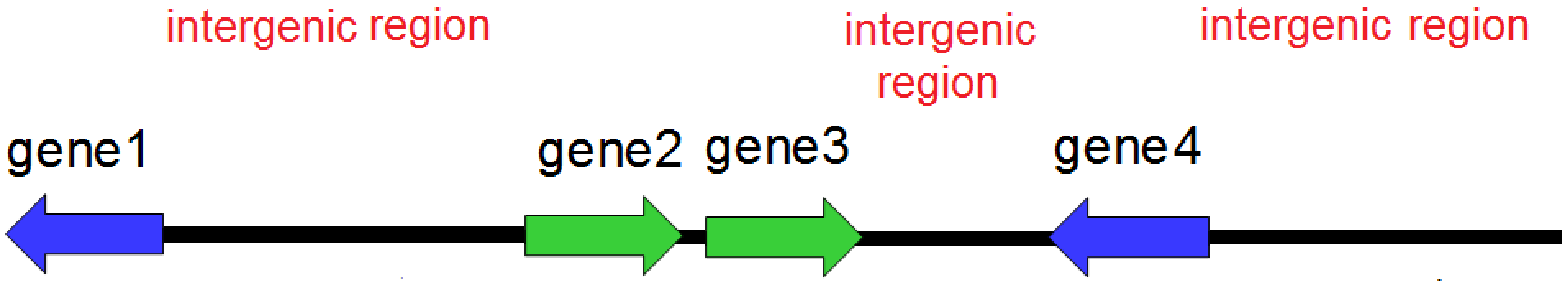
**The layout**. Experimental evidence (measurements, observations etc.) related to molecular biological processes is extracted from public data and analyzed to find

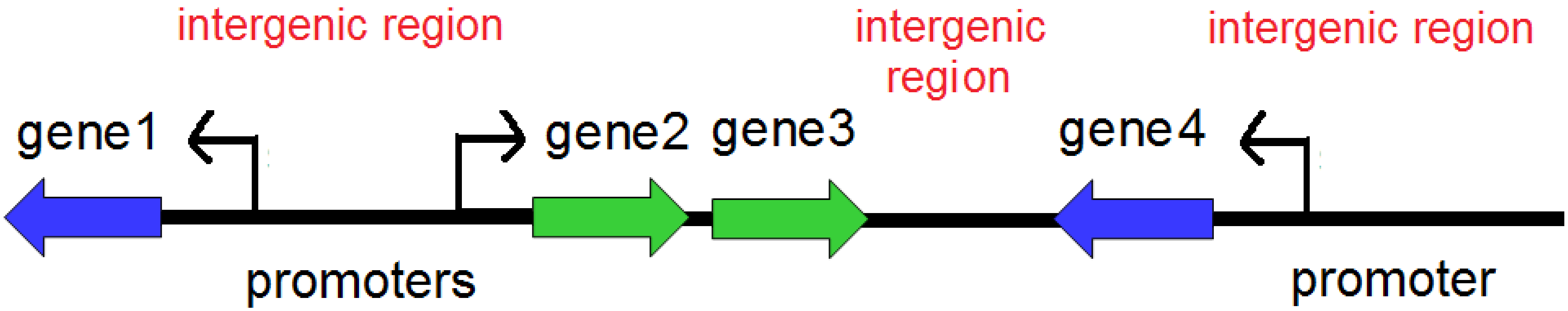
a “mathematical explanation” = a “model”.

Rules of "molecules' behavior" are sought for that best describe the experimental data and predict yet not obtained measurements. Otherwise, the model can be purely mathematical to optimize a certain functional to describe and predict biological objects.

This research includes: **1)** accurate formulation of the model and its computer verification to reproduce the known measurements and predict some unknown; **2)** mathematical studies of the model. We will exemplify some of our models, for which point 1) is true, but point 2) is uncertain. The latter is a general problem!

**Research subject: Given** is a **sequence** of typically 3 millions - 6 billions of characters in the 4-letter alphabet {A,C,T,G}. It contains many **genes, which are shorter segments** $[a_i, b_i]$, each with a **direction** (i.e., **vectors**)





In intergenic regions there are **promoters,** which are also **segments** $[c_j, d_j]$ of a **certain type**, each with a **direction** (i.e., **vectors** as well). Promoter examples:

**human case CAAACCCCAAAGACA**

**bacterial caseTTGACA -17..18- TATAAT -4..7- A(or G)**

Thus, all genes and all promoters are **vectors**.

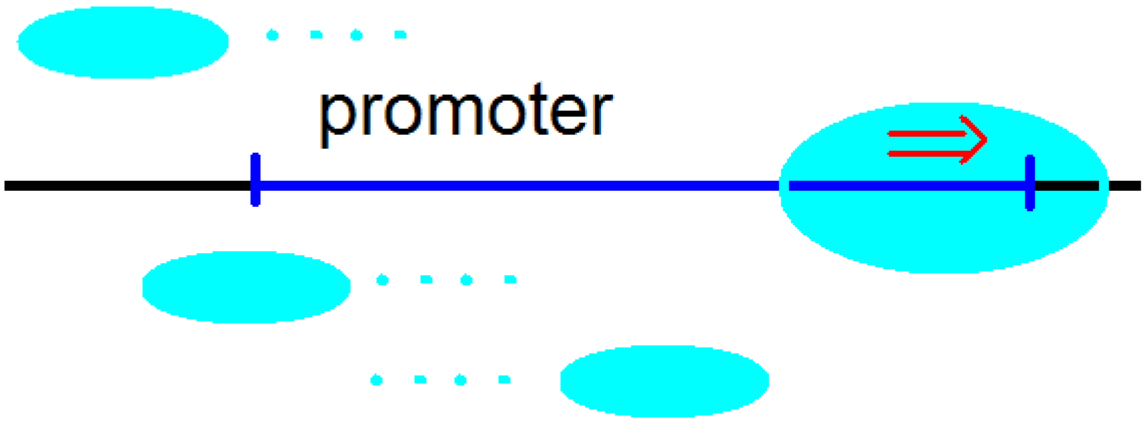
So, a **<u>system</u> of vectors on a fixed <u>sequence</u> is given**

Specific molecular machines (=**polymerases**) first **bind** to the sequence only at **promoters** and then **slide** along the sequence.

A polymerase after **binding to its promoter** **moves** to the **direction** of the promoter and **reads** **only genes co-directed** with the promoter (as co-directional vectors). Similar to a drive read head

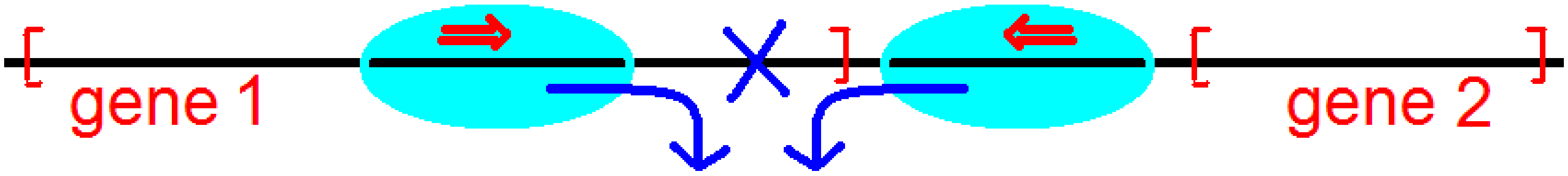
**Attempts** to bind a promoter are allowed to form a Poisson process, with a polymerase moving at a predetermined rate fixed for each type (e.g., 42 letters/sec) **until colliding** with another polymerase

The promoter is ***available*** if **none of polymerases overlap with it**. **Binding occurs** only if the promoter is available:



Each promoter, for each polymerase type, is characterized by the ***intensity* of binding attempts**

If two polymerases moving in **the same direction** **collide**, their rates become equal to that of the leading polymerase until it is attached to the sequence. In case of a **front collision** both polymerases **detach**:

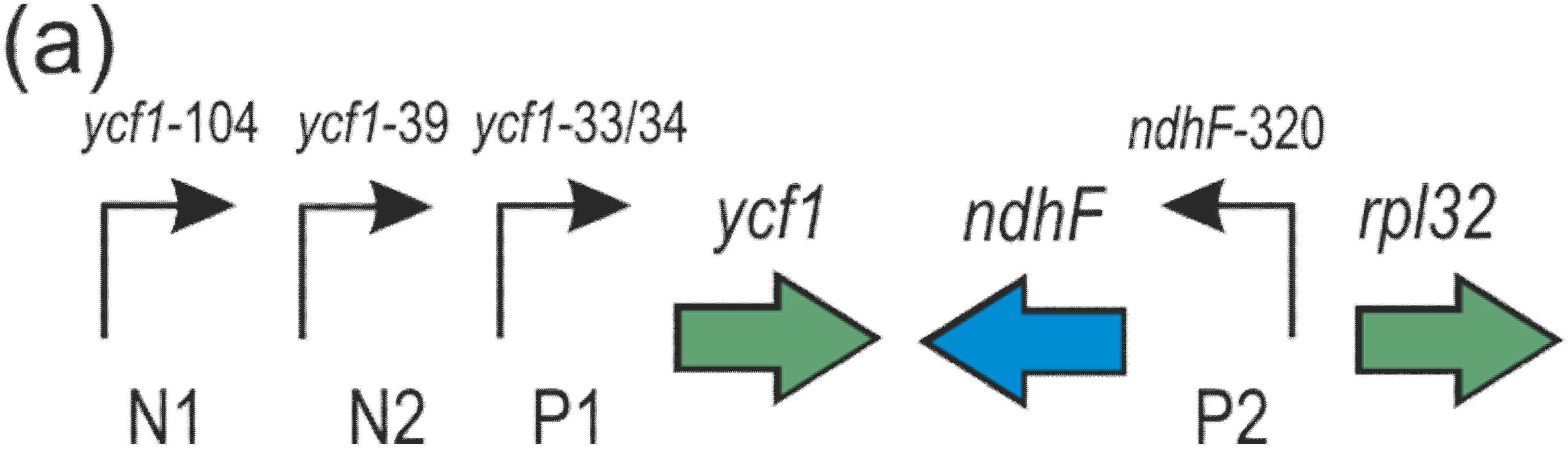


**Many polymerases concurrently** bind the sequence **and move** each in **its direction**

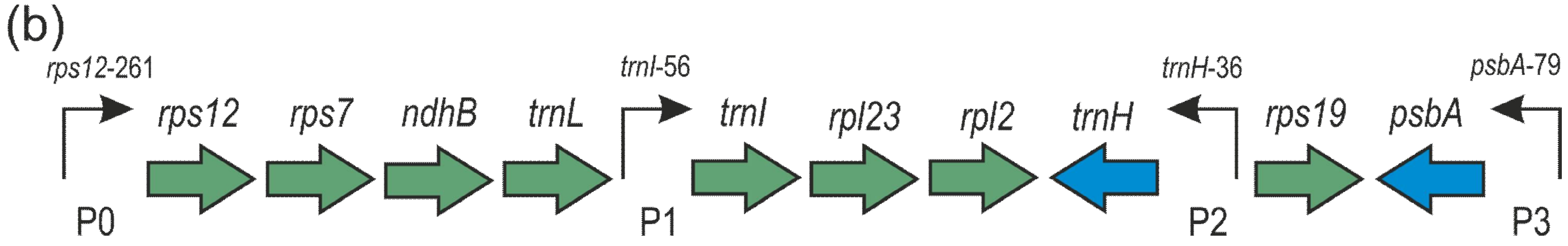
**Example: 3 genes and 4 promoters**.

The mutual arrangement of promoters and genes

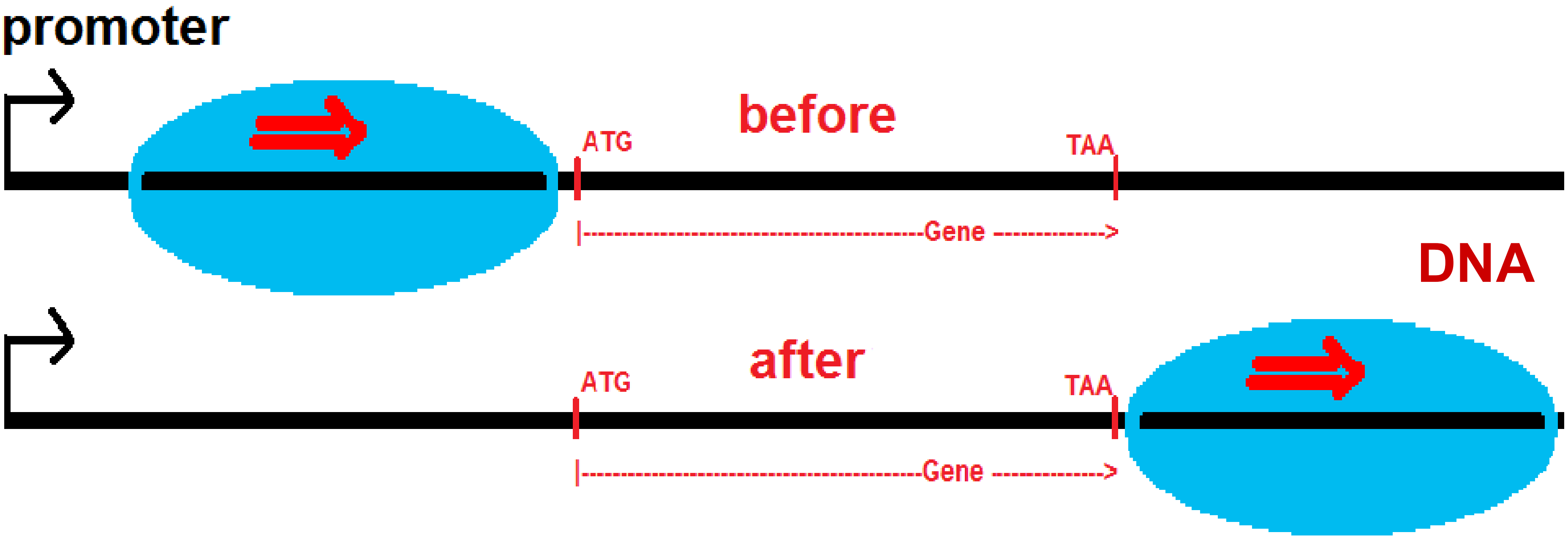
is important and varies widely



# Another example: 10 genes and 4 promoters



The gene is "**read**" if a polymerase **moved** from its **beginning** to the **end**. The gene's reading frequency is called the ***transcription level*** of this gene

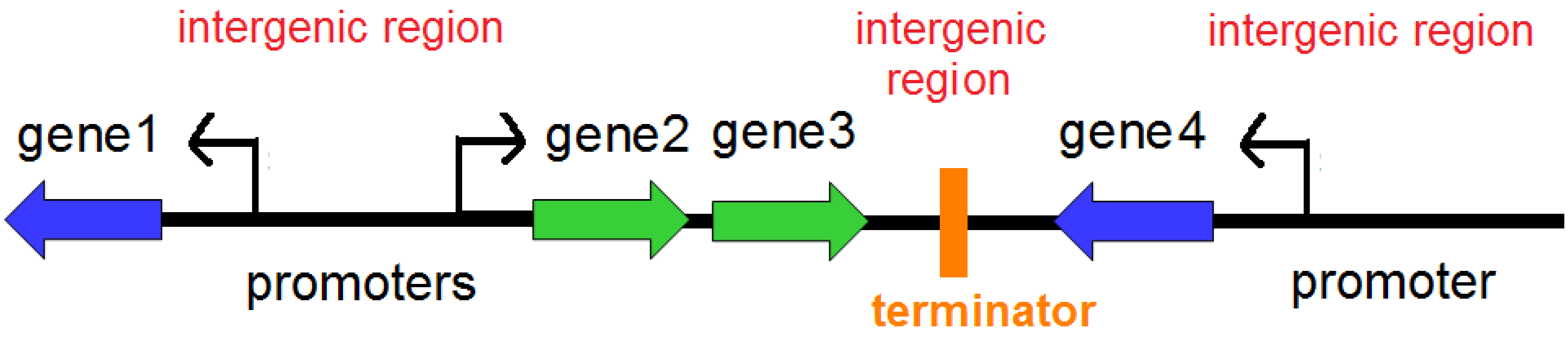


**Problem 1: the dynamics of this process, including bifurcation points, is to be described**.

There are many tasks here: for example, **1)** inferring transcription levels of all genes given the binding attempt intensities of all promoters; **2)** inferring binding attempt intensities that best approximate given gene transcription levels; **3)** inferring binding attempt intensities that best approximate known changes of gene transcription levels under wide fluctuations of temperature and polymerase rates (described by simple combinations of affine functions);

**4)** investigate for a more realistic case of the stochastic movement of polymerases.

Many particular questions remain, such as inferring the average length of the polymerase run, asymptotic distribution of the lengths, etc



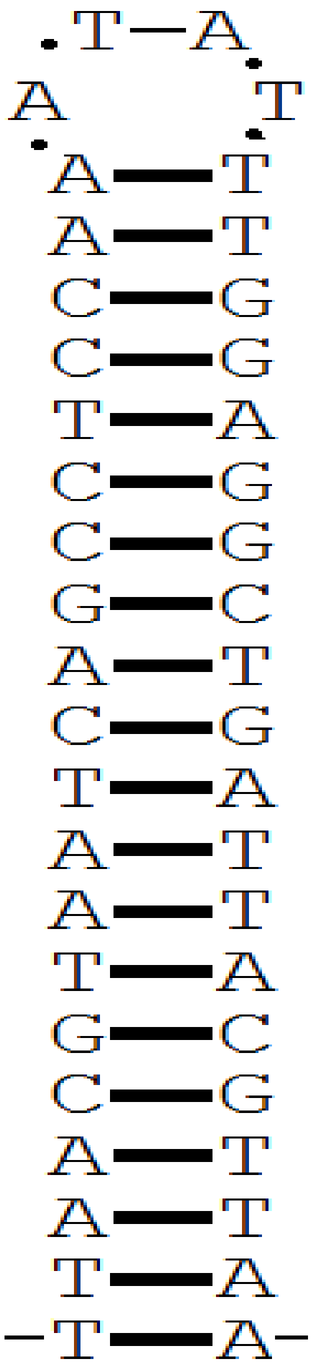
The “**terminators**” are **regions** $[e_k, f_k]$ that allow through a certain average amount of polymerases in each direction.

Thus, a **system of vectors on a fixed sequence is given**: **genes, promotors, terminators**

Note a great practical value: e.g., changes in characters leading to terminators misfunction may cause severe human health disorders

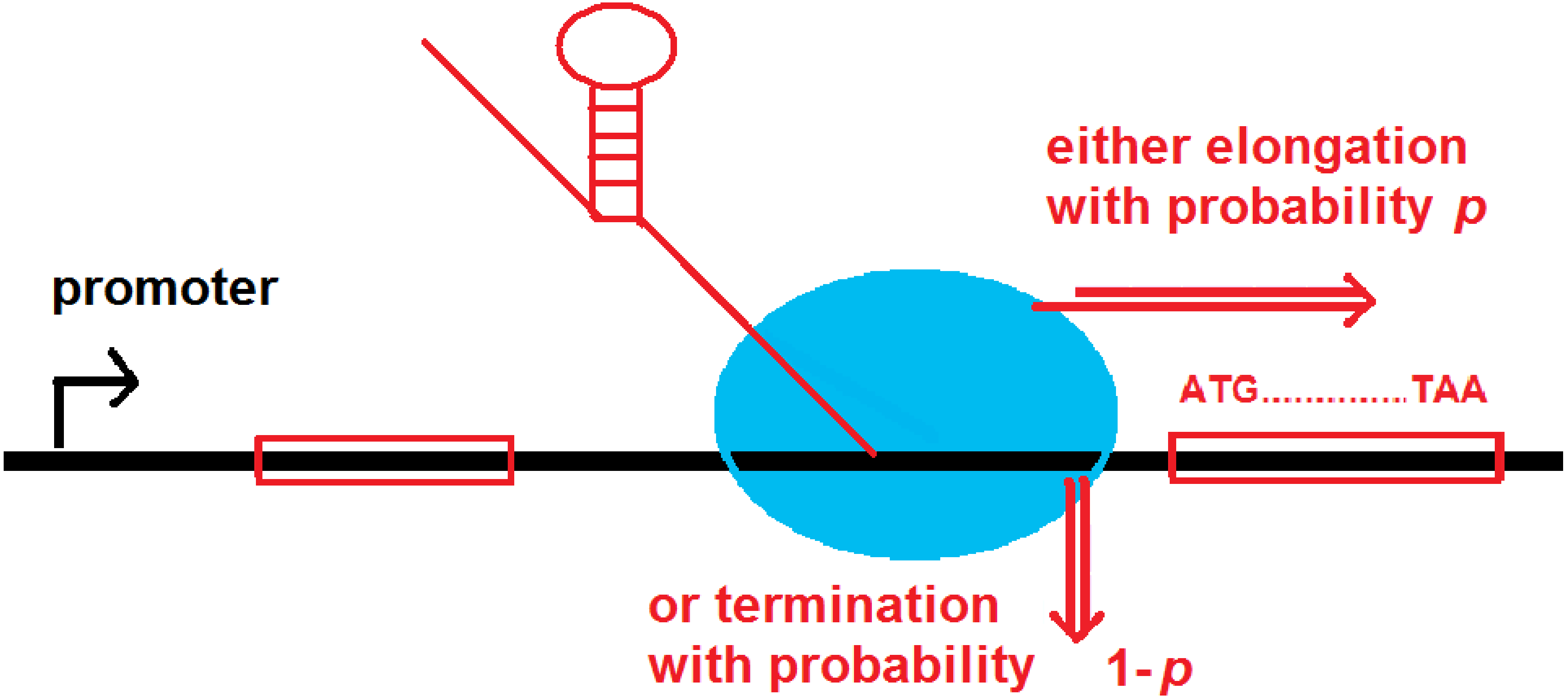
# How a terminator works?

**Terminator** forms a «**helix**»
(in yellow is its left shoulder, in blue – the right shoulder).
Paired are G to C, and A to T



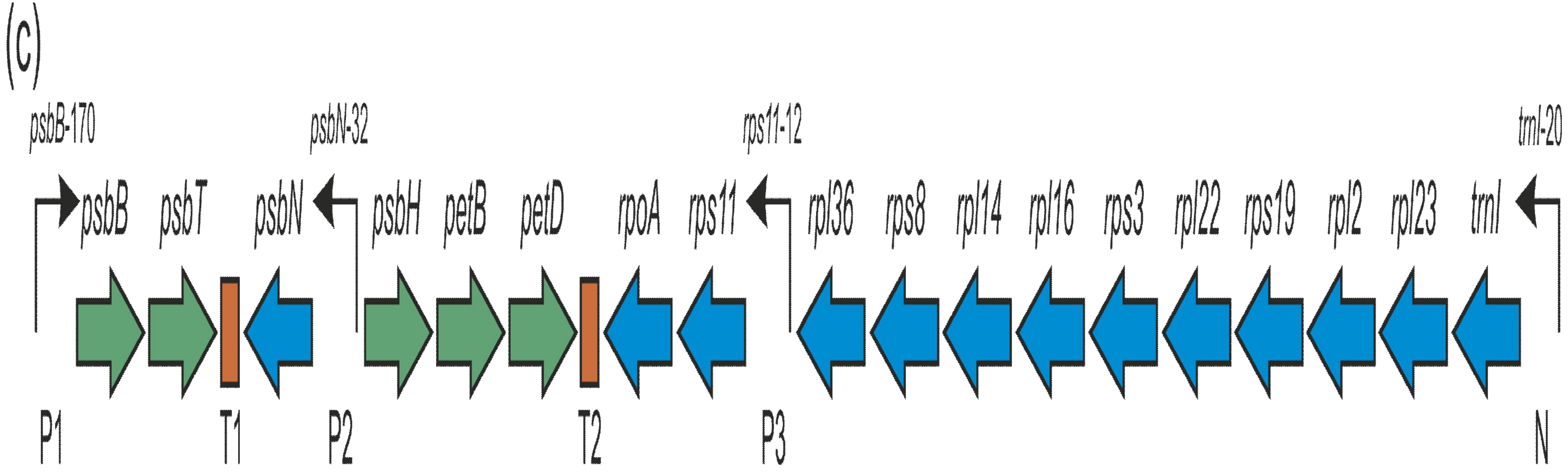
TTAACGTAATCAGCCTCCAAATATTTGGAGGCTGATTACGTTAA

# Each terminator has a certain intensity **1-*p*** of the polymerase **detachments**



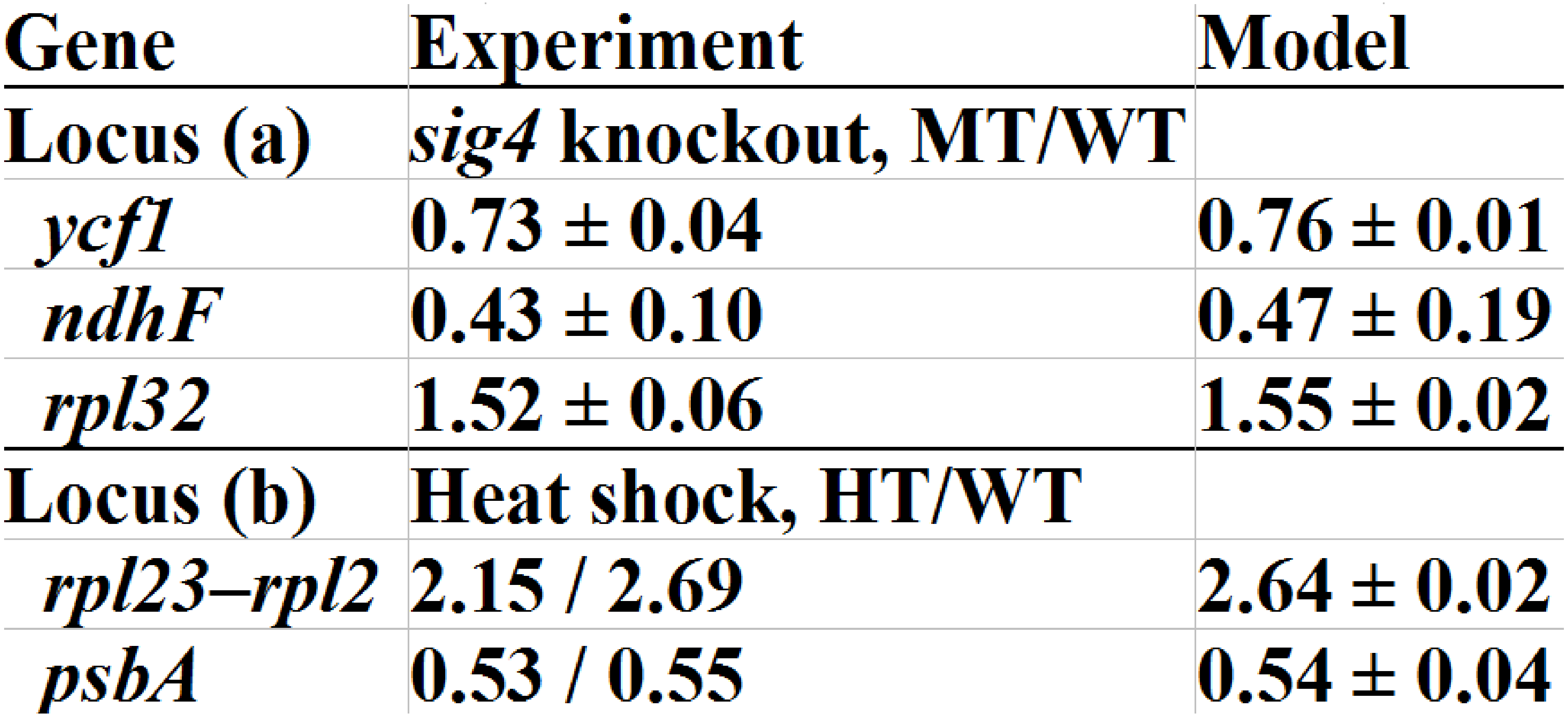
**Example with terminators**: 18 genes, 4 promoters and 2 **terminators** designated **T1**, **T2**.

The "**terminators**" are **regions** that allow through a certain average amount of polymerases in each direction



# Comparison of gene transcription levels obtained in the model and experiment

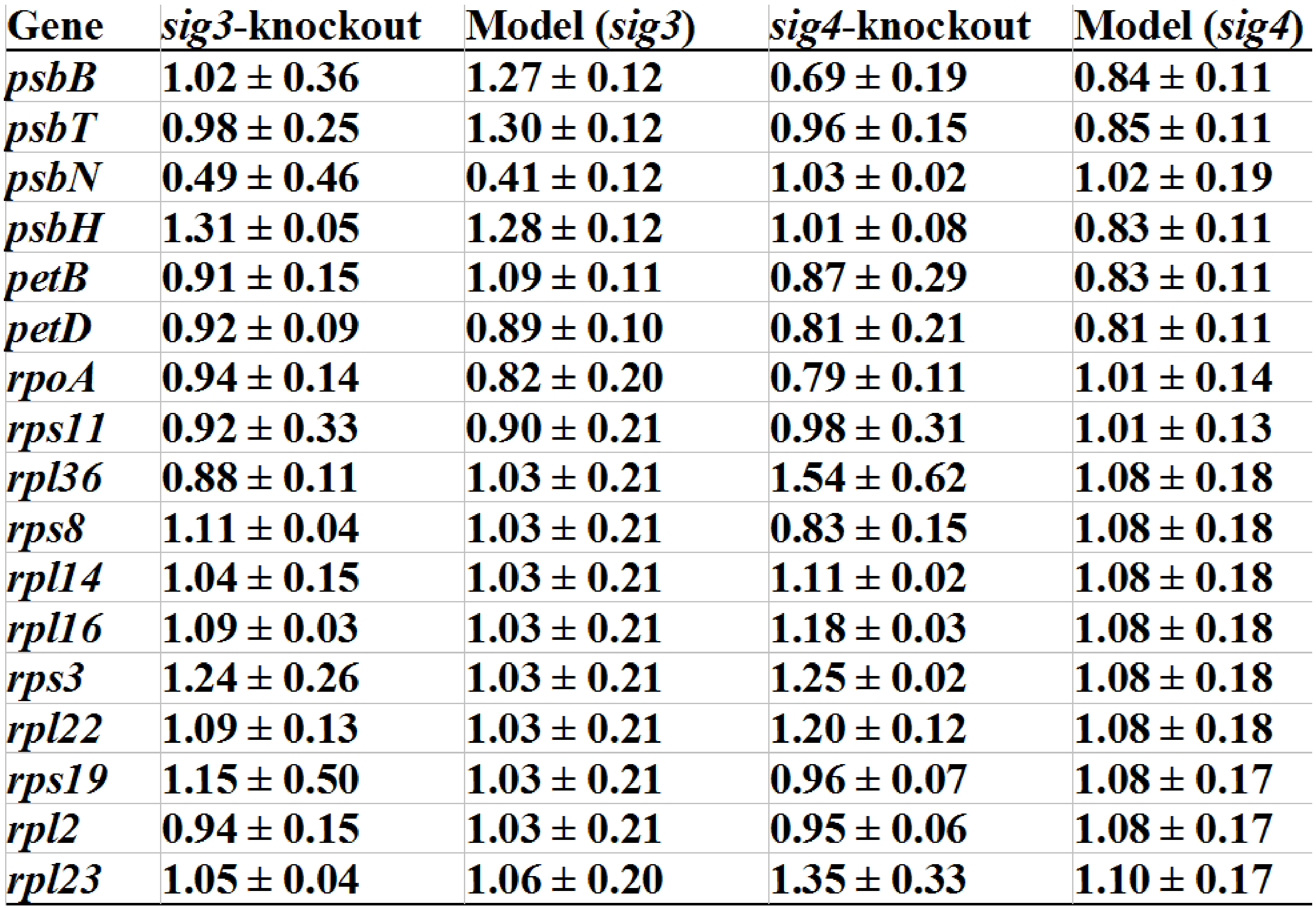
for Locus (a) in *Arabidopsis* and Locus (b) in *Hordeum.*
Standard deviations are provided where applicable. Values separated by a “/” in the second column for Locus 2 are the results of two independent heat shock studies



| Gene | Experiment | Model |
|---|---|---|
| **Locus (a)** | ***sig4* knockout, MT/WT** | |
| *ycf1* | 0.73 ± 0.04 | 0.76 ± 0.01 |
| *ndhF* | 0.43 ± 0.10 | 0.47 ± 0.19 |
| *rpl32* | 1.52 ± 0.06 | 1.55 ± 0.02 |
| **Locus (b)** | **Heat shock, HT/WT** | |
| *rpl23–rpl2* | 2.15 / 2.69 | 2.64 ± 0.02 |
| *psbA* | 0.53 / 0.55 | 0.54 ± 0.04 |

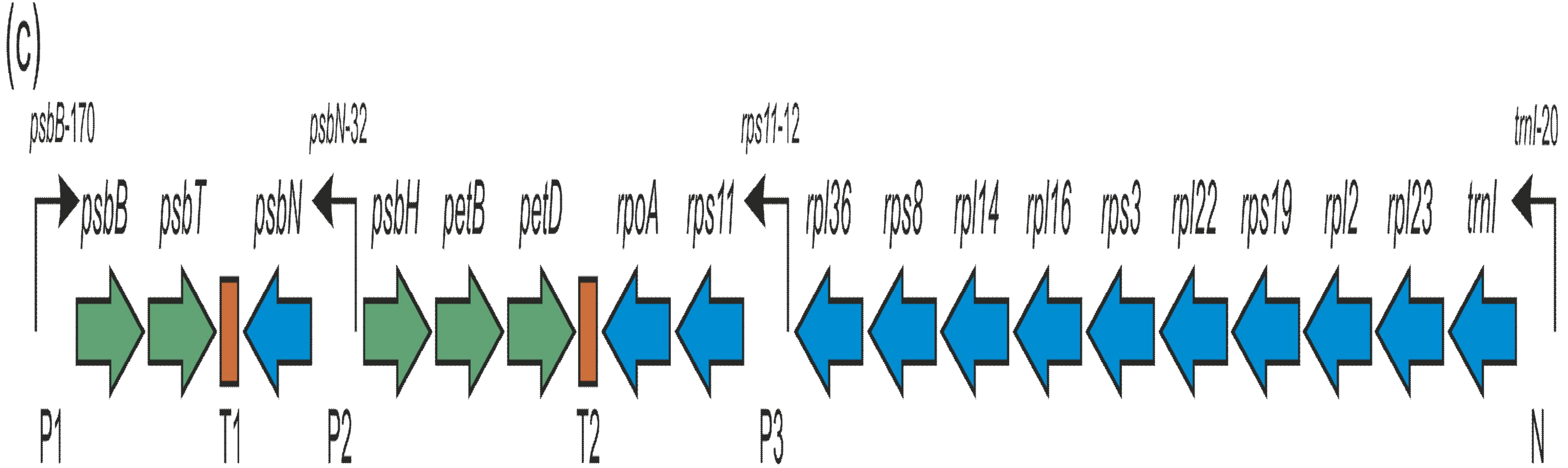
# Comparison of gene transcription levels obtained in the model and experiments

MT/WT *sig3* and *sig4* gene knockout for Locus (c)

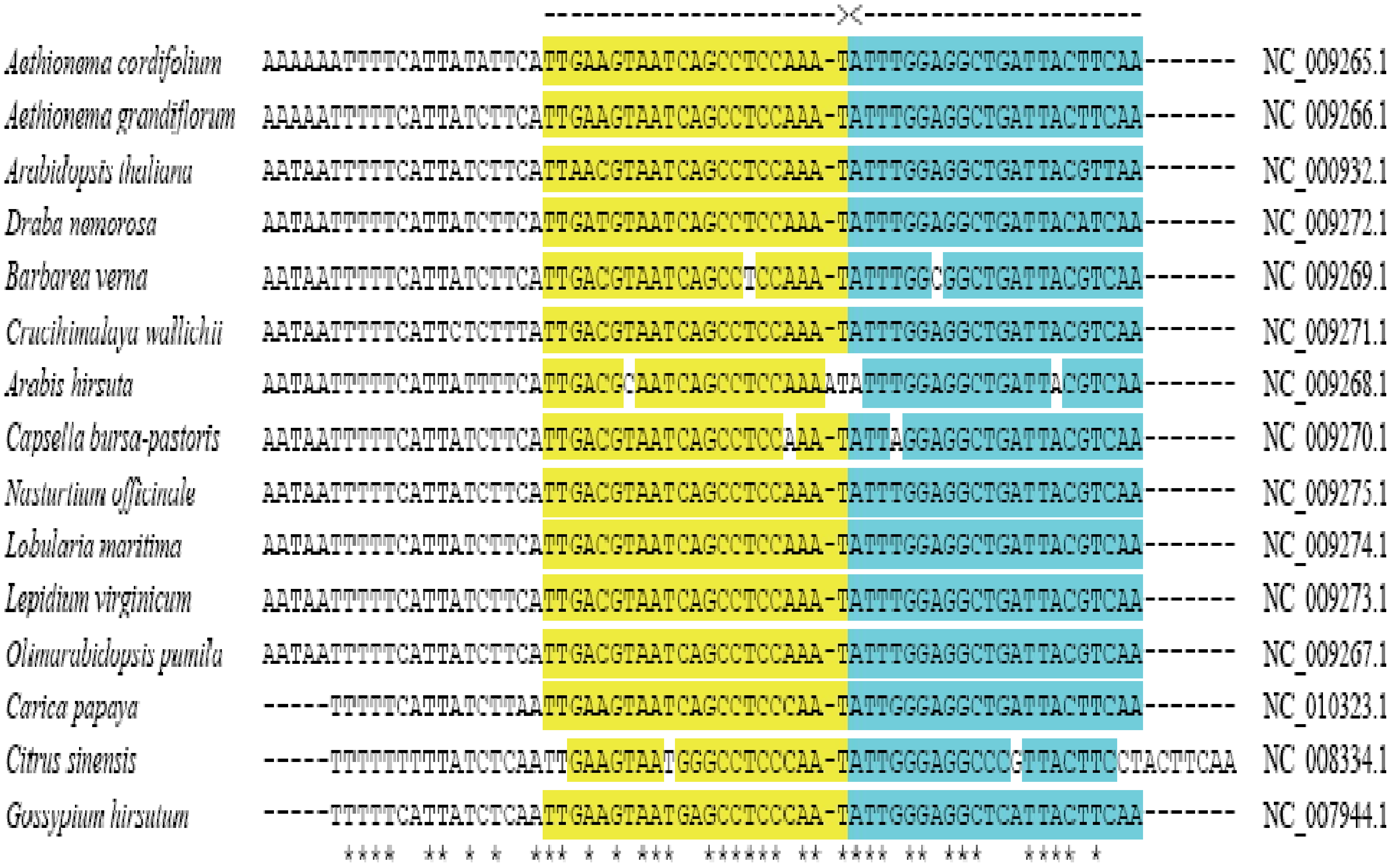


| Gene | *sig3*-knockout | Model (*sig3*) | *sig4*-knockout | Model (*sig4*) |
|---|---|---|---|---|
| *psbB* | 1.02 ± 0.36 | 1.27 ± 0.12 | 0.69 ± 0.19 | 0.84 ± 0.11 |
| *psbT* | 0.98 ± 0.25 | 1.30 ± 0.12 | 0.96 ± 0.15 | 0.85 ± 0.11 |
| *psbN* | 0.49 ± 0.46 | 0.41 ± 0.12 | 1.03 ± 0.02 | 1.02 ± 0.19 |
| *psbH* | 1.31 ± 0.05 | 1.28 ± 0.12 | 1.01 ± 0.08 | 0.83 ± 0.11 |
| *petB* | 0.91 ± 0.15 | 1.09 ± 0.11 | 0.87 ± 0.29 | 0.83 ± 0.11 |
| *petD* | 0.92 ± 0.09 | 0.89 ± 0.10 | 0.81 ± 0.21 | 0.81 ± 0.11 |
| *rpoA* | 0.94 ± 0.14 | 0.82 ± 0.20 | 0.79 ± 0.11 | 1.01 ± 0.14 |
| *rps11* | 0.92 ± 0.33 | 0.90 ± 0.21 | 0.98 ± 0.31 | 1.01 ± 0.13 |
| *rpl36* | 0.88 ± 0.11 | 1.03 ± 0.21 | 1.54 ± 0.62 | 1.08 ± 0.18 |
| *rps8* | 1.11 ± 0.04 | 1.03 ± 0.21 | 0.83 ± 0.15 | 1.08 ± 0.18 |
| *rpl14* | 1.04 ± 0.15 | 1.03 ± 0.21 | 1.11 ± 0.02 | 1.08 ± 0.18 |
| *rpl16* | 1.09 ± 0.03 | 1.03 ± 0.21 | 1.18 ± 0.03 | 1.08 ± 0.18 |
| *rps3* | 1.24 ± 0.26 | 1.03 ± 0.21 | 1.25 ± 0.02 | 1.08 ± 0.18 |
| *rpl22* | 1.09 ± 0.13 | 1.03 ± 0.21 | 1.20 ± 0.12 | 1.08 ± 0.18 |
| *rps19* | 1.15 ± 0.50 | 1.03 ± 0.21 | 0.96 ± 0.07 | 1.08 ± 0.17 |
| *rpl2* | 0.94 ± 0.15 | 1.03 ± 0.21 | 0.95 ± 0.06 | 1.08 ± 0.17 |
| *rpl23* | 1.05 ± 0.04 | 1.06 ± 0.20 | 1.35 ± 0.33 | 1.10 ± 0.17 |

**Terminators** T1 and T2 were postulated to bring the model predictions in agreement with the experiment. Introducing the two terminators in these specific regions allowed to reach the congruence. The terminators and their location were independently proved in the experiment



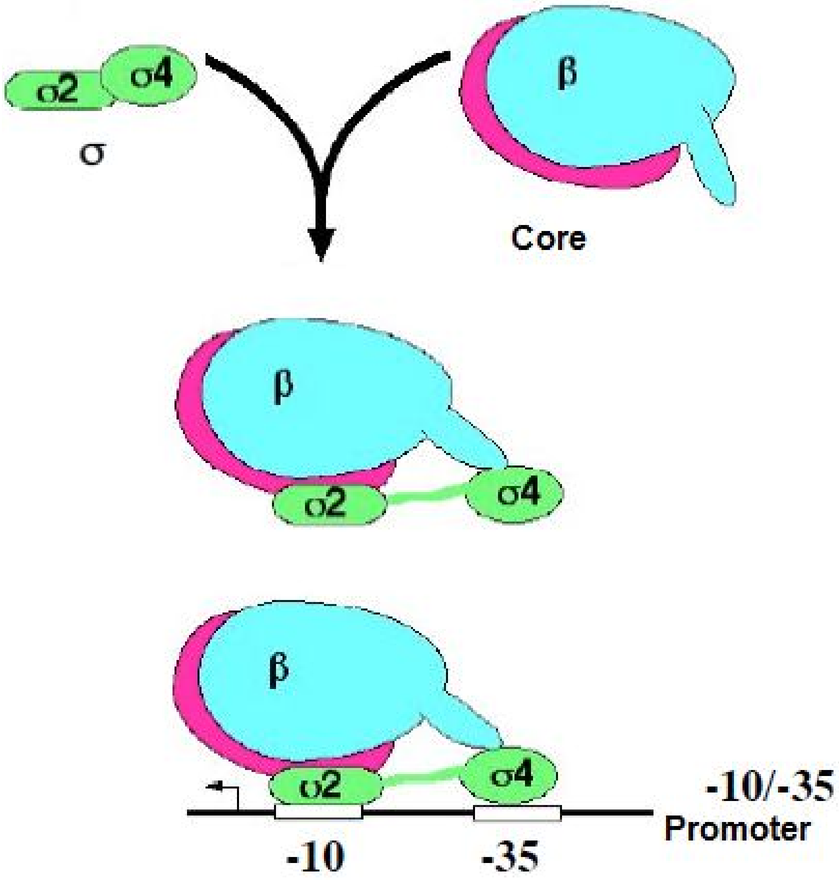
# **The terminators were verified with the alignment:** the example of terminator T1 in different species

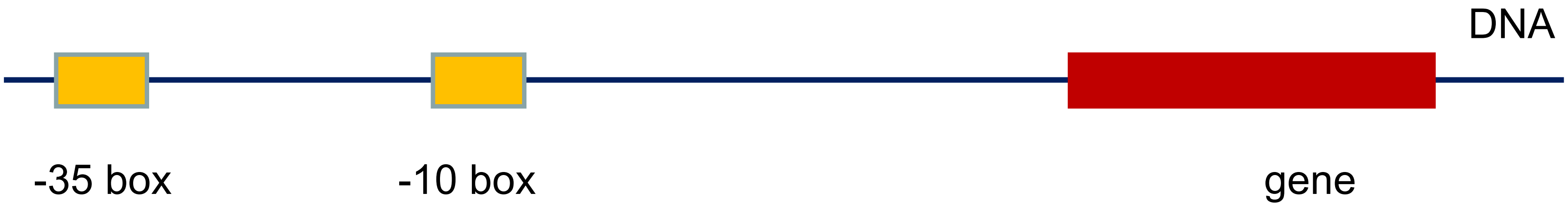


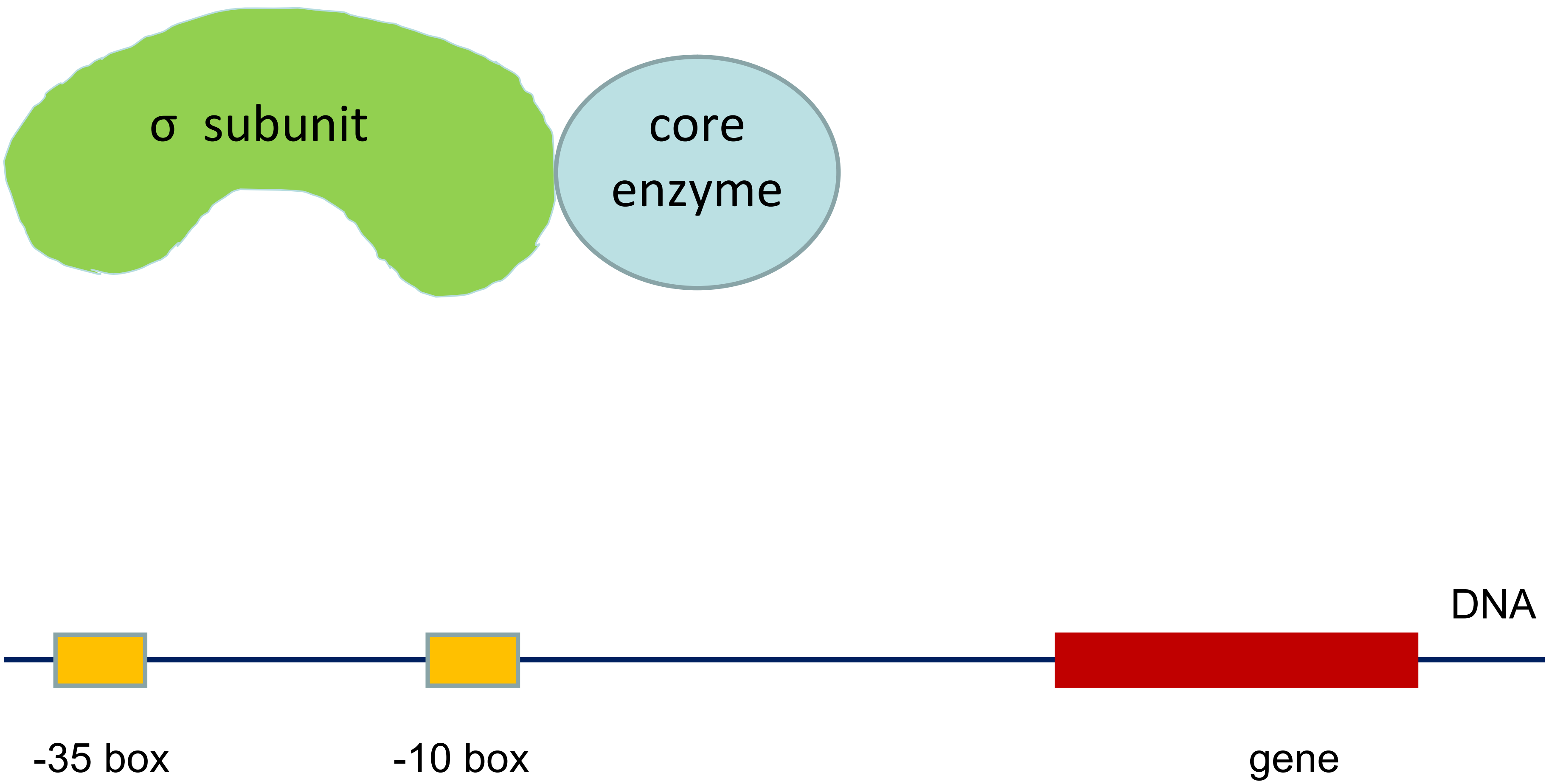
```
                                                ----------------------><---------------------
Aethionema cordifolium       AAAAAATTTTCATTATATTCATTGAAGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACTTCAA-------   NC_009265.1
Aethionema grandiflorum      AAAAATTTTTCATTATCTTCATTGAAGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACTTCAA-------   NC_009266.1
Arabidopsis thaliana         AATAATTTTTCATTATCTTCATTAACGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACGTTAA-------   NC_000932.1
Draba nemorosa               AATAATTTTTCATTATCTTCATTGATGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACATCAA-------   NC_009272.1
Barbarea verna               AATAATTTTTCATTATCTTCATTGACGTAATCAGCCTCCAAA-TATTTGGCGGCTGATTACGTCAA-------   NC_009269.1
Crucihimalaya wallichii      AATAATTTTTCATTCTCTTTATTGACGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACGTCAA-------   NC_009271.1
Arabis hirsuta               AATAATTTTTCATTATTTTCATTGACGCAATCAGCCTCCAAAATATTTGGAGGCTGATTACGTCAA-------   NC_009268.1
Capsella bursa-pastoris      AATAATTTTTCATTATCTTCATTGACGTAATCAGCCTCCAAA-TATTAGGAGGCTGATTACGTCAA-------   NC_009270.1
Nasturtium officinale        AATAATTTTTCATTATCTTCATTGACGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACGTCAA-------   NC_009275.1
Lobularia maritima           AATAATTTTTCATTATCTTCATTGACGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACGTCAA-------   NC_009274.1
Lepidium virginicum          AATAATTTTTCATTATCTTCATTGACGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACGTCAA-------   NC_009273.1
Olimarabidopsis pumila       AATAATTTTTCATTATCTTCATTGACGTAATCAGCCTCCAAA-TATTTGGAGGCTGATTACGTCAA-------   NC_009267.1
Carica papaya                -----TTTTTCATTATCTTAATTGAAGTAATCAGCCTCCCAA-TATTGGAGGCTGATTACTTCAA-------   NC_010323.1
Citrus sinensis              -----TTTTTTTTTATCTCAATTGAAGTAATGGGCCTCCCAA-TATTGGAGGCCCGTTACTTCCTACTTCAA   NC_008334.1
Gossypium hirsutum           -----TTTTTCATTATCTCAATTGAAGTAATGAGCCTCCCAA-TATTGGAGGCTCATTACTTCAA-------   NC_007944.1
                                   ****   **  *  *   ***  *  *  ***    ******  **  ****  **  ***     ****  *
```

# Two further pitfalls

**Binding is followed by the abort process**: an alternation of movement at a fixed finite rate in the corresponding direction of promoter at an arbitrary (e.g., exponentially distributed) distance and instantaneous return to the initial position. Such alternations occur an arbitrary (e.g., geometrically distributed) number of times until the polymerase reaches at a threshold distance from the promoter. At this instance the polymerase detaches from the promoter, its size instantaneously decreases by a fixed value and movement continues in the same direction


σ2
σ4
σ
β
Core
β
σ2
σ4
β
σ2
σ4
-10/-35
Promoter
-10
-35


DNA
-35 box
-10 box
gene



σ subunit

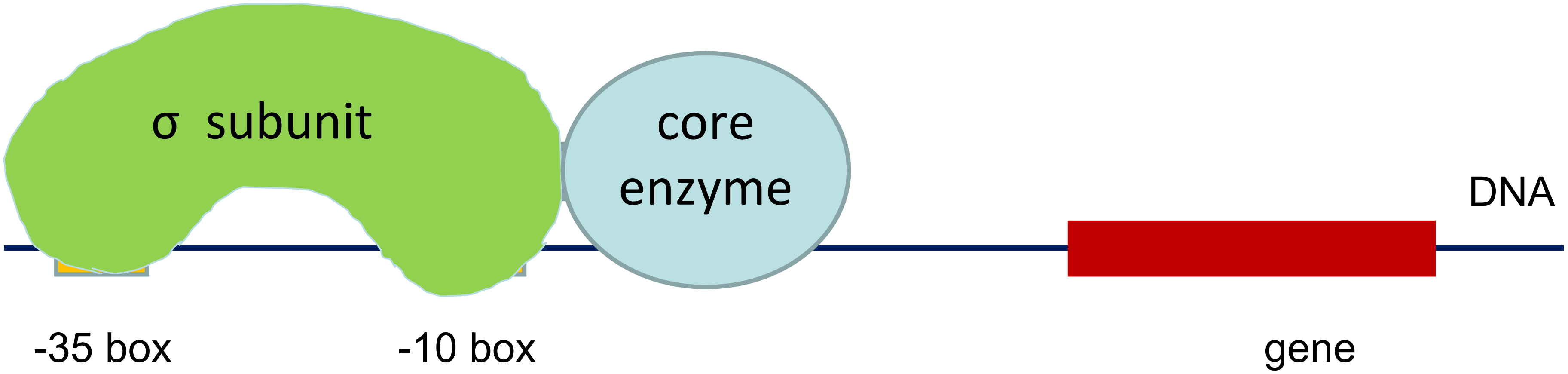
core enzyme

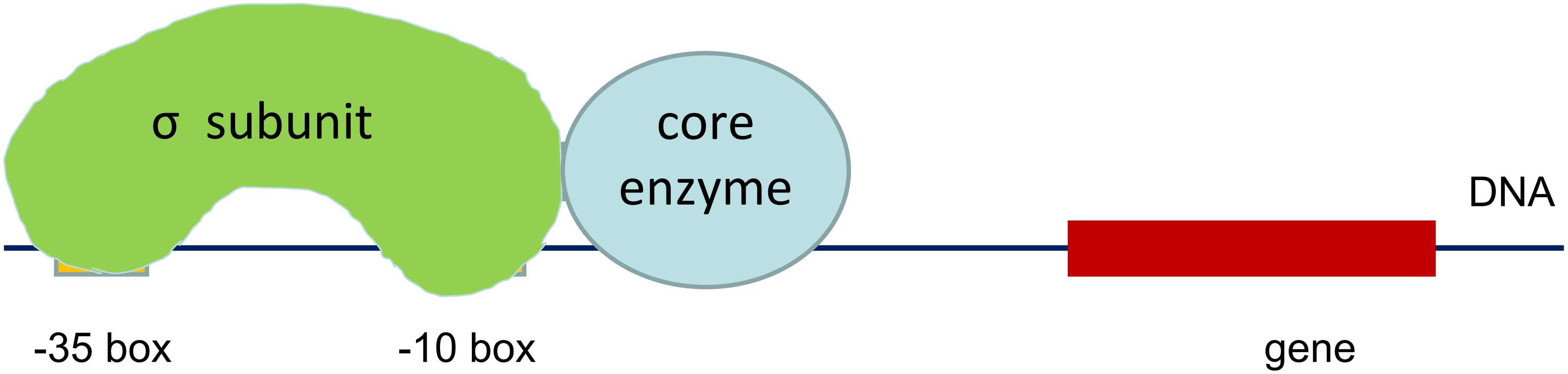
DNA

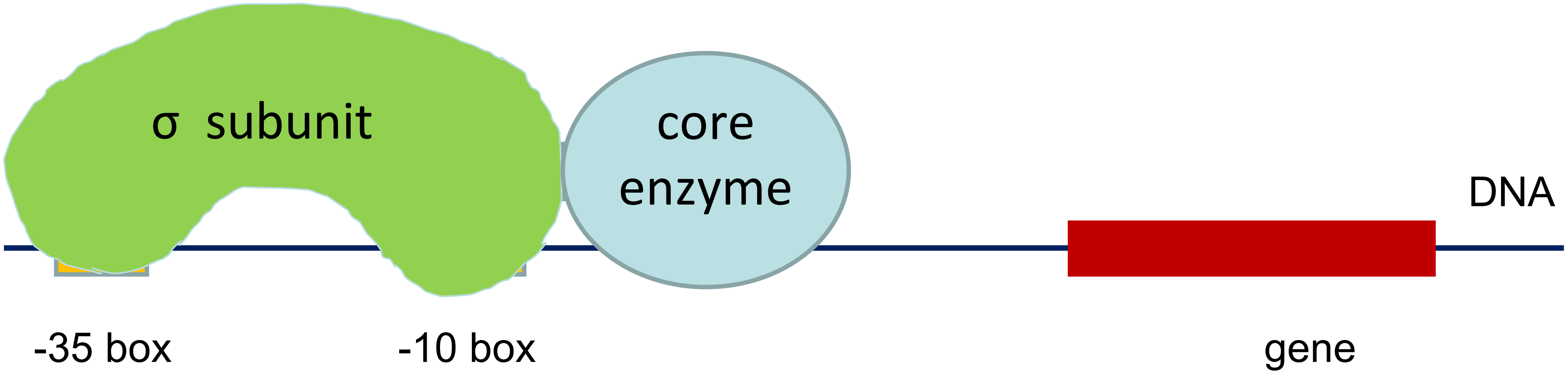
-35 box

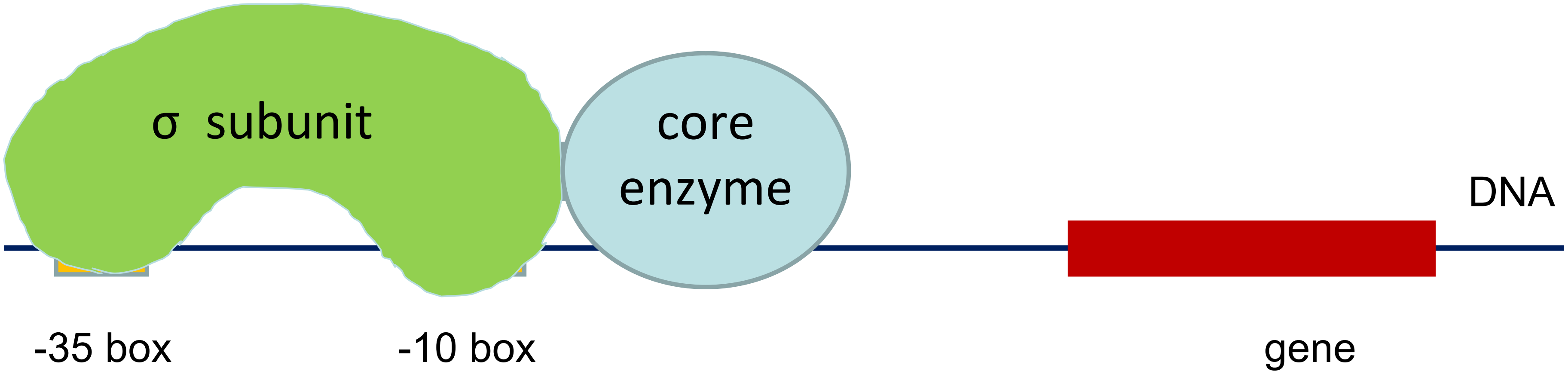
-10 box

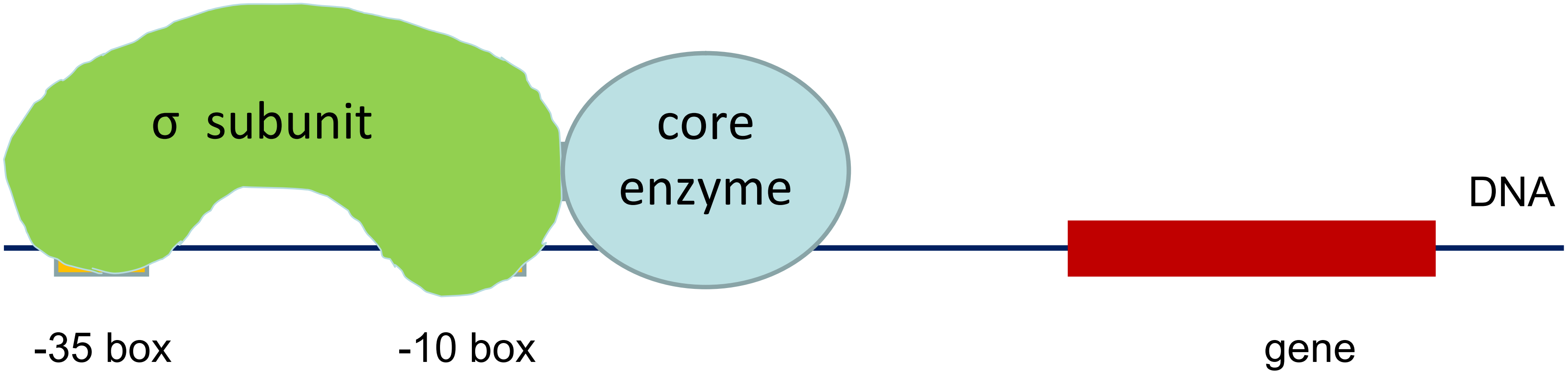
gene

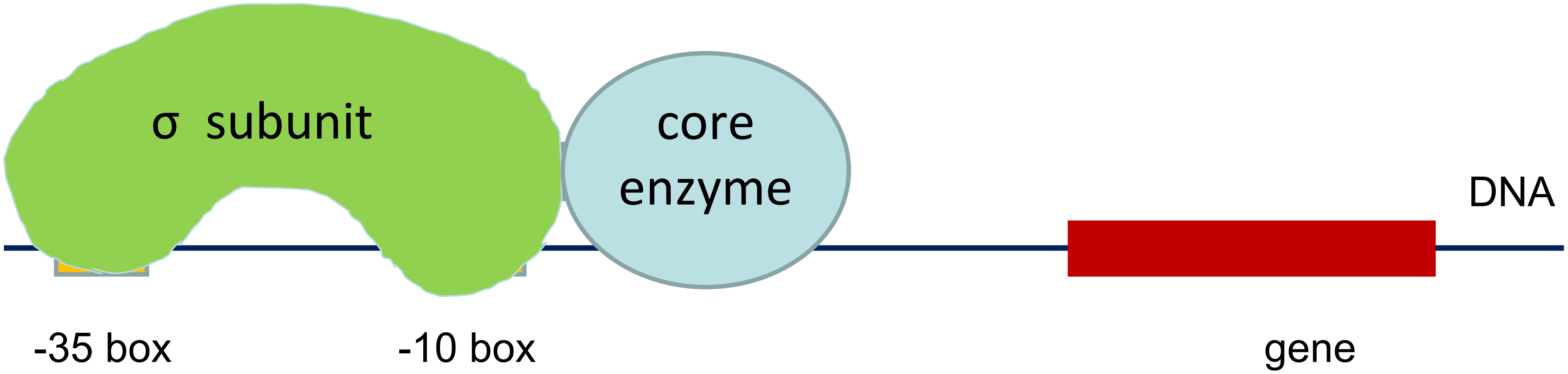

σ subunit
core enzyme
DNA
-35 box
-10 box
gene

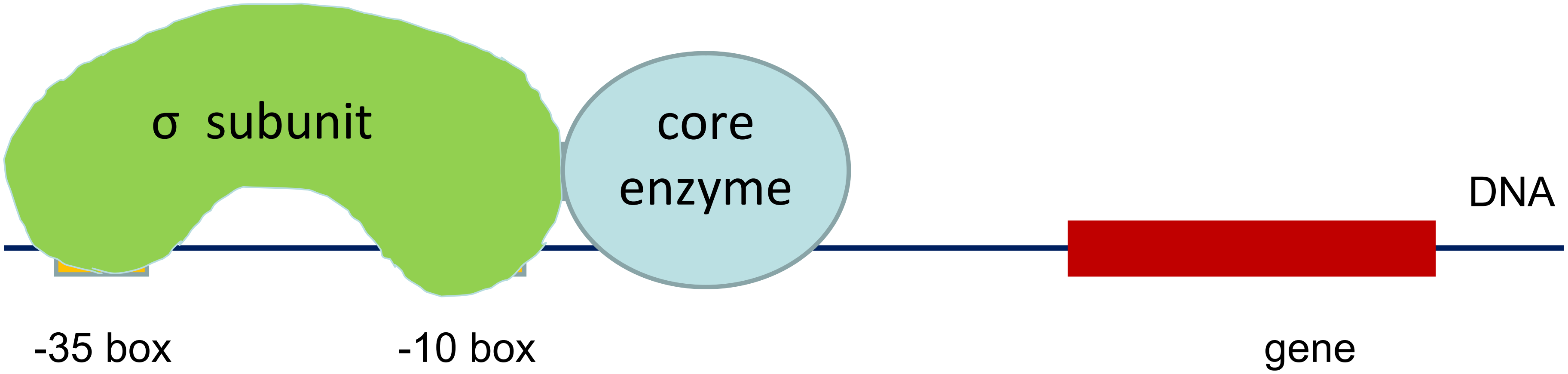

σ subunit
core enzyme
DNA
-35 box
-10 box
gene

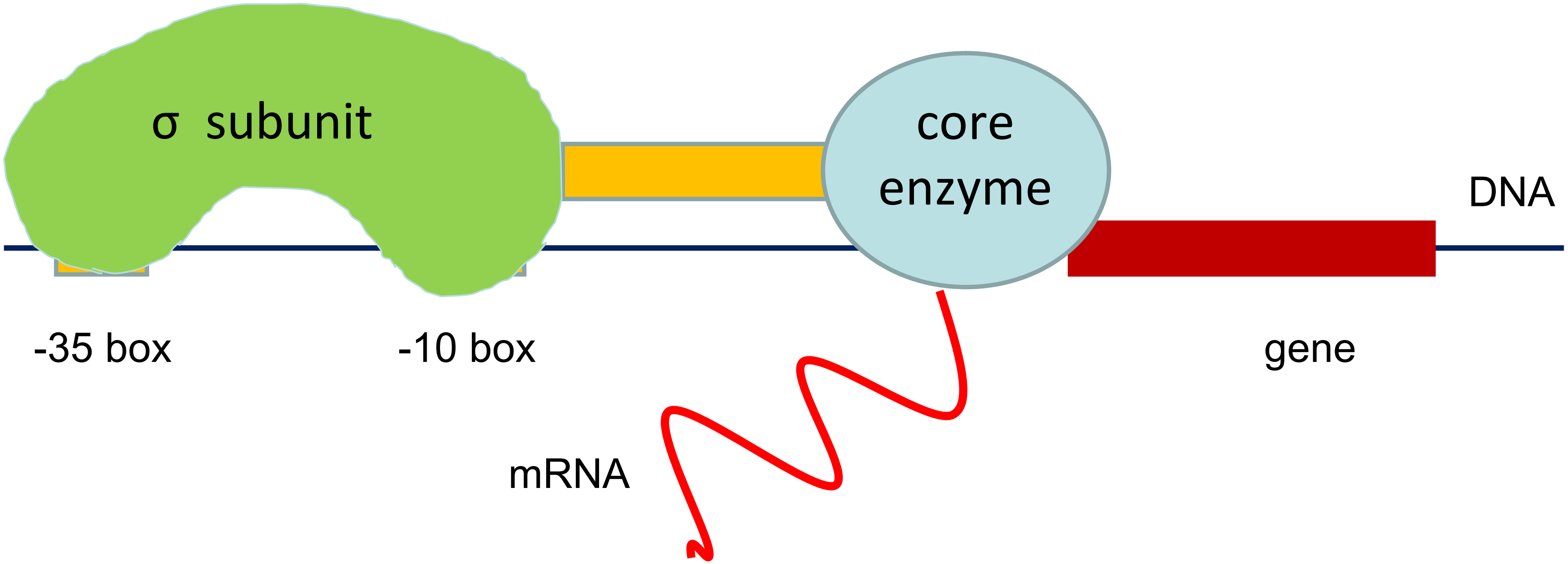

σ subunit
core
enzyme
DNA
-35 box
-10 box
gene

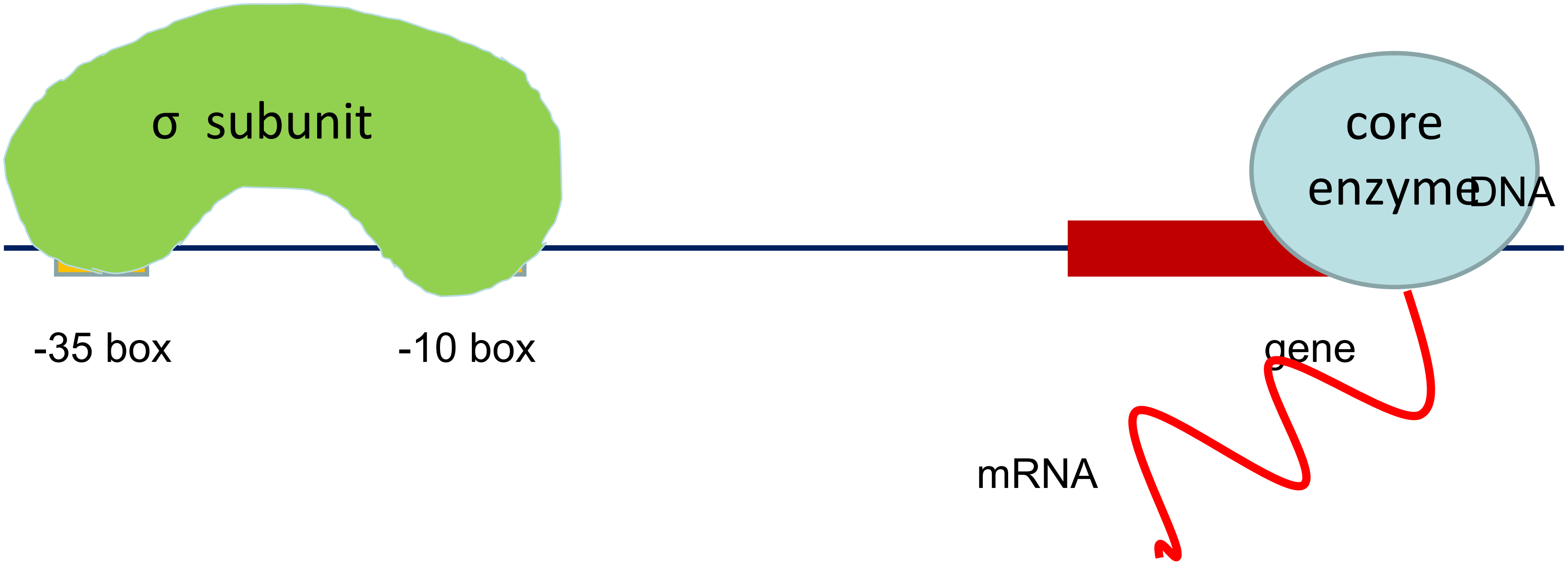

σ subunit
core enzyme
DNA
-35 box
-10 box
gene


σ subunit
core enzyme
DNA
-35 box
-10 box
gene

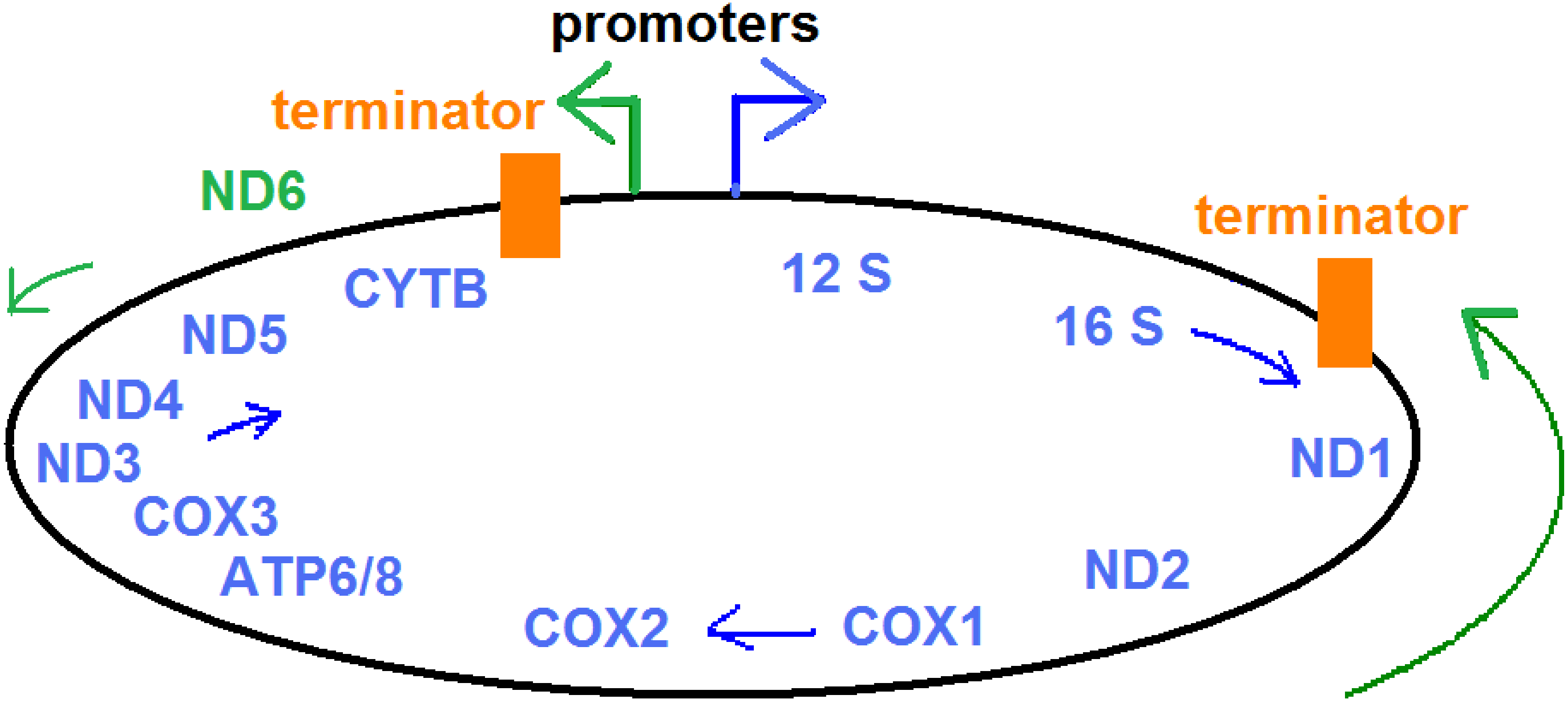

σ  subunit
core
enzyme
DNA
-35 box
-10 box
gene


σ subunit
core enzyme
DNA
-35 box
-10 box
gene


σ subunit
core enzyme
DNA
-35 box
-10 box
gene
mRNA


σ subunit
core enzyme
DNA
-35 box
-10 box
gene
mRNA

# RNA Polymerase Competition in the **circle case** (mitochondrial DNA)



Initially, polymerases do not complete the circle, their counter-flows from the two promoters collide and the polymerases detach. Genes distant from the promoters have nearly zero expression levels, which contradicts biological observations. This is an unstable state: one of the promoters realizes by 10 more bindings, the extra polymerases avoid collisions and complete the full circle including the initial promoter. It **simulates the increasing number of successful bindings** and increases the number of  polymerases completing the circle in one direction. If another promoter also receives enough bindings, the movement in **opposite direction may become more successful**. The directions are **rarely swapped several times**, and a winning direction rapidly establishes

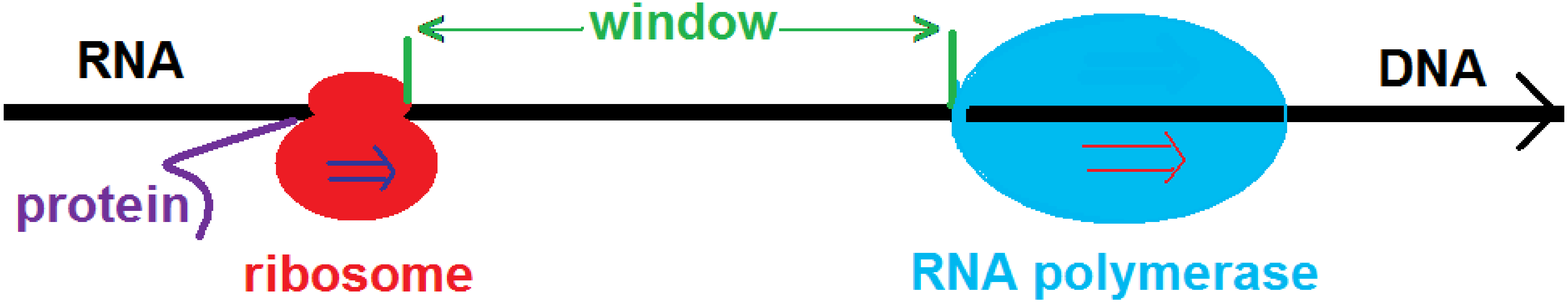
**Thus, Problem 1. Describe the process: multiple machines (polymerases) simultaneously attempt to bind different regions (if those are unattended at the instance of binding) of a long sequence. When bound, the machines slide along the sequence not affecting each other, OR collide in opposite directions and slip. Slippage can also be caused by scattered terminators. The fate depends on the local arrangement of objects in the sequence.**

**To estimate frequencies of pre-defined regions (genes).**

**Among particular questions: what is an average distance that machine cover before collision?**

Other problems (detailed in the proceedings):

2. Two **molecular machines follow** each other at a certain distance (a "**window**").

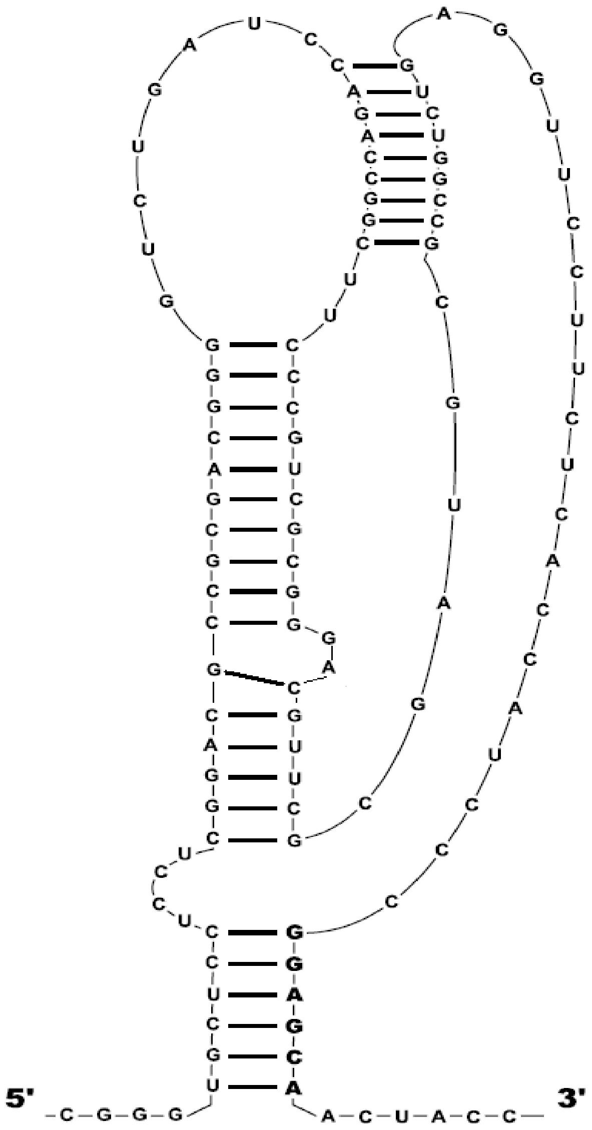


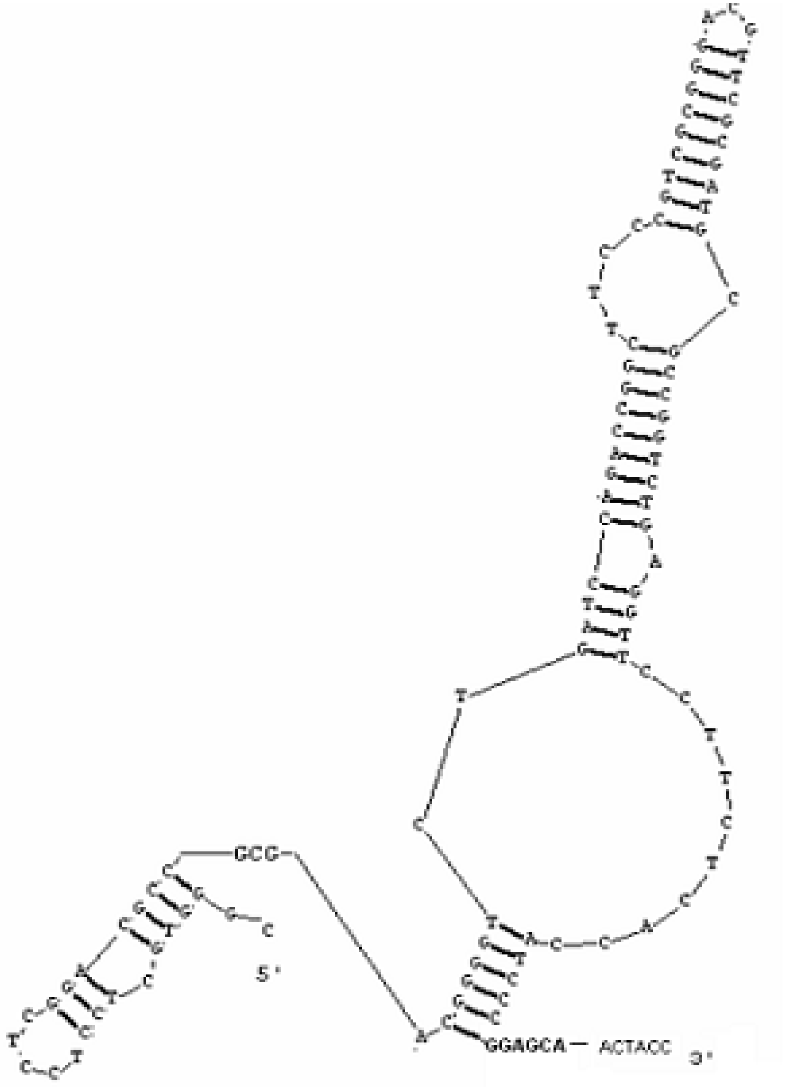
The machines’ behavior is controlled by a

**secondary structure**

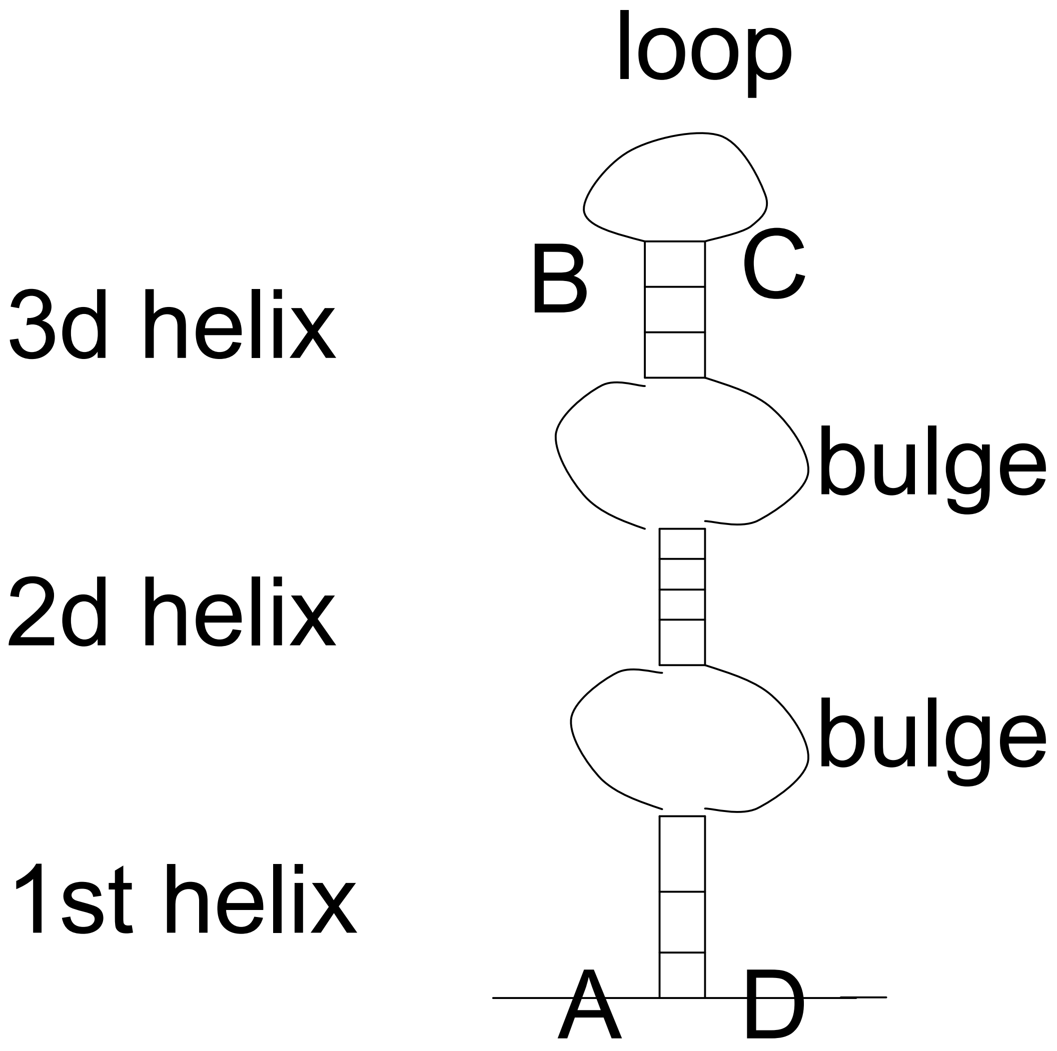
**with minimal energy** formed in the window

**Secondary structures** are composed of helices. The example of a very simple structure. How to classify such structures and estimate their energies? We offered some solutions





A **hairpin** is a linear chain of **helices**:



Thus we estimated the **hairpin energy** as the sum of the **bond energy** $\frac{1}{RT}\cdot\sum_i E_i$

and **loop energy** $\sum_i\left(1.77\cdot\ln(l_i+1)+B+\frac{C}{l_i}\right)$

where *i* varies over all **helices** of the hairpin and $E_i$ is the energy of the *i*-th **helix** determined from the experimentally known hydrogen bonds and stacking energies; $l_i$ is the loop length of the *i*-th **helix**; and *B*, *C* are constants

All elements can be accurately described here.

**Thus, problem 2**:

to **describe the process dynamics**

**Problem 3**. The behavior is described by a **Gibbs functional with nonlocal interaction**.

To **find are its global minima**

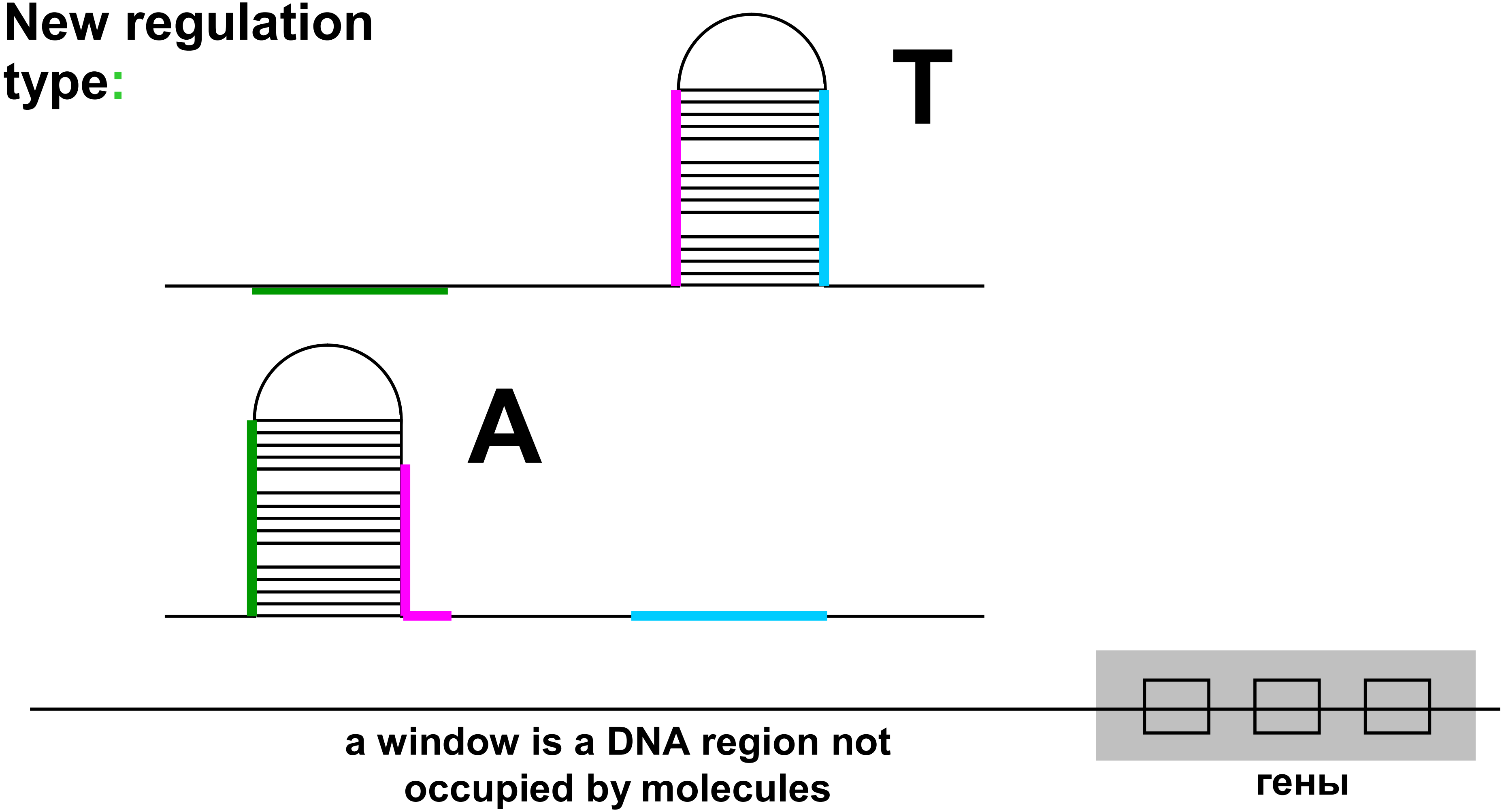
**Problem 4**. A **set of trees** is given.
To find is the **average tree**.
The problems begin with defining an **“average” tree**.

**Problem 5**. We described co-evolution of a large number of long sequences (genomes).
**Is there a time point when sequences with similar characteristics form clusters**, i.e., species?

We are thankful to L. Rusin, K. Gorbunov, L. Rubanov, S. Pirogov, E. Zhizhina for cooperation and publishing the results

Thank You

## New regulation type:



a window is a DNA region not occupied by molecules

**Two signal states. The outcome** depends on which **alternative** structure is formed: «**T**» –«**termination**» (polymerase detaches) or «**A**» –«**antitermination**» (polymerase continues moving and reading downstream genes)

# Transitions allowed in the model for this regulation:

(1) Right border y of the window **moves** at one character to the right or is fixes or signal “T” is received **(“slippage”)**. Alternatively: right border ***y*** reaches the gene start and signal “A” is received**.**
**Decision between T and A** is determined by <u>the secondary structure formed in the window</u>;

(2) Left border ***x*** of the window **moves** at three characters to the right **or is fixed**, depending on frequency c of **<u>prior gene reading</u>**;

(3) The secondary structure transforms in the window, i.e. current structure ω transforms into new structure ω', very fast!

In reality, border x is the right border of one molecular machine (“ribosome”), and y is the left border of another machine (the already familiar polymerase). Thus, the window corresponds to a gap between the ribosome and polymerase.

Both machines move to the right

Each of the four transitions is described as a Poisson flow with rate constants *k1, k2, k3, k4*:

**polymerase *shift*:**

$$k1 = -\left[40 - F(\omega)\right]$$

**polymerase *slippage*:**

$$k2 = -\frac{1}{4}\frac{\delta}{L_1^{\,2}\cdot(p(\omega)-p_0)^2+1}\cdot\exp\left(-\frac{r}{r_0}\right)$$

**ribosome *shift***: $$k3 = -\frac{45 \cdot c}{c_0 + c}$$

**Secondary structure *rearrangement*** from *state* ω into state ω' within the window:

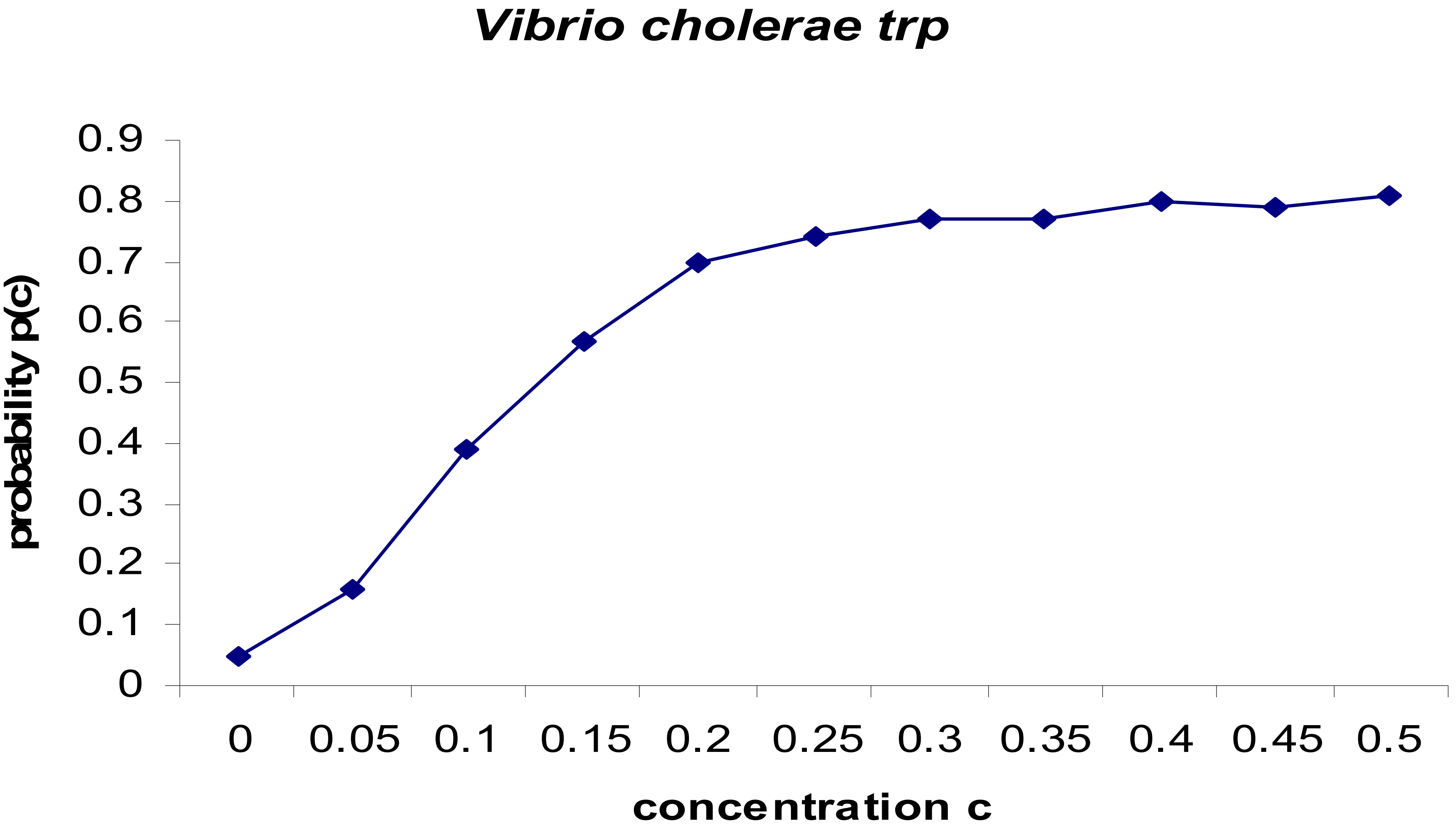
$$k4 = -\left[\kappa \cdot \exp\left(\frac{1}{2}\left(\left(G_{loop}(\omega) + G_{hel}(\omega)\right) - \left(G_{loop}(\omega') + G_{hel}(\omega')\right)\right)\right)\right]$$

where $G_{loop}(\omega)$ - loop energy of ω, >0;

$G_{hel}(\omega)$ - bond energy of neighboring pairs in ω (stacking), <0

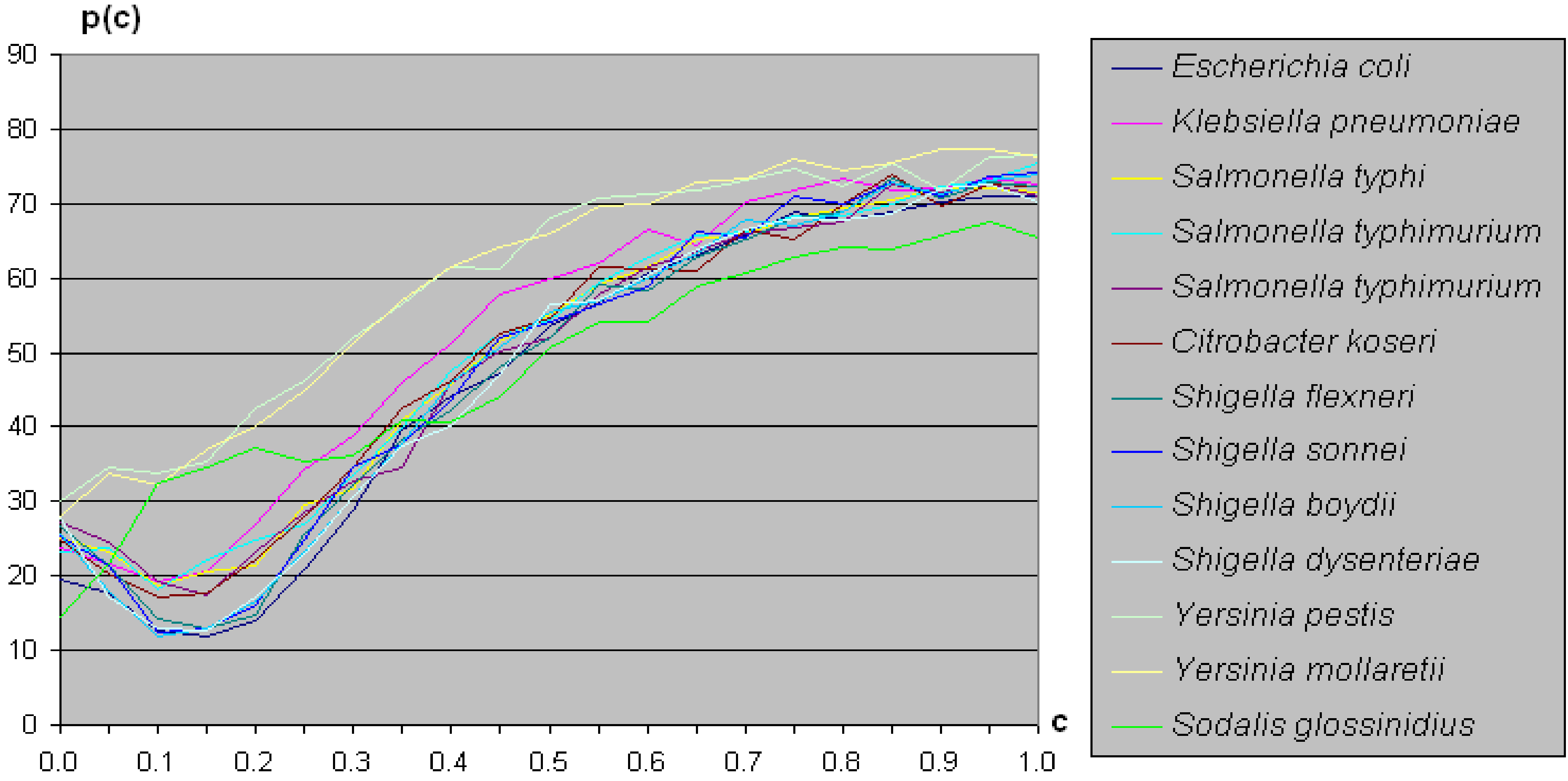
**To find is frequency *p*(*c*) of occurrence of state “T” (failure to read genes, i.e. polymerase slippage) at time *t*+d*t* depending on <u>reading frequency</u> c at time t**

# An example model prediction (the case of tryptophan biosynthesis regulation in *Vibrio cholerae*):



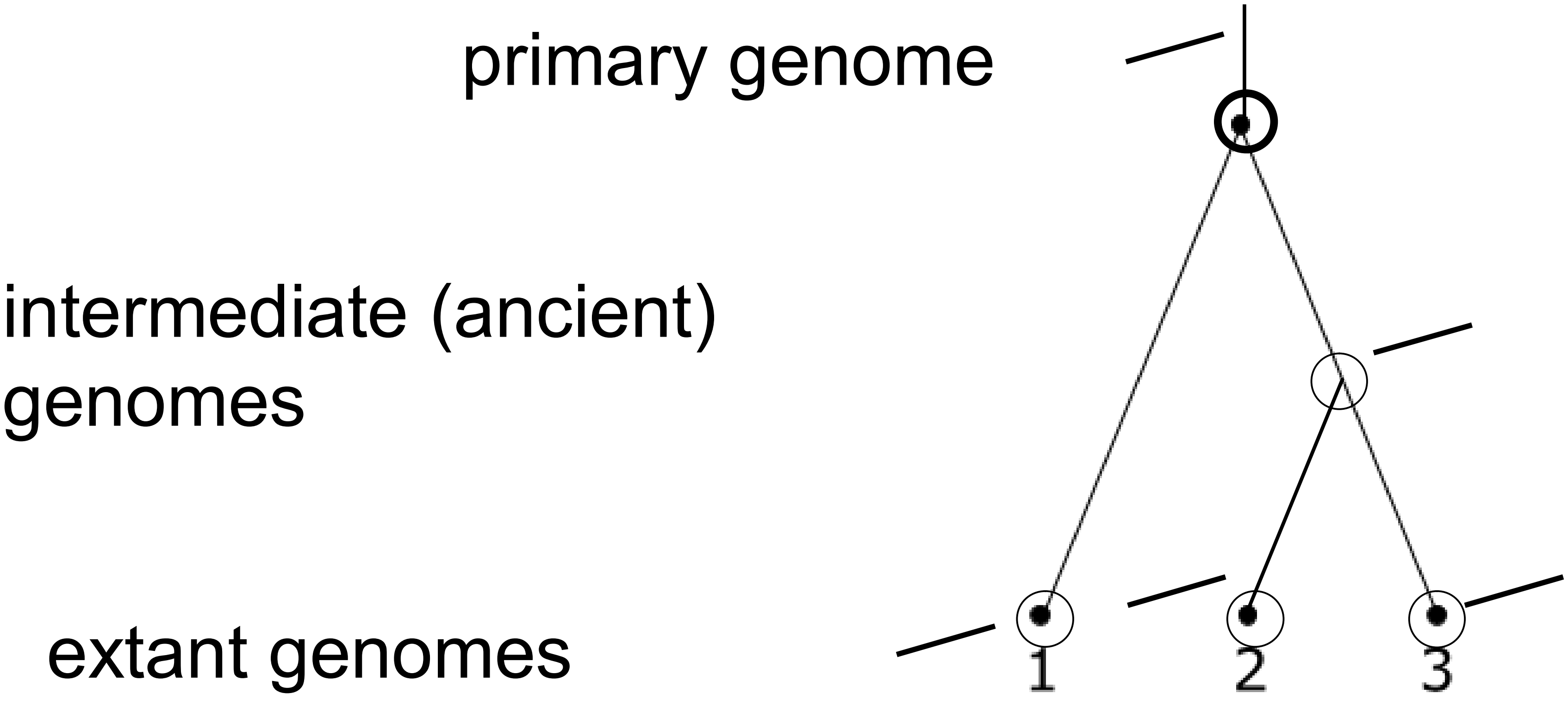
# The model conforms well with known evidence and has high predictive capacity

**for most leader regions of amino acid operons and aminoacyl-tRNA synthetases. Shown below are *thrA* operons in gamma-proteobacteria**



# PROBLEM III

At each node of the organism (species) tree a genome is duplicated (=speciation event). Thus, the primary genome generates intermediate (ancient) and ultimately modern genomes. The tree corresponds to discrete time



primary genome

intermediate (ancient) genomes

extant genomes

Species tree $S$

A gene undergoes three types of changes: continuous character substitution, insertions and deletions of blocks of characters. Thus, an instant gene is a sequence, and a gene sampled over time is a cluster of similar sequences (a function of time).

Dynamic in case of character **substitution**. Let a gene be sequence σ that transforms into sequence σ′ of the same length in time t, with the i-th position transition rate $\gamma_i$ Given the transition rate matrix R, we estimate the transition probability trivially:

$$\ln \prod_i \left(e^{\gamma_i t R}\right)(\sigma_i, \sigma_i')$$

If insertions and deletions are allowed, sequences σ and σ′ may differ in length. Their subsequent alignment produces new sequences, $\bar{\sigma}$ and $\bar{\sigma}'$ E.g., primary sequences are

**GGGTTTCAAACCATTGGCCCAATGGG** σ

**TGGTTTCAAACCAATTTGG** σ′

and their alignment (new sequences) are:

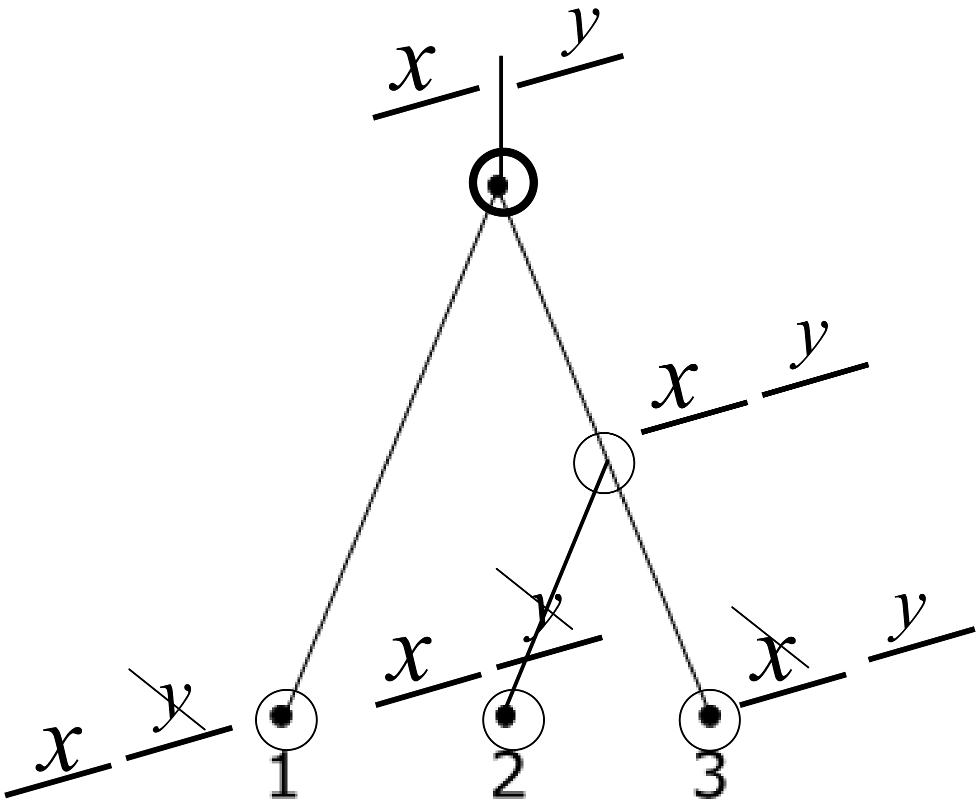
**GGGTTTCAAACCA-T-TGGCCCAATGGG** $\bar{\sigma}$

**TGGTTTCAAACCAATTTGG---------** $\bar{\sigma}'$

Designate the lengths of empty strings as *lm.* Estimate such transition probability:

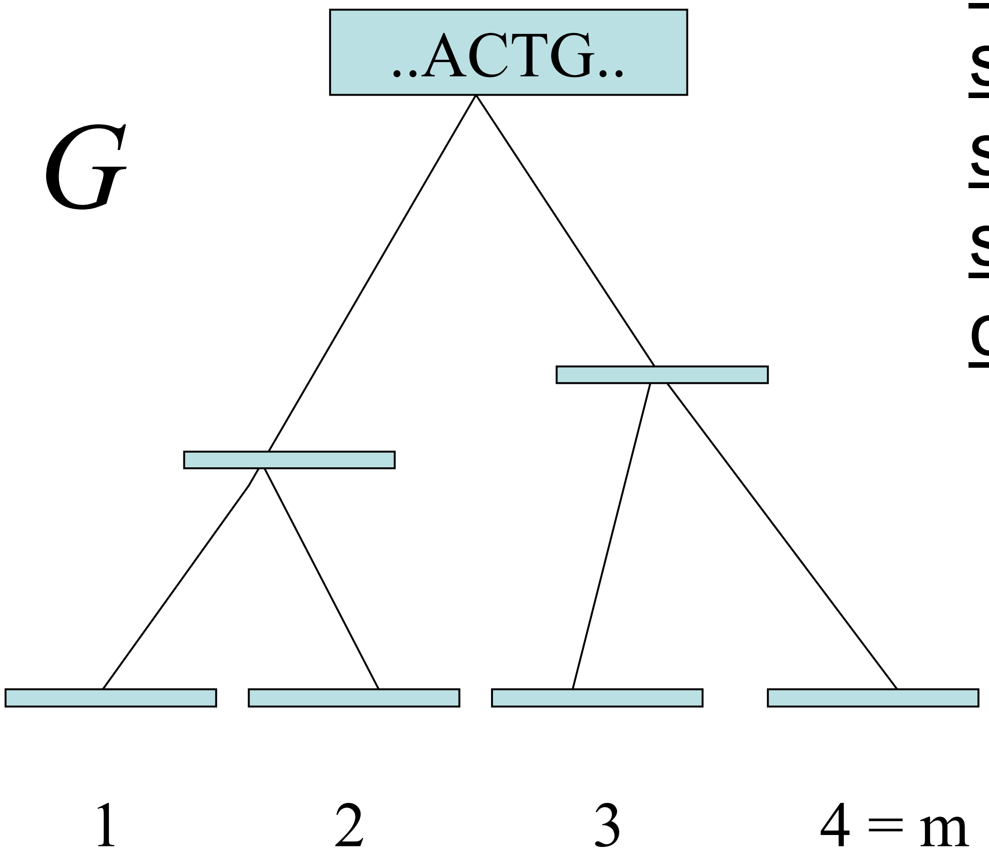
$$\ln \prod_{i}{}' \left(e^{\gamma_i t R}\right)\left(\overline{\sigma}_i, \overline{\sigma}'_i\right) - 10 \cdot \sum_{m} \ln\left(l_m + 1\right)$$

Some genes, apart from speciation, undergo the events of duplications, losses, etc.
Here the gene duplicated into x and y in the root, and some of its copies were lost in the leaves:



Species tree $S$

**Given** are a gene tree and modern sequences; edge lengths are times of transition from ancestors to descendants.



We search for all ancient sequences and secondary structures in all sequences; name this set configuration *σ*

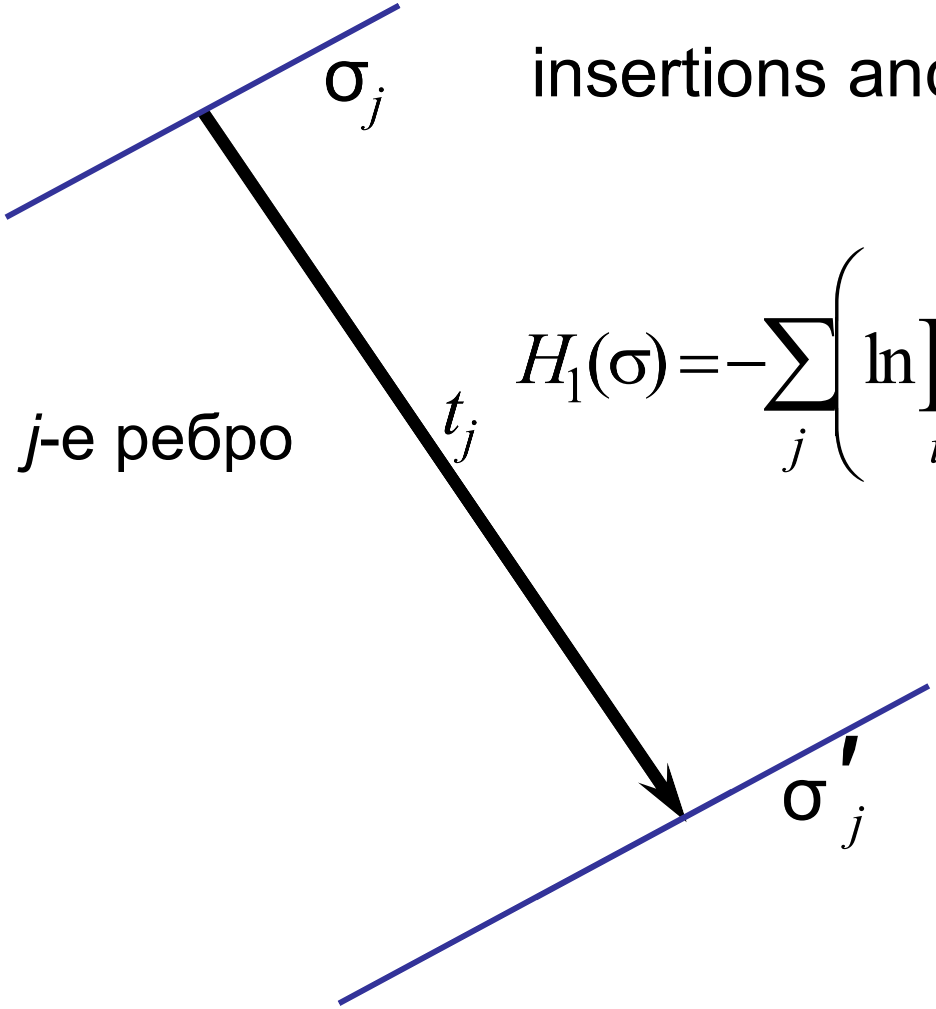
## In our model the desired configuration is defined by the global minimum of functional:

$$H(\sigma) = H_1(\sigma) + H_2(\sigma)$$

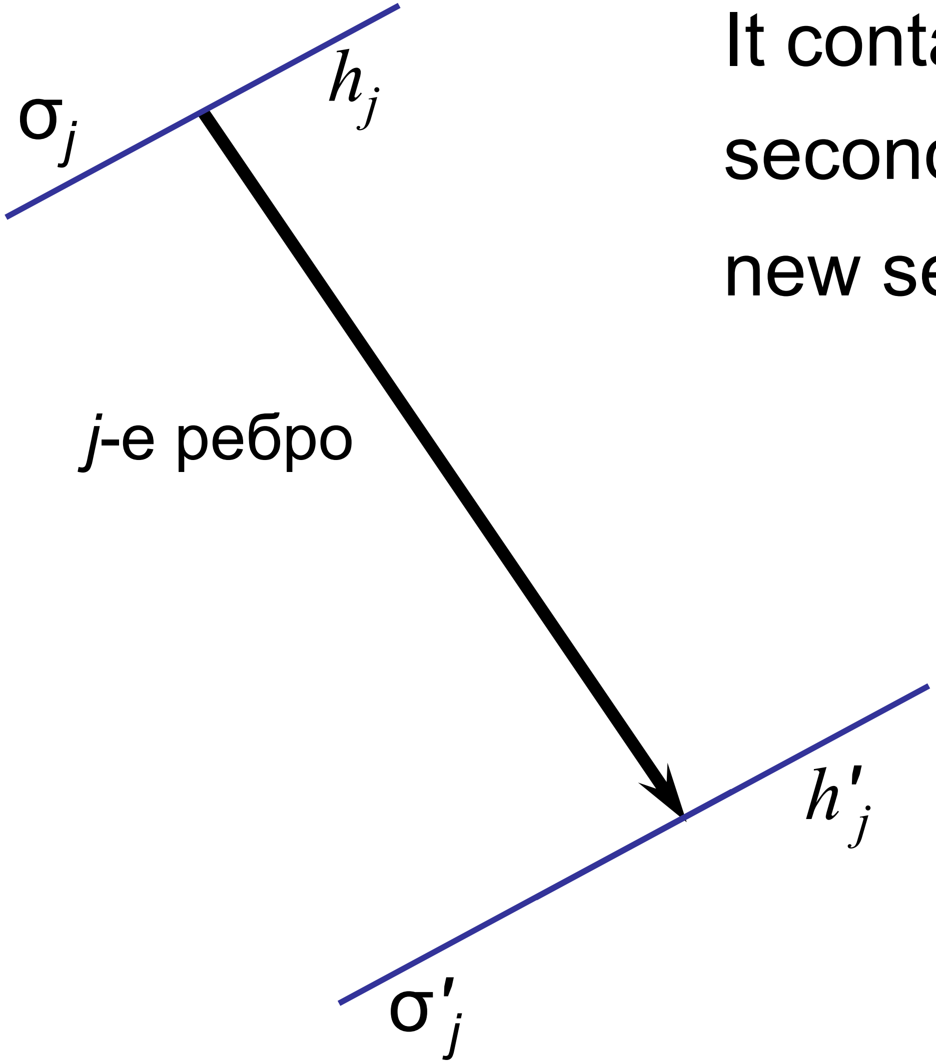
**Слагаемое**
**$H_1(\sigma)$ в $H$**

Consider an edge from certain configuration σ. Over time $t_j$ it contains character substitutions with rates R, insertions and deletions. Define as above:

$$H_1(\sigma) = -\sum_j \left( \ln \prod_{i=1}^{n_j}{}' \left(e^{\gamma_i t_j R}\right)\left(\bar{\sigma}_{ji}, \bar{\sigma}'_{ji}\right) - 10 \cdot \sum_m \ln\left(l_{jm}+1\right) \right)$$



# Слагаемое $H_2(\sigma)$ в $H$



Another edge from configuration σ. It contains a transition from secondary structure $h_j$ in $\sigma_j$ to new secondary **structure** $h'_j$ in $\sigma'_j$.

Тогда:

$$H_2(\sigma) = -\sum_j \Phi\left(h_j, h'_j\right)$$

We minimize the functional with annealing. At each algorithm step current configuration $\sigma$ is replaced by new configuration $\tilde{\sigma}$ from a set of candidates with probability

$$q(\sigma,\tilde{\sigma}) = \exp\left\{-\beta_m \cdot \left[H(\tilde{\sigma}) - H(\sigma)\right]^+\right\}$$

or is kept unchanged with probability $(1-q)$

Convergence to the global minimum is proved under

$$\lim_{m\to\infty} \frac{\log m}{\beta_m} > const$$

# Solution (partly shown): evolution of the ancient signal

```
gGTTGGGGCGGGCcgctgtcttcgaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc       N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa       N03: H3=-45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa       N12: H3=-61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa       N13: H3=-47.5
tGTTGGGGCAGGCTgctgagcgcaaaatttcacAAAAAAGGCCTGTATCCCAACcGATACAGGCCTTTTTTta        VC:  Σ=-234.3
```

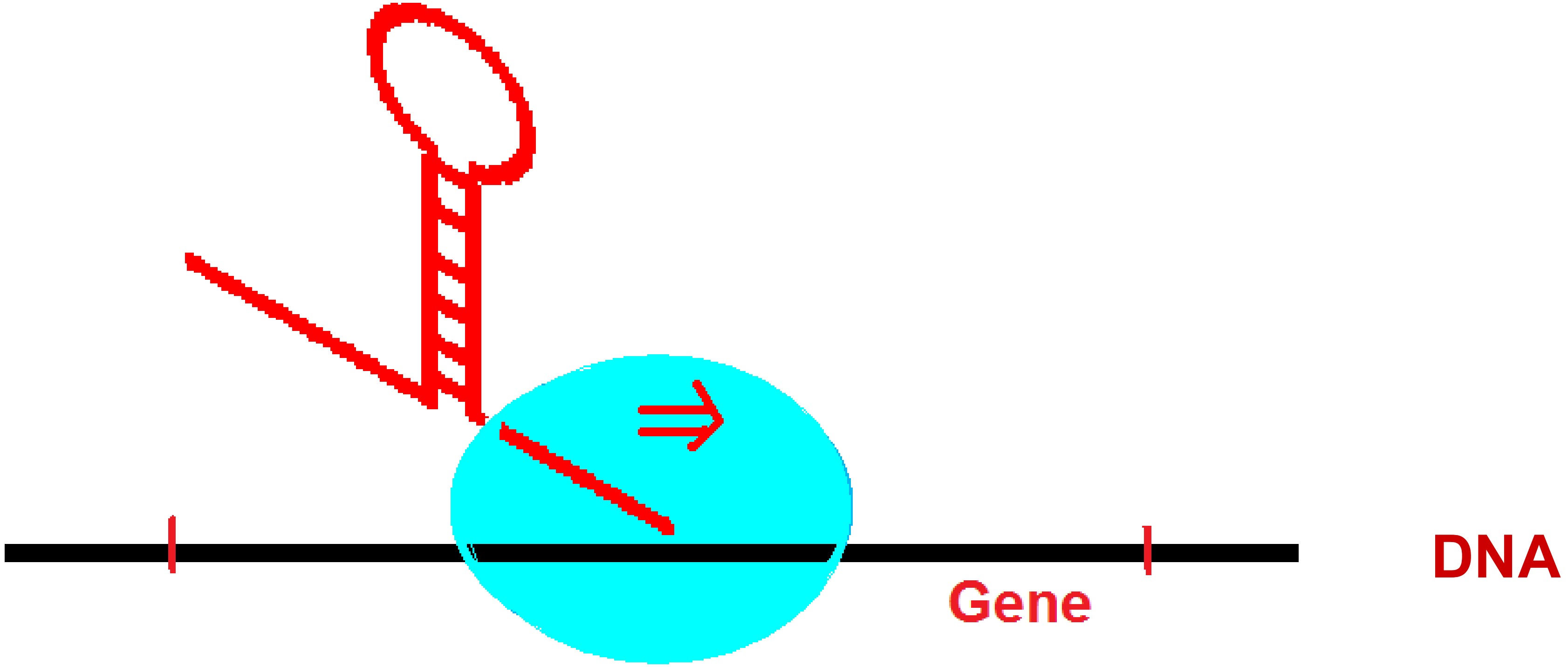
```
gGTTGGGGCGGGCcgctgtcttcgaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc       N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa       N03: H3=-45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa       N12: H3=-61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa       N13: H3=-61.3
TGTTGGGGCAGGCTgctgagcgaaagaacaaatttcAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa    VV:  Σ=-248.1
```

```
gGTTGGGGCGGGCcgctgtcttcgaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc       N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa       N03: H3=-45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa       N12: H3=-57.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcacAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTta        VP:  Σ=-182.6
```

```
gGTTGGGGCGGGCcgctgtcttcgaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc       N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa       N03: H3=-39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa       N04: H3=-24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata       N09: H3=-39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata       N10: H3=-51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata        N11: H3=-6.2
aaTGGGGCGGGctagtgcgttgaagaatagaattcatGAACCCGCaTTTCCCGAGaGCGGGTTTttttatg          AB:  Σ=-240.5
```

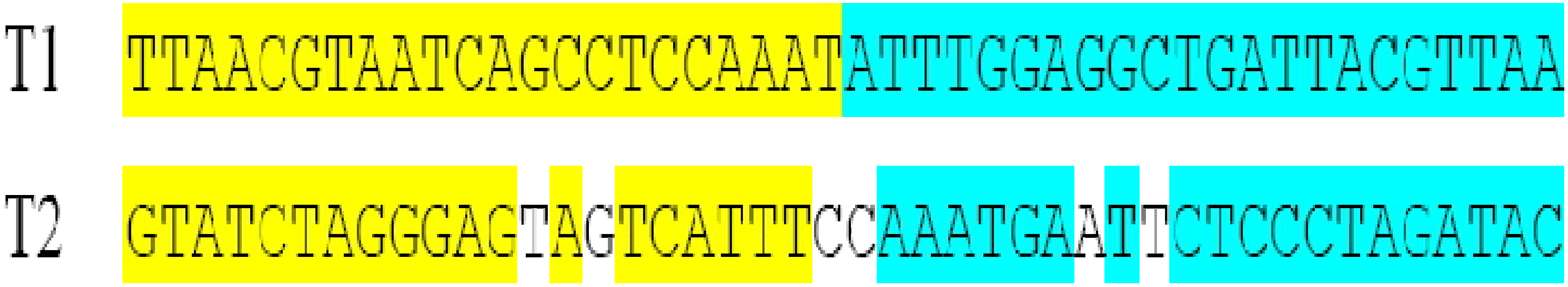
```
gGTTGGGGCGGGCcgctgtcttcgaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc       N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa       N03: H3=-39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa       N04: H3=-24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata       N09: H3=-39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata       N10: H3=-51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata        N11: H3=-35.0
aaTGGTGCGGGTTagtgcagcaaaaacaagatacAGAAAACCCGCGATTCAactGAATaGCGGGTTTTTTtata       HI:  Σ=-269.3
```

**The polymerase** is a machine that slides along DNA in a certain direction and reads a gene if reaches it (similar to a drive read head)



The polymerase **can detach** from DNA, e.g., after encountering such a DNA helix

A stack array of two or more sequences that maximizes their similarity is the “**alignment**”:

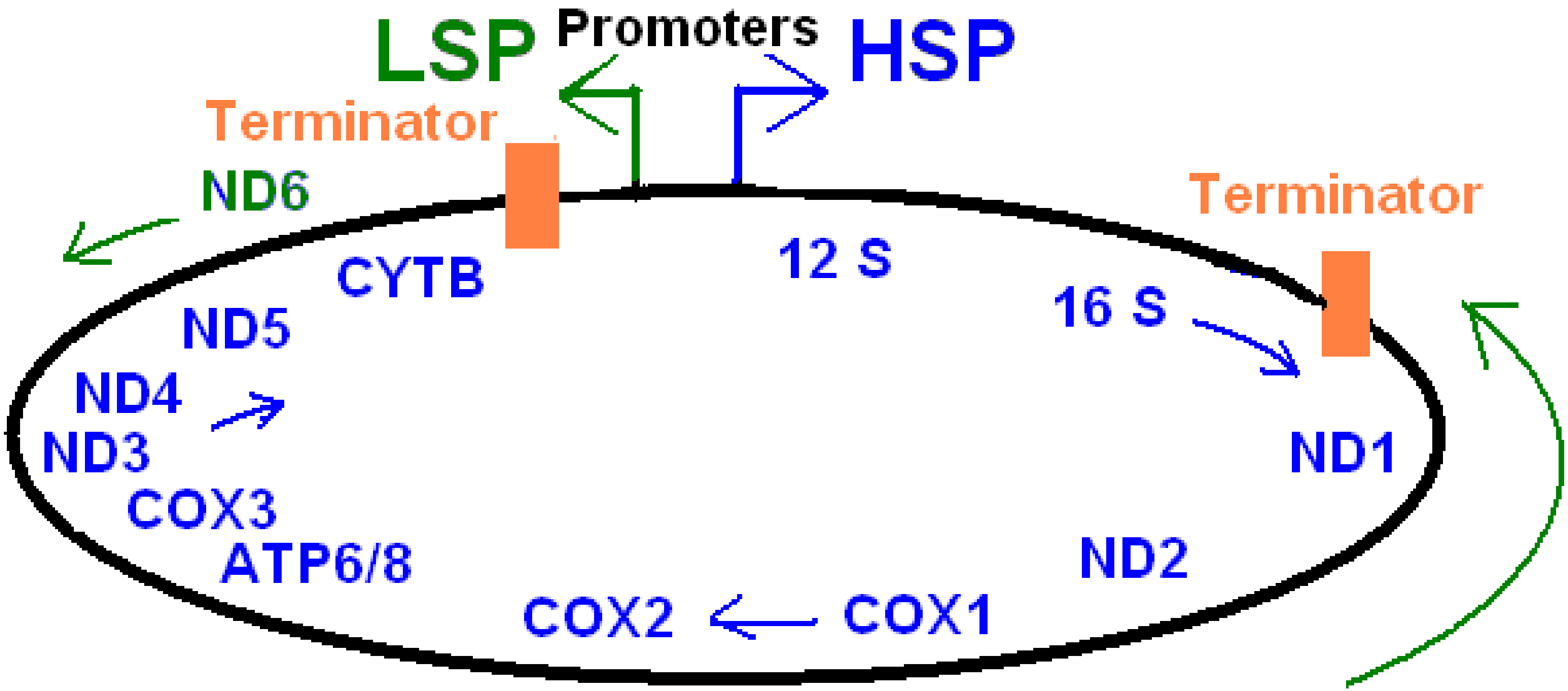


```
T1   TTAACGTAATCAGCCTCCAAATATTTGGAGGCTGATTACGTTAA

T2   GTATCTAGGGAGTAGTCATTTCCAAATGAATTCTCCCTAGATAC
```

Evidently, structures T1 and T2 are similar in folding into helices

# RNA Polymerase Competition in the circle case (mitochondrial DNA)


LSP
Promoters
HSP
Terminator
ND6
Terminator
12 S
CYTB
16 S
ND5
ND4
ND3
ND1
COX3
ATP6/8
ND2
COX2
COX1