IMSV: bacteria simulator specifications and architecture Version 0.0

M. Dinh and S. Fischer

October 22, 2015

Abstract

xxx

1 Introduction

The objective of the project is to develop a simulator for a generic bacteria. The project is performed in parallel to the creation of an ontology and the associated database that describes living systems. This document presents the specifications of the simulator.

2 Specifications

F10 The simulator shall be able to simulate the following scenario:

(a) different kind of (constant) extracellular conditions;

(b) upshift and downshift (change of nutritional condition) and also some stress scenarios;

(c) change of temperature;

(d) entry in stationary phase.

F20 It shall be possible to kill a cell among others.

F30 It shall be possible to initialize the state of the simulation. (A nice to have feature: moreover some aggregation or division of the state may be needed for such an initialization. For instance, if a cell has been simulated with one volume, a state division is needed to initialize a simulation with a cell simulated on several volumes.)

F40 Several stopping criterion shall be implemented:

- at the end of cellular division;
- simulated time exceeds a certain value;
- simulation time exceeds a certain value.

Any combination of criteria may be active.

F50 Typical values to be simulated are:

- 3e6 proteins (including proteins included in ribosomes);
- $[1e31e4]$ mRNA;
- $[4e3 * nbcopiesDNA]$ genes, each gene corresponding in mean to $[430 * 3basis]$;
- time cycle duration typically belongs to [20240] min.

F60 A simulation of the lifetime (cell reproduction) of a cell (Bacillus Subtilis) shall last 8 hours at most.

F70 The simulator shall output an ASCII logfile containing at least:

- the operating system;
- the version of the simulator and softs used;
- the date, the computer and the user;
- the localization of the data and results files.

F80 The simulator shall be composed of 3 modules:

- the first module outputs the description of the simulation from queries of the database. This description is contained in an ASCII file that uses SBML tags to the maximum;
- the second module is the simulation itself. From the ASCII file provided by the first module, it outputs the simulation results in a set of ASCII files. These files shall be readable from Excel 2013 or Matlab 2012a;
- the third module is the analysis and visualization of the simulation results provided by the second module. (A nice to have feature is a visual representation of the simulation.)

F90 The functionalities of the simulator shall be separated from its data and results.

F100 It shall be possible to launch a simulation interactively or by batch.

F110 It shall be possible to select the data to be recorded. By default all data shall be recorded.

F115 An option of log data / fill shall save the data necessary to restart the simulation at any given time point. Consequently, the simulator initialization has to be compatible to this data log.

F120 The cellular processes described in WholeCell shall be simulated. Beyond the simulation of a large set of cell sub-systems, a key achievement of the simulator is to be able to manage the so-called growth rate management, allowing to handle the resource balance problem. That leads to integrate in a suitable way the stringent control loop (i.e. RelA/GTP/ppGpp).

F125 xxx 1 partie moyenne (grossier) et 1 partie stochastique (fine) xxx

F130 The simulator shall be able to simulate different kind of process modeling. Typically, 4 kinds of modelling are envisaged for the process $S \longrightarrow P$

- $P = f(S)$: this is a static (steady state) relation. Typically corresponding to the equilibrium manifold of a fast (and attractive) dynamics as e.g. enzymatic dynamics;
- $\dot{P} = f(S)$: a deterministic and continuous dynamical systems, e.g., a detailed model of a subpart of the metabolic network (the model is deterministic but its parameters could be stochastics as e.g. the enzyme levels in a metabolic network);
- $P(nk+1) = f(P(nk), ...) = P(nk) + 1$: a deterministic and discrete model in time and state i.e. the states are integer (number of). No averaging, we are at the level of one bacteria. The event (time) is assumed to be deterministic;
- $P(t+) = f(P(t-), ...) = P(t-) + 1$: a stochastic and discrete model in time and state see previous description. By contrast to the previous one, the time event is assumed to be stochastics (mainly exponential distributed)

Furthermore, the global simulator is then a composite set of different kinds of this modeling (among the four), however a particular sub-cellular process is modeled only by one of this modeling.
Furthermore, some of the previous modeling are only possible if some states are available and then considered in the simulator. That is the case e.g. with respect to the metabolite concentrations.

F140 ~~This specification assumes that it is possible to obtain metabolite pool during the simulations. The interface between the different modeling granularities is the pool of metabolites.~~

F140 The set of cell-subsystems shall be defined and has to include the definition of interfaces with the the WholeCell, the possible kind of modeling (including compatibility of this modeling with the WholeCell, etc.).

F150 It shall be possible to impose the output of a particular process, typically for debug and validation (coupled with the fact that the simulation can be initialized at any time point with respect to a given datalog of a previous simulation).

D10 The development of the simulator shall be versioned.

D20 The simulator shall be documented with:

- a user manual;
- a developer manual.

D30 The simulator shall be validated on Bacillus Subtilis.

3 Data and states

It is recommended to implement the simulator with the following architecture

Simulation object

1. Time (state)
2. Interface between extracellular conditions (processes)
3. Extracellular conditions 1 (object):
   (a) concentration of entities (glucose, metabolites, AA or whatever) (state)
   (b) evolution of the concentration (process)
4. Extracellular conditions 2
5. ...
6. Extracellular conditions n
7. Cell 1 (object)
   (a) Interface between volumes (which volume interact with which other volume and the corresponding surface) (object)
   (b) Volume 1 (object)
      i. Extracellular conditions (processes, numbers, surfaces) (should allow for interaction between cells through a virtual extracellular conditions)
      ii. Volume (state)
      iii. Number of ribosome in different states or specific ribosome and its state (Ribosome i, free occupied) (state)
      iv. Number of mRNA in different states or specific mRNA and its state (state)
      v. ...
      vi. Number of metabolites in different states (state)
      vii. Volume Chromosome 1: table of the same size describing the state of the chromosome (state)
      viii. Volume Chromosome 2
      ix. ...
      x. Volume Chromosome n (the same number as for the cell) (state)
      xi. Internal (to the volume) processes
   (c) Volume 2
   (d) ...
   (e) Volume n
   (f) Cell Chromosome 1 (object):
      i. Table of mre / filles (state)
      ii. Table (size = number of structure (gene or codon or groups of codons) x maximal number of brindilles) with the number of the cell containing the structure. By convention, 0 = non active/does not exist; -1 = to fork (go the chromosomes filles); -2 = from fork (go the chromosome mre) (state)
      iii. Description of the structures (which AA needed), description of the genes (which structures, to produce what) (data)

(g) Cell Chromosome 2 if needed
(h) ...
(i) Cell Chromosome n if needed
(j) Internal (to the cell) processes, typically replication or DNA movement from a volume to another one

8. Cell 2
9. ...
10. Cell n

Basically, cell only communicate with extracellular conditions. Each cell is divided into volumes that can communicate either with extracellular conditions or other volumes of the same cell. The chromosome is divided into 'cell chromosome' which handles the continuity of the chromosome and 'volume chromosome' which handles transcription and translation.

3.1 Chromosome

There are 'cell chromosome' and 'volume chromosome' because we wish that only a higher level object may access to the data and state of a lower object; this distinction is otherwise not needed and the chromosome object should be placed at cell level. 'Cell chromosome' mainly handles the continuity of the chromosome of the cell through all the volumes. 'Volume chromosome' directly handles the biological process internally to each volume. 'Cell chromosome' and 'volume chromosome' must be coherent.

Representation A chromosome is represented in a condensed way in the simulator. In fact depending on the aggregation performed, the representation can be more or less fine.

NTP 1	NTP aggregation 1
NTP 2 NTP 3 NTP 4	NTP aggregation 2
NTP 5 NTP 6 NTP 7 NTP 8 NTP 9	NTP aggregation 3
NTP 10 NTP 11	NTP aggregation 4
NTP 12 NTP 13 NTP 14 NTP 15 NTP 16 NTP 17 NTP 18 NTP 19 NTP 20 NTP 21	NTP aggregation 5
NTP 22 NTP 23 NTP 24 NTP 25 NTP 26 NTP 27 NTP 28 NTP 29	NTP aggregation 6
NTP 30	NTP aggregation 7

Note that with this aggregation, there are some conflicts between DNA replication / transcription and DNA damage / repair / mutation... Unless some rules are enforced to solve the conflicts, it is not possible to have different kind of process be performed the same time on a NTP aggregation.

Assumption 3.1.
1. To avoid change of a volume state to another one due to, for instance, a polymerase changing of volume, it is assumed that a gene is necessary contained in one volume;

2. The decomposition into NTP aggregations applies to any biological process related to DNA (replication, transcription, translation and manipulation).

This assumption should not too drastic as a gene is much smaller than a chromosome.

3.1.1 Cell Chromosome

Evolution state For DNA replication purpose, it is needed to represent fork. It is assumed that the maximal number of forks n in known and that the considered chromosome is represented by m NTP aggregations. Then, the chromosome can be represented by 3 tables:

1. a main table of size $m \times \sum_{i=0}^{n} 2^i$ containing the number of the volume where it is. Moreover, -1 means go to 'mother', -2 means go to 'daughters' and -3 means go both to 'mother' and 'daughters';
2. a 'mother' table of size $1 \times \sum_{i=0}^{n} 2^i$ representing one strand of the fork;
3. a 'daughters' table of size $2 \times \sum_{i=0}^{n} 2^i$ representing the two last strands of the fork.

We now illustrate how it works. Assume that the chromosome (with $n = 2$ and $m = 10$) does not have any fork yet and is contained in the volume 2, then the tables are

Main table	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
	2	0	0	0	0	0	0
'Mother' table	0	0	0	0	0	0	0
'Daughters' table	0	0	0	0	0	0	0
	0	0	0	0	0	0	0

Now assume that the NTP aggregation 1 to 5 and 7 to 10 were replicated, one strand being in the volume 2 and the other one in the volume 1, the tables are

Main table	0	2	1	0	0	0	0
	0	2	1	0	0	0	0
	0	2	1	0	0	0	0
	0	2	1	0	0	0	0
	-2	2	1	0	0	0	0
	2	-1	-1	0	0	0	0
	-2	2	1	0	0	0	0
	0	2	1	0	0	0	0
	0	2	1	0	0	0	0
	0	2	1	0	0	0	0
'Mother' table	0	1	1	0	0	0	0
'Daughters' table	2	0	0	0	0	0	0
	3	0	0	0	0	0	0

Assume that there is another fork in the volume 1 for the NTP aggregation 1 only contained in volume 1 also while a part of the strand moved to volume 3, then the tables are

Main table	0	2	-2	1	1	0	0
	0	2	1	-1	-1	0	0
	0	2	1	0	0	0	0
	0	2	1	0	0	0	0
	-2	2	1	0	0	0	0
	2	-1	-1	0	0	0	0
	-2	2	3	0	0	0	0
	0	2	3	0	0	0	0
	0	2	3	0	0	0	0
	0	2	3	-1	-1	0	0
'Mother' table	0	1	1	3	3	0	0
'Daughters' table	2	0	4	0	0	0	0
	3	0	5	0	0	0	0

Finally assume that the last fork is performed on NTP aggregations 1 to 5 (even if it is a weird as it would mean that the replication is performed only in one direction, see the continuation of this example below) and that the resulting strands are contained in volume 3 even if the 'mother' strand is contained in volume 2, then the tables are

Main table	0	-2	-2	1	1	3	3
	0	0	1	-1	-1	3	3
	0	0	1	0	0	3	3
	0	0	1	0	0	3	3
	-2	-2	1	0	0	3	3
	2	-1	-1	0	0	-1	-1
	-2	2	3	0	0	0	0
	0	2	3	0	0	0	0
	0	2	3	0	0	0	0
	0	2	3	-1	-1	-1	-1
'Mother' table	0	1	1	3	3	2	2
'Daughters' table	2	6	4	0	0	0	0
	3	7	5	0	0	0	0

Original chromosome data This state contains other needed information for transcription and translation purposes. Moreover, they do not change as often that the evolution state but it can be the case whenever something happens to the DNA. The needed information are

- Transcription: the location of the genes, which bases are needed, kinetic parameters (typically how long does it takes to replicate or transcript it);
- Translation: the needed AA, the protein produced which is the property of the gene;
- Manipulation: is it an original NTP aggregation?

With a 10 NTP aggregations chromosome, the following original information state table

Gene						Base				AA				
number	oriC	terC	Start	Stop	Nb of NTPs	A	T	C	G	AA 1	...	AA i	...	AA n
1	1	0	0	0	1	1	1	0	1	0	...	0	...	0
2	0	0	1	0	1	1	0	1	1	0	...	0	...	0
2	0	0	0	0	5	5	4	5	1	1	...	2	...	2
2	0	0	0	0	4	2	1	5	4	1	...	1	...	0
2	0	0	0	1	1	2	1	0	0	0	...	0	...	0
3	0	1	0	0	1	0	1	1	1	0	...	0	...	0
4	0	0	1	0	1	1	0	1	1	0	...	0	...	0
4	0	0	0	0	5	1	1	12	1	0	...	0	...	1
4	0	0	0	0	3	3	3	0	3	0	...	3	...	0
4	0	0	0	1	1	2	1	0	0	0	...	0	...	0

indicates

- oriC as gene 1, encoded in NTP aggregation 1. The corresponding codon is any combination of bases ATG;
- terC as gene 3, encoded in NTP aggregation 6. The corresponding codon is any combination of bases TCG;
- 2 real genes (2 and 4) encoded between NTP aggregations 2 and 5 and between NTP aggregations 7 and 10;
- both genes have the same start and stop codon (coded by any combination of ACG and AAT) and the do not need any AA;
- the translation elongation of NTP aggregation 3 would need 4 AA 1, ..., 1 AA i, ... and 8 AAn. Its transcription needs the equivalent of 5 bases A, 4 bases T, 5 bases C and 1 base G whereas its replication would need the double of bases;
- in my little head, a codon is composed of 3 bases and each codon encodes one AA. The column 'Number of codons' should be consistent with the 'Base' and 'AA' columns. The AA coded bu aggregation 3 are all displayed in the table whereas it is not the case for codon aggregation 4 (it lacks 2 AA).

Note that with this kind of modeling, the oriC, terC, start and stop codons can be aggregated with others. *This is not recommended however*. Another way would be to delete the oriC and terC property and reserve some values (1 and -1 for example) in the Gene number to them.

Note, for gene 4 and more generally in this part of the chromosome, the replication and the translation go in reverse way =¿ conflict.

In fact, not gene but rather operons are described for transcription. Delete the part from AA on the chromosome description and add something about the the operons. Dissociate for each operons the number of mRNAs possible that are created from on operon, and for each mRNA the number for proteins that are created.

In fact, it might be easier to not use the codon aggregations and directly code every codon...

Current chomosome state It is the same as the original data except that it indicates the current state of the DNA. The original table allows to retrieve the original DNA when a DNA repair has been performed on a damaged DNA. In that sense, the current table also contains infirmation about the DNA damages.

Gene state Should just be something that tells for each gene what it produces. As I see only proteins at the moment as output, is there something else potentially? How do we store this information?

3.1.2 Volume Chromosome

Volume chromosome contains information about the gene and codon aggregations that are inside the volume. They also contains information necessary for an independent processing of transcription and translation, but not DNA replication (which is a cell process).

The information needed for modeling are

- what are the different state, more precisely, how other entities can bind to DNA for transcription;
- in what step is the elongation (all the bases are in place?);
- in order to manage conflict with replication, a state about occupation by replication is added.

The volume chromosome is organized by operons with the same codon aggregations. Also adds an information about the tracking of the strand of the operon.

Mind backtrack: per operon, the number, the strand it comes from.

3.2 mRNA

A mRNA has the same kind of structure as an operon. Add binding site (and a stop site?) to Chromosome characteristics.

Codon	AA 1	AA 2	AA 3	AA 4	Aggregation	Binding site	Stop site	AA 1	AA 2	AA 3	AA 4
1	0	0	0	0	1	1	0	0	0	0	0
2	0	0	1	0							
3	1	0	0	0	2	0	0	1	0	2	0
4	0	0	1	0							
5	0	0	0	1							
6	1	0	0	0	3	0	0	1	0	0	1
7	0	0	0	0	4	0	1	0	0	0	0

So an mRNA is represented by

1. what it produces (how this information is stored?);
2. the aggregated table of information;
3. per aggregation, the state. Basically, for a brief description, we assume that the translation is performed in the following steps
   (a) binding with 30S;
   (b) binding with 50S. Here we assume that the 50S binds only if the 30S is already binded to form a 70S;
   (c) elongation: binding of tRNA (charged or not), binding of ATP and step forward (one step). This step is repeated until the end;
   (d) stop (not described here in term of state because on't know how it works).

Follows an example. Assume that there is already a 70S at the beginning of the aggregation 2.

Aggregation	30S	50S	non charged tRNA	tRNA 1	tRNA 2	tRNA 3	tRNA 4	AA 1	AA 2	AA 3	AA 4
2	0	1	0	0	0	0	0	0	0	0	0

Now assume that a tRNA charged with an AA 1 is binded.

Aggregation	30S	50S	non charged tRNA	tRNA 1	tRNA 2	tRNA 3	tRNA 4	AA 1	AA 2	AA 3	AA 4
2	0	0	0	1	0	0	0	0	0	0	0

Then elongation is performed. Of course an ATP would be consumed but it is not here to be modeled.

Aggregation	30S	50S	non charged tRNA	tRNA 1	tRNA 2	tRNA 3	tRNA 4	AA 1	AA 2	AA 3	AA 4
2	0	1	0	0	0	0	0	1	0	0	0

Now assume that a 30S binds to the binding site, and then a 50S also binds to form a 70S. The state would evolve like follows.

Aggregation	30S	50S	non charged tRNA	tRNA 1	tRNA 2	tRNA 3	tRNA 4	AA 1	AA 2	AA 3	AA 4
2	0	1	0	0	0	0	0	1	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0

Aggregation	30S	50S	non charged tRNA	tRNA 1	tRNA 2	tRNA 3	tRNA 4	AA 1	AA 2	AA 3	AA 4
2	0	1	0	0	0	0	0	1	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0

Now we assume that we allow the 70S in aggregation 1 to go to aggregation 2 because we are very happy with this happening (!).

Aggregation	30S	50S	non charged tRNA	tRNA 1	tRNA 2	tRNA 3	tRNA 4	AA 1	AA 2	AA 3	AA 4
2	0	1	0	0	0	0	0	1	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0

It is possible to be less detailed about the tracking of each ribosome.

3.3 Metabolites and proteins

The difference between metabolites and proteins (and macro-proteins) is delimited by properties:

- metabolites do not have properties;
- proteins do have properties (such as folded, phosphorilized...).

Moreover, the distinction between protein and macro-protein is defined here as:

- a macro-protein is the association of two or more proteins for which the properties can be applied indifferently on each of the protein;
- whatever else proteinic entity. Any entity that comes directly from the translation is considered to be a protein and not a macro-protein.

How to store weird things? Unfinished proteins, degradated ones...

3.3.1 Metabolites

Of course, only concerns the ones that are not already described elsewhere. No problem of representation: as there is not any property, the state is the pool of metabolite. A proposed way is to list all the metabolites (let's say there are m types metabolites), to number them and to represent the pool of metabolites by a vector of size m containing the pool of each metabolites.

3.3.2 Proteins and macro-proteins

Of course, only concerns the ones that are not already described elsewhere.

Information needed for proteins are:

degradation the composition of the 'standard' protein (no property) and what more entities a property comes with;

allowed processes the properties that can take place with this protein (maybe some proteins cannot be phosphorized!!!).

Depends on how degradation works (or rather what do we want to model). If degradation of a macro-protein is modeled at the level of the proteins or of the AA, the data and state are not the same.

Consider

$$P_1 + P_2 \rightleftharpoons P_1P_2$$

where P_1 and P_2 are proteins. Assume that:

- there are 10 AA, P_1 is composed of 10 AA1, 2 AA4 and 4 AA9 and P_2 is composed of 37 AA7;
- 4 properties are defined for the proteins; for P_1 processes 1, 2 and 4 are allowed whereas only process 2 and 3 are allowed for P_2;
- there are 5 'standard' proteins of each P_1 and P_2. No P_1P_2.

One would have as data and state

P_1 is a protein

composition is a vector $compP1$ of size 10 (the number of AA) and it contains the number of AA of the composition

AA1	AA2	AA3	AA4	AA5	AA6	AA7	AA8	AA9	AA10
10	0	0	2	0	0	0	0	4	0

allowed processes is a vector $procP1$ of size 4 (the number of processes) and it contains

	Property			
booleans	1	2	3	4
	1	1	0	1

state is a matrix $stateP1$ of size 5×4 (number of P_1 proteins $\times$ number of processes) and it contains booleans representing the state of *each* protein. Here 5 'standard' (no property) proteins P_1

Protein	Property			
	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0

P_2 is a protein

composition is a vector $compP2$ of size 10 (the number of AA) and it contains the number of AA of the composition

AA1	AA2	AA3	AA4	AA5	AA6	AA7	AA8	AA9	AA10
0	0	0	0	0	0	37	0	0	0

allowed processes is a vector $procP2$ of size 4 (the number of processes) and it contains booleans

	Property			
	1	2	3	4
	0	1	1	0

state is a matrix $stateP2$ of size 5×4 (number of P_2 proteins $\times$ number of processes) and it contains booleans representing the state of *each* protein. Here 5 'standard' (no property) proteins P_2

Protein	Property			
	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0

P_1P_2 is a macro-protein

composition is a matrix $compP1P2$ of size $1 \times 10 \times 2$ (10 for the number of AA and 2 for the number of protein composing the macro-protein) and it contains the detailed composition with $compP1P2(1,:,1) = compP1$ and $compP1P2(1,:,2) = compP2$;

allowed processes is a matrix $procP1P2$ of size $1 \times 4 \times 2$ with $procP1P2(1,:,1) = procP1$ and $procP1P2(1,:,2) = procP2$;

state is a matrix $stateP1P2$ that is void.

Basically with the representation, the 'size' of the state depends on the pool of proteins and the number of properties. The state contain booleans only.

As a quick benchmark, with Matlab 2012a on Biosys1:

- initializing both a false matrix and a true matrix of size $3e6 \times 20$ takes 0.13 seconds;
- changing all the boolean in a matrix, using linear indexing in a loop, takes 2 seconds;
- changing all the boolean in a matrix, using 2 loops on row and column indexes, takes 3.3 seconds;
- there are around 3e6 proteins, each proteins having around 200 AA each. Subtilis cycle is 20 minutes, so there are at least 3e6*200/(20*60) = 5e5 events per second. Changing 5e5 booleans in a matrix, using row and column indexing, takes 5.5e-3 seconds.

We thus hope that a fine description of the DNA and proteins is possible.

Need to decide what is modeled as a protein and what is modeled as a macro-protein.

4 Biological processes

The minimum requirement is to be able to simulate a cell cycle during. Thus it should model

1. DNA: replication and segregation;
2. mass production: transcription, translation and metabolism. This includes the stringent response via Rhs;
3. cell division: cytokinesis and FtsZ ring.

4.1 Simple reaction

By simple, it is understood that the reaction is performed in one step, instantaneously. The result of a reaction is to change the pools of the reactants and products whenever the reaction happens. *When the reaction happens is dealt with in another process.* To describe a reaction, the needed information is thus the pools of the reactants and the enzyme as well as the stoichiometry of the reaction.

Let us consider the reaction written in the conventional direction:

$$\underbrace{\sum_{l\in\mathcal{I}_S} m_l^S S_l}_{Substracts\ S} \overset{E}{\rightleftharpoons} \underbrace{\sum_{l\in\mathcal{I}_P} m_l^P P_l}_{Products\ P}.$$

Note that if this is an enzymatic reaction, then the reaction should not be possible without the enzyme. So that the process is defined by:

- inputs:
  - the direction of the reaction;
  - the pool(s) S_l^{pool} of the substract(s);
  - the pool(s) E^{pool} of the enzyme(s);
  - the pool(s) X_l^{pool} of the product(s);
- parameters: the stoichiometry m_l^S and m_l^P;
- processes: if $E^{pool} > 0$ (otherwise nothing changes) then change
  - if the reaction is in the forward direction: $S_l^{pool} = S_l^{pool} - m_l^S$ and $P_l^{pool} = P_l^{pool} + m_l^P$;
  - if the reaction is in the backward direction: $P_l^{pool} = P_l^{pool} - m_l^P$ and $S_l^{pool} = S_l^{pool} + m_l^S$.

The above mechanism for process modeling works well as long as every possible 'state' of the reactants are states of the simulator. For instance, we consider the simple reaction (without enzyme for the sake of simplification)

$$P_1 + P_2 \rightleftharpoons P_1P_2 \longrightarrow Q_1Q_2.$$

In reality, both P_1 and P_2 have a property, say phosphorized or not. We then have the possible reactions

$$\begin{cases} P_1 + P_2 \rightleftharpoons P_1P_2 \longrightarrow Q_1Q_2 \\ P_1^* + P_2 \rightleftharpoons P_1^*P_2 \longrightarrow Q_1^*Q_2 \\ P_1 + P_2^* \rightleftharpoons P_1P_2^* \longrightarrow Q_1Q_2^* \\ P_1^* + P_2^* \rightleftharpoons P_1^*P_2^* \longrightarrow Q_1^*Q_2^* \end{cases}$$

In addition, we can also have the reactions

$$\begin{cases} P_1 + X \rightleftharpoons P_1^* \\ P_2 + X \rightleftharpoons P_2^* \\ P_1P_2 + X \rightleftharpoons P_1^*P_2 \\ P_1P_2 + X \rightleftharpoons P_1P_2^* \\ P_1^*P_2 + X \rightleftharpoons P_1^*P_2^* \\ P_1P_2^* + X \rightleftharpoons P_1^*P_2^* \\ Q_1Q_2 + X \rightleftharpoons Q_1^*Q_2 \\ Q_1Q_2 + X \rightleftharpoons Q_1Q_2^* \\ Q_1^*Q_2 + X \rightleftharpoons Q_1^*Q_2^* \\ Q_1Q_2^* + X \rightleftharpoons Q_1^*Q_2^* \end{cases}$$

Of course, if one puts all the states as described above, the process described above is fine. In this case, the simulator state would be the pools of

$$X,\ P_1,\ P_1^*,\ P_2,\ P_2^*,\ P_1P_2,\ P_1^*P_2,\ P_1P_2^*,\ P_1^*P_2^*,\ Q_1Q_2,\ Q_1^*Q_2,\ Q_1Q_2^*,\ Q_1^*Q_2^*$$

that is 13 states. We want to decrease the number of state. The idea is that the intermediate state are can be computed from the information on P_1, P_1^*, P_2 and P_2^* if we also know the number that are used for these intermediate reactants. Assume, we need the pools of

$$X,\ P_1,\ P_1^*,\ P_2,\ P_2^*,\ Q_1Q_2,\ Q_1^*Q_2,\ Q_1Q_2^*,\ Q_1^*Q_2^*.$$

One would then split in two pools (free and used for P_1 and P_2 type reactant) the ones of P_1, P_1^*, P_2 and P_2^* leading to also 13 states. Moreover, there is a need for computation to recover the pools of P_1P_2, $P_1^*P_2$, $P_1P_2^*$, and $P_1^*P_2^*$, which is really bad. The relations would have been

$$P_{1used}^{pool} + P_{1used}^{*pool} = P_{2used}^{pool} + P_{2used}^{*pool} = (P_1P_2)^{pool} + (P_1^*P_2)^{pool} + (P_1P_2^*)^{pool} + (P_1^*P_2^*)^{pool}$$

along with

$$\begin{cases} (P_1P_2)^{pool} = \min(P_{1used}^{pool}, P_{2used}^{pool}) \\ (P_1^*P_2^*)^{pool} = \min(P_{1used}^{*pool}, P_{2used}^{*pool}) \\ (P_1^*P_2)^{pool} = P_{2used}^{pool} - (P_1P_2)^{pool} = P_{1used}^{*pool} - (P_1^*P_2^*)^{pool} \\ (P_1P_2^*)^{pool} = P_{1used}^{pool} - (P_1P_2)^{pool} = P_{2used}^{*pool} - (P_1^*P_2^*)^{pool} \end{cases}$$

Now assume that we only need for the reactant of type Q_1Q_2 only the pool of Q_1Q_2 separately and that the other pools are not necessary:

$$X,\ P_{1free},\ P_{1used},\ P_{1free}^*,\ P_{1used}^*,\ P_{2free},\ P_{2used},\ P_{2free}^*,\ P_{2used}^*,\ Q_1Q_2$$

We only have 10 states but there is a loss of information as it is now impossible to distinguish between P_1P_2 and Q_1Q_2 type of reactants. The relation are now

$$\begin{cases} (P_1P_2)^{pool} = \min(P_{1used}^{pool}, P_{2used}^{pool}) \\ (P_1^*P_2^*)^{pool} + (Q_1^*Q_2^*)^{pool} = \min(P_{1used}^{*pool}, P_{2used}^{*pool}) \\ (P_1^*P_2)^{pool} + (Q_1^*Q_2)^{pool} = P_{2used}^{pool} - (P_1P_2)^{pool} = P_{1used}^{*pool} - \min(P_{1used}^{*pool}, P_{2used}^{*pool}) \\ (P_1P_2^*)^{pool} + (Q_1Q_2^*)^{pool} = P_{1used}^{pool} - (P_1P_2)^{pool} = P_{2used}^{*pool} - \min(P_{1used}^{*pool}, P_{2used}^{*pool}) \end{cases}$$

Not worth the bother (imagine for 2 properties on P_1 and P_2 instead of only 1!!!). Use the full information case or limit the number of properties or use the following intermediate solution with the state

$$X,\ P_1,\ P_1^*,\ P_2,\ P_2^*,\ Q_1Q_2,\ Q_1^*Q_2,\ Q_1Q_2^*,\ Q_1^*Q_2^*$$

and with

$$(P_1P_2)_{all},\ X_{used}$$

$(P_1P_2)_{all}$ standing for the whole P_1P_2, $P_1^*P_2$, $P_1P_2^*$ and $P_1^*P_2^*$ and X_{used} standing for X used in $(P_1P_2)_{all}$. A condition for feasibility is then $X_{used}^{pool} \leq 2(P_1P_2)_{all}^{pool}$. One would have the following possible reactions

$$\begin{cases} P_1 + P_2 \rightleftharpoons (P_1P_2)_{all} \longrightarrow Q_1Q_2 \\ P_1^* + P_2 \rightleftharpoons (P_1P_2)_{all} + X_{used} \longrightarrow Q_1^*Q_2 \\ P_1 + P_2^* \rightleftharpoons (P_1P_2)_{all} + X_{used} \longrightarrow Q_1Q_2^* \\ P_1^* + P_2^* \rightleftharpoons (P_1P_2)_{all} + 2X_{used} \longrightarrow Q_1^*Q_2^* \end{cases}$$

In addition, we can also have the reactions

$$\begin{cases} P_1 + X \rightleftharpoons P_1^* \\ P_2 + X \rightleftharpoons P_2^* \\ (P_1P_2)_{all} + X \rightleftharpoons (P_1P_2)_{all} + X_{used} \\ Q_1Q_2 + X \rightleftharpoons Q_1^*Q_2 \\ Q_1Q_2 + X \rightleftharpoons Q_1Q_2^* \\ Q_1^*Q_2 + X \rightleftharpoons Q_1^*Q_2^* \\ Q_1Q_2^* + X \rightleftharpoons Q_1^*Q_2^* \end{cases}$$

The difficulty would be to decide when these reactions can occur. An even 'more' intermediate solution would be to use the states (assuming that $Q_1^*Q_2$ is the reactant of interest since otherwise it is too simple!)

$$X,\ P_1,\ P_1^*,\ P_2,\ P_2^*,\ Q_1^*Q_2$$

and with

$$(P_1P_2)_{all},\ X_{Pused},\ (Q_1Q_2)_{all},\ X_{Qused}$$

with the same constraint as above. The set of reactions are then

$$\begin{cases} P_1 + P_2 \rightleftharpoons (P_1P_2)_{all} \longrightarrow (Q_1Q_2)_{all} \\ P_1^* + P_2 \rightleftharpoons (P_1P_2)_{all} + X_{Pused} \longrightarrow Q_1^*Q_2 \\ P_1 + P_2^* \rightleftharpoons (P_1P_2)_{all} + X_{Pused} \longrightarrow (Q_1Q_2)_{all} + X_{Qused} \\ P_1^* + P_2^* \rightleftharpoons (P_1P_2)_{all} + 2X_{Pused} \longrightarrow (Q_1Q_2)_{all} + X_{Qused} \end{cases}$$

In addition, we can also have the reactions

$$\begin{cases} P_1 + X \rightleftharpoons P_1^* \\ P_2 + X \rightleftharpoons P_2^* \\ (P_1P_2)_{all} + X \rightleftharpoons (P_1P_2)_{all} + X_{used} \\ (Q_1Q_2)_{all} + X \rightleftharpoons Q_1^*Q_2 \\ (Q_1Q_2)_{all} + X_{Qused} \rightleftharpoons Q_1^*Q_2 + X \\ (Q_1Q_2)_{all} + X \rightleftharpoons (Q_1Q_2)_{all} + X_{Qused} \end{cases}$$

Again, the difficulty would be to decide when these reactions can occur.

Finally, it could also be possible to loose even more information by deciding that there only an active form and some non-active forms (one non-active form per type: phosphorized or non-folded – hum mauvais exemple, pas grave – for instance). Only the active form can go forward in the pathway; the non-active form must go back into its previous state. Going back to our example with the pathway being:

$$P_1 + P_2 \rightleftharpoons P_1P_2 \longrightarrow Q_1Q_2.$$

One would consider with this pathway the non-active forms phosphorized and non-folded:

$$\begin{cases} P_1^{nonFold} \overset{C_1}{\rightleftharpoons} P_1 \\ P_2^{nonFold} \overset{C_2}{\rightleftharpoons} P_2 \\ P_1 + X \rightleftharpoons P_1^{phosphorized} \\ P_2 + X \rightleftharpoons P_2^{phosphorized} \\ (P_1P_2)^{nonFold} \overset{C}{\rightleftharpoons} P_1P_2 \\ P_1P_2 + X \rightleftharpoons (P_1P_2)^{phosphorized} \end{cases}$$

Would be able to model assembly of 30S, 50S and tRNA charging (hum...) if each pool is correctly encoded somewhere as a state.

4.2 DNA processes

4.2.1 DNA replication

During exponential growth, the cell needs to duplicate its DNA content in order to proceed to division. Replication of DNA needs to be well coordinated with cell growth and division to ensure viability of the daughter cells. There are three important phases: replication initiation, elongation and termination. These phases are controlled by several (partly unknown) mechanisms to adapt to external conditions and adjust growth rate. For example, several bacteria, including *E. coli* and *B. subtilis*, are able to initiate replication several times during one cell cycle, so that the cell cycle can actually be shorter (down to 20 minutes) than the time needed to replicate the full chromosome (approximately 40 minutes) (**?**).

Replication initiation Replication initiation of the chromosome is an event that appears to be precisely controlled. Replication should only be initiated if growth conditions are favorable. What is more, replication should generally be triggered exactly once during cell cycle. Even in excellent growth conditions, when cells inherit a chromosome already engaged in replication, the two origins of replication fire at exactly the same time (**?**). This implies the existence of mechanisms that inhibit replication initiation once replication has already started.

Initiation is mainly controlled by DnaA, a protein that can bind DNA in its activated form, DnaA-ATP. There are numerous DnaA binding boxes along the chromosomes, but only a few of them are essential for replication initiation. The latter are located next to the *oriC* locus (Figure 1), where the replisomes are loaded and replication actually starts. Interestingly, *oriC* and the *dnaA* gene are colocated in numerous bacteria (Briggs et al., 2012), so that the binding of DnaA inhibits its own expression, autoregulating the levels of DnaA available.

In the most simple model, DnaA polymerizes along the DnaA binding boxes, unwinding the DNA around *oriC*. It is probable that this unwinding is not necessarily performed by DnaA itself. For example, in *B. subtilis*, DnaA binds with DnaD, which is mainly responsible for untwisting (Briggs et al., 2012). Once the DNA is sufficiently untwisted, a neighboring AT-rich region, termed DNA-Unwinding Element (DUE), opens slightly, enabling loading of the first elements of the replisomes, the helicase and helicase loader (DnaC-DnaI for *B. subtilis*, DnaB-DnaC for *E. coli*). The position of DnaA binding boxes is not exactly conserved in different bacterial species, the loading mechanism is poorly understood. Because a loop is observed during replication initiation of *B. subtilis*, Briggs et al. (2012) propose that DnaD is used for loop-forming and that the loop enables synchronous loading of the two replisomes (one for each direction) through DnaB (Figure 2, top).

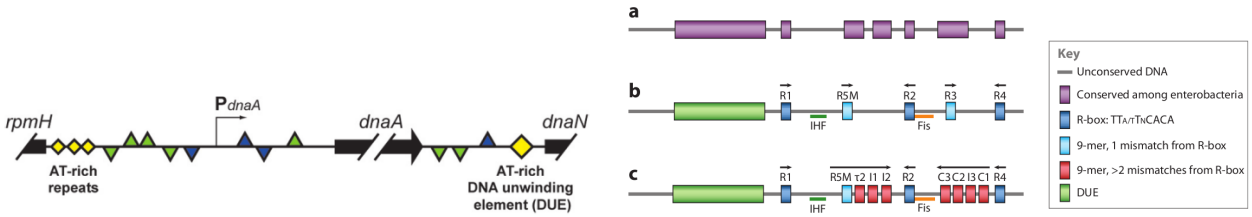

Figure 1: Around *oriC*, several DnaA binding sites (blue and green triangles) allow for the polymerization of DnaA and, eventually, opening of the DNA at the AT-rich DNA-Unwinding Element (DUE) (left). A more detailed view shows that there are three high affinity boxes (R1, R2 and R4) that serve as nucleation points and a lot of low-affinity boxes that guide polymerization (right). Figures from Leonard and Grimwade (2011) (right) and Briggs et al. (2012) (left).

Closer inspection shows that even if DnaA-ATP oligomerizes as a helix *in vitro*, it may only be capable of binding ssDNA *in vivo* and that the actual complex responsible for DNA unwinding also contains DnaA-ADP (Leonard and Grimwade, 2011). In this more advanced model for *E. coli*, DnaA may cooperate with DNA-binding proteins IHF (integration host factor) and Fis (factor inversion stimulation) to change the 3D conformation of the origin. Several interesting points highlight that the conformation changes originate from subtle interplay between the three proteins DnaA, Fis and IHF (**?**). (i) If none of these proteins binds DNA, the DUE opens spontaneously, while if the 3 main DnaA boxes are occupied by DnaA, it is closed. (ii) In the presence of Fis and IHF, up to two high-affinity binding boxes (R1, R2 and R4) can be deleted without impacting cell viability, while the cell dies if Fis or IHF is also removed. (iii) Fis can inhibit DnaA binding to a distance of up to 100 bp. Given these elements, it seems that the origin of replication is rather compacted, allowing all the proteins that bind in its vicinity to interact in a nucleosome-like structure (Fig. 2, bottom). The following ideas have been proposed (Leonard and Grimwade, 2011; **?**): (i) Fis binds to the origin and compacts it, inhibiting DnaA binding at first, (ii) when DnaA levels are high enough, DnaA manages to bind its high affinity boxes, dislodging Fis and quickly stabilized by IHF binding and DnaA oligomerization along weak affinity boxes, (iii) the DUE opens, DnaA-ATP can bind along the ssDNA, stabilizing the opening and recruiting the helicase and helicase-loader.

How initiation is controlled is largely unknown and may be species specific. For example, *E. coli* does not have homologs for *B. subtilis* DnaB and DnaD proteins (*E. coli* DnaB and DnaC are the homologs of *B. subtilis* DnaC and DnaI, respectively). As a result, the initiation is strongly dependent on DnaA-ATP levels in *E. coli* but less in *B. subtilis*, where DnaB and DnaD seem more critical (Briggs et al., 2012). In *E. coli*, it is estimated that around 20 DnaA-ATP are needed for intiation, the cell containing 1000 to 2000 DnaA molecules (Leonard and Grimwade, 2011). DnaA-ATP levels are controlled by several mechanisms (Leonard and Grimwade, 2011):

- Newly synthesized DnaA has a higher affinity for ATP. If levels of ATP are high compared to ADP, there is high probability that it will associate with ATP.
- Regulatory Inactivation of DnaA (RIDA): the replisome contains processivity β-clamps (see below), that bind to proteins that hydrolize DnaA-ATP (in *B. subtilis*, a similar role is played by YabA, except it does not hydrolyze DnaA-ATP, it favors colocalization of DnaA-ATP with the replisome).
- Once replication has started, the number of DnaA binding sites rapidly increases, diluting

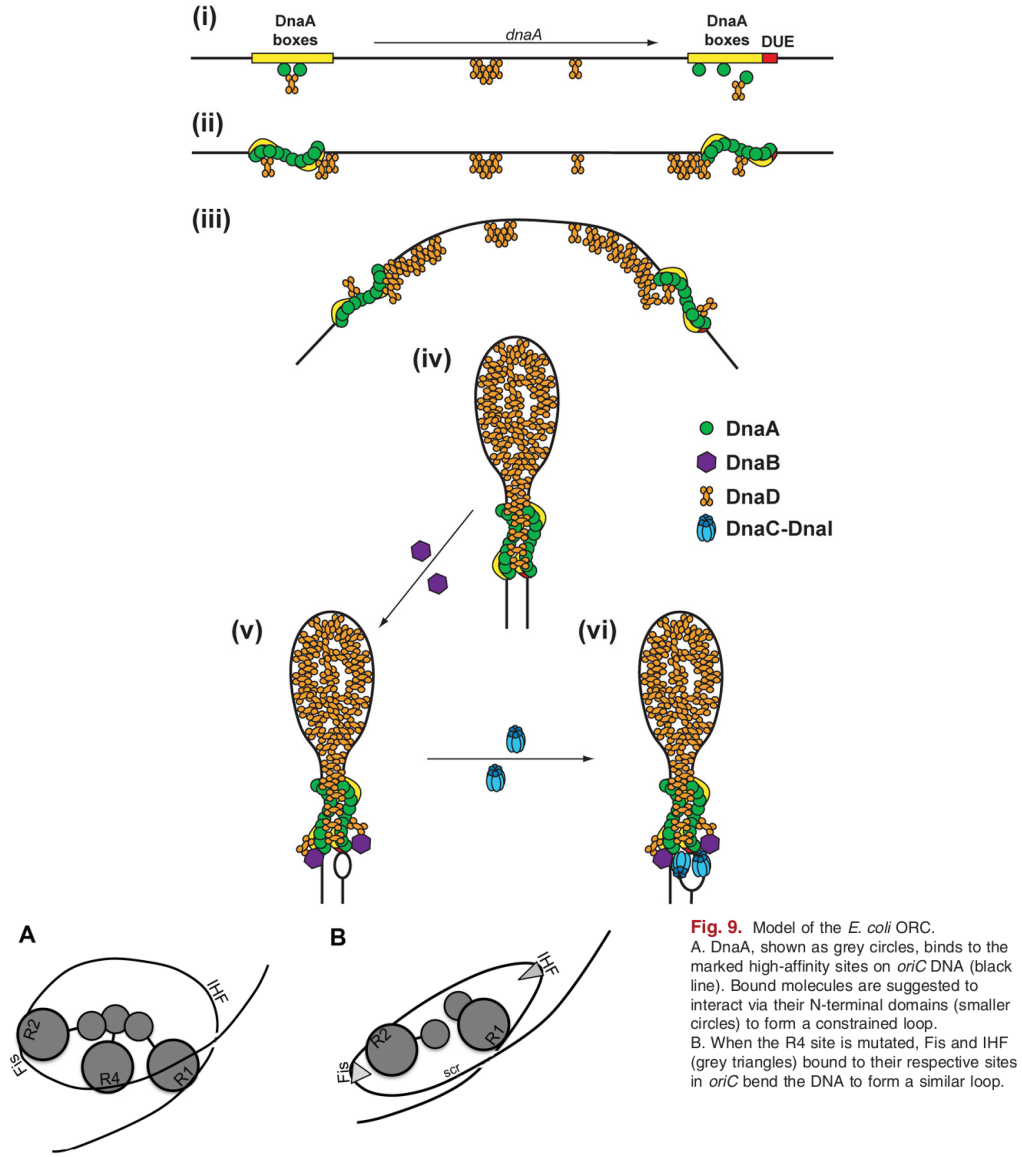

Fig. 9. Model of the *E. coli* ORC.
A. DnaA, shown as grey circles, binds to the marked high-affinity sites on *oriC* DNA (black line). Bound molecules are suggested to interact via their N-terminal domains (smaller circles) to form a constrained loop.
B. When the R4 site is mutated, Fis and IHF (grey triangles) bound to their respective sites in *oriC* bend the DNA to form a similar loop.

Figure 2: Control of DNA replication through DnaA proteins and other helper proteins in *E. coli* and *B. subtilis*. Top: Model for replisome loading by helper proteins in *B. subtilis*. DnaD accelerates DnaA binding and its unwinding and polymerization activities form a loop. Finally DnaB is recruited on each end on the loop, each DnaB loading a helicase DnaC on one of the strands with the help of helicase-loader DnaI. Bottom: Similar model for *E. coli*, where loop forming occurs cooperatively between DnaA and DNA binding proteins Fis and IHF. Figures from Briggs et al. (2012) (top) and (**?**) (bottom).

remaining DnaA-ATP, notably the *datA* locus, which contains 3 high affinity DnaA binding sites. 60 to 300 proteins can bind to this locus which is replicated at approximately 1/3rd of the cell cycle.

- DnaA Recharging Sites (DARS): loci which bind 3 DnaA-ADP and converts them to DnaA-ATP by an unknown mechanism (only 2 DARS loci have been uncovered). Membrane based processes could also be involved in DnaA recharging.

What is more, a sequestration protein SeqA may bind the low-affinity DnaA binding boxes during one third of the cell cycle after initiation, avoiding DnaA rebinding. In mutant cells lacking

SeqA, replication reinitiates immediatly (Leonard and Grimwade, 2011) A similar mechanism for *B. subtilis* could be the ParAB-*parS* system (**?**). This system, probably responsible for chromsome segregation (see below), binds to the *parS* locus through parB. It seems that ParA binds DnaA and influences replication initiation, maybe because ParB hydrolyzes ParA, complexing ParB-*parS* to the *oriC*, this association being stabilized by SMC proteins. The complex can then migrate towards one of the cell poles and it is possible that it may be unavailable for reinitiation during that time. Taking into account all these elements, Leonard and Grimwade (2011) propose a model for DnaA titration that could partly regulate replication initiation (Fig. 3).

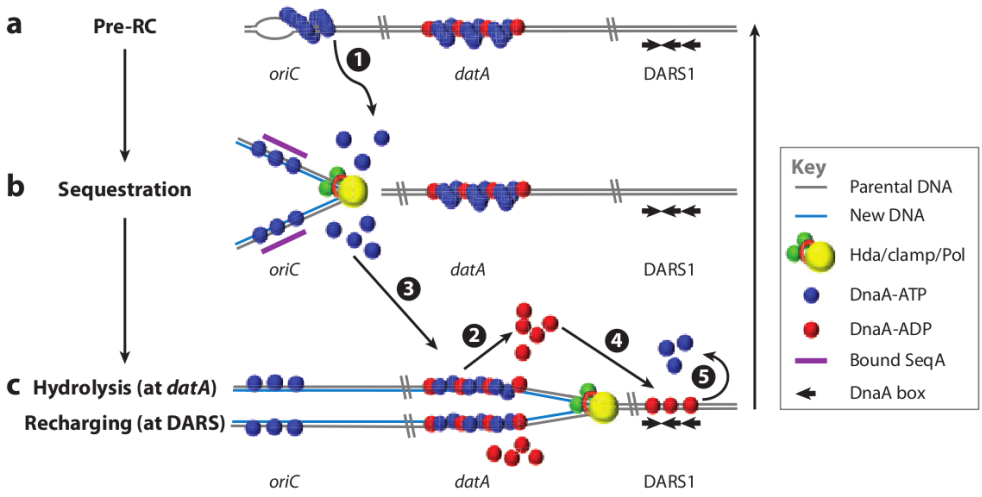

Figure 3: Titration of DnaA levels. Once Dna-ATP levels are high enough next to the *oriC*, replication initiates and displaces most of the DnaA molecules (a). SeqA binds on the low affinity binding sites next to *oriC*, so displaced DnaA-ATP cannot rebind and maybe clusters around *datA* (b). Hda protein, associated with the β-clamp of the replisome, may hydrolyze DnaA-ATP when the replisome goes through *datA*. Dna-ATP is regenerated thanks to DARS loci and slowly becomes available again for replication initiation (c). Figure from Leonard and Grimwade (2011).

Even though our understanding of the initiation of replication on the chromosome is somewhat limited, experiments show that it is strictly controlled by the cell. On the other hand, the replication of other DNA elements such as plasmids is not as severely controlled, as their exact number within the cell can vary. Initiation is not triggered by DnaA but by other proteins or RNAs that are plasmid-specific. Shortly, there are two main types of plasmids in a bacterial cell: large plasmids present in low copy numbers and small plasmids present in high copy numbers (Figure 4). The large plasmids have a similar initiation system as the chromosome, triggered by a DNA binding protein coded by the plasmid itself (generically called Rep). Initiation is probably regulated by the copy number of plasmids, either through a RNA coded by the plasmid that blocks the synthesis of Rep, or simply because Rep is able to polymerize at high concentrations, becoming unavailable for initiation. In small plasmids, initiation is probably controlled by a RNA and DNA polymerase I. The RNA (termed RNAII) slightly opens the double helix and binds to DNA, acting as a primer for DNA polymerase I, which further opens DNA. A single replisome can then be loaded on the other strand (in this case, replication is unidirectional). Copy number is probably controlled by another RNA (termed RNAI), which interfers with RNAII at sufficiently high concentrations.

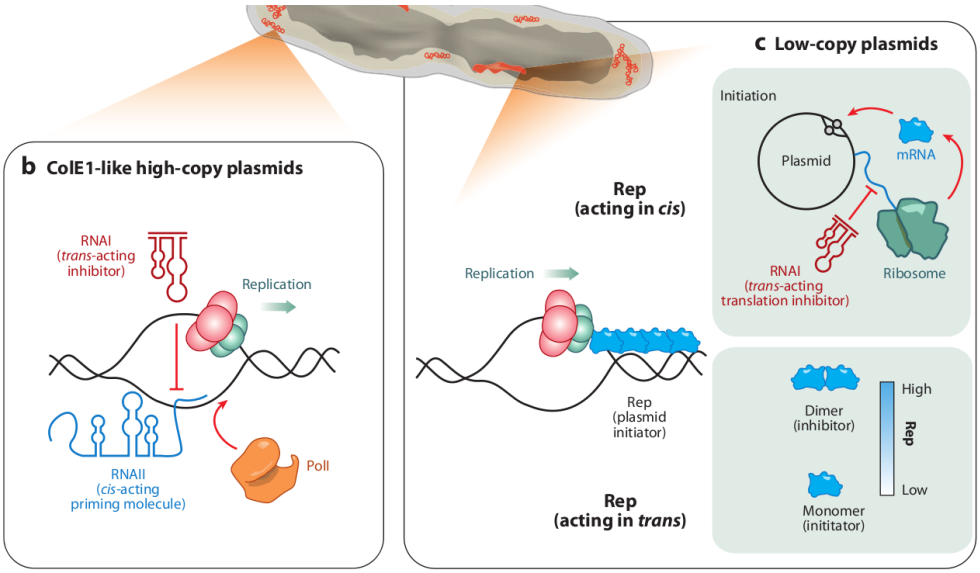

Figure 4: Plasmid replication follows similar principles as chromosome replication (it uses the same replisome). However, initiation is regulated by plasmid specific elements. Small plasmids probably use RNAs to initiate (RNAII) and inhibit (RNAI) replication. Large plasmids use a protein that acts similarly to DnaA and which is regulated by the plasmid copy number. Figure from **?**

.

Replication elongation Once replication is initiated, helicases are loaded upon the DNA strands, one in each direction (see above). From the helicases, the whole replisome complex can be loaded and start polymerizing DNA. The replisome ensures processivity and rapid replication of the whole chromosome, as well as synchronous replication of the leading strand and the lagging strand. Indeed, DNA polymerases are only able to assemble DNA in the 5' to 3' sense, which corresponds to the direction the leading strand is processed, but antisense to lagging strand processing. Therefore, the leading strand is easily handled while replication of the lagging strand is handled by loop forming that allows for fragment-wise replication of a few kbs of DNA (termed Okazaki fragment).

The composition of the replication reflects these different tasks (Figure 5). The helicase DnaB (DnaC for *B. subtilis*) is formed of a hexamer that separates the DNA, forming the replication fork. Bound to the helicases, three primases DnaG stabilize the helicase structure and cooperate for synthesis of short RNA sequences on the lagging strand called primers used to initiate Okazaki fragments. Finally, a clamp loader bound to the helicase is responsible for recruiting DNA polymerases and β-clamps. The number of DNA polymerases can vary depending on the number of subunits τ in the clamp loader (**??**): replisomes with 2 or 3 associated DNA polymerases have been observed, the latter seemingly more efficient. The nature of DNA polymerases is also unclear: it seems that in *E. coli* Pol III is preferentially recruited because of its high fidelity, while in *B. subtilis* two different polymerases may be used for the leading and the lagging strand (**??**). The recruitment of β-clamps is also essential as it binds to DNA polymerases and increases their processivity from a few nucleotides to several tenths of kb (**?**). Because these elements are bound to the clamp loader, they are efficiently recycled, allowing for very fast replication.

Coordination of leading and lagging strand replication is not fully understood. As replication occurs continuously on the leading strand, it seems intuitive that replication should occur more

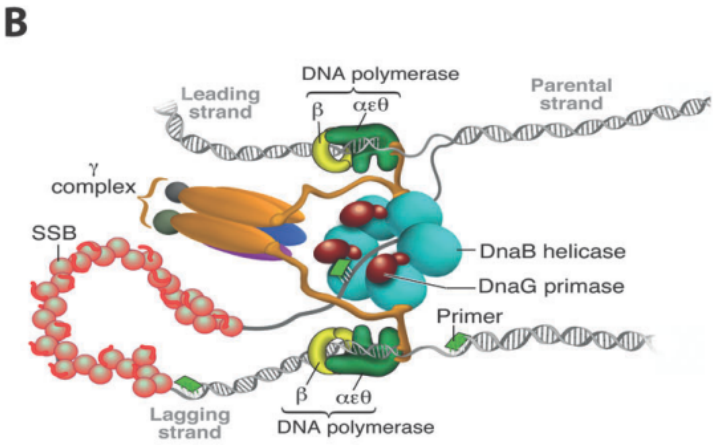

Figure 5: Standard bacterial replisome stucture (as reconstructed from *E. coli*). Figure from **?**.

rapidly than on the lagging strand or that the loop forming/relase of Okazaki fragments has to be particularly efficient. Two models have been proposed for the loop forming process: the collision process and the signaling process (Figure 6, left). In both models, DNA polymerase starts from a primer, progressively forming a loop as ssDNA (protected by SSB) and recently polymerized DNA accumulate between the fork and the DNA polymerase. In the collision model, the DNA polymerase proceeds until the next Okazaki fragment, stalling and becoming available for elongation from the next primer. If a primer is ready before the current fragment is finished, the fork could pause, making this process quite inefficient. In the signalling model, the DNA polymerase is reused as soon as a primer is ready, so that a lot of fragments remain incomplete, containing large portions of ssDNA. In fact, the presence of 3 polymerases indicates that the reality might be in-between, explaining how the lagging strand synthesis can be as efficient as the leading strand synthesis (Figure 6) (**??**). With 3 polymerases, 2 polymerases might be synthesizing at the same time, one finishing the previous fragment while the other one is available for recruitment on a new primer. As a matter of fact, it is also highly probable that the replisome is very dynamic, with frequent recruitment and detachment of primases and polymerases. In this way, even a 2 polymerase replisome could operate in a similar manner, by periodically detaching polymerases to finish a fragment and recruiting a new polymerase at the replisome. This remains to investigate but seems pretty likely as numerous polymerases gravitate around the replisome and such exchanges have been observed on the leading strand (**?**).

Some details of elongation remain unclear. RNA primers have to be removed, resynthesized and ligated by specific polymerases, but the coordination with the replisome has not been investigated (to our knowledge). The cooperation with SMC molecules or obstacle management is known to exist but very little is known. Some elements are presented in the repair section, where stalling of the replisome is handled by RecA and restarting of replication is handled by a specific primosome complex. However, it is unknown if the replisome collapses, helps recruiting repair proteins or lower fidelity polymerases that can bypass some specific DNA damages.

Replication termination Termination of replication occurs in the *terC* region thanks to Tus proteins. These proteins are able to block a replisome along one direction. Replisomes will thus meet along a small segment delimited by two Tus proteins in opposite directions, probably

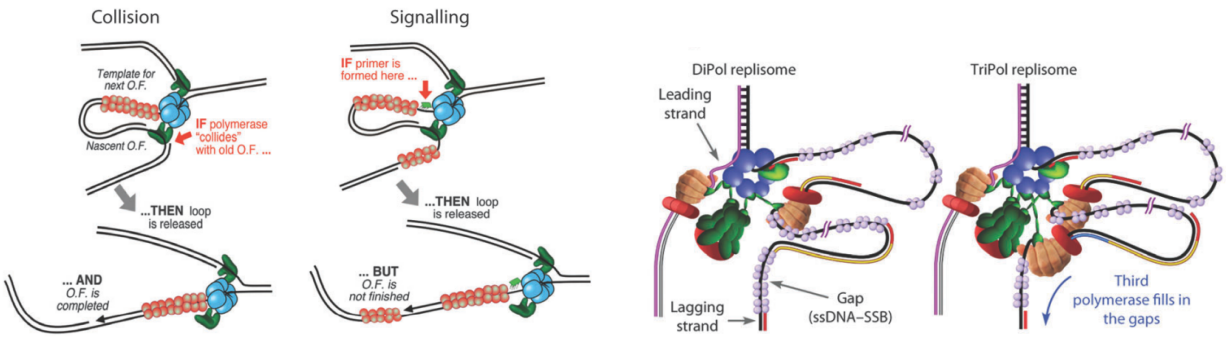

Figure 6: Elongation models for the lagging strand (left). Likely elongation according to recent experiments that mixes the two previous models (right). Figure from **?**.

even at the site of one of these proteins. The length of the two replicons is thus not totally fixed but limited to a certain range.

Computational representation See illustration how the cell chromosome is used. Some more information may be needed TBD. When a chromosome is fully duplicated (when the first column of the cell chromosome only contains -2), the process consists of deleting the current chromosome and creating 2 new ones with the correct initialization. TBD do we clean the chromosomes? Typically, when the chromosome was manipulated.

Needs also other things but from the DNA point of view, there seems to have enough information.

4.2.2 DNA movement

What do we do when a gene change of volume due to condensation or segregation for example. Normally, a matter of changing the number of the volume in the cell chromosome and the corresponding volume chromosomes. But assume that anything that is currently binded to the DNA is the property of the DNA and was 'erased' from everywhere else. Also assume that volume chromosome only contains the strictly necessary information about the DNA in the volume and that it is a state with changing size.

4.2.3 DNA manipulations

Codon aggregation damage Gap site, Abasic site, Sugar-phosphate, Base, Intrastrand cross link, Strand break, Holliday junction DNA is subject to numerous forms of damage that can be either endogenous or exogenous. They can result in chemical modifications of some bases, in single strand breakage (missing of one base on one of the strands), double strand breakage or cross links between DNA strands. Chemical modifications can originate from different type of radiations (UV, x-rays, etc.), drugs or reactants naturally present in the cell leading to alkylations, oxydations, deaminations, etc. Another important source of DNA mismatches is the replication machinery itself which can make use of the wrong dNTP (dUTP for example).

DNA modification is one of the prerequisites to evolution. DNA mutations lead to the development of novel functions, regulatory systems, etc. However, replication fidelity is also essential for selection and conservation of important existent functions. There is a trade-off between these two aspects that is well illustrated in *B. subtilis* by the existence of efficient repair mechanism on the one hand and some DNA polymerases that favor propagation of some types of damage (such as DnaE) on the other hand.

Codon aggregation insertion Insertion of one (or several) lines in the DNA states.

Codon aggregation deletion Putting 0 in the corresponding places and not deletion of the line(s) because a codon aggregation can be deleted in a fork.

Codon aggregation repair There are several pathways employed for DNA repair corresponding to the type of damage undergone.

Mismatch Repair (MMR) This pathway is dedicated to reparation of base mismatches. In *E. coli*, repairing is initiated by MutS (Sensor) which detects the mismatch and MutL (Linker) which recruits further proteins. The endonuclease MutH then nicks the DNA next to the mismatch enabling the helicase and exonuclase UvrD to remove bases around the mismatch. The resulting gap is filled in by DNA polymerase III and DNA Ligase. The newly synthesized strand is specifically targeted by these proteins thanks to the methylase Dam used for marking the original strand. In *B. subtilis*, only MutS and MutL seem to be conserved (MutL having an additional endonuclease activity). Recognition of the newly synthesized strand could be linked to a strong coupling of MMR with replication and colocalization with the replisome.

Base excision repair (BER) The BER pathway repairs most non-bulky base modifications such as oxydations, deaminations, UTP incorporation, etc. Schematically, glycosylases detect the lesion, remove the damaged base, endonucleases then nick the DNA next to the missing base so that exonucleases remove some bases on the strand around the missing base. The small gap is then closed by a repair DNA polymerase (such as Polymerase I) and ligated by a DNA ligase. For example, in *B. subtilis*, the glycosylases MutM and MutY (part of the GO system) detect oxidized Guanine to avoid its pairing up with dATP. Another example is Uracil DNA-glycosylase, which removes dUMP from DNA.

Nucleotide excision repair (NER) The NER pathway is very similar to the BER pathway, except that it repairs bulky lesions caused by UV radiations or drugs. This pathway is highly conserved and partly regulated by the SOS response (mediated by RecA). The UvrABC complex is responsible for detecting the damaged base and nicking the DNA at surrounding bases. Helicase UvrD removes the nicked segment. Finally, DNA Pol. I and DNA ligase restore the missing segment.

Alkylation damage There are specific pathways that address alkylation (such as methylations). *B. subtilis*, as a soil-living bacterium, is particularly exposed to alkylating agents. There are at least three pathways responsible for repairing alkylated bases: two pathways based on glycosylases (one being constitutive, the other inducible) and one pathway based on alkyltransferases (enzymes that suicide by transferring the alkyl group onto themselves).

Homologous recombination (HR) During replication, double strand breaks (DSB) can be repaired by using the other copy of the chromosome (Figure 7). First, the DSB is stabilized by RecN and digested by protein complexes (AddAB and RecQSJ in *B. subtilis*) that create hangovers of single stranded DNA (ssDNA) on the 3' strands. RecA binds to the ssDNA (probably with the help of the RecFOR complex that prevents SSB from binding). RecA, activated by ATP or dATP, enables invasion of the sister chromosome at a homologous sequence, creating a D-loop where one of the broken 3' strands is inserted and forming Holliday

Junctions between the two chromosomes. DNA elongation occurs from the 3' strand and the Holliday Junctions are cleaved by RecU (or a homologous protein).

There is a variation if the DSB caused the replication to stall (Figure 7, right). This situation may occur if a single-strand break is present on the original chromosome which becomes a DSB after passage of the replication fork. In this case, there is only one DNA end to digest and the invasion leads to only one Holliday Junction. The primosome complex, composed of Pri and Dna proteins, detects the D-loop and helps loading the replisome to resume replication. It seems that several Structural Maintenance of Chromosome (SMC) proteins are involved throughout the process (such as RecN), but their role is not clearly elucidated.

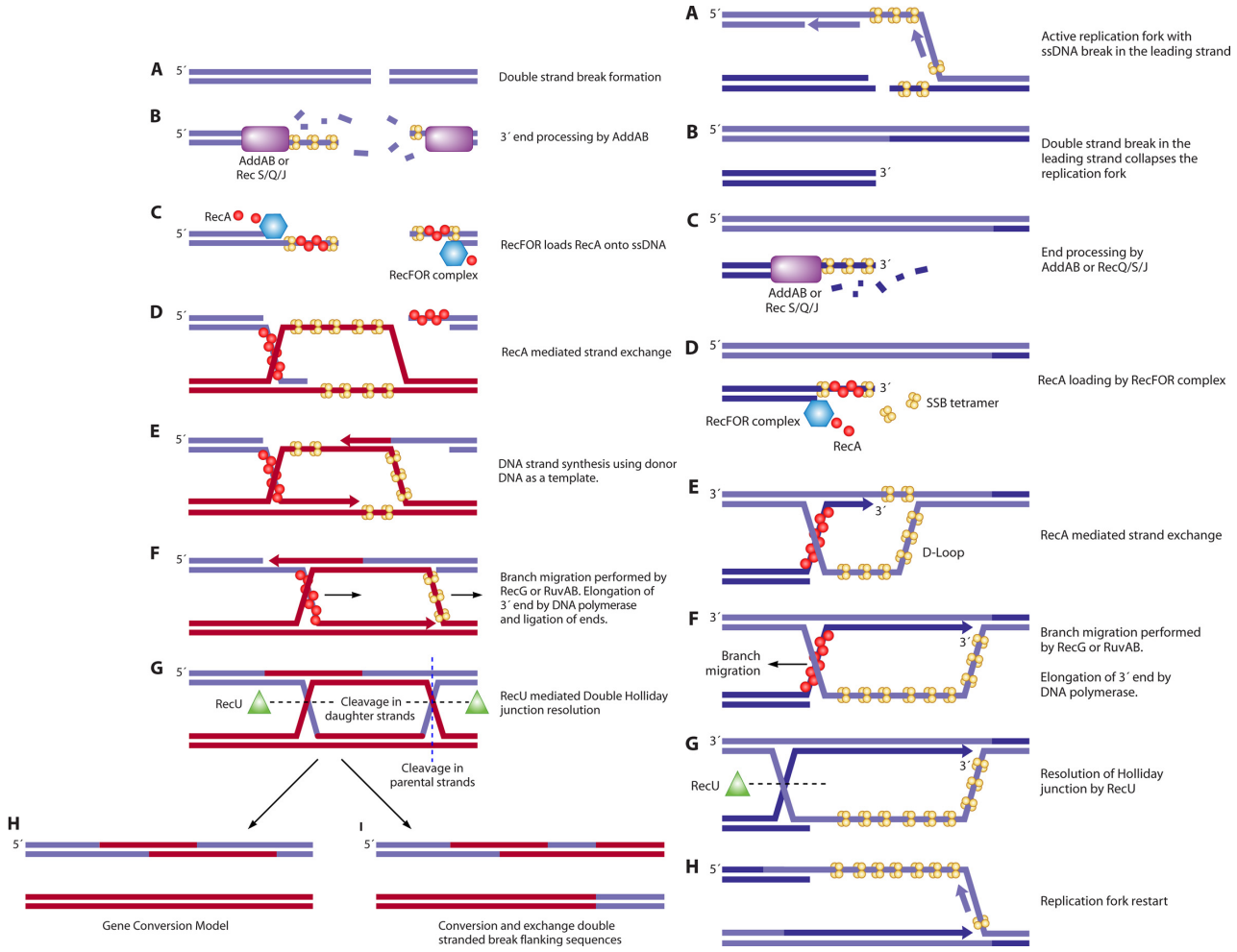

Figure 7: Homologous recombination and repair of DSB in *B. subtilis* in the general case (left) and when the DSB appears in the replication fork (right). From Lenhart *et al.*.

Nonhomologous end joining (NHEJ) This pathway is also responsible for DSB repairing but it is less efficient than homologous recombination. It is used when there is no other copy of the chromosome present in the cell. As in HR, a protein (probably YkoV for *B. subtilis*) binds the DSB and favors recruitment of another protein (LigD like, probably YkoU for *B. subtilis*) that is able to perform exonucleation, polymerisation and ligation. The mechanisms are not totally clear but it seems that because LigD is able to perform these 3 functions, no other protein is needed. However, some bases need to be deleted and repolymerized during the reparation, possibly leading to DNA losses or insertions, making NHEJ a low-fidelity repair

mechanism.

4.2.4 DNA compaction

DNA compaction occurs through various proteins that are able to clamp the DNA together, forming high-densitiy bundles, leading to a compact form within the cell called a nucleoid. The compaction is due to supercoiling generated by DNA gyrase and topoisomerases, the action of histone-like proteins (HU, IHF, Fis H-NS) and SMC proteins (SMC-ScpA-ScpB for *B. subtilis*, MukBEF for *E. coli*)(Fig. 8). According to several studies, the nucleoid adopts a large helical structure composed of two intertwined branches, at least during G1 phase (Fig. 8) (**???**). This structure is dynamical, with specific loci moving of about 10% of cell length within a few seconds (**??**) and seems to be ATP-dependent (**??**). **?** identify these variations as waves that could be used to unbind some of the DNA binding proteins, avoiding overcondensation, unwanted linkage between loci and facilitating segregation. In general, it seems that the chromosome is linearly organized around the *oriC*, meaning that the bundling occurs progressively, so that the position within the cell reflects the position along the chromosome (**?**). Condensation, clamping of the DNA by structural maintenance of chromosome (SMC) proteins, supercoiling, macromolecular crowding, charge neuralization? If spatialization is used, condensation and segragation might be modelled directly. Supercoiling needs another state. Compactation should also impact the accessibility of the chromosome.

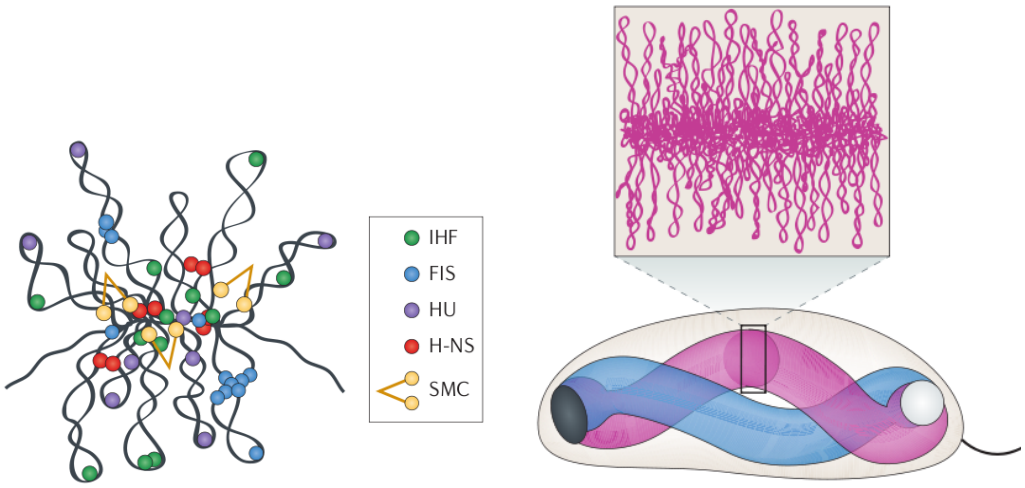

Figure 8: Proteins responsible for chromosome compaction (left) and model for nucleoid organization in *C. crescentus* (right). Figures from **?**.

4.2.5 DNA segregation

As replication proceeds, sister chromosomes/plasmids have to be separated in order to allow proper cell division. This process is also largely unknown, but several models based on experiments have been proposed, it seems that there is no universal solution valid for all bacteria. There are two main challenges: unlinking the chromosomes and driving them to opposite poles of the cell.

Unlinking is generally done at the replisome level. Supercoiling accumulates at the front of the helicase due to its unwinding activity. This supercoiling can be physically propagated to the back of the replisome, creating an entanglement between sister chromosomes. The DNA

gyrase limits this propagation by diminishing supercoiling at the front of the replisome, while Topoisomerase IV disentangles the chromosome copies (**?**).

Segregation can be done according to several mechanisms, particularly for plasmids. For high copy number plasmids, it is possible that segregation occurs purely through diffusion. For other plasmids, elements of the cytoskeleton can be used to separate the copies (Figure 9ab). ParR proteins may bind to *parC* loci on the plasmid and serve as a basis for actin-like ParM that polymerizes between the copies, progressively seperating them. Similarly, TubR might bind to *tubC* and migrate along filaments of tubulin-like TubZ. The last mechanism may target plasmids as well as the chromosomes (Figure 9c). It is also composed of a DNA binding protein ParB, binding to *parS* (close to *oriC*), and a motor protein ParA. However, ParA attracts ParB only in its activated and DNA-binding form ParA-ATP, probably located along the nucleoid. ParB hydrolises ParA-ATP, releasing it in the cytosol until it gets reactivated and rebinds DNA away from ParB. In this way the two ParB-*parS*-*oriC* complexes migrate in opposite directions until steady-state is reached, with equivalent ParA-ATP pools located on each side of each complex (approximately at the quarter of each pole). It seems that this system works cooperatively with SMC proteins, but the details are yet unknown (**?**). Another possibility is radial stress (**?**). Sister chromatides are bundled separately but hold together by some tether. They accumulate at the center of the cell, along with the mother DNA, which is bundled separately. The mother DNA pushes the two growing chromatides aside with a force that gets stronger for sterical reasons until the tethers break and the chromatides migrate towards the poles. According to **?**, this phenomenon happens up to four times during segregation. The main idea here is that segregation results from efficient bundling of neosynthesized DNA, so that mother DNA and sister chromatides form 3 very distinct structures that repel each other but are linked by tethers.

Similar to initial segregation, final segregation includes decatanation of two chromosomes and migration of the *ter* region to the two poles, but it is assisted by a new protein and associated with the formation of the FtsZ ring at the septum. The FtsZ ring cannot polymerize as long as the nucleoid is located at the center of the cell. Cytokinesis thus begins when the initial migration of DNA copies is already advanced. Once FtsZ ring forms, the DNA translocase FtsK (SftA in *B. subtilis*) is recruited by the divisome to the membrane next to the FtsZ ring. FtsK seems to coordinate several actions in the final segregation. FtsK binds to *dif* loci next to the *ter* regions, aligning the *ter* region with the septum. FtsK can also translocate remaining DNA at the final stages of septum closing. FtsK also cooperates with Xer proteins to separate chromosomes copies that might have merged due to recombination by creating a Holliday junction.

4.2.6 DNA transformation

DNA transformation is the incorporation of extracellular DNA into the cell. The role of transformation is not totally understood. Most probably, it allows for generation of bacterial sex and generation of an enhanced genetic diversity in changing environments or stress situation. The master regulator of competence ComK induces more than 150 genes, 30 of which only are known to be directly involved in DNA transformation.

Transformation uses the recombination machinery to integrate the DNA into its own genetic material. Numerous proteins of the SOS response are therefore involved but it seems that the induction of these proteins is mediated by ComK, the master regulator of competence state, and is LexA-independent (**?**). DNA transformation is only activated in specific physiological conditions by quorum sensing. ComX and the Competence and Sporulation Factor are secreted,

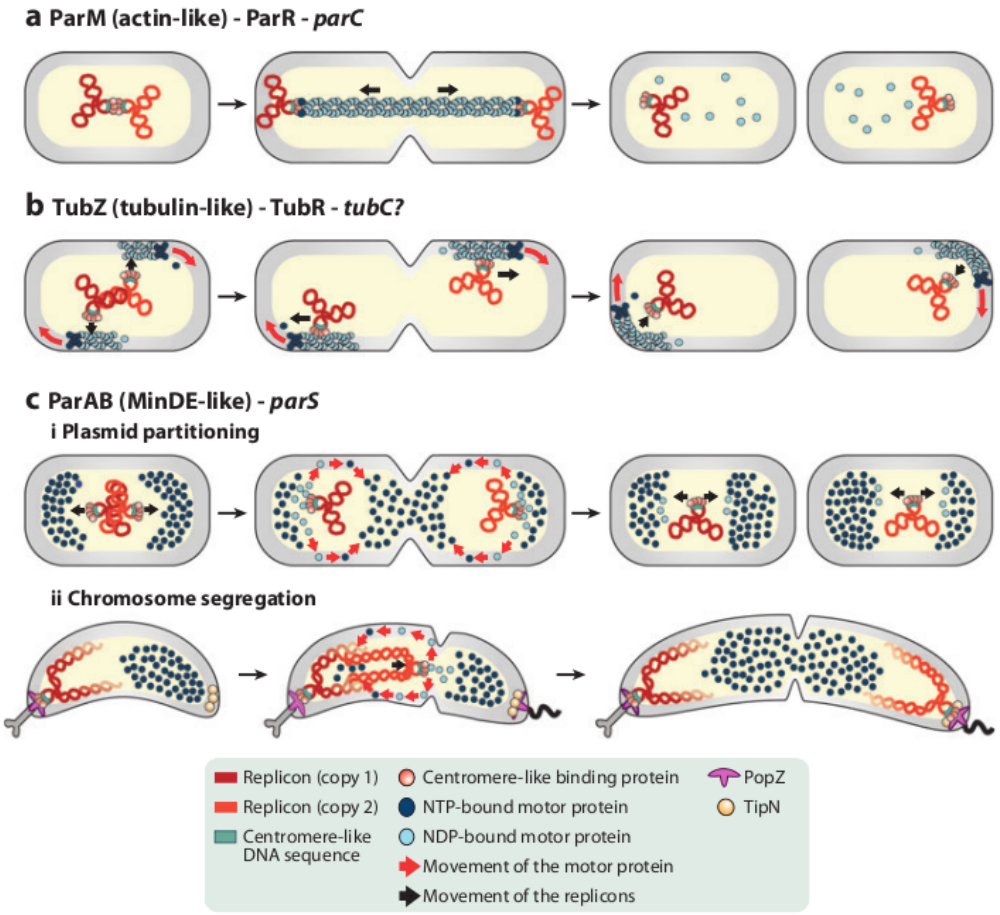

Figure 9: Migration of plasmids and chromosome can be mediated by different systems. It can be based on elements of the cytoskeleton: separation through polymerization of actin-like proteins (a), migration along tubulin-like proteins (b). Alternatively, migration may based on an oscillatory system, where activated proteins (ParA-ATP) attracts another protein (ParB) linked to the plasmid or chromosome (*parS* loci) that hydrolises it (c). DNA migration is then controlled by the location of pools of activated proteins. Figure from **?**.

activating ComP-ComA transduction. More than 150 genes are then directly or indirectly induced, including the DNA translocation machinery and the recombination machinery (**?**).

The DNA translocation machinery assembles towards the cell pole and is composed of three groups (**?**):

- The first group is responsible for binding dsDNA and transforming it into ssDNA. Known proteins are from the ComG family (ComGA, ComGC, ComGD, ComGE, ComGG) that form a pseudopilus able to drag DNA toward the membrane and NucA, which fragments dsDNA, allowing it to be digested to ssDNA by an unknown enzyme.

- The second group translocates ssDNA into the cell. It is composed of proteins from the ComE and ComF families (ComFA, ComEA, ComEC, maybe ComFB, ComFC, ComEB). The junction with the first group is done by ComFA and ComGA.

- The last group binds DNA inside the cell and performs its integration. It is composed of the induced proteins DprA, SsbA, SsbB, CoiA, RecA and the constiutive proteins RecU and RecX. Its association with the rest of the translocation machinery is highly dynamic.

Once the ssDNA is translocated within the cell, the third group protects it from degradation and recruits protein that are also involved in classical homologous recombination:

- SsbA and SsbB bind and protect ssDNA from degradation.
- RecN and SsbE stabilize the ends of the DNA.
- DprA and RecO destabilize the SSB proteins and allow RecA loading.
- RecA loading is regulated by RecF, RecX and RecU.

Once RecA is loaded along the ssDNA, it forms a compact filament that is driven towards the center of the cell. It may be integrated within the genome if it is homologous to an endogeneous sequence (with a probability of up to 40% (**?**)) or may be transformed into a plasmid if it has a primosome assembly site (*pas*) (**?**). If a homologous sequence is found, an invasion similar to Homologous Recombination occurs (see DNA repair process). However, branch migration is not fully understood and HJ resolution may not be mediated by RecU, contrary to classical homologous recombination.

Formation of new plasmid is a little more complex as it may occur according to different mechanisms that are species specific. **?** list three possibilites that could contribute to plasmid creation for ssDNA containing a *pas* sequence:

- Plasmid facilitation: the ssDNA transiently binds to the chromosome, forming a loop that is replicated using the free ends of the ssDNA as primers, closing the loop by adding chromosomal DNA to the ssDNA.
- Plasmid transformation (necessitates internal homologies): in the absence of homology, RecA unbinds and the ssDNA is converted to dsDNA thanks to his *pas*. It is then circularized based on its internal homologies.
- Monomeric activation: poorly understood process that could work even without internal homologies. Does not seem to exist in *B. subtilis*.

4.3 Transcription

4.3.1 Promoter search

Before binding to DNA, the RNA polymerase has to find it: it is called promoter search. It was generally admitted that a 3D-diffusion search was not enough as some search rate measure were above the 3D-diffusion possible rate. Some other mechanisms were proposed: 1D sliding along the DNA, 1D hoping and DNA intersegment jump. In (Halford, 2009), it is suggested however that the promoter search rate is consistent with a 3D diffusion rate.

4.3.2 Initiation

The purpose of initiation is to locate precisely the Pribnow's box (TATAAT for σ^{70} sigma factor) so that the RNA polymerase can bind to it. This binding is favoured by a sigma factor. A non-bounded RNA polymerase is composed of 4 sub-units (2α, β, β' and ω): it is a core-enzyme. The initiation ends with the promoter clearance and eventually the release of the sigma factor. It follows the steps (Saecker et al., 2011):

1. a sigma factor binds to the core-enzyme becoming a holo-enzyme:

$$pRNA + \sigma^i \rightleftharpoons pRNA\Sigma^i$$

There are several types of sigma factor and each of them increases the affinity of the holo-enzyme RNA polymerase to the specific promoter.

2. The holo-enzyme then binds to a DNA promoter and forms a closed complex. This triggers a series of conformational changes collectively called 'izomerization':

   - opens 13 bp from the -10 elements beyond the transcription start site for the sigma factor and -35 for the α sub-unit while the complex protects a 'footprint' of around 30 bp from nuclease digestion (von Hippel, 1998);
   - creates the initiation 'bubble' and a stable open complex after an unstable one.

   It is summarized with the reactions:

$$\begin{cases} pRNA\sigma^i + DNA \longrightarrow cDpRNA\sigma^i \\ cDpRNA\sigma^i \longrightarrow oDpRNA\sigma^i \end{cases}$$

3. The open complex then synthesizes nascent RNA and tries to leave the promoter site (promoter clearance). However, around 10 abortive initiations happens (Goldman et al., 2009) in mean before promoter clearance is really performed:

$$\begin{cases} oDpRNA\sigma^i + \sum m_i rNTP^i \longrightarrow oDpRNA\sigma^i + nRNA \\ oDpRNA\sigma^i + \sum m_i rNTP^i \longrightarrow DpRNA\sigma \end{cases}.$$

   The first reaction models the abortive initiation as the synthesis of the nascent RNA being apart from the DNA - RNA polymerase complex. Physicality, the nascent RNA is still inside the complex and goes away from the complex when the sigma factor is released. The second reaction models the promoter clearance with the nascent RNA still attached to the complex.

4. The sigma factor is released. Physically, the sigma factor can be either released during the promoter clearance or during the beginning of elongation.

$$DpRNA\sigma \longrightarrow DpRNA$$

   The trigger is not clear however it is released typically when the nascent RNA reaches a length of 12-15 nt.

4.3.3 Elongation

The elongation consists in binding the ribose NTP to each other and step forward:

1. recruitment of the ribose NTP corresponding:

$$DpRNA + rNTP^i \rightleftharpoons DpRNA^i.$$

   A 'wrong' ribose NTP could be recruited at this step, or even a deoxyribose NTP [reference needed];

2. binding of the recruited ribose NTP to the former one, catalyzed by a pair of $Mg_2{}^+$:

$$DpRNA^i + \mathrm{H_2O} \xrightarrow{\mathrm{Mg^{2+}}} DpRNA^+ + PP_i;$$

3. translocation where the RNA polymerase step one base forward:

$$DpRNA^+ \longrightarrow DpRNA.$$

The movement of the polymerase forms a Brownian ratchet motion. The elongation rate is around 6.2-20 bp/s. In competition to the step forward motion, there are:

- pausing consists in the RNA polymerase to pause during some time but the mechanics are unclear. Promoter proximal pause also happens. (Larson et al., 2014) proposes a consensus sequence of 11 nt length for pausing detection. An early release of the nascent RNA transcript may happen during pausing;
- backtracking consists in the cleavage of 2 or more nucleotides. It seems to be a kind of proofreading. It is not clear when and how it happens;
- stalling consists in the RNA polymerase to wait, especially for rare ribose NTP to be recruited. Stalling is thought to help the folding of the nascent polypeptide chain [reference needed].

4.3.4 Termination

For prokaryotes, there are two types of termination:

1. Rho-independent or independent termination (Gusarov and Nudler, 1999; Wang and Greene, 2011): nascent RNA forms a rich G-C hairpin followed by 7-9 U bases. This weakens the DNA - RNA polymerase complex. The force due to the hairpin is not enough though and some other mechanisms, which is performed by the binding of NusA, is needed (Herbert et al., 2008);
2. Rho-dependent termination: a rho co-factor helps the termination. Two major models exist: (*i*) the rho factor binds to the rut (rho utilization site of about 70 - 100 nucleotides) or rho binding site and then moves forward towards transcription stop point (tsp) region creating an hairpin; (*ii*) the rho factor binds to the polymerase (Epshtein et al., 2010) and an hairpin is created while the polymerase elongates. In the tsp region, there are (potentially) several pause positions for the RNA polymerase in the absence of rho co-factor, the termination occurs at these stop positions.

4.3.5 Others

Transcript slippage The phenomenon is illustrated in Figure 10 in the case where the RNA polymerase idles; forward and backward slippage are also described in Anikin et al. (2010). It mainly occurs on homopolymeric tracts. Slippage can occur during initiation, elongation and termination, as well as during replication and translation. Slippage during transcription initiation plays a role in transcription regulation by abortive transcripts Turnbough (2011). Bacteria can take advantage of slippage in few ways as described in Anikin et al. (2010):

- use a single mRNA to encode two proteins;
- restore a reading frame that would be otherwise be interrupted by the addition or deletion of a nucleotide.

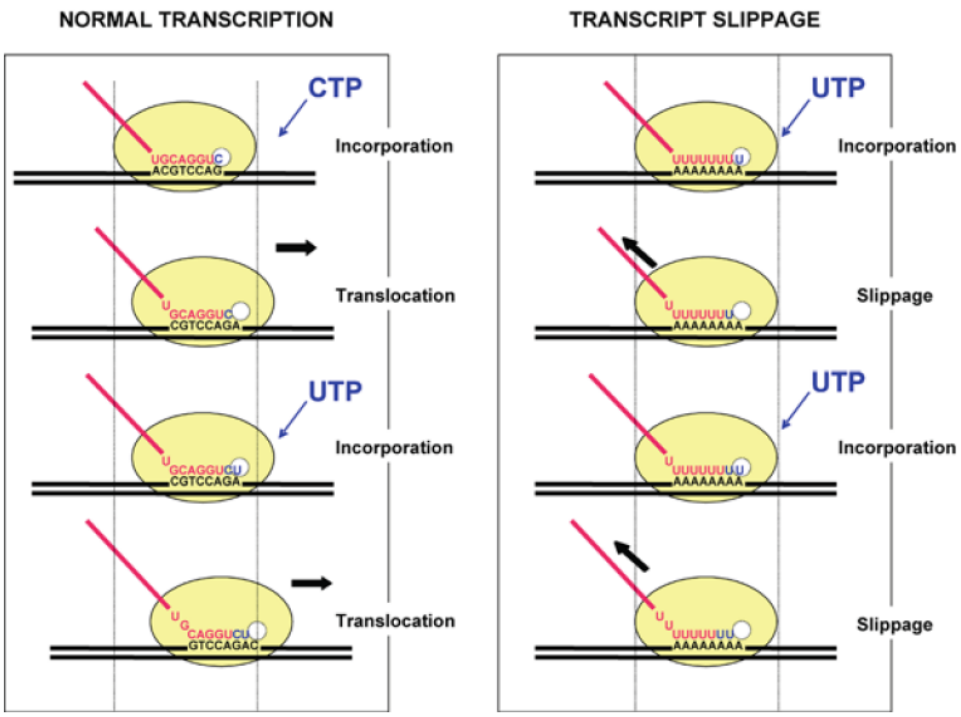

Figure 10: Illustration of transcript slippage, from Anikin et al. (2010)

4.4 RNA processes

4.4.1 RNA processing

Some polycistronic RNAs have to be cleaved before they become functional. This may be the case for some mRNAs, but is typical for sRNAs, tRNAs and rRNAs. The 30S RNA transcript of *E. coli* undergoes several rounds of cleaving performed by different ribonucleases yielding a 9S, pre 16/17S and pre-23S intermediate before obtaining the 5S, 16S and 23S rRNA (Arraiano et al., 2010). tRNAs also have to be cleaved at 3' 5' ends by a similar process involving several ribonucleases (Arraiano et al., 2010). As for sRNA, their processing is globally unknown, except for a few examples. What level of detail do we want ?

4.4.2 RNA modifications

According to Karr, some bases of the RNAs are modified for proper folding and better coding recognition, but we have little data about it.

4.4.3 RNA decay

There are two sides to RNA decay: the degradation machinery and possible determinants for targetting specific mRNA degradation. Let us start with the degradation machinery. There are numerous proteins that are able to cleave RNAs, some are more involved in RNA processing, some in degradation, but there is a large overlap in these activities (Arraiano et al., 2010). We are only going to focus on the main ribonucleases generally considered as being responsible for global RNA degradation.

In *E. coli*, the main enzyme is RNase E, an endonuclease that initiates degradation of RNAs. It seems to have two recognition pathways. Either the RNA has been modified on its 5'-end by

the pyrophosphohydrolase RppH from 5' PPP to 5'P, enabling docking of the RNA by RNase E, or it may be recognized by associating with the C-terminal domain of RNase E, possibly using intermediates (**?**). The latter case is not totally understood, numerous cofactors such as sRNAs, riboswitches or, more generally, ribosome depletion of the RNA could help recognition by RNase E. RNase E binds to the inner membrane, localizing RNA degradation at the cell periphery and is part of a structure called degradosome (**???**). Its C-terminal domain is able to recruit another ribonuclease called PNPase that can slice RNAs in smaller parts through a 3'-5' exonuclease activity and other helper proteins such as the RNA helicase B (RhlB) and the glycolytic enzyme enolase. Once a RNA has been recognized by the degradosome and cleaved by RNase E, its degradation occurs very quickly. RhlB and PNPase or RNase II (with the help of the polyadenylation polymerase) can efficiently degrade secondary structures and an oligoribonuclease degrades remaining parts into single nucleotids (Fig. 11). The process seems so efficient that degradation intermediates are not observed experimentally (**?**).

In *B. subtilis*, RNase E is not present. It has therefore been hypothesized that a similar ribonuclease is responsible for the initial cleaving of RNAs. RNase Y seems to fulfill this task (**???**). Due its recent discovery, some questions remain open. RNase Y seems to be able to bind to the membrane or to be part of a degradosome, but these asumptions have to be completely validated (**?**). At the same time, another important ribonuclease was discovered, namely RNase J1, which is a 5'-3' exonuclease, which seems to be able to degrade RNAs alone, particularly 5' monophosphate RNAs (Fig. 11) (**???**). RNase J1 probably also participates in degradation of fragments along with the PNPase as part of the degradosome. In other bacteria, any combination of RNase E, Y and J can be found but it seems that at least one of them is present (**?**). The degradosome can also vary quite largely in composition (**?**). For example, different specific ribonucleases or protein chaperones GroEL and DnaK can be recruited. It does also not necessarily bind to the membrane, e.g. in *C. crescentus*, where DNA degradation seems to be located around the nucleoid, not at cell periphery.

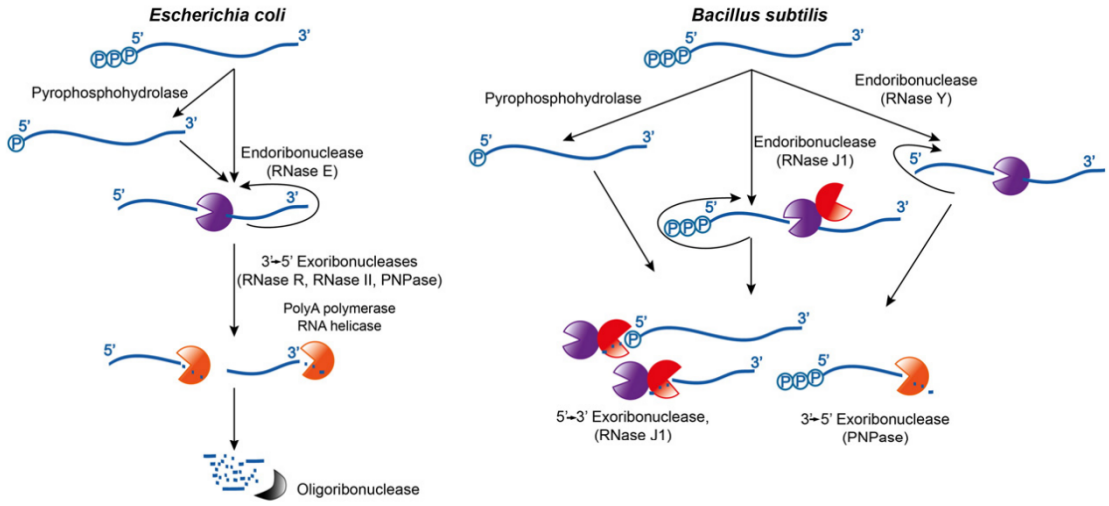

Figure 11: RNA degradation. Figure from **?**.

RNA half-lives depend on the stability of the RNA, thus on its recognition by the degradosome or other ribonucleases. However, this stability is difficult to predict from the function of a gene, its sequence or even its secondary structure (**?**). It is believed that ribosomes actively translating an mRNA protects it from being degraded by binding it, so that the RNA becomes unavailable for cleavage (**?**). Recent results shows that, while this is the case, it is

not necessary to have ribosomes along the whole RNA, a ribosome can protect large parts of the mRNA "at distance" by making the docking of a specific part of the RNA impossible (**?**). Whole genome studies yield distributions of RNA half-lives, showing that in *B. subtilis*, most RNAs have a very short half-life (80% shorter than 7 minutes) and a few have a relatively long half-life (3% longer than 15 minutes) (**?**). These half-lives probably vary depending on the physiological state of the cell. For example, targetting mRNAs with sRNAs is an efficient way to induce/inhibit translation (Fig. 12).

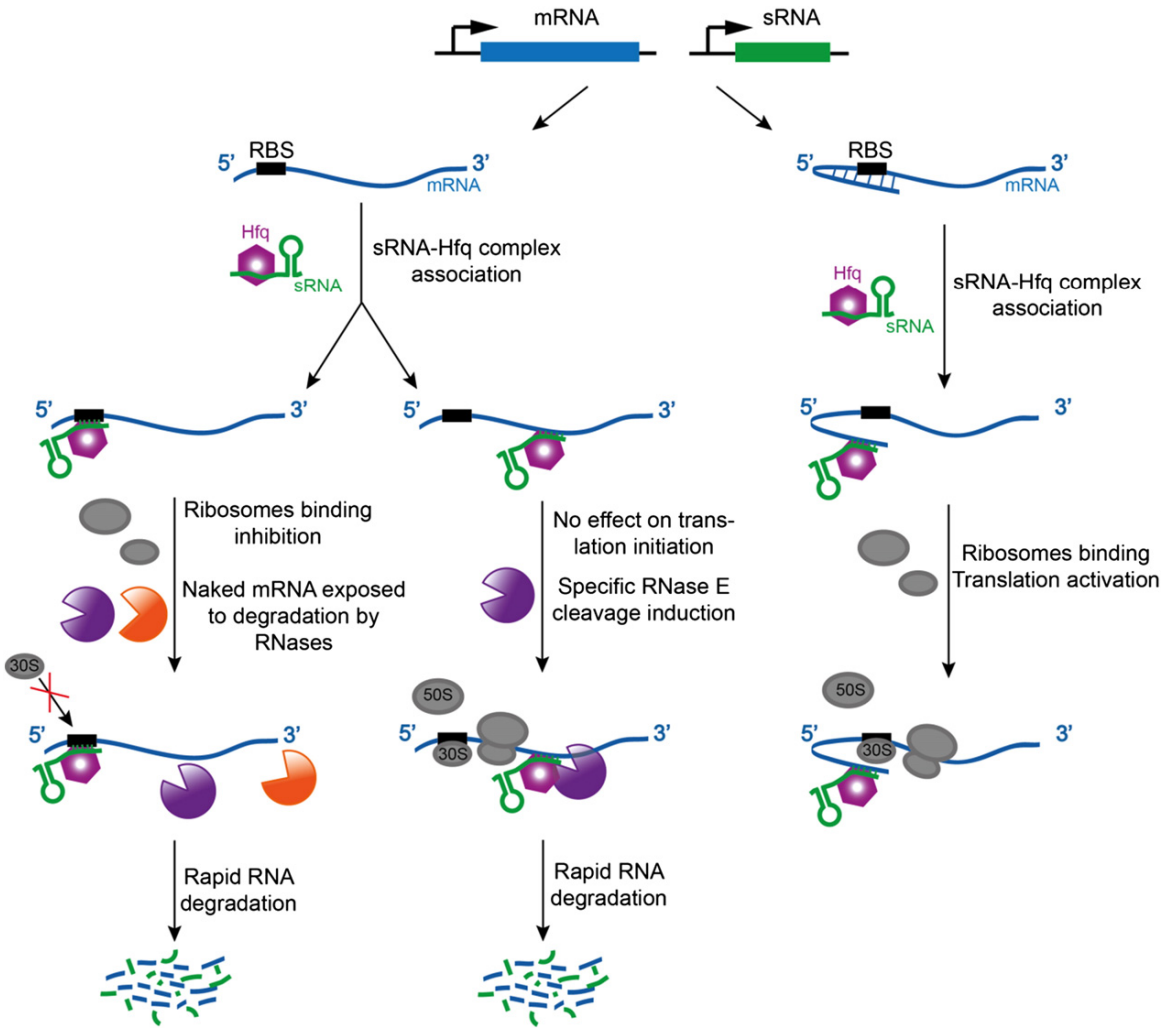

Figure 12: RNA control mediated by sRNAs. Figure from **?**.

4.5 Translation

It is assumed here that the formation of the 30S and 50S complexes are described and modeled somewhere else as well as for the metabolites.

4.5.1 Transfer RNA charging or loading

A tRNA has an anticodon which corresponds to a specific amino acid. There are 64 possibilities for a codon for twenty or so amino acids. Let's number the possibilities with i. Each one of these possibilities leads to a type of tRNA, $tRNA^i$. The mechanism of charging the tRNA is

performed by an enzyme, an aminoacyl tRNA synthetase (aaRS), which is also specific to an anticodon sequence. The mechanism is described below:

$$\begin{cases} aaRS^i + ATP \rightleftharpoons aaRS^i_t \\ aaRS^i_t + AA^i \longrightarrow aaRS^i_{cm} + PP_i + \mathrm{H}^+ \\ aaRS^i_{cm} + tRNA^i \longrightarrow aaRS^i + tRNA^i_c + AMP \end{cases}$$

with the pyrophosphate anion $PP_i = {\mathrm{P_2O_7}}^{4-}$. Pyrophosphate anion is unstable in aqueous solution and hydrolyzes into inorganic phosphate $P_i = {\mathrm{HPO_4}}^{2-}$. In short:

$$\begin{cases} AA^i + ATP + tRNA^i \xrightarrow{aaRS^i} tRNA^i_c + AMP + PP_i + \mathrm{H}^+ \\ PP_i + \mathrm{H_2O} \longrightarrow 2P_i \end{cases}.$$

Finally, note that during the elongation, a $\mathrm{H_2O}$ is consumed. If we pair both process, it is possible to write

$$AA^i + ATP + tRNA^i \xrightarrow{aaRS^i} tRNA^i_c + AMP + 2P_i + \mathrm{H}^+$$

4.5.2 Initiation

The purpose of the initiation is to locate precisely the initiation $fMet$ amino acid at the Ribosome Binding Site where is located the start codon of the mRNA. The initiation is a complex mechanism that starts from the association of the initiation factors to the $30S$ complex and ends with the formation of the $70S$ complex via the binding of the $50S$ complex to the mRNA. The active steps are illustrated in Figure 13. The steps are detailed hereafter:

1. in parallel, the $30S$ complex binds to the $mRNA$ when the IF_3 and IF_1 have already binded to it whereas IF_2 binds with a co-factor and to a $tRNA^{fMet}$ (an $fMet$ charged tRNA):

   - the $30S$ complex binds to the $mRNA$:
     (a) the binding of the initiation factors IF_3 and IF_1 to a $30S$ complex:

     $$\begin{cases} 30S + IF_3 \rightleftharpoons 30S_3 \\ 30S_3 + IF_1 \rightleftharpoons 30S_{31} \end{cases},$$

     (b) the binding of formed complex to the $mRNA$:

     $$30S_{31} + mRNA \rightleftharpoons m30S_{31}$$

   - the binding of the initiation factor IF_2 with a cofactor and to a $tRNA^{fMet}$:
     - IF_2: the initiation factor without cofactor. It is assumed that IF_2 does not bind with a $tRNA^{fMet}$,
     - IF_{2a}: the active form coming from the GTP cofactor

     $$\begin{cases} IF_2 + GTP \rightleftharpoons IF_{2a} \\ tRNA^{fMet} + IF_{2a} \rightleftharpoons tRNA^{fMet}_{2a} \end{cases},$$

     - IF_{2d}: an inactive form coming from the GDP cofactor

     $$\begin{cases} IF_2 + GDP \rightleftharpoons IF_{2d} \\ tRNA^{fMet} + IF_{2d} \rightleftharpoons tRNA^{fMet}_{2d} \end{cases},$$

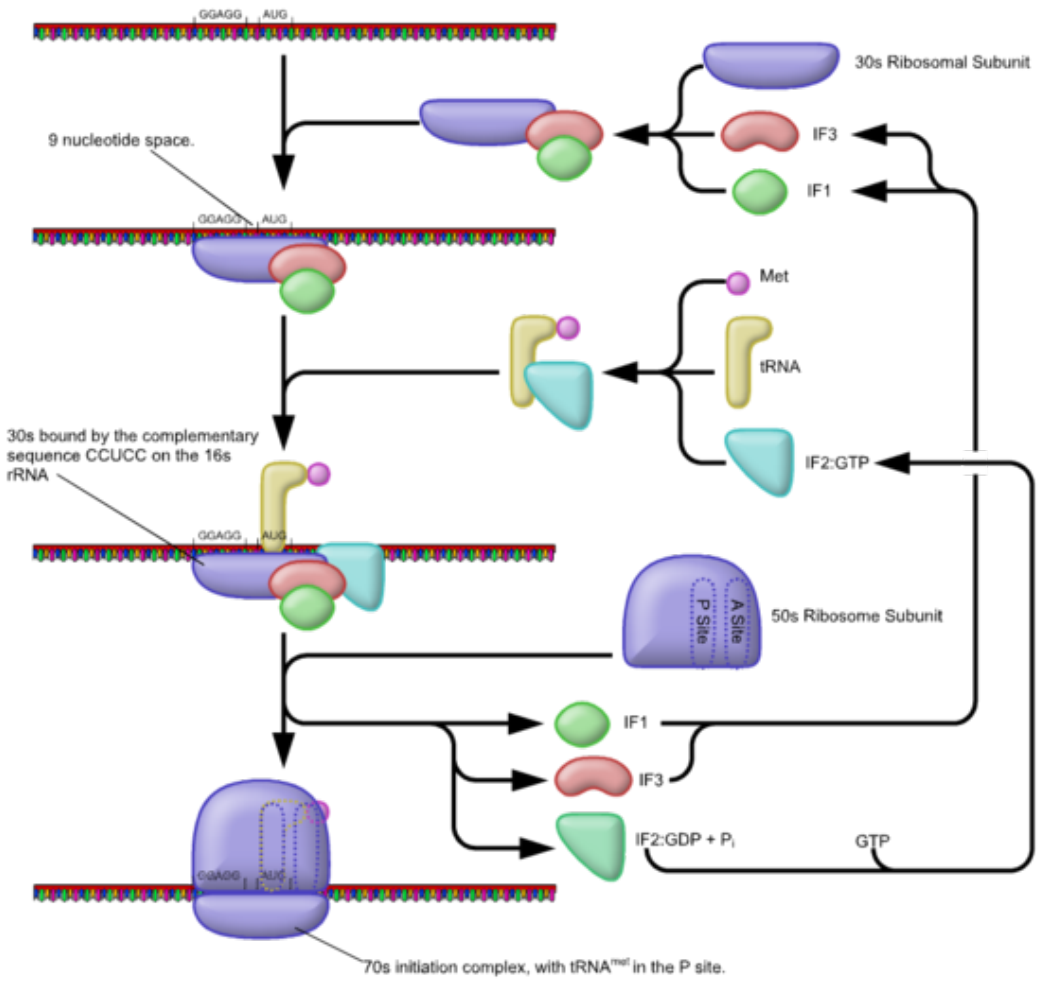

Figure 13: Initiation of translation @Wikipedia

– IF_{2p}: an inactive form coming from the *ppGpp* cofactor

$$\begin{cases} IF_2 + ppGpp \rightleftharpoons IF_{2p} \\ tRNA^{fMet} + IF_{2p} \rightleftharpoons tRNA_{2p}^{fMet} \end{cases};$$

2. the $m30S_{31}$ complex has a high affinity with charged tRNA with $fMet$ (denoted by $tRNA^{fMet}$). It is assumed that $tRNA_{2p}^{fMet}$ does not binds:

$$\begin{cases} m30S_{31} + tRNA_{2a}^{fMet} \longrightarrow mt30S_{312a}^{fMet} \\ m30S_{31} + tRNA_{2d}^{fMet} \rightleftharpoons mt30S_{312d}^{fMet} \end{cases};$$

3. IF_3 is then released:

$$mt30S_{312a}^{fMet} \longrightarrow mt30S_{12a}^{fMet} + IF_3;$$

4. the $50S$ complex binds to form the $70S$ complex and IF_1 and IF_2 are released. Only an active form allows this binding:

$$mt30S_{12a}^{fMet} + 50S \longrightarrow 70S + IF_1 + IF_{2d} + P_i + \mathrm{H_2O}.$$

Note that the $tRNA^{fMet}$ is already at P-site of the $70S$ complex so that it is ready for elongation.

In summary, we have the possible overall reactions:

$$\begin{cases} 30S + IF_3 + IF_1 + mRNA \rightleftharpoons m30S_{31} \\ IF_2 + GTP \rightleftharpoons IF_{2a} \\ IF_2 + GDP \rightleftharpoons IF_{2d} \\ IF_2 + ppGpp \rightleftharpoons IF_{2p} \\ IF_{2p} + tRNA^{fMet} \rightleftharpoons tRNA_{2p}^{fMet} \\ m30S_{31} + IF_{2d} + tRNA^{fMet} \rightleftharpoons mt30S_{312d}^{fMet} \\ m30S_{31} + IF_{2a} + tRNA^{fMet} \longrightarrow 70S + IF_1 + IF_{2d} + IF_3 + P_i + \mathrm{H_2O} \end{cases}$$

4.5.3 Elongation

Elongation is the process of creating the protein by putting amino acid one by one. Roughly, the elongation is illustrated in Figure 14 and follows the steps:

1. a charged tRNA binds to the A-site of the 70S complex;
2. a peptide bond is formed between the new amino acid and the already here one at P-site;
3. the ribosome translocate and the tRNA at E-site is released.

The steps are explained hereafter, with the current (at A-site) codon being i_0:

1. a tRNA (charged or uncharged) binds to the A-site of the 70S ribosomal complex. The affinity for cognate tRNA is higher the one for near-cognate tRNA. This tRNA is carried by the GTPase EF-Tu whose hydrolysis allows for the decoding between the mRNA codon and the tRNA anti-codon on the A-site. EF-Tu is released:

   (a) an EF-Tu with GTP ($EFTu_a$) binds to the tRNA (charged or uncharged)

$$\begin{cases} tRNA_c^i + EFTu_a \rightleftharpoons EtRNA_{ca}^i \\ tRNA^i + EFTu_a \rightleftharpoons EtRNA_a^i \end{cases}$$

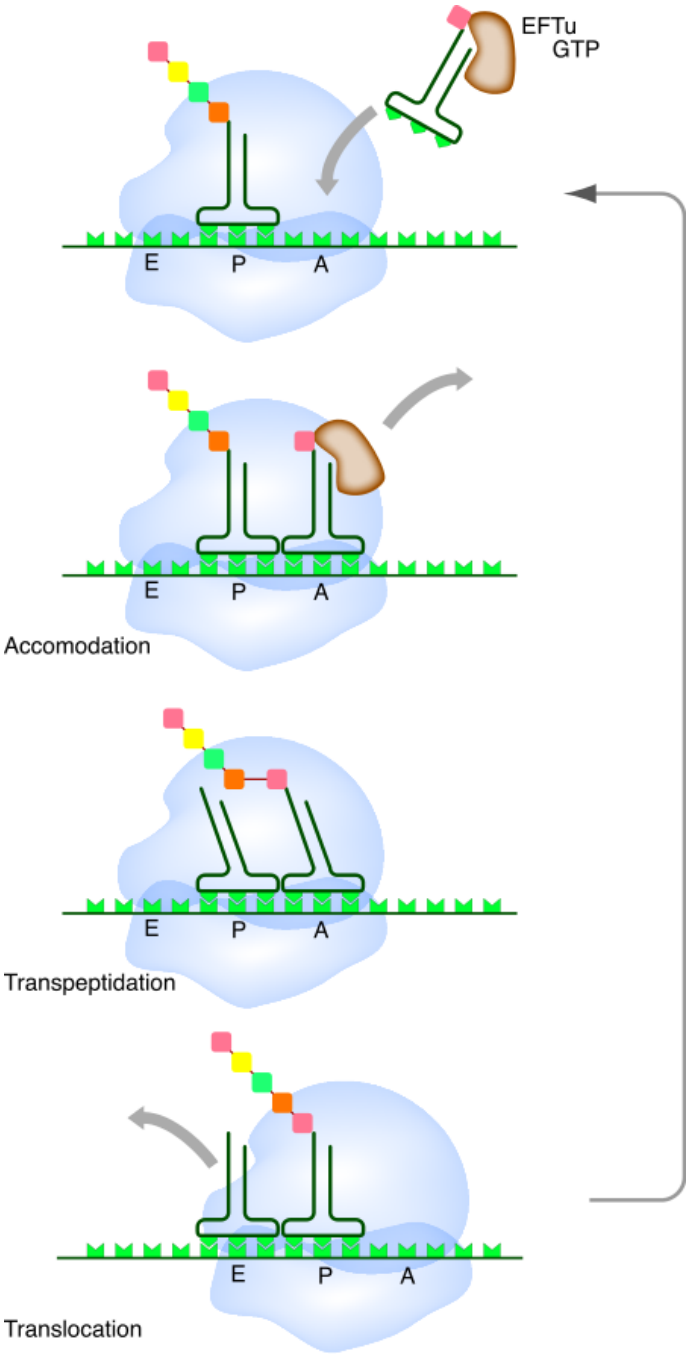

Figure 14: Elongation during translation @Wikipedia

(b) this tRNA then binds to the A-site of the 70S complex and EF-Tu is released (with GDP denoted $EFTu_d$)

$$\begin{cases} EtRNA_a^i + 70S \rightleftharpoons Et70S_a^i \\ EtRNA_{ca}^i + 70S + \mathrm{H_2O} \longrightarrow t70S_c^i + EFTu_d + P_i \end{cases}$$

It s possible that the charged tRNA that comes to the A-site does not correspond to the anti-codon of the mRNA. The protein produced may still functional due to the redundance of the amino acid coding, or that amino acid location is not crucial for the folding and function of the protein; Proofreading (Olivier PhD)

(c) as an aside, $EFTu_d$ is transformed back to $EFTu_a$ via a GTPase EF-Ts:

$$EFTu_d + GTP \xrightarrow{EFTs} EFTu_a + GDP$$

2. the peptide bond is formed via the elongation factor EF-P

$$\begin{cases} EFP + GTP \rightleftharpoons EFP_a \\ t70S_c^i + EFP_a \longrightarrow t70S^{+i} + EFP_d + \mathrm{H_2O} + P_i \end{cases}$$

with EFP_d being the GDP bounded form of EFP.

3. translocation takes place via a elongation factor EF-G, a GTPase:

$$t70S^{+i} + GTP + \mathrm{H_2O} \xrightarrow{EFG} 70S^+ + tRNA^i + GDP + P_i.$$

4.5.4 Termination

When the ribosome encounters a stop codon, it release the mRNA and the protein. In more details, when after the translocation step of elongation the ribosome encounters at its A-site a stop codon, a release factor (either RF_1 or RF_2 depending on the stop codon) binds. Note that tRNA are unable to recognize a stop codon. A final GTPase RF_3 hydrolysis allows the release of the protein. It can be summarized as follows:

$$\begin{cases} RF_3 + GTP \rightleftharpoons RF_{3a} \\ 70S + RF_1 \rightleftharpoons 70S_1 \\ 70S + RF_2 \rightleftharpoons 70S_2 \\ 70S_1 + RF_{3a} + \mathrm{H_2O} \longrightarrow 30S + 50S + protein + RF_1 + RF_{3d} + P_i \\ 70S_2 + RF_{3a} + \mathrm{H_2O} \longrightarrow 30S + 50S + protein + RF_2 + RF_{3d} + P_i \end{cases}$$

with RF_{3d} being GDP bounded form of RF_3.
From Olivier PhD, last AA not in the protein.

4.5.5 Translation termination on a broken mRNA

tmRNA coded by ssrA.

4.6 Protein processes

After being translated, a protein undergoes several processes before becoming mature. 5 important post-translational processes can be isolated (Figure 16), even though they may occur in parallel. 3 processes are common to cytosolic, membrane and secreted proteins, namely

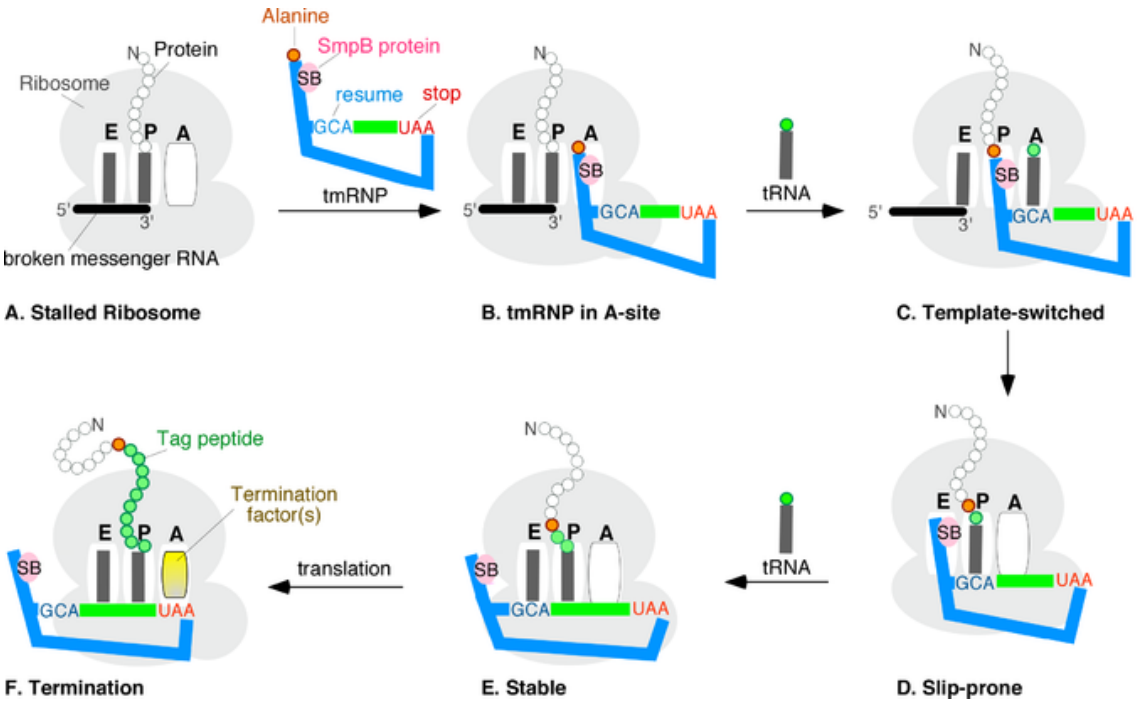

Figure 15: Termination on stalled ribosome @Wikipedia

- Deformylation and demethylation.
- Folding.
- Residue modifications.

Moreover, there are 2 processes that specifically target membrane and secreted proteins

- Translocation.
- Cleavage of the peptide signal and lipoprotein diacylglyceryl adduction.

4.6.1 Deformylation and demethylation

For every protein, translation begins with a formylmethionine (fMet) residue. Several authors suggest that using exclusively fMet to start protein translation may be a way to regulate protein synthesis, as fMet is one of the most expensive amino acids to synthesize. Thus global protein synthesis levels would depend on global cell health.

This residue is then partially removed, but the role of deformylation and demethylation remains poorly understood. It may help in recycling fMet and influence protein degradation. According to the N-end rule, some N-terminal residues will lead to enhanced degradation. By changing the nature of the N-terminal residue, deformylation and demethylation allows for the N-end pathway to catalyze protein degradation.

Peptide deformylases (PDF) remove the formyl residue from fMet for most proteins. In *B. subtilis*, YkrB is the main PDF, even though Def is known to have a similar action. It remains unclear whether these enzymes act during protein translation or after translation is completed. Methionine cleavage is performed by Methionine Aminopeptidases (MAP) but, contrarily to deformylation, it only targets a minority of proteins.

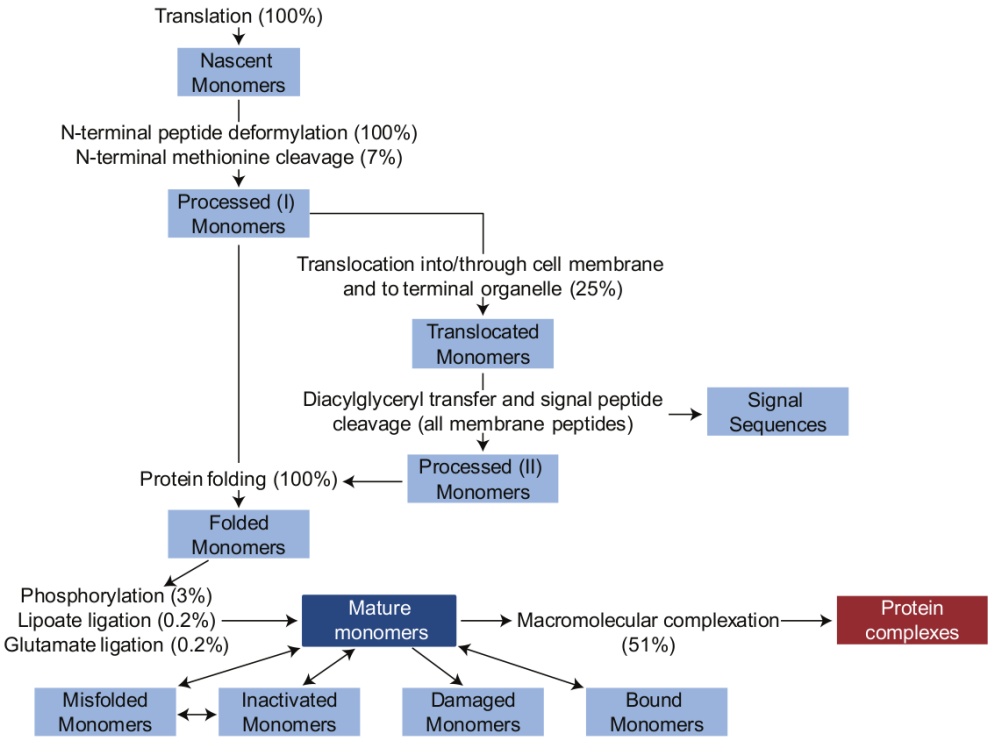

Figure 16: Processes undergone by protein following translation (from Karr *et al.*)

4.6.2 Protein folding

After translation, proteins can be viewed as a one-dimensional chain of amino acids. In order to acquire their functional form, proteins need to adopt a specific three-dimensional folding. Even though this folding is energetically favorable, numerous proteins need the help of another protein to adopt the proper conformation. These helper proteins are called chaperones. Bacteria use several chaperones. For example, in *B. subtilis*

- The trigger factor Tig is known to assist early folding.
- GroEL and its co-enzyme GroES assist late folding of short proteins.
- DnaK, DnaJ and GrpE assist late folding of longer proteins.
- FtsH may assist membrane protein folding.
- Some proteins have specific chaperones (for example the translocated proteins TorA and NiFe are chaperoned by TorD and HybE respectively).

The trigger Factor (abbreviated Tig or TF) is an ATP-dependent protein responsible for folding assistance of nascent proteins by associating with the ribosome. It seems that it does not associate during the first steps of translation (maybe because of deformylation, demethylation or peptide signal recognition). 60 to 70% of proteins may be in their native conformation after translation and binding by TF **?**. Despite its large action spectrum, TF is not an essential protein as it does not seem to be responsible for folding of vital proteins and its action can be supplemented by other chaperones.

DnaK is part of the family of Heat Shock Proteins (it is a homolog of Hsp70). It is an ATP-dependent chaperone responsible for the folding of 5 to 18% of newly synthesized proteins

(?). As part of the heat shock response it is also responsible for desaggregating and refolding proteins. This is particularly true for large proteins: 80% of proteins ¿70kDa aggregate during heat shocks and even at 37C in the absence of TF and DnaK (GroEL only folds proteins ¡60kDa) (?). Roughly speaking, DnaK is composed of a hydrophobic pocket composed of β-strands, an arch of a few amino acids that facilitates substrate binding, a "lid" composed of α-helices and an ATP-binding domain (???). When ATP is bound to DnaK, the protein is in an "open" state with a higher dissociation constant than when it is bound to APD. Cycles where DnaK associates and dissociates from ATP could therefore lead to capture-recaptures of the substrate that may induce its folding (?). Schematically, when DnaK is bound to ATP, it is in an acceptor state. Binding to a substrate with the help of DnaJ leads to dramatically increased hydrolysis rates and trapping of the substrate. GrpE then binds to the nucleotide binding and triggers ADP release when a new ATP binds (Fig. 17, top)(?). In heat shock conditions, the cycle is similar, except there is a competition between GrpE and ClpB, another ATP-dependent chaperone that is similar to AAA+ domains of proteases (see Protein degradation below). ClpB-DnaK is supposed to trigger desaggregation and partial refolding, while GrpE may help in completing folding if necessary (Fig. 17, bottom) (?). DnaK is not an essential protein, in its absence the cell remains viable but does not grow optimally. In the absence of DnaK and TF, the cell dies except if GroEL is strongly overexpressed (?).

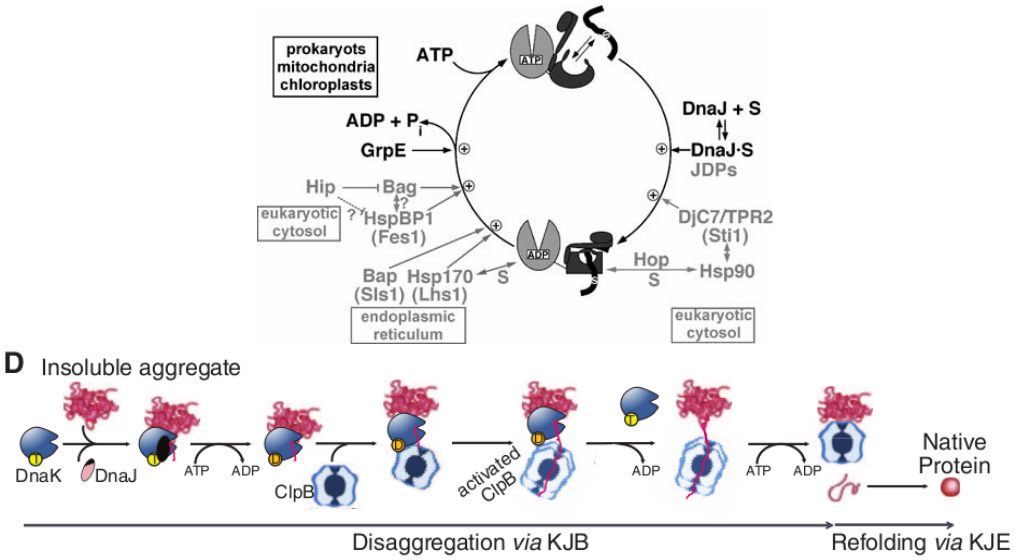

Figure 17: One cycle of protein folding by DnaK/DnaJ/GrpE (top). Protein desaggregation and refolding by DnaK/DnaJ assisted by ClpB for desaggregation and GrpE for refolding (bottom). Figures from ? and ?.

As DnaK, GroEL is responsible for folding and desaggregation. It folds 5 to 10% of newly synthesized proteins but it is only able to fold proteins whos size is in the range of 20kDa to 60 kDa (?). It is composed of two back-to-back heptameric cavities that can each bind 7 ATP molecules (?). When the substrate is not in excess, the cavities work in turn at a rhythm dictated by the hydrolysis of the 7 ATP molecules (Fig. 18). The two cavities may influence each other. According to ?, one of the cavities is always occupied, binding of substrate to the entrance of the empty cavity releases ADP from that cavity, which enables ATP hydrolysis in the other cavity. Once ATP is hydrolysed in the occupied cavity, ATP may bind to the empty cavity and starts drawing the substrate inside and drives closing of the cavity by GroES and release of substrate in the other cavity. Folding could be linked to both substrate elongation

and compaction when it enters the cavity (?) or strongly charged surfaces within the cavity (?). More precisely, ? show that GroEL captures/recaptures one of its substrate at a rate that depends on the temperature (higher temperature leads to stronger ATP consumption and quicker cycling). The higher the temperature, the quicker the folding. At more than 30C, the folding occurs quicker than pure sequestration in a mutated GroEL that does not release the substrate (and does not consume ATP), indicating a potential functional role to the capture mechanism. On the contrary, ? suggest that sequestration may be enough for a substrate that contains a TIM barrel composed of α-helices. By positioning some of the helices through interactions with the cavity, the folding is increased 50-fold and becomes quicker than the synthesis of the protein. Through sequestration, GroEL also participates to removing protein aggregates. Although it is more specific than DnaK and TF, GroEL is an essential protein, probably because it folds essential proteins that can not be folded properly by other chaperones.

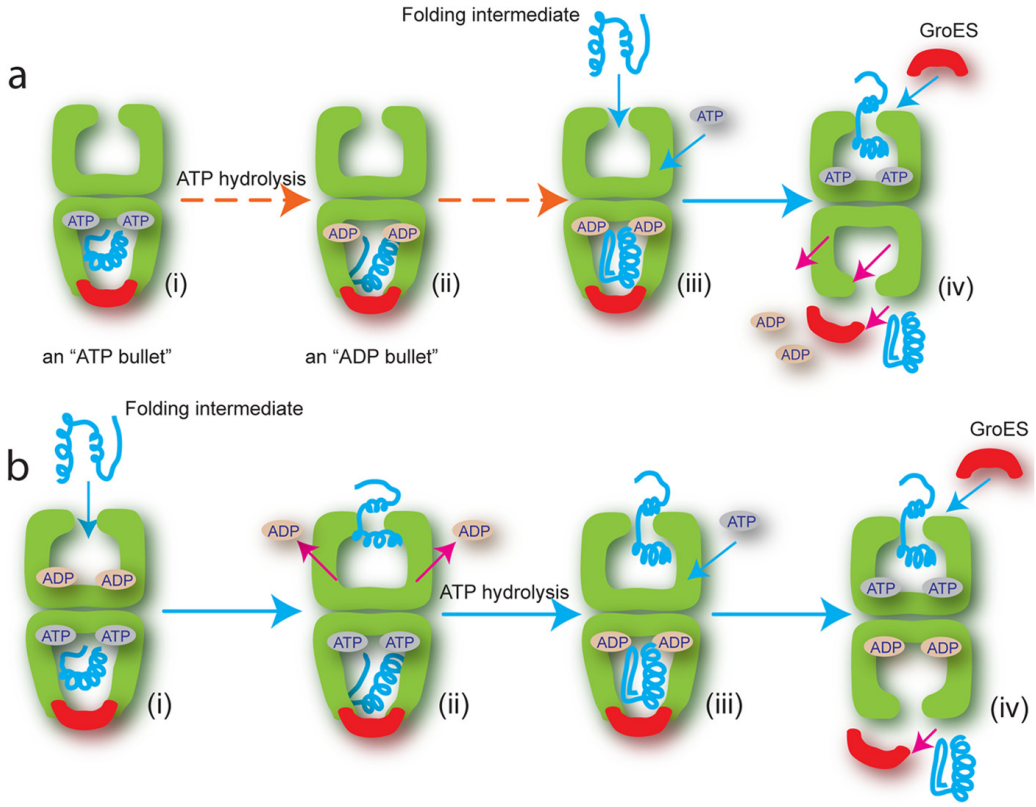

Figure 18: Models for cycles of protein folding by GroEL and GroES. Figure from ?.

4.6.3 Translocation, peptide signal cleavage and diacylglyceryl adduction

During its life cycle, a bacterium heavily interacts with the extracellular medium. These interactions are enabled by proteins that are embedded in the membrane (nutrient transporters, ATP synthase, receptors, transducers) or secreted (virulence factors, cell wall components). Proteins that need to undergo translocation are identified by their N-terminal and C-terminal signal sequences (petide signals). Their embedding and secretion through the hydrophobic membrane is assisted by a specific machinery.

Two important pathways are involved in protein translocation: Tat and SecA. They both involve a docking and a translocation step. Docking is mediated by peptide signals that can be recognized by signal recognition proteins (SRP) or directly by membrane proteins involved in

the translocons (such as SecA or TatC). Translocation then occurs through a pore of variable size consisting of several proteins (SecYEG translocase or TatA like proteins). Depending on the nature of the proteins, translocation can occur simultaneously to translation (integral membrane proteins) or after translation (lipoproteins and secreted proteins).

Following translocation, lipoproteins are anchored to the membrane through ligation of diacylglyceryl mediated by diacylglyceryl transferase and their peptide signal is removed by a signal peptidase. (what about the cleavage of peptide signals of secreted proteins and integral membrane proteins in *B. subtilis* ???)

4.6.4 Residue modifications

Some proteins undergo post-translational chemical modifications. These modifications may serve different objectives such as favoring alternative conformations or regulating protein activity. Examples of residue modifications include phosphorylation, lipoyl transfer to lysine and α-glutamate ligation, all catalyzed by specific enzymes.

4.6.5 Protein degradation

Protein degradation is assured by proteases in a process called proteolysis. Proteolysis has several roles within the cell (**??**):

- Proteolysis may degrade destructured proteins to avoid toxicity and enable amino acid recycling. However, **?** argue that refolding may be the usual recycling mechanisms, as a typical *E. coli* cell contains roughly 1600 and 10000 copies of chaperones GroEL and DnaK, respectively, against 50-150 copies of proteases ClpXP, ClpAP and Lon. Molecular competition for denatured substrates would therefore largely favor chaperones.

- Proteolysis may be used to downregulate specific protein pools. Gene and mRNA regulation does not directly impact these pools, targeted degradation allows rapid decrease of protein concentration.

- Proteolysis may only affect a small part of the protein. Typical examples are found in transduction, where a protein is cut in two at the level of the membrane, yielding a cytosolic part that may activate gene transcription.

There is a family of well-characterized proteases that are called AAA+ proteases (ATPases Associated with diverse cellular Activities) which are thought to act on a very large spectrum of proteins. The members are schematically composed of two parts: an AAA+ module (often a ring-shaped hexamer) that is able to unfold and translocate proteins, and a protease module (a cavity-shaped oligomer), that cuts the protein into small pieces, in an ATP independent manner (Fig. 19) (**?**). Folded proteins cannot enter the protease chamber, the association with the AAA+ module is essential for proteolysis. This association enables targeted substrate recognition.

Figure 20 gives an overview of the main AAA+ proteases found in bacteria. The AAA+ module can be on the same protein as the protease (FtsH, Lon) or translated independently (HslU+HslV, ClpX+ClpP, ClpA/C+ClpP, PAN+20S). In the latter case, it can have a translocation activity that is independent of proteolysis, as has been shown for ClpX, or ClpA homolog ClpB (which does not associate with a protease) (**?**). The cost of translocation is difficult to evaluate. **?** remind that, depending on the conformation, the translocation of the Tit protein can cost from 100 ATPs to 600 ATP. On ClpX 6 ATP can theoretically bind, even though it

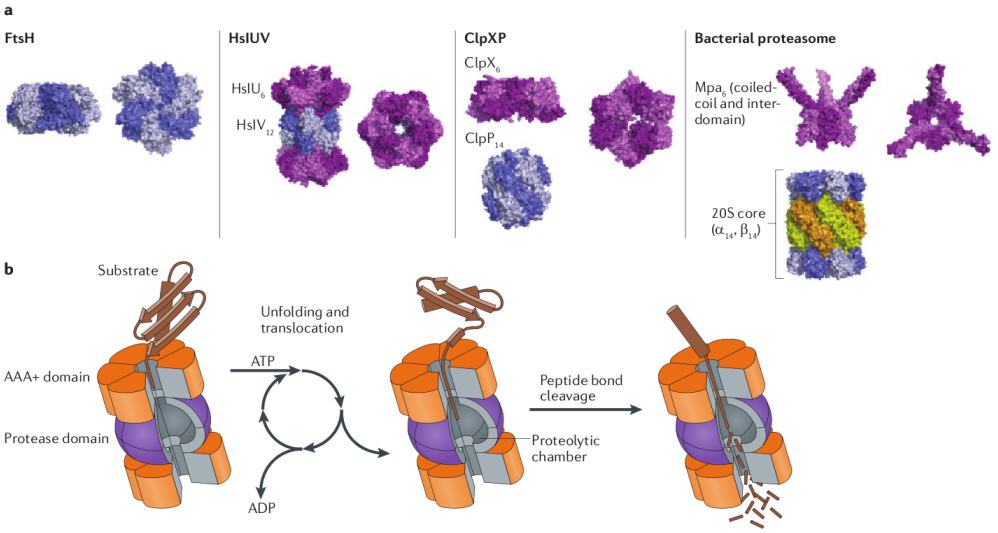

Figure 19: Structure of AAA+ proteases. Top: AAA+ hexamers are shown in purple, proteolytic cavities in other colors. Bottom: schematic representation of protein degradation by an assembled protease. Figure from **?**.

seems that only 4 can bind at the same time (because of conformation changes, 2 sites cannot be accessed) and a typical cycle alternates between 4 ATP and 3 ATP + ADP. The authors also distinguish three translocation states that influence ATP consumption: (i) protein spontaneously unfolds with traction, translocation is cooperative, (ii) protein resists to traction, leading to translocation "slippage" and wasted ATP, (iii) protein resists to traction and dissociates from the translocase. The latter case might sound dramatic at first, but it may be a functional way to distinguish between native and destructured proteins, native proteins having a much higher probability of unbinding, so only unfolded proteins are efficiently degraded.

A protein may be degraded only if it is recognized by the AAA+ module. The recognition is mediated by a degradation tag (often referred to as degron). **?** distinguish four essential mechanisms in substrate recognition:

- **Cryptic degron**: a sequence of amino acids within the protein that is accessible only in specific conditions, e.g. the protein is unfolded.
- **Amino acid appendage**: a sequence of amino acids that is recognized by a protease is added to the protein. A typical example is the SsrA tag: when a ribosome stalls, a tmRNA is loaded and the ribosome translates the tmRNA instead of the original mRNA. The amino acid sequence coded by the tmRNA is termed SsrA tag and is specifically recognized by ClpXP.
- **N-end rule**: the mechanism here is not totally understood, but the presence of some amino acids at the N-terminal end influence degradation, namely leucine (L), phenylalanine (F), tryptophane (W) and tyrosine (Y) (**??**). This process is mediated by ClpS and allows recognition by ClpAP, even though the details are unclear (**???**). Protein could be tagged for degradation by cleaving the initial methionine or by changing the N-terminal residue by specific enzymes (which target specific amino acids, called secondary targets, see Fig. 21).
- **Adaptator proteins**: these proteins serves as intermediates between the substrate and the protease, e.g. ClpS connects N-end rule degrons to ClpAP, SspB connects SsrA

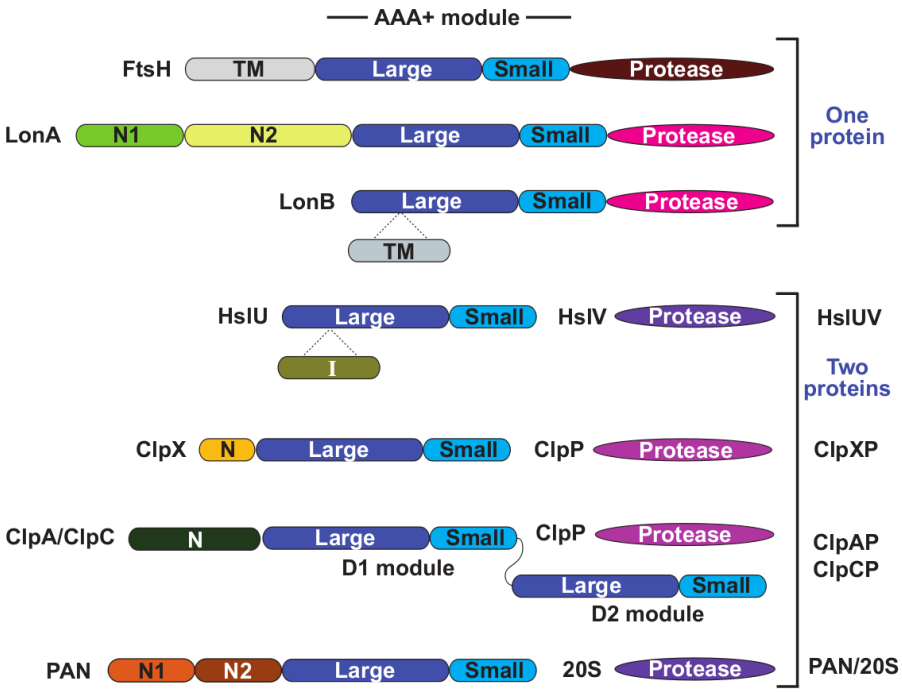

Figure 20: Phylogeny of the main AAA+ proteases. Top bracket: proteases where the AAA+ module and the proteolytic cavity are within a single protein. Bottom bracket: proteases where AAA+ module(s) and proteolytic cavity are separated. Colors indicate homologous domains (AAA+ modules are particularly well conserved). Figure from **?**.

degrons to ClpXP. This allows for numerous regulation possibilities (competition between adaptators for a protease, antiadaptators).

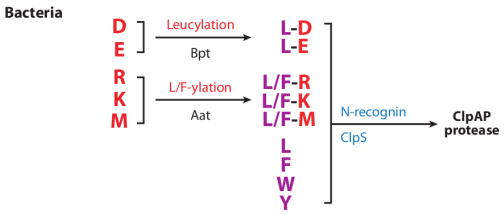

Figure 21: Primary targets (purple) and secondary targets (red) of the N-end rule pathway. The main proteins responsible for each process is indicated next to each arrow. Figure from **?**.

4.7 Metabolism

Joker!

4.8 Regulation

We only describe the mechanics of (transcriptional and translational) regulations. For specific regulations, see somewhere else. The stringent response is a form of regulation but it is so central to the bacteria growth that we preferred to separate it from the remaining regulations and detail it.

4.8.1 Transcription

Transcription factor Transcription factors are cis-acting regulatory effectors. They bind near the promoter region and can have an inhibition or activation effect. Generally, transcription factors with an inhibition effect binds after the promoter region whereas the ones with an activator generally binds before the promoter region:

- Inhibition transcription factors inhibits the transcription by physically preventing the polymerase RNA to bind to the promoter or to move forward (roadblock). A more complex inhibition mechanism occurs when one or several effectors binds to the DNA in order fold around the promoter;
- Activation transcription factors either increases the affinity of the promoter to the pRNA or stabilizes the DNA - RNA complex.

Reiterative transcription As described in the initiation of transcription, several abortive initiation may occur before promoter clearance. This regulatory phenomenon is correlated to slippage during initiation Turnbough (2011):

- direct control is performed by the competition between the repeated nucleotide and the next normally templated one;
- indirect control is performed by other transcription factors: the molecular mechanisms are unclear.

Initiating NTP Unlike for most operons, the open holoenzyme for rrn operons are unstable ones Gaal et al. (1997); Krsn and Gourse (2004); Rojo et al. (1993). In E. Coli, this unstability is mediated by the binding of DksA to the pRNA (Paul et al., 2004). The initiating NTP is then determinant for the stability of the DNA-RNA complex.

Riboswitch Riboswitches are cis-acting effects that control the transcription via a termination / antitermination mechanism. Riboswitch includes S-box, L-box, A-box and G-box

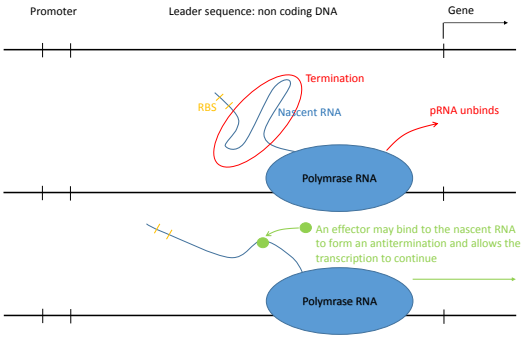

Figure 22: Illustration of a riboswitch regulation: during the transcription of a non coding sequence, a termination is formed. When an effector binds to a cognate location on the nascent RNA, an antitermination is formed instead of the termination.

mechanism. We also include the T-box mechanism is the same as the one of riboswitches even if they are generally not classified as the effector is not a metabolite but a (specific to the nascent RNA) uncharged tRNA. The effectors are listed in Table 1. In the review of Winkler and Breaker (2005), at least 9 effectors are listed with several more putative ones.

Mechanism	Effector
S-box	s-adenosyl-l-methionine
L-box	lysine
A-box	adenine
G-box	(mostly) hypoxanthine and guanine
T-box	specific uncharged tRNA

Table 1: Effectors of riboswitch mechanisms

Regulation by RNA cleavage / degradation Post transcriptional regulation were evidenced on the *gabA* operon by Ludwig et al. (2001) for which the mRNA is cleaved between *cggR* and *gapA* which allows a fine regulation of *gapA* expression in *Bacillus Subtilis*. Due to the structure of the mRNA, an RNase E could be the actor of the cleavage but RNase E does not exist in *Bacillus Subtilis* so that the cleavage actor is unknown.

4.8.2 Translation

In addition of the role of thr stringent response via the alarmones described in Section 4.9, we describe here other mechanisms of translation regulations.

Sequestration of ribosome binding site This sequestration can be performed by the mRNA itself or by an effector:

- in the case of the mRNA itself, the mechanism is similar to riboswitch in the sense that the nascent RNA forms a motif by folding itself around the RBS;
- an effector (metabolite, protein or small RNA) binds to the nascent mRNA and sequesters the RBS.

In both cases, the RBS of the mRNA is made physically inaccessible to the ribosome.

Unsequestration of ribosome binding site In (Lovett and Rogers, 1996), an terminator is present when no stalling ribosome. Presence of stalling ribosome insequesters the second RBS termed as RBS-C. The stalling location of the first ribosome also determines the accessibility of the RBS-C.

Small RNA Small RNAs are non coding and can bind to mRNA in the RBS region and sequesters it. xxx Can be also regulation on elongation and decay xxx

tmRNA xxx For stalled ribosome. Add a tag on the peptide chain for protease, see Section 4.5.5 xxx

4.8.3 Transcription and Translation coupled

Peptide leader In prokaryotes, transcription and translation can occur at the same time on a nascent RNA. The peptide leader regulation mechanism is a coupled transcription / translation one and is illustrated in Figure 24

4.8.4 Allosteric regulation

???

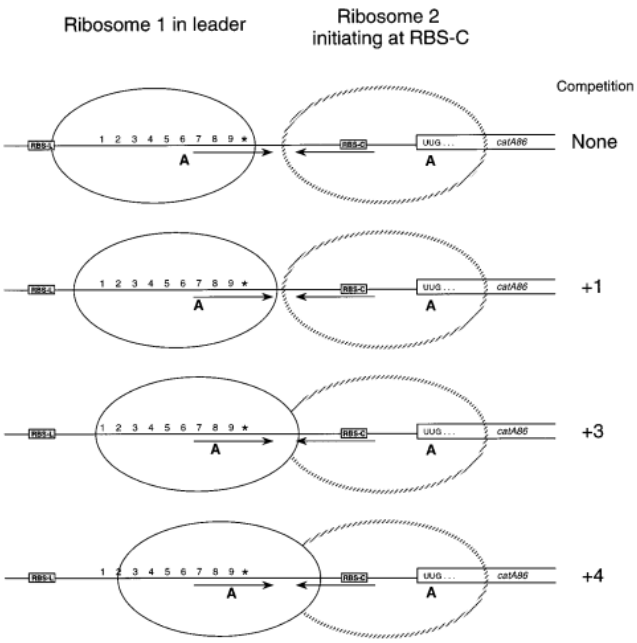

FIG. 4. Proposed effect of stalling the aminoacyl site of a ribosome at different leader codons on translation initiation at RBS-C. From the experiments of Alexieva et al. (1) and Gu and Lovett (44), we propose that a ribosome stalled with its A site at leader codon 6 can destabilize the stem-loop structure and not compete with entry of a ribosome at RBS-C. When the leader ribosome A site occupies leader codon 7, 8, or 9, the leader ribosome progressively encroaches on the space needed for a second ribosome to initiate at RBS-C. Consequently, a ribosome stalled at leader codons 3′ to leader codon 6 will impede or prevent the entry of a ribosome at RBS-C. This model explains why ribosome stalling at leader codon 6 is the ideal site for ribosome stalling that results in induction of *cat* translation. The model further predicts that basal (uninduced) *cat* expression is largely the result of initiation at RBS-C which occur as the leader ribosome translates through leader codon 6.

Figure 23: From (Lovett and Rogers, 1996)

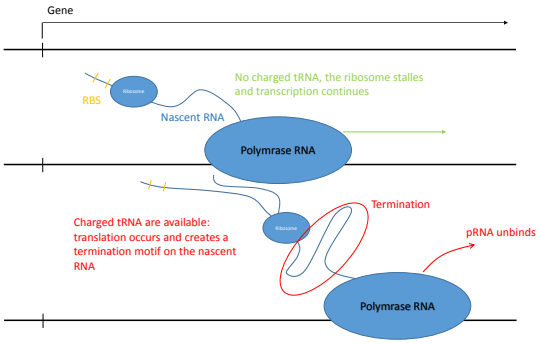

Figure 24: Illustration of peptide leader regulatory mechanism

4.8.5 The tryptophan regulation pathway: an example of several mechanisms

This section is heavily based on (Gollnick et al., 2005) and the references therein. Except for Figure 25, all of them are from Gollnick et al. (2005). The regulation network of the tryptophan biosynthesis in *Bacillus Subtilis* is illustrated in Figure 25 and several mechanisms described above are used.

The major regulatory effector is TRAP (trp RNA-binding Attenuation Protein) coded by mtrB and is a 11-mer protein. mtrB is within a two genes operon that is not regulated by tryptophan or any intermediate of the biosynthesis pathway. The active TRAP forms when it binds to tryptophan; it then literally wraps itself with the nascent mRNA which has a double effect on the expression of the trp operon: *(i)* it acts as a riboswitch effector and creates a terminator hairpin instead of the antiterminator one as illsutrated in Figure 26 ; *(ii)* another hairpin is created that sequesters the SD (see Figure 27). Two positions (U107 and U144) were identified as pause position stimulated by NusA. These are thought to be 'dynamical'

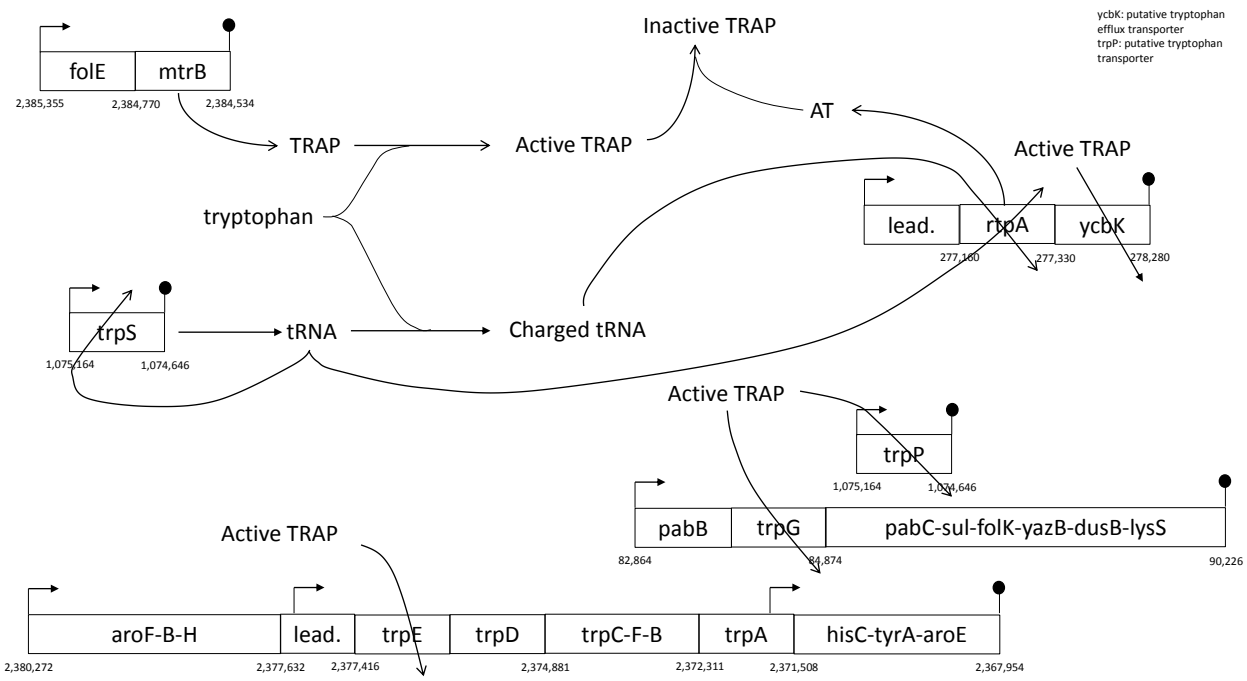

Figure 25: Summary of tryptophan biosynthesis regulation network in *Bacillus Subtilis*

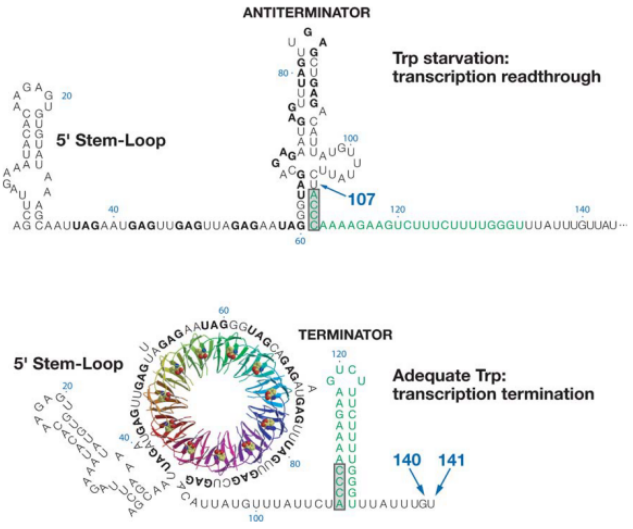

Figure 26: TRAP effect on transcription initiation

pauses that let an active TRAP protein to bind to forms the transcription terminator hairpin and the SD sequestering hairpin. trpE regulation is coupled to trpE translation.

Active TRAP also inhibits trpG and trpP translation initiation by (competitively with the ribosome) binding to the mRNA and sequesters the SD region. While bounded, it also inhibits the translation of ycbK but it seems that this inhibition occurs during elongation (roadblock) and not initiation.

The *at* operon is regulated by the ratio of uncharged / charged tRNA. The organization of the *at* operon and the regulations observed are presented in Figure 28. In this figure, the leader region is a T-box where an antiterminator is promoted when uncharged tRNA binds. The regulations observed may be explained by an interaction between the synthesized small

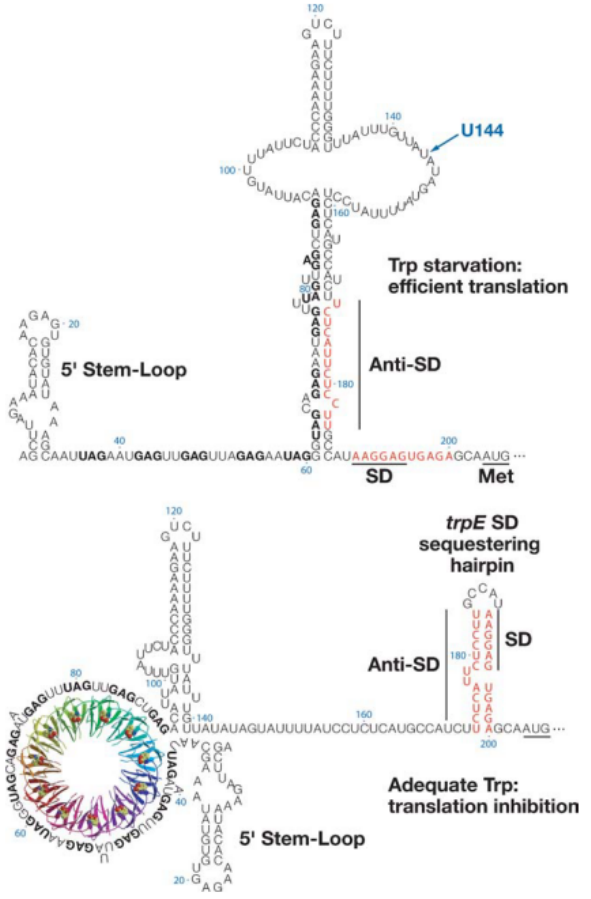

Figure 27: TRAP effect on translation initiation

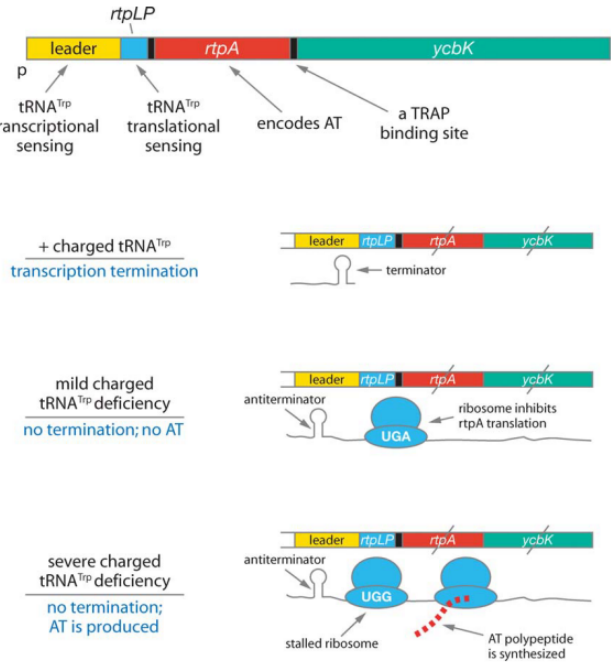

Figure 28: Organization of *at* operon and observed regulation

protein and the nascent mRNA (Chen and Yanofsky, 2003) (binding of the small protein on the start codon?) but they are also compatible with the model of peptide leader mechanism described in Section 4.8.3.

Finally, trpS encoding the tryptophan tRNA is regulated by a T-box riboswitch.

4.9 Stringent response

The stringent response is a fundamental regulation of the bacteria that regulates its growth rate and acts on practical all aspects of the cell. Discovered when amino-acid starving a bacteria (stringent response) compared to a rich nutriment response (relaxed response), the bacteria produces alarmones (p)ppGpp that induces the stringent response. Now stringent response means any response linked to the presence of the alarmones, including other stresses than the amino-acid starvation. The stringent response is different between *Escherichia Coli* and *Bacillus Subtilis* in its mechanism. The following text is principally based on the reviews of Kriel et al. (2012) and of Wolz et al. (2010).

Alarmones synthesis / hydrolysis Figure 29 summarizes the current knowledge. Rsh

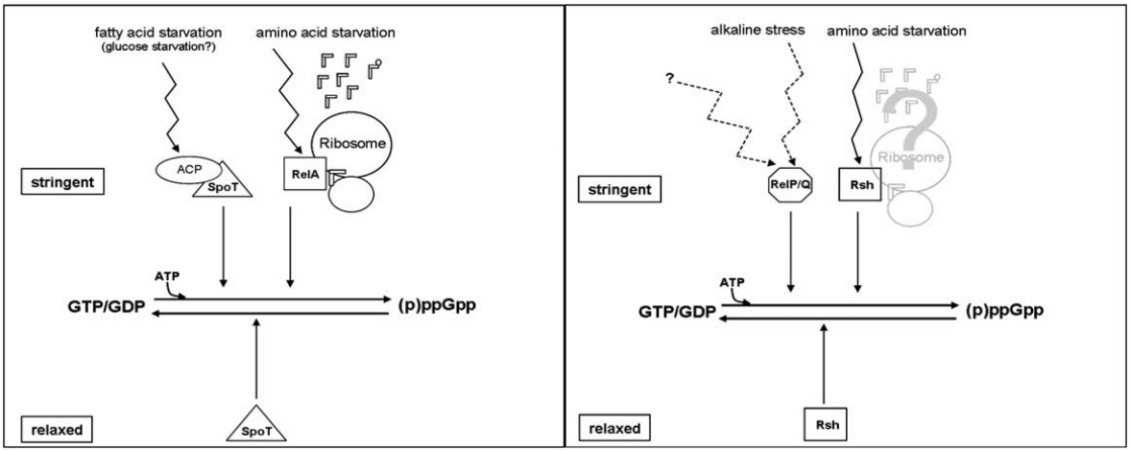

Figure 29: Comparison of alarmones synthesis and hydrolysis between Escherichia Coli (left) and Bacillus Subtilis (right) (Wolz et al., 2010)

(which is also referred to as RelA) in *B. Subtilis* and SpoT in *E. Coli* are bifunctional (the others being monofunctional) enzymes that can perform synthesis and hydrolysis. They have two conformations ((p)ppGpp-hydrolase-OFF/(p)ppGpp-synthase-ON and hydrolase-ON/synthase-OFF) that is regulated by the C-terminal domain as its truncation enhances the synthase activity. It is not known if, as in the case of *E. Coli*, Rsh binds to a ribosome to sense the amino-acid starvation and more precisely the presence of uncharged tRNA. In *E. Coli*, RelA binds to a ribosome and is thought to unbind in the presence of an uncharged tRNA in the A-site. This unbinding activates the synthetase function of RelA and inhibits the rebinding during a period of time. In *B. Subtilis*, it is knwon that there are only one bifunctional enzyme (Rsh) and only two monofunctional ones (RelP seems to be connected to competence and is encoded by yjbM whereas RelQ seems to be connected to persistence and virulence and is encoded by ywaC) that lack the Mn^{2+}-dependent hydrolase domain (Gaca et al., 2015).

Overall effect of alarmones It is estimated that, in exponential gorwth conditions, the concentration of alarmones is less than 20 pmol per optical-density unit by rises to millimolar

levels in response to adverse growth conditions, such as amino acid starvation (Ababneh and Herman, 2015). The effect of the alarmones on proteobacteria (*E. Coli*) and firmicutes (*B. Subitlis*) is summarized in Table 2. Alarmones affects the majors process of the bacteria: replication, transcription and translation. The means are however different. More generally, it

(p)ppGpp-related effects in proteobacteria and firmicutes.

	Proteobacteria	Firmicutes	Reference
RNAP	(p)ppGpp interaction	No indication for (p)ppGpp interaction	(Barker et al., 2001; Krasny and Gourse, 2004; Vrentas et al., 2008)
DksA cofactor for RNAP-(p)ppGpp interaction	Yes	DksA not present	(Magnusson et al., 2007; Paul et al., 2004, 2005)
rrn promoters	GC-rich discriminators	No discriminators	(Haugen et al., 2006)
Rrn transcription	*rrn* promoter insensitive to initiator nucleotide	Indirect effect on transcription via lowering of the GTP pool, G essential as +1 nucleotide of *rrn*	(Haugen et al., 2006; Krasny and Gourse, 2004)
Sigma factor competition	(p)ppGpp facilitates RNAP interaction with alternative sigma factors	Only indirect interaction with e.g. sigma factor B	(Costanzo et al., 2008; Laurie et al., 2003; Magnusson et al., 2005; Zhang and Haldenwang, 2003)
CodY	CodY not present	GTP activates CodY repressor function in some firmicutes	(Sonenshein, 2005)
Replication	Inhibition of initiation	Inhibition of elongation, inhibition of primase	(Levine et al., 1991; Wang et al., 2007)
Translation	Inhibition of protein synthesis	ppGpp binding to translation factor IF2	(Milon et al., 2006; Svitil et al., 1993)
SOS response	Induced	Unclear	(Durfee et al., 2008)
Essential Small GTPases	Cgt/Der not essential in relA-negative background, Interaction with SpoT	pppGpp co-crystallized with Obg, Obg essential also in Rsh-negative background	(Buglino et al., 2002; Hwang and Inouye, 2008; Kuo et al., 2008; Raskin et al., 2007)
Nucleotide pools	(p)ppGpp conc. < ppGpp conc., ppGpp more potent for growth rate regulation	(p)ppGpp conc. > ppGpp conc., (p)ppGpp more potent for primase inhibition	(Potrykus and Cashel, 2008; Wang et al., 2007)
Inhibition of IMP dehydrogenase	Yes	Yes	(Gallant et al., 1971; Lopez et al., 1981)

Table 2: Comparison of stringent response between Escherichia Coli and Bacillus Subtilis (Wolz et al., 2010)

is suggested in (Kanjee et al., 2012) that the alarmones can bind in place of GTP in the three GTPase categories: translation co-factors (EF-G, EF-Tu, IF2), cell-signalling and cell-division (Obg) and protein translocation.

Alarmones effect on ATP and GTP pool in *B. Subtilis* This effect is illustrated in Figure 30 and in more details in Figure 31 for GTP. The regulation of GTP pool via the

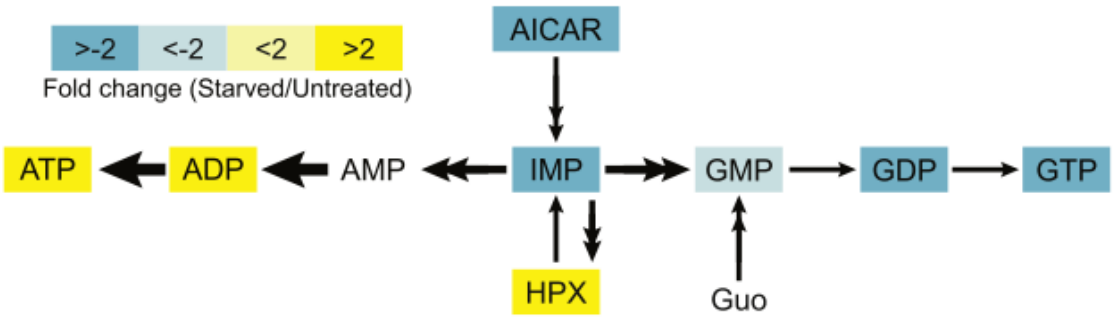

Figure 30: Bacillus Subtilis: regulation of ATP and GTP pool by the alarmones (Kriel et al., 2012)

alarmones is more importantly performed by Gmk and HprT than by GuaB (Kriel et al., 2012). The IC_{50} are: $\approx 20\mu M$ for Gmk, $\approx 11\mu M$ for HprT and $\approx 0.3 - 0.5mM$ for GuaB.

4.10 Volume growth

When? Where?

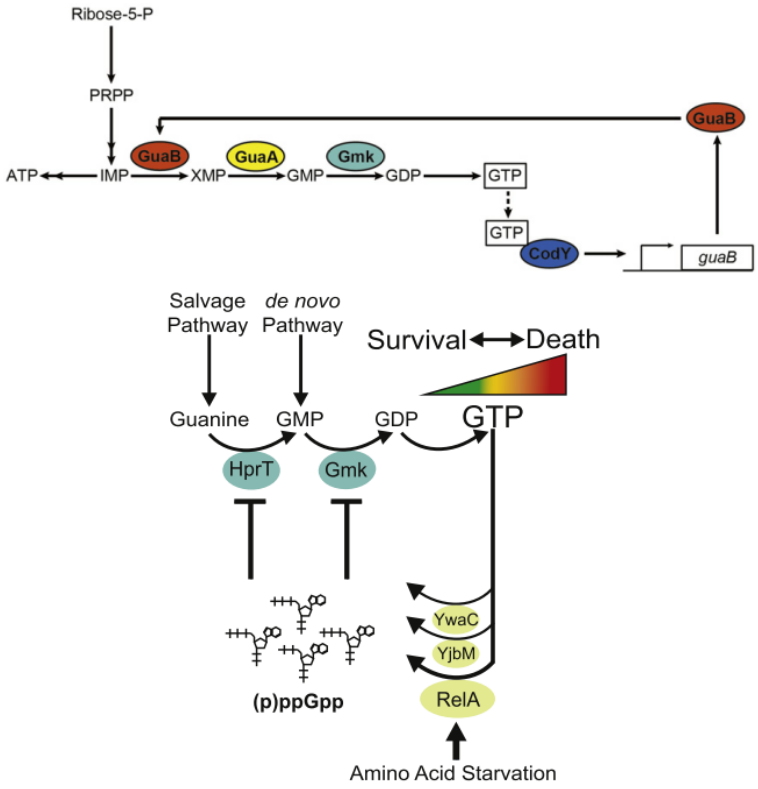

Figure 31: Bacillus Subtilis: GTP pathway regulation (Kriel et al., 2012)

4.11 Exchanges

4.11.1 Exchange with extracellular conditions

Not described but a matter of changing the pool (concentrations probably for extracellular conditions) in volume and extracellular conditions. Works in both ways if needed.

4.11.2 Exchange between volumes

Not described but a matter of changing the pool in volumes. Works in both ways if needed. Just remind that for 'complex' entities, it is easier if everything is contained in one entity only. I mean, if a mRNA binded with a 70S and an AA 1 charged tRNA moves, then if only one entity contains all these information, it is much easier. Here we would have chosen the mRNA.

4.12 Volume division

Need to obtain a storage for interface between volume. But otherwise, volume division is a matter of deleting a volume, creating 2 (3? 4?) volumes, adjusting the interfaces and pools. Needs to be a cell process since it can change the chromosome location.

4.13 Cytokinesis, cell fission?

Depending on how it is modeled. The part on geometry is 'simple' to model:

- volume division;

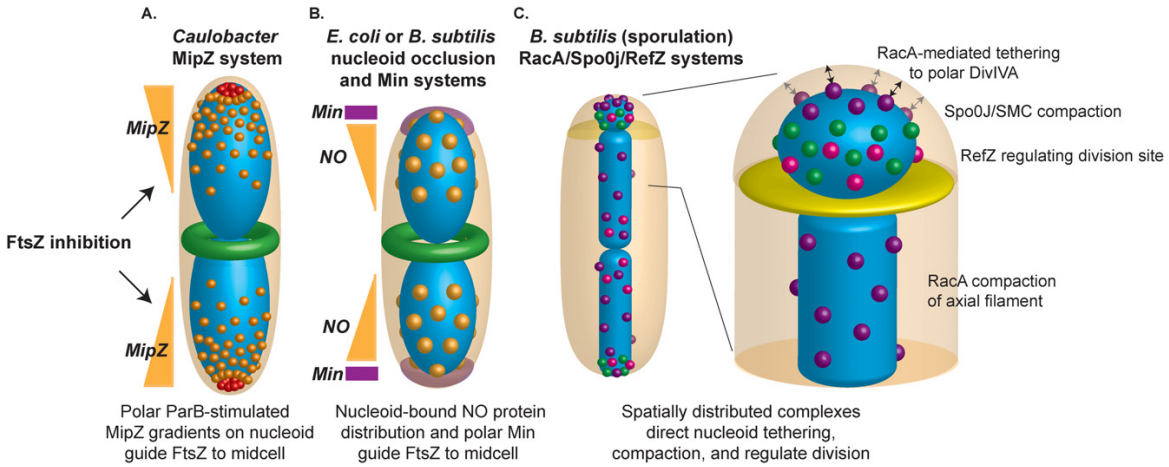

Figure 32: Location of the FtsZ ring could be mediated by Nucleoid Occlusion (NO) molecules. These proteins bind on the first two thirds of the chromosome (depleted in the *ter* region) and inhibit FtsZ polymerization. As long as segregation has not started, NO molecules inhibit FtsZ polymerization. Once chromosomes have migrated and the center of the cell only contains the unreplicated *ter* region, FtsZ polymerization and septum formation start. Figure from **?**

- changing the characteristics of the concerned volumes to adjust the surface of exchange for example.

Other part? Needs to be a cell process.

4.14 Processes from wholeCell

Police barrée = un process de wholeCell que j'ai remis en haut, sans pour autant l'avoir modéliser ou avoir les informations nécessaires dans les états.

- ~~Chromosome condensation: DNA clamping by SMC complexes~~
- ~~Chromosome segregation~~
- ~~Cytokinesis: pinching of the cell membrane~~
- ~~DNA damage: Gap site, Abasic site, Sugar-phosphate, Base, Intrastrand cross link, Strand break, Holliday junction~~
- ~~DNA repair~~
- ~~DNA supercoiling~~
- FtsZ polymerization
- Host interaction: des trucs mais en quoi c'est utile ?
- Macromolecular complexation: models the formation of all macromolecular complexes except the 30S and 50S ribosomal particles, the 70S ribosome, the FtsZ ring, and the oriC DnaA complex
- ~~Metabolism: models the import of extracellular nutrients and their conversion into macromolecule building blocks~~

- Protein activation: implements a Boolean model of their effects on the functional state enzymatically competent or incompetent of mature proteins
- ~~Protein decay: models the degradation of protein monomers, macromolecular complexes, cleaved signal sequences, and prematurely aborted polypeptides as well as the misfolding and refolding of protein monomers and complexes~~
- ~~Protein folding~~
- ~~Protein modification: models protein covalent modification including phosphorylation, lipoyl transfer, and -glutamate ligation~~
- ~~Protein processing I: models N-terminal formylmethionine deformylation and N-terminal methionine cleavage, the first steps in post-translational processing. What's that???~~
- ~~Protein processing II: models the third step of post-translational processing: lipoprotein diacylglyceryl adduction and lipoprotein and secreted protein signal peptide cleavage. What's that?~~
- ~~Protein translocation: models membrane and extracellular protein localization, the second step in post-translational processing~~
- ~~Replication~~
- ~~Replication initialization: determines when during the cell cycle chromosome duplication begins. Uses the protein DnaA (MG469)~~
- ~~Ribosome assembly: models the enzyme-catalyzed formation of 30S and 50S ribosomal particles~~
- ~~RNA decay: decays all species of RNA, and at all maturation states including aminoacylated states~~
- ~~RNA modification: the exact role of rRNA modification is unknown. This process models tRNA and rRNA modification~~
- ~~RNA processing: models operonic RNA cleavage into individual RNA gene products. Something about operons.~~
- Terminal organelle assembly: models the assembly of the protein content of the terminal organelle
- ~~Transcription: For simplicity, our model doesnt represent this phenomenon, allowing translation only of completed mRNAs~~
- Transcription regulation: models the binding of transcriptional regulators to promoters and the fold-change effect of transcriptional regulators on the affinity of RNA polymerase for individual promoters.
- ~~Translation~~
- ~~tRNA aminocylation~~

References

Qutaiba O. Ababneh and Jennifer K. Herman. RelA Inhibits Bacillus subtilis Motility and Chaining. *Journal of Bacteriology*, 197(1):128–137, January 2015. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.02063-14. URL http://jb.asm.org/content/197/1/128.

M. Anikin, V. Molodtsov, D. Temiakov, and W.T. McAllister. Transcript Slippage and Recoding. In J.F. Atkins and R.F. Gesteland, editors, *Recoding: Expansion of Decoding Rules Enriches Gene Expression*, number 24 in Nucleic Acids and Molecular Biology, pages 409–432. Springer New York, 2010. ISBN 978-0-387-89381-5 978-0-387-89382-2. URL http://link.springer.com/chapter/10.1007/978-0-387-89382-2_19.

Ceclia M. Arraiano, Jos M. Andrade, Susana Domingues, Ins B. Guinote, Michal Malecki, Rute G. Matos, Ricardo N. Moreira, Vnia Pobre, Filipa P. Reis, Margarida Saramago, Ins J. Silva, and Sandra C. Viegas. The critical role of RNA processing and degradation in the control of gene expression. *FEMS Microbiology Reviews*, 34(5):883–923, September 2010. ISSN 1574-6976. doi: 10.1111/j.1574-6976.2010.00242.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1574-6976.2010.00242.x/abstract.

Geoffrey S. Briggs, Wiep Klaas Smits, and Panos Soultanas. Chromosomal Replication Initiation Machinery of Low-G+C-Content Firmicutes. *Journal of Bacteriology*, 194(19): 5162–5170, January 2012. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.00865-12. URL http://jb.asm.org/content/194/19/5162.

Guangnan Chen and Charles Yanofsky. Tandem Transcription and Translation Regulatory Sensing of Uncharged Tryptophan tRNA. *Science*, 301(5630):211–213, November 2003. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1084902. URL http://www.sciencemag.org/content/301/5630/211.

V. Epshtein, D. Dutta, J. Wade, and E. Nudler. An allosteric mechanism of rho-dependent transcription termination. *Nature*, 463:245 – 249, 2010. ISSN 0028-0836. doi: http://dx.doi.org/10.1038/nature08669.

Tamas Gaal, Michael S. Bartlett, Wilma Ross, Charles L. Turnbough, and Richard L. Gourse. Transcription Regulation by Initiating NTP Concentration: rRNA Synthesis in Bacteria. *Science*, 278(5346):2092–2097, December 1997. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.278.5346.2092. URL http://www.sciencemag.org/content/278/5346/2092.

Anthony O. Gaca, Cristina Colomer-Winter, and Jos A. Lemos. Many means to a common end: the intricacies of (p)ppgpp metabolism and its control of bacterial homeostasis. *Journal of Bacteriology*, 2015. doi: 10.1128/JB.02577-14. URL http://jb.asm.org/content/early/2015/01/14/JB.02577-14.abstract.

S.R. Goldman, R.H. Ebright, and B.E. Nickels. Direct detection of abortive rna transcripts in vivo. *Science*, 324(5929):927–928, 2009. doi: 10.1126/science.1169237. URL http://www.sciencemag.org/content/324/5929/927.abstract.

Paul Gollnick, Paul Babitzke, Alfred Antson, and Charles Yanofsky. Complexity in Regulation of Tryptophan Biosynthesis in Bacillus subtilis. *Annual Review of Genetics*, 39(1):47–68, 2005. doi: 10.1146/annurev.genet.39.073003.093745. URL http://dx.doi.org/10.1146/annurev.genet.39.073003.093745.

I. Gusarov and E. Nudler. The mechanism of intrinsic transcription termination. *Molecular Cell*, 3(4):495 – 504, 1999. ISSN 1097-2765. doi: http://dx.doi.org/10.1016/S1097-2765(00)80477-3. URL http://www.sciencedirect.com/science/article/pii/S1097276500804773.

S.E. Halford. An end to 40 years of mistakes in dna-protein association kinetics? *Biochemical Society Transactions*, 37(2):343 – 348, 2009. doi: 10.1042/BST0370343. URL http://www.sciencemag.org/content/344/6187/1042.abstract.

K.M. Herbert, W.J. Greenleaf, and S.M. Block. Single-molecule studies of rna polymerase: Motoring along. *Annual review of biochemistry*, 77:149 – 176, 2008. doi: 10.1146/annurev.biochem.77.073106.100741. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2854675/.

Usheer Kanjee, Koji Ogata, and Walid A. Houry. Direct binding targets of the stringent response alarmone (p)ppGpp. *Molecular Microbiology*, 85(6):1029–1043, September 2012. ISSN 1365-2958. doi: 10.1111/j.1365-2958.2012.08177.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2012.08177.x/abstract.

Allison Kriel, Alycia N. Bittner, Sok Ho Kim, Kuanqing Liu, Ashley K. Tehranchi, Winnie Y. Zou, Samantha Rendon, Rui Chen, Benjamin P. Tu, and Jue D. Wang. Direct Regulation of GTP Homeostasis by (p)ppGpp: A Critical Component of Viability and Stress Resistance. *Molecular cell*, 48(2):231–241, October 2012. ISSN 1097-2765. doi: 10.1016/j.molcel.2012.08.009. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3483369/.

Libor Krsn and Richard L. Gourse. An alternative strategy for bacterial ribosome synthesis: Bacillus subtilis rRNA transcription regulation. *The EMBO Journal*, 23(22):4473–4483, November 2004. ISSN 0261-4189, 1460-2075. doi: 10.1038/sj.emboj.7600423. URL http://emboj.embopress.org/content/23/22/4473.

M.H. Larson, R.A. Mooney, J.M. Peters, T. Windgassen, D. Nayak, C.A. Gross, S.M. Block, W.J. Greenleaf, R. Landick, and J.S. Weissman. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, 344(6187):1042–1047, 2014. doi: 10.1126/science.1251871. URL http://www.sciencemag.org/content/344/6187/1042.abstract.

Alan C. Leonard and Julia E. Grimwade. Regulation of DnaA Assembly and Activity: Taking Directions from the Genome. *Annual Review of Microbiology*, 65(1):19–35, 2011. doi: 10.1146/annurev-micro-090110-102934. URL http://dx.doi.org/10.1146/annurev-micro-090110-102934.

P. S. Lovett and E. J. Rogers. Ribosome regulation by the nascent peptide. *Microbiological Reviews*, 60(2):366–385, January 1996. ISSN 1092-2172, 1098-5557. URL http://mmbr.asm.org/content/60/2/366.

Holger Ludwig, Georg Homuth, Matthias Schmalisch, Frank M. Dyka, Michael Hecker, and Jrg Stlke. Transcription of glycolytic genes and operons in Bacillus subtilis: evidence for the presence of multiple levels of control of the gapA operon. *Molecular Microbiology*, 41(2):409–422, July 2001. ISSN 1365-2958. doi: 10.1046/j.1365-2958.2001.02523.x. URL http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2958.2001.02523.x/abstract.

Brian J. Paul, Melanie M. Barker, Wilma Ross, David A. Schneider, Cathy Webb, John W. Foster, and Richard L. Gourse. DksA: a critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell*, 118(3):311–322, August 2004. ISSN 0092-8674. doi: 10.1016/j.cell.2004.07.009.

F. Rojo, B. Nuez, M. Menca, and M. Salas. The main early and late promoters of Bacillus subtilis phage phi 29 form unstable open complexes with sigma A-RNA polymerase that are stabilized by DNA supercoiling. *Nucleic Acids Research*, 21(4):935–940, February 1993. ISSN 0305-1048.

R.M. Saecker, M.T. Record Jr., and deHaseth P.L. Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *Journal of Molecular Biology*, 412(5):754 – 771, 2011. ISSN 0022-2836. doi: http://dx.doi.org/10.1016/j.jmb.2011.01.018. URL `http://www.sciencedirect.com/science/article/pii/S0022283611000350`.

Jr. C.L. Turnbough. Regulation of gene expression by reiterative transcription. *Current Opinion in Microbiology*, 14(2):142–147, April 2011. ISSN 1369-5274. doi: 10.1016/j.mib.2011.01.012. URL `http://www.sciencedirect.com/science/article/pii/S1369527411000245`.

P.H. von Hippel. An integrated model of the transcription complex in elongation, termination, and editing. *Science*, 281(5377):660 – 665, 1998. doi: http://dx.doi.org/10.1126/science.281.5377.660. URL `http://www.sciencemag.org/content/281/5377/660.full.pdf?sid=f01d12ba-26d1-4ad4-96ff-1ce63ff5c415`.

F. Wang and E.C. Greene. Single-molecule studies of transcription: From one RNA polymerase at a time to the gene expression profile of a cell. *Journal of Molecular Biology*, 412(5):814 – 831, 2011. ISSN 0022-2836. doi: http://dx.doi.org/10.1016/j.jmb.2011.01.024. URL `http://www.sciencedirect.com/science/article/pii/S0022283611000532`.

Wade C. Winkler and Ronald R. Breaker. Regulation of Bacterial Gene Expression by Riboswitches. *Annual Review of Microbiology*, 59(1):487–517, 2005. doi: 10.1146/annurev.micro.59.030804.121336. URL `http://dx.doi.org/10.1146/annurev.micro.59.030804.121336`.

Christiane Wolz, Tobias Geiger, and Christiane Goerke. The synthesis and function of the alarmone (p)ppGpp in firmicutes. *International Journal of Medical Microbiology*, 300(23):142–147, February 2010. ISSN 1438-4221. doi: 10.1016/j.ijmm.2009.08.017. URL `http://www.sciencedirect.com/science/article/pii/S1438422109001106`.