UNIVERSITÀ DEGLI STUDI DI PISA

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

CORSO DI LAUREA MAGISTRALE IN BIOLOGIA
MOLECOLARE E CELLULARE

Quasi-Cellular Systems: Stochastic Simulation Analysis at Nanoscale Range

Anno accademico 2011/2012

RELATORI:

PROF. ROBERTO MARANGONI
DR. PASQUALE STANO

CANDIDATO:

LORENZO CALVIELLO

Per Aspera Ad Astra...

CONTENTS

ABSTRACT

The artificial creation of the simplest forms of life (minimal cells) is a challenging aspect in modern synthetic biology. Quasi-cellular systems able to produce proteins directly from DNA can be created by encapsulating a *cell-free* transcription/translation system (PURESYSTEM™) in liposomes (10^{-5} – 10^{-7} m diameter). It is possible to detect the overall protein production inside these compartments using DNA encoding for GFP and monitoring the fluorescence emission over time.

The entrapment of solutes in lipid compartments is a complex process that creates a population of vesicles with different internal compositions of molecular species, which affects the final protein production. A complete understanding of the distribution of solutes inside the different compartments and on its effect on the course of internal reactions are two relevant and still open issues in the field.

Stochastic simulation is a valuable tool in the study of biochemical reaction at nanoscale range; QDC (Quick Direct-Method Controlled), a stochastic simulation software based on the well-known Gillespie's SSA algorithm, was used.

A translation model of the PURESYSTEM™ previously built in our laboratory was improved to describe in detail a coupled transcription/translation system with simultaneous elongation events on the same molecule. The dynamical coupling between the transcription and translation systems was assessed using logical formulations allowed in QDC's syntax, thus creating sequentially dependent processes in the concurrent-only environment of Gillespie's algorithm. Stochastic simulations were performed in order to globally fit, by sigmoid curves, the entire experimental dataset for protein production, with the aim to describe how the different composition of species affects the overall translation process.

To the best of our knowledge, the present work is the first one describing in detail the stochastic behavior of the PURESYSTEM™. Thanks to our results, an experimental approach is now possible, aimed at recording the GFP production kinetics in very small compartment, and inferring, by using the simulation as a hypotheses test benchmark, the internal solutes distribution, and shed light on the still unknown forces driving the entrapment phenomenon.

INTRODUCTION

General theories for the origin of Life

Life is without doubt the most tangible example of complexity: every day we unravel new different kinds of elegant regulation of biological phenomena as modern technology advances at exponential rate. Despite this rapidly evolving scenario, there are several engaging challenges that modern biology is nowadays facing. The topic regarding the Origin of Life represents one of the most interesting field of biological research. Two fundamental hypotheses regarding the origin of life in Earth were proposed: the Abiogenesis hypothesis, which describes the formation of living matter from a molecular evolution of inorganic compounds (“α – βἰος” = non - life”); the Panspermia hypothesis, stating that life itself, or its primary precursors, is present throughout the Universe (“πᾶν” = everything, “σπἐρμα” = seed), and it is carried to different planets by space vectors as meteorites, asteroids and so on.

Alexander Oparin, a Russian biochemist who first developed an abiogenetic theory, stated in 1922: “*There is no fundamental difference between a living organism and lifeless matter. The complex combination of manifestations and properties so characteristic of life must have arisen in the process of the evolution of matter.*”

Later, in 1952, Stanley Miller’s experiment tested Oparin’s theory of chemical origin of life, reproducing in lab the hypothetical conditions thought at the time to be present on early Earth, and observing the production of amino acids from simple inorganic chemicals: water, methane, ammonia and hydrogen [4]; this experiment proved that environmental condition of ancient Earth (3.5 billion years ago) allowed the formation of most of the molecular precursors of life.

The Panspermia theory, from the other hand, states that life came to planet Earth from outer space; anyway, modern theories in this field comprise different formulations of this general concept [5]: the “Strong Panspermia” theory simply removes the problem of the origin of life from our planet to some other unknown

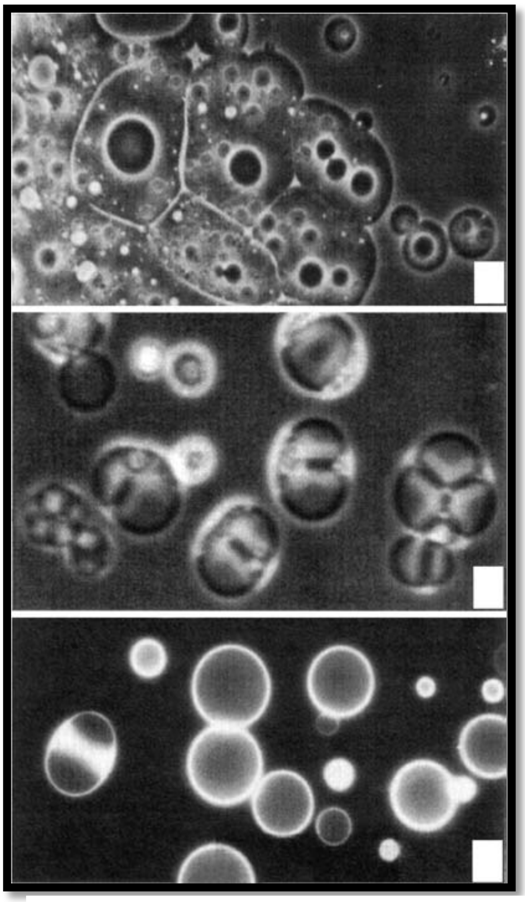


Fig. 1: Membrane formation by compounds from Murchison meteorite[1]: the vesicular nature of such structures is clearly visible

place, drawing from the idea of Exogenesis (the theory that suggests life formation outside the planet Earth), which is scientifically very limited and gives us no answers about the chemical formation of life, representing a more suitable idea for a sci-fi movie…

"Pseudo-panspermia" hypothesis (also called "Weak Panspermia") deals instead with the delivery of complex organic compounds from space. This notion has become widely accepted, as it takes into account the Earth environmental conditions in its first billion year of existence (the Haedean and the first part of the Archaean aeons), which was a period of massive meteoritic impacts on the planetary surface due to the absence of a shielding ozone layer. Thus, different molecules were brought from space to our planet, and probably also different prebiotic compounds. Data supporting this theory comes from studies of the well-known Murchison meteorite, which is proven to harbor several organic compounds, such as aminoacids and even nucleobases (xanthine and uracil) [6]; moreover, components extracted from the meteorite can form vesicular structures [1], as shown in **Fig. 1**, strengthening the "Weak Panspermia" hypothesis for the extraterrestrial origin of prebiotic structures. However, both the compositions of the mixture extracted from the Murchison meteorite and that of Miller's experiment comprise many of the biochemical "bricks" that form living matter; thus, the origin of Life problem seems to transcend the different explanations about the exact spatial origin of the single chemical species, but invites us to focus the attention on their organization and on the different properties that emerge from their complex interactions.

The Minimal Cell

Autopoiesis

The real challenge in the study of the origin of Life aims to understand the primary, simple mechanisms underlying the emergence of the minimal life form, starting from few, simple chemical components, regardless of their origin. Speaking of primordial life form we clearly refer to the fundamental unit of all known organisms, the cell. A cell is an autopoietic (i.e. self-producing) enclosed system, that means it is capable of generating its own components via a network process that is internal to its boundary [7]. Equally important, but more subtle, is the definition of “living cell”. Autopoiesis is the primal property exhibited by living organisms, but we need to define other attributes to draw a line between the “living” and the “non-living”. To be considered “alive”, a cell must exert three fundamental properties: self-maintenance, self-reproduction and evolvability [8]. The first two are summarized by the definition of autopoiesis (which also includes the concept of physiological homeostasis), and strictly depends on the composition of the single cells, while Darwinian evolution is a property observable when taking into account a population over time. The study of the evolution of a population of different protocells is an argument of outstanding interest, and the efforts of future research in synthetic biology will be probably addressed towards this topic.

One of the major goals of scientific research is the synthesis of artificial life in the lab, trying to put into practice all the aforementioned theoretical concepts; the first critical step in this *synthetic biology* approach is the definition of the minimal cell and its components.

Self-organization

A general definition of “minimal cell” is not a quite simple task: even the simplest known living organism presents an incredible level of complexity, encompassing hundreds of genes and proteins; during billion years of evolution a series of redundancies and metabolic loops arose, continuously adding chemical complexity and fine molecular regulations for disparate functions, from signal transduction to DNA replication; therefore, it is important to consider the nature of the molecular ensemble a candidate living cell is comprising.

The condition of autopoietic enclosed system implies the presence of a physical boundary that confines the network of process that permits the self-maintenance of the system. In living cells this boundary is a lipid bilayer that acts as a semi-permeable membrane, allowing the uptake of some substances but also acting as an impenetrable barrier for other compounds; the lipid molecules composing the membrane are amphipathic molecules, such as POPC (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphatidylcholine), formed by an hydrophilic head and long aliphatic tails; thanks to this peculiar structural dichotomy, lipid molecules can self-organize in different stable molecular structures according to the chemical environment; despite the local increase in order, the overall formation process of these aggregates remains thermodynamically favorable: the lipid molecules organize themselves in closed structures that negate disadvantageous interactions between water molecules and the long aliphatic tail of fatty acids, thus maximizing the entropy increase for solvent molecules; moreover, this process is also auto-catalytic for some kind of aggregates: the organized structures speed up their own self-assembly once a membrane 'seed' has formed, resembling the phenomenon of nucleation in crystals [9]. Some examples of stable structures that lipid molecule can spontaneously form are micelles, liposomes, reverse micelles and monolayers; liposomes (or, in general, vesicles) are the most studied lipid aggregates, and take the shape of a spherical compartment with an internal cavity defined by a lipid bilayer membrane.

Liposomes are prepared mainly with two procedures: by pouring a lipid-in-ethanol solution into an aqueous medium (injection method) or by rehydrating a previously dried lipid film (film rehydration).

By changing experimental parameters (concentrations of species, pH etc...) it is possible to obtain liposomes with different size and shape, spanning from 20 nm to 2 μm of diameter. Thus, the range of internal volumes available goes from attoliters (10^{-18}) to picoliters (10^{-12}); it is also possible to obtain giant vesicles up to 100 μm of diameter using other techniques, such as electroformation [10].

Giant vesicles are often used in different areas of supramolecular chemistry as biomimetic models for the study of mechanical properties of lipid membranes.

Thanks to their significant size, they allow a direct visualization of the reactions and transformation phenomena they carry inside by using common optical microscopy. Giant Unilamellar Vesicles (GUVs) have been extensively used also in the field of artificial life synthesis; however, GUVs do not form spontaneously, and the available preparation methods are generally troublesome [11]; indeed, GUVs represent an excellent model for studying cell-scale phenomena, but smaller systems, like liposomes, can spontaneously form without any need of external treatment.

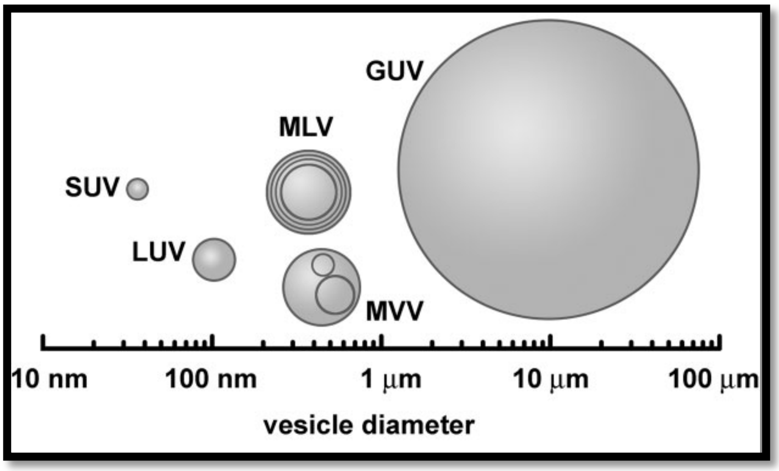


Fig. 2: Different types of vesicles classified according to lamellarity and size[3]: Small unilamellar vesicles (SUVs), large unilamellar vesicles (LUVs), multilamellar vesicles (MLVs), multivesicular vesicles (MVVs), and giant unilamellar vesicles (GUVs).

Liposomes self-assembly is a fundamental property for the creation of artificial candidates for minimal synthetic life, as it is represents a completely spontaneous self-organization process, arising from few interactions between simple molecules.

During their preparation, vesicles can comprise one or even more concentrically nested lipid bilayers, being called uni- or multi-lamellar vesicles (**Fig. 2**); liposome size and lamellarity can be controlled in two ways:

1) using a polycarbonate membrane and forcing liposomes extrusion through fixed size nanopores or by sonication;
2) acting on the chemical conditions (lipid concentration, buffer, salts, pH, etc...).

The extrusion methods allows us to obtain a narrowly distributed population of unilamellar vesicles, but the obtained liposomes are not reflecting the results of the completely spontaneous self-organizations process of liposomes formation, probably biasing the outcoming observations.

Considering the philosophical strategy in the search for the spontaneous mechanisms for the emergence of minimal cells, the second method is then preferred.

During their formation process, liposomes encapsulate the different molecules present in solution; moreover, the filled vesicles can fuse with cell membranes and release their inner content into the cytoplasm. Thanks to these useful properties, liposomes are massively used in wide areas of biotechnology and nanomedicine [12], as they represent excellent carriers for the most diverse molecules, from DNA to small chemicals.

Besides their pharmaceutical applications, vesicles are valid biomimetic systems; liposomes are not living organisms, but the auto-organizing behavior of such structures is indeed a fundamental property in defining the mechanism of formation of the first precursors of living cells. In the search for the minimal synthetic living organism, liposomes technology can be applied to enclose the minimal biochemical machinery sufficient to assess self-maintenance and replication.

In fact, by encapsulating enzymes and reagent molecules it is possible to carry biochemical reactions inside lipid compartments

A Semi-Synthetic Approach

Creating *de novo* (from scratch) an artificial living systems using simple biomolecules is a challenging goal; the self-organizing behavior of lipid molecules represents indeed a great advancement in the field, but it is not sufficient to define an “alive” system able to exhibit the three fundamental properties already mentioned (self-maintenance, self-reproduction and evolvability); unfortunately, up to date, it has been impossible to create a biological systems able to exhibit full living properties.

It is possible to attempt the creation of minimal cells with different approaches; a *top-down* philosophy aims at defining the minimal set of molecular species starting from known complex living organism. Many studies have investigated the minimal genetic information required for sustaining life [13, 14] , or the minimal size that the system must reach to enclose all the molecular species, which is strongly related to the amount of DNA content [15]. All these attempts greatly improved our knowledge about the essential "ingredients" a system requires to be alive; but even comprising a minimal genome, the biological complexity of these organism is still high, they enclose different hundreds of molecular species, resulting in a practical unfeasibility for a wet-lab approach.

Inversely, the *bottom-up* approach starts by describing simple systems as they increase in complexity, by adding new species and creating new interactions, with the aim to observe how emerging properties arise from the few elementary interactions of the whole system; anyway, different possible pathways can bring inorganic matter into simple organic molecules, and different pathways can bring simple organic molecules into different unknown prebiotic biochemical catalysts.

One of the major assumptions regarding the presence of prebiotic catalytic molecules is given by the "RNA world" theory: some models describes as it is possible to achieve cellular life using very few RNA molecules as catalytic agents (ribozymes) involved in elementary reactions for continuous self-renewal of membrane and ribozymes themselves [16, 17].

These very interesting attempts, together with other studies involving self-assembling biological devices [18], absolutely deserve attention, as they show how protocellular systems could comprise very few molecular species thus moving closer to the definition of minimal living organism.

Unfortunately, these approaches are often strictly theoretical, due to the fact that such species, as they are described (i.e. lipid-synthesizing ribozymes), do not exist in nature, and so they are of little use in the experimental realm of life synthesis in laboratory.

Trying to overcome the disadvantages of both the *top-down* and *bottom-up* philosophies, a mixed strategy has been proposed to create the possibility to master the problem of the artificial life synthesis from an experimental point of view. This strategy aims at targeting extant molecular species which can carry out determined cellular functions, and trying to carry these reactions "*in lipo*" (inside a lipid compartment), with the purpose to obtain an adequately complex, tangible biological entity (a "quasi-cellular" system) able to exhibit interesting properties, having incorporated a network of biochemical reactions involving the real molecular agents therein.

This approach is called *semi-synthetic* because it makes use of liposome technology for the synthesis of artificial systems not present in nature, but uses molecular constituents isolated from extant living organism, which are not artificial molecules but represent the outcome of billions of years of evolution. In the past two decades, a lot of experimental work following this *semi-synthetic* approach produced a significant advancement in the study of biochemical reactions inside the confined medium of lipid systems.

Initially the attention was drawn towards the production of nucleic acids in liposomes: the polymerization of short RNA and DNA sequences and even the Polymerase Chain Reaction (PCR) were carried inside vesicles[19-21]. In a very recent work, compartmentalized DNA amplification molecules via PCR was chemically linked with the self-reproduction of the vesicle itself, showing a chemical synergy between the two processes [22, 23]. This work represents indeed an important step in the construction of a minimal artificial cell, even if sharing the same fundamental problem with the previous given examples: the enzymes which catalyze the different reactions (i.e. polymerases) are not regenerated *in situ*, and after some division cycles many of the newly-formed vesicles lack the biochemical machinery they need to assess their physiological functions (the so-called "death by dilution" effect). To assess a complete ("core-and-shell") reproduction of the entire system all the components must be regenerated from within, implying the presence of a biochemical apparatus able to regenerate both internal and membrane molecules. The scientific attention thus shifted to accomplish protein production inside vesicles from DNA/RNA sequences.

The evolutionary process greatly awarded the use of nucleic acid sequences to store genetic information decipherable with the use of a reading code, ultimately resulting in protein production from nucleic acids molecules with a process called *translation*.

The discussion about the origin of the genetic code and its relevance for the emergence of the first proto-cells is not simple and goes outside the purpose of this thesis, but it is important, again, to clarify the general strategy: artificial systems are created by encapsulating known biological processes (as translation) in vesicles, to investigate the behavior of such sufficiently complex networks in a compartmentalized environment, trying to infer the link between such processes and

vesicle behavior, in terms of biochemical functionality and possible evolution dynamics.
It is possible to carry the translation process inside lipid compartment by using cell-free systems constructed with cell lysates from different sources, as wheat germ, rabbit reticulocytes or *E. coli* [24]. However, these systems lack a full control of the translation reaction: only a minority of the molecular components present in the cellular extracts participates in the translation process, and many species (i.e. proteases, nucleases) greatly affect the final protein production, acting with protein modification/degradation reactions and on the overall energy availability.
Considering the philosophy of the *semi-synthetic* approach, it is fundamental to use a completely controllable system with a low level of complexity, where the single components are known and can be easily manipulated; due to this necessity and to the incrementing use of such *in vitro* systems, a novel *cell-free* translation system was created in 2001 that found great applications in biotechnology and synthetic biology studies[25] [26].

The PURESYSTEM™

“Protein synthesis using recombinant elements” (PURE) system is the name of the cell-free translation system created in 2001 by Ueda and collaborators [28]; they individually overexpressed in *E. coli* all molecular species involved in the prokaryotic translation process, adding a His-tag to each protein for easy purification. The total ensemble contains (including tRNAs) 83 species (Table 1) representing the minimal collection of components able to afford protein production from a DNA sequence. The translation process, intended as the physical movement of the ribosome on the RNA molecule while incorporating aminoacids in the elongating peptide, involves the use of the translation factors and the ribosome. With the aim to assist and improve the protein production process three additional processes where included:

Table 1: The PURESYSTEM, from [27]

Translation factors
2.7 µM IF1
0.40 µM IF2
1.5 µM IF3
0.26 µM EF-G
0.92 µM EF-Tu
0.66 µM EF-Ts
0.25 µM RF1
0.24 µM RF2
0.17 µM RF3
0.50 µM RRF
Aminoacyl-tRNA synthetases
1900 U/ml AlaRS
2500 U/ml ArgRS
20 mg/ml AsnRS
2500 U/ml AspRS
630 U/ml CysRS
1300 U/ml GlnRS
1900 U/ml GluRS
5000 U/ml GlyRS
630 U/ml HisRS
2500 U/ml IleRS
3800 U/ml LeuRS
3800 U/ml LysRS
6300 U/ml MetRS
1300 U/ml PheRS
1300 U/ml ProRS
1900 U/ml SerRS
1300 U/ml ThrRS
630 U/ml TrpRS
630 U/ml TyrRS
3100 U/ml ValRS
Other enzymes
4500 U/ml MTF
1.2 µM ribosomes
4.0 µg/ml creatine kinase
3.0 µg/ml myokinase
1.1 µg/ml nucleoside-diphosphate kinase
2.0 units/ml pyrophosphatase
10 µg/ml T7 RNA polymerase
Energy sources
2 mM ATP, GTP
1 mM CTP, UTP
20 mM creatine phosphate
Buffers
50 mM Hepes–KOH, pH 7.6
100 mM potassium glutamate
13 mM magnesium acetate
2 mM spermidine
1 mM DTT
Other components
0.3 mM 20 amino acids
10 mg/ml 10-formyl-5,6,7,8-tetrahydrofolic acid
56 A260/ml tRNAmix (Roche)

1) Transcription: with the addition of the T7 RNA polymerase it is possible to accomplish protein production directly from the corresponding cDNA sequence.

2) Aminoacylation of tRNAs: which allows the elongation of the nascent peptide directly with the added aminoacids, incorporating the native reactions for tRNA charging and ARSs aminoacylation.

3) Energy recycling: myokinase, nucleoside-diphosphate kinase, creatine kinase and creatine-phosphate were added to the system in order to regenerate the tri-phosphate nucleotides (ATP, GTP) during the translation process.

The procedure for protein production is very simple: a plasmid encoding for the desired protein must include a T7 promoter, a Shine-Dalgarno sequence upstream the ORF and a T7 terminator sequence downstream the stop codon. It is also possible to include chaperones or other agents to ensure the correct folding of the protein product [27]; different functional proteins were successfully produced *in vitro* using the PURESYSTEM™, starting from their respective cDNA sequences.

By using the PURESYSTEM™ it is possible to accomplish protein production with a totally defined set of molecular reagents, minimized to produce proteins with a minimal set of molecular factors. This feature is really appreciated considering the philosophy of the *semi-synthetic* approach and the general tendency to achieve complete biological functions with a minimal molecular ensemble; accordingly to this observation, the PURESYSTEM™ was successfully encapsulated in vesicles [29, 30].

The use of a DNA sequence encoding for GFP (Green Fluorescent Protein) together with the transcription/translation system allows to detect and monitor the protein production over time. Based on the fluorescent emission, quantification the protein production of the encapsulated PURE system is calculable; the total protein yield in vesicles is relatively high [29, 31] considering that only a fraction of liposomes has entrapped all the molecular species the transcription/translation system comprises.

Using the PURE system it is possible to produce functional enzymes, as reported in different studies aiming at synthesizing candidate minimal protocells [32, 33].

Of particular relevance is a paper [34] which describes the *in lipo* production of two membrane proteins involved in the lipid biosynthesis pathway: *sn*-glycerol-3-phosphate acyltransferase (GPAT) and lysophosphatidic acid acyltransferase (LPAAT). The coupled activity of this two enzymes resulted in the production (from soluble precursors) of phosphatidic acid, an amphipathic molecule which becomes part of the liposome membrane. For the first time vesicle membrane components were formed using catalysts produced from the liposome itself, revealing again the importance of using a defined, full-controllable protein synthesis device in the synthesis of quasi-cellular systems.

The possibility to obtain liposomes capable of genetic expression represented a critical step towards the synthesis of the first minimal cell: it is crucial to point out that RNA and protein metabolisms represent the major part of the essential process for bacterial life, constituting more than 50% of the essential genes in a minimal genome [8, 35]. Consequently, protein-producing liposomes do represent a good model for a full viable cell, thus opening the way for considerations about the generic features of minimal cells.

A recent work [31] used the PURESYSTEM to obtain GFP-expressing POPC liposomes as cell models, with the purpose of enlightening the discussion about the minimal size of biological entities, to compare the observed data with other predictions coming from different approaches to the definition of minimal-sized organisms [15]. A remarkable, unusual observation came out considering the size of such protein-synthesizing vesicles: GFP fluorescence was measured in separated bulk solution containing liposomes with different sizes; unexpectedly, liposomes with a radius of 100 nm were able to produce functional GFP protein.

POPC vesicles cannot fuse between each other and their lipid membrane is impermeable to nucleic acids, proteins, and small charged particles: this means that all the over 80 molecules species of the PURESYSTEM have been successfully entrapped in such tiny volume (<10^{-19} liters).

Classical statistics for encapsulation (also known as "*entrapment*") events gives zero or negligible probability for the co-encapsulation of the entire molecular ensemble in such small volume, showing how the entrapment process exhibits unexpected interesting properties in the nanoscale range.

The "Entrapment Conundrum"

The observation of GFP-expressing liposomes with 100 nm of diameter is in strong contrast with the hypotheses that defines the encapsulation of molecules as a purely random phenomenon. The solute distribution of water-soluble molecules in micrometric lipid vesicle has previously been assumed to follow as a Poisson distribution [30, 36-38], which describes the entrapment event as a discrete stochastic variable, in a pure probabilistic fashion.

Entrapment as a random event (adapted from Souza, 2009)

The average number *a* of internalized molecules is calculated by the *a priori* probability of entrapment, which is given by the concentration of the solution multiplied by the internal volume (vesicles are supposed to be spherical):

Eq. 1:

$$\boldsymbol{a = C_a \times \frac{4}{3}\pi\left(\frac{d}{2} - \rho\right)^3}$$

where C_a is the concentration of the specimen is solution, *d* is the vesicle diameter and ρ is the bilayer thickness (3.8 nm for POPC vesicles). According to the Poisson distribution, the probability to find a vesicle with *n* molecules of a certain specimen is given by the formula:

Eq. 2:

$$\boldsymbol{P(\alpha, n) = e^{-a}\frac{a^n}{n!}}$$

Using the Eq. 2 we can easily estimate the probability to find a vesicles which has enclosed at least one molecule, by calculating the difference between 1 (total probability) and the probability to find an "empty" vesicle (if $n = 0 \rightarrow \frac{\mathrm{a}^0}{0!} = 1$):

Eq. 3:

$$\boldsymbol{P(\alpha, n \geq 1) = \ 1 - e^{-a}}$$

Taking into account the presence of co-entrapment events it is possible to use Eq. 3 to define the probability to find a vesicle which has enclosed at least one molecule of each of the k species of the PURESYSTEM (N = 83); the entrapment events are assumed to be independent, so the equation takes the form of a product of sequences:

Eq. 4:

$$\boldsymbol{P(\alpha_k, n_k \geq 1) = \prod_{k=1}^{N} (1 - e^{a_k})}$$

As reported in **Eq. 1** the *a priori* average number a of entrapped molecule is calculated by taking into account the internal volume and the concentration of the species.
Then Eq. 4 can be transformed by writing a_k explicitly:

Eq. 5:

$$\boldsymbol{P(\alpha_k, n_k \geq 1) = \prod_{k=1}^{N} \left(1 - e^{-\frac{4\pi C_k}{3}\left(\frac{d}{2} - \rho\right)^3}\right)}$$

It is clearly visible that such probability heavily depends on the vesicle diameter d and on the concentrations of the different species (C_k); considering that all the PURESYSTEM macromolecular species are present in solution (Table 1) with a concentration <10 μM, the probability to find a 100 nm (inner diameter) liposome capable to afford protein production is absolutely negligible (**Fig. 3**). Even considering a possible aggregation of the PURESYSTEM species into macromolecular clusters; the probability to find a 100 nm filled vesicle reaches 1% only by assuming an increase of concentration of all species of a factor 20, which is unquestionably hard to justify.

This great deviation from the expected Poisson distribution fiercely questions the assumption of random and independent entrapment events, revealing the presence of still unknown phenomena which drive the encapsulation of biomolecules at nanoscale range. As noted above, fluorescence signal was measured in bulk solution, representing the total fluorescence emitted by the totality of GFP-expressing liposomes.

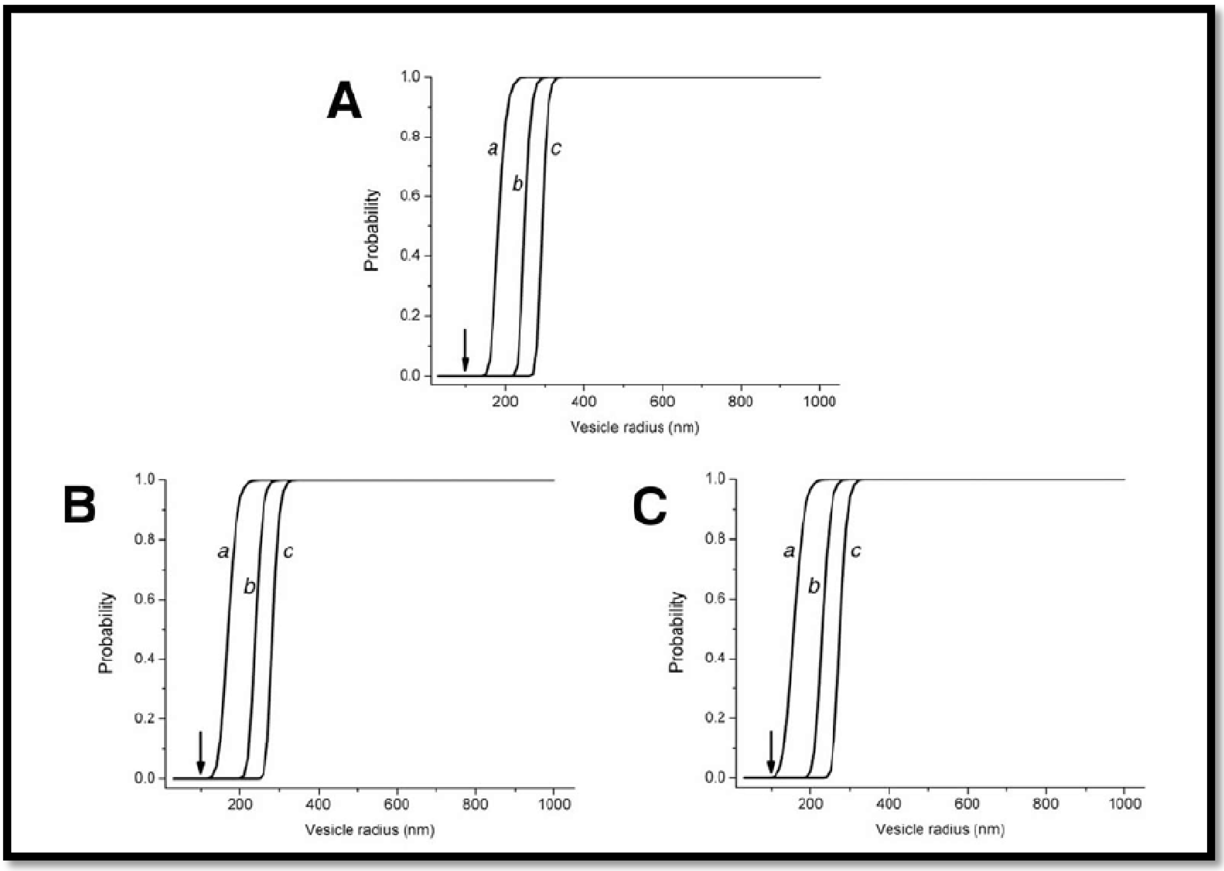


Fig. 3: Probabilities of co-encapsulation of the PURESYSTEM species vs. vesicle radius; the probability for a 100 nm liposome (black arrow) is vanishingly small, even if considering the PURESYSTEM species as organized into 50(A), 20(B) or 10(C) macromolecular clusters; the three curves a,b,c represents the chances to entrap 1, 2 or 3 copies of each species. From Souza, 2009[31]

Unfortunately, it is impossible to quantify the emission signal from single liposomes due to experimental limitations at such nanoscale range: flow cytometry, for example, has a limit of detection around 500 nm, making it impossible the detection of GFP production in 100 nm vesicles.

Every attempt using classical optical microscopy cannot be used to investigate the behavior of such small liposomes, because the resolution limit of a light microscope using visible light is about 300 nm, being unfeasible even to distinguish two different 100 nm-radius liposomes in the medium.

Trying to better clarify this curious phenomenon, a different approach was used to allow a direct visualization of nanoscale liposomes with their inner content: Cryo-Electron-Microscopy (Cryo-EM). It uses a beam of electrons coming form an excited source (generally a tungsten filament) to illuminate a thin target sample at cryogenic temperature.

Using electron microscopy the limit of detection is highly extended: electrons have wavelengths about 100,000 times shorter than photons (visible light has a wavelength from 740 to 380 nm), making it possible to achieve a resolution limit even below the nanometer range.

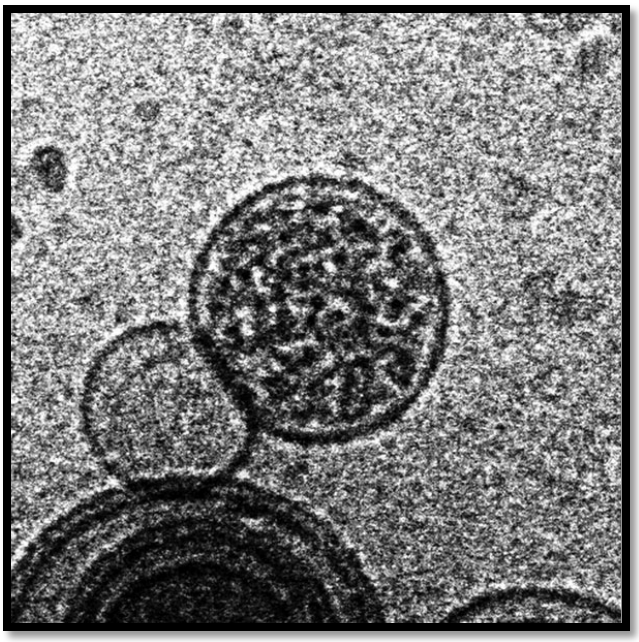


Fig. 4: Cryo-TEM image of liposomes with their encapsulated content, from Souza 2011[2]

Another great advantage of Cryo-EM is that there is no need to embed the sample in particular resins which can alter its macromolecular structure; the sample has to be frozen solid, in general by using liquid nitrogen (-191 C°). The immersion in a such low temperature solution causes the water to turn instantaneously in a vitreous state, avoiding the formation of ice crystals which can destroy the sample. Negative staining solution as uranyl acetate are often used in the sample preparation for transmission cryo-electron microscopy (Cryo-TEM) experiments, since molecules as uranium or lead have excellent scattering properties, ultimately producing clearer images with high contrast when visualized at the microscope. Electron microscopy was used to visualize and count the number of entrapped molecules inside liposomes, trying to understand if there is some general process of concentration enrichment at small volume level.

By using big macromolecular species or molecules rich of heavy-metal ions enclosed in a liposome, it is possible to directly count the entrapped molecules skipping the staining process (**Fig. 4**), that could cause some problems by destabilizing liposomes structure or biasing the resulting images. As "reporter" molecule, ferritin was often used to directly count entrapped molecules in liposomes [39], due to its iron phosphate core composed by circa 4500 iron atoms.

Recently, two papers used ferritin [40] or ribosomes [2] as entrapped species to estimate the distribution of solutes inside nanoscale POPC liposomes, and they found in both cases a spontaneous regular deviation from the Poisson model, showing results much far from the assumption of randomness in entrapment process.

Power-law distribution of entrapped solutes

Results from the paper involving ferritin-containing liposomes [40] showed that the solute distribution in nanoscale liposomes follows the same behavior, even changing the initial solute distribution or the liposomes preparation method: the vast majority of liposomes is nearly as empty, with no entrapped molecules, while a small percentage of the vesicle has enclosed an unexpectedly high number of solutes: the occupancy distribution of the internalized solutes strictly follows the Zipf-Mandlebrot law, that is a *power-law* distribution:

Eq. 6:

$$\boldsymbol{f(N) = A(N+1)^{-q}}$$

As shown in Eq. 6 the probability to find a vesicle with *N* entrapped particles is equal to the inverse of its power, with q<0 as the exponent of the power law distribution (*A* is a normalization constant); this means that it is very likely to find a vesicle with zero or few *N* entrapped molecule, but the probability to find a vesicle with a high *N* number of particles is still sufficient to be measurable, considering the high number of vesicle examined in every experiment (more than 7700 liposomes were individually examined [40]).

The solute concentration in the filled vesicles over-exceeds the eternal one, showing a strong super-concentration effect in the few, but yet measurable, internally crowded liposomes; moreover, this interesting behavior seems to be independent by the initial ferritin concentration (**Fig. 5**).

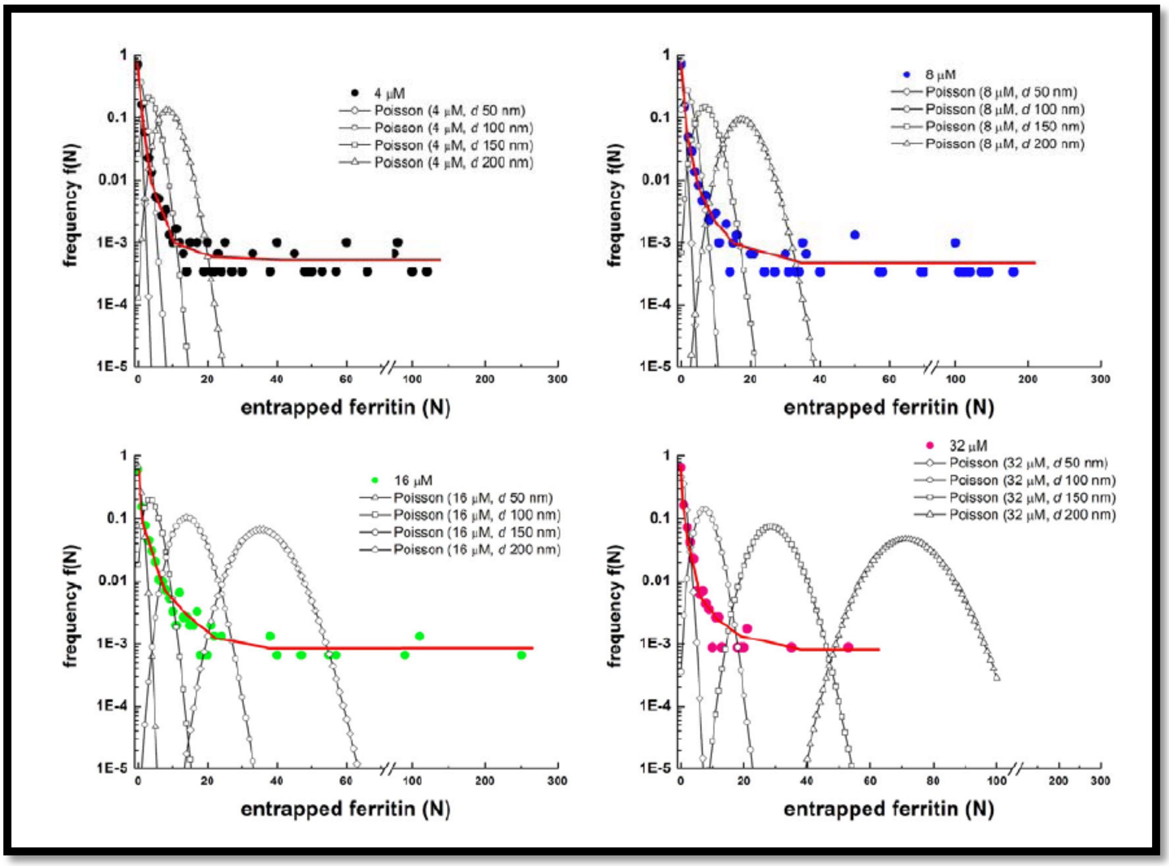


Fig. 5: Solute frequency distribution in 100 nm liposomes with different initial ferritin concentration, according to the Poisson probability (open symbols) *vs*. experimentally observed solute distributions in 100 nm liposomes(filled symbols). The red line is a Power law curve, which nicely fits the experimental results (from Luisi, 2010[40])

As said before, this strong deviation from the expected Poisson behavior has been observed at nanoscale level, while the encapsulation of biomolecules in femtoliter volumes (inside liposomes with 1 μm of diameter) acts according to Poisson statistics [36]; the volume dependency of this super-crowding phenomenon seems to follow a power law too, resulting to be extremely marked as the vesicle diameter decreases [40].

Another recent paper [2] confirmed the strong deviation from the Poisson distribution of entrapped species in nanoscale liposomes by using ribosomes as reporter molecule: ribosomes were chosen for their biological relevance and for their sufficient electron-dense properties, allowing, as for ferritin, a direct visualization using Cryo-TEM without the need for a staining process; the power-law trend, both considering the occupation frequency of ribosomes in liposomes and the size dependence of the super-concentration effect, is once again independent from the initial concentration of solutes. It is interesting to note that the final ribosome concentration inside the super-crowded vesicles is similar to the concentration measured in *E. coli*.

From these results a quite universal observation seems to emerge, which describes a spontaneous accumulation of biomolecules during the liposome formation, which acts in a size-dependent manner according to a power-law behavior; regarding the origin of life scenario, this super-crowding effect has a great relevance in the study of the formation of the first protocells. Lipid compartments could have enclosed several biomolecules, starting from a diluted but highly heterogeneous chemical environment, resulting in a functional vesicle containing an ensemble of different molecular species, with a local concentration high enough to permit biochemical reactivity.

The compartmentalization of biomolecules is indeed the critical step in the formation of first presumable quasi-cellular systems, and very unusual event of super-concentration become clearly visible in the nanoscale range; a quick discussion about the parameters which can affect the encapsulation in liposomes is thus required, trying to search for a possible link between the physical parameters affecting the entrapment process and the generative models for power-law distributions.

Physical parameters and power-law generative models

The entrapment phenomenon in vesicles has always been studied in terms of microencapsulation yield (entrapped molecules / total number of molecules in solution), since liposomes are used most as drug delivery vectors; several factors affect encapsulation of drugs in liposomes, such as liposome size and composition, charge on the liposome surface, bilayer rigidity, preparation method and other biophysical parameters [41]. However the correlation between some parameters (such as vesicle size) and the microencapsulation efficiency has not been thoroughly discussed, as they are aspects of minor importance concerning the pharmaceutical use of liposomes; on the contrary, one of the “hot” topics regarding the search for the simplest life form is about the minimal size of cellular systems [15].

As previously stated, the super-crowding effect seems to act in a size-dependent manner, according to a power law behavior; this means that there is a direct relation with volume availability, suggesting a strong role of entropic forces in driving the entrapment phenomenon.

Volume exclusion is strongly related to macromolecular crowding, which is known to greatly affect the encapsulation efficiency in lipid compartments [42]: polymeric crowding agents are volume excluders, and can enhance biomolecule encapsulation by reducing their hydrodynamic radius. Volume exclusion affects in a greater extent bigger molecules, driving their condensation in crowded solutions and thus greatly reducing their total radius. However this enhancement phenomenon has been observed using protein concentrations below micromolarity, and it is not clear if a high concentration of ferritin or ribosomes in the micromolar order can affect their encapsulation: ferritin and ribosomes are big complex molecules (they have a radius around 6 and 9 nm, respectively), and so they can presumably act as macromolecular agents in concentrated solutions enhancing their self-encapsulation.

The contribution of macromolecular concentrations in determining the anomalous entrapment can be related to one of the generative models of power-law distributions [43-45]; the percolation model, where a systems undergoes a self-organization process when reaching a critical state (Note: a proper explanation of these mathematical models goes out the aim of this thesis, but it is important to evaluate how these models can point out which physical parameters can represent a key problem to understand the underlying unknown phenomena that drive the anomalous entrapment process).

The percolation model is often represented by the presence of clusters on a lattice square, where each square has probability to be included in a cluster of squares; when this probability reaches a certain critical value, the distribution of the clusters on the lattice follows a power-law, independently by the size of the lattice, revealing the *scale-free* behavior typical of power-law distributions [46].

The percolation model, however, does not give a plausible explanation for real systems which exhibit a power-law behavior, but the notion of criticality as fundamental requirement for power-law generation allowed the formulation of an another generative model for dynamical systems, the *self-organized-criticality* (SOC) behavior: in this model the system oscillates around a critical point, encompassing multiple cycles of evolution, but ultimately developing a power-law distribution of the species of the system, as explained in the well-known forest fire model [47]; with respect to the encapsulation processes, criticality can be represented by the thermodynamic state of the small enclosing liposome system, where a sufficiently high concentration of macromolecules in the surrounding environment determines a certain amount of volume exclusion, driving anomalous encapsulation of biomolecules in the forming small lipid vesicles. To test this hypothesis more experiment must be performed, using lower concentration of macromolecular solutes, to understand if the anomalous entrapment occurs also when the solutes are present in submicromolar concentration, thus lowering the crowding effect on the system.

Power-law distribution may rise also when the system acts according to a *Random Walk* behavior: a Random Walk is a trajectory traced out by taking subsequent random steps. Many phenomena in the more disparate fields, from economy to physics, behave in a Random Walk-like manner: for example, Brownian motion, which is the path traced by a molecule in an aqueous solution, can be modeled as a Random Walk process.

Brownian motion acts by distributing molecules in a random manner, causing local superconcentration effects that can become important when approaching the nanoscale range. This random behavior can presumably result in a marked inequality in solute distribution, which is subsequently reflected in the power-law distribution of liposome content. In the literature there are examples reporting the contribution of Brownian motion in biological processes [48], mainly for the actin-myosin molecular motor[49], showing how the stochastic behavior of particle trajectory can determine the velocity of transition between different molecular states and the direction along the filaments. However, at this level of magnification there are strong experimental restrictions that limit a complete understanding of the force-generating process; at nanoscale level stochastic fluctuations are absolutely fundamental, and so even an apparently non-significant mechanism as the Brownian motion can add its contribution in shaping the anomalous entrapment process.

Maybe the most common generative models for power-law distribution is the Yule process, also known as preferential attachment [46], which is most studied in graph theory, as it explains how, following a stochastic random growth, small graphs turn into a network organized according to a power-law distribution, where the majority of nodes has got few connections and few extremely important nodes (called "*hubs*") are highly connected with the rest of the network. Preferential attachment was discovered studying the speciation process in higher organisms and it is considered a possible candidate for the generation of the most diverse phenomena, as the distribution of the wealth among individuals (the *"rich get richer"* process), the sizes of cities, the number of links to pages on the World Wide Web, which are all power-law distributed [44, 46, 50].

The distribution of the internalized molecules in the superfilled vesicle can be determined by a preferential attachment behavior of the enclosing liposome membrane. Theoretically, lipid molecules can take contact with the solutes, attracting other biomolecules in the enclosure process; the more biomolecules are contacting the lipid membrane, the more other molecules will probably take contact with the enclosing liposome, resulting in a supercrowded vesicle. However, in all the Cryo-EM experiments, no solute aggregates were found and neither ferritin or ribosomes have been observed as adsorbed to the lipid membrane. Notwithstanding this, interactions between the lipid membrane and solutes play indeed an important role in the encapsulation process, as reported for DNA encapsulation using anionic lipid molecules [22, 36] and for internalization of drugs in niosomes [51], which are non-ionic surfactant-based liposomes. UV spectroscopy showed how the encapsulation efficiency in niosomes is strongly influenced by the formation of hydrogen bonds between the solutes and the membrane. Moreover, interactions between proteins and membrane probably affect the kinetics of liposome formation by slowing down the enclosure of vesicle [2], allowing the internalization of more biomolecules.

The presence of co-operative mechanism is utterly deducible; anyway, more experiments has to be performed to study the behavior of single vesicles in different chemical environments and with different membrane composition, also trying to evaluate the exact kinetics of vesicle formation. Single vesicle experiments shows a size dependency of entrapped solute concentration [52] and permits to evaluate the encapsulation efficiency; for example, by using an optical trap to immobilize single vesicles, it is possible to evaluate their entrapped content after a photolysis process using an high energy laser beam [53]. Anyhow, these powerful techniques require some preparation steps for the vesicles, as addition of chemicals (as sucrose) aiding the optical manipulation, or multiple freeze and thaw cycles after vesicle formation, that definitely spoil the spontaneous process of liposomes self-organization and influence the resultant entrapment process; the adaptation of single vesicle detection techniques to the semi-synthetic approach requirements will surely improve our knowledge about the vesicle behavior at nanoscale sizes.

All these different theories try to speculate on the still obscure phenomena that influence the encapsulation of macromolecules, and different experimental procedures aim at the detection of encapsulation efficiency in single vesicles.
However, back to the synthetic biology scenario, the outstanding relevance of this property (the superconcentration effect) of nanoscale vesicles lies in the ability to create a real entity with a moderate level of biological complexity in one single step, that turns unorganized molecules into a functional biological architecture which approximates the definition of a real cell, able to afford complex biological processes; therefore, more efforts in the study of protein-producing liposomes will certainly guide the discovery of important phenomena in the emergence of biological complexity.

The unknown processes that drive the unusual entrapment phenomenon for ferritin and ribosomes are presumably the same which permit the co-encapsulation of the over 80 PURESYSTEM species in 100 nm liposomes; the occupancy distribution of the translational mixture however is not easily verifiable, as Cryo-EM allows a direct clear visualization for some molecules but not for proteins as translation factors or polymerases, which anyway remain indistinguishable in the resulting image. Using the PURESYSTEM in Cryo-EM experiments, a small fraction of internally dark liposomes was observed, suggesting, once again, the presence of few (0.1%) super-crowded vesicles [2], but a correct measurement of the internal distribution of the numerous different molecular species remains an inconceivable task.
To overcome all the experimental limitations, it is indeed useful to characterize in detail the behavior of encapsulated biochemical networks such as the PURESYSTEM. Computational modeling of biochemical processes can help understand the properties of complex networks of biochemical reactions, providing different solutions as many useful features.

Computational modelling of biochemical systems

Computational models have been used as useful investigation tools for the most diverse disciplines, from economics to astrophysics, encompassing social and life sciences; mathematical representations of processes are used to perform computer-assisted (*in silico*) simulations of the modeled systems of interest under different circumstances, trying to describe its global properties and make reasonable predictions for possible future evolutions.

Biochemical processes can be modeled according to different criteria, and each way provides some fundamental assumptions which gives different insights into the system; thus the mathematical formalism should be adequately chosen according to the final purpose, to give a realistic description of the phenomenon we are investigating.

Historically, biochemical systems has been represented in a deterministic way, using differential equation and mass-action kinetics to obtain the time course of species concentrations;

Eq. 7:

$$\boldsymbol{\frac{dx_i}{dt} = f(x, p, t)}$$

Following **Eq.** 7, the concentration of a specie x_i is represented as a function of the different concentration of species *x*, the parameters *p* and time *t*; considering a simple reaction a):

$$\text{a) } S + R \overset{k}{\rightarrow} P$$

$$\text{b) } \frac{dS}{dt} = -k[S][R] \qquad \text{c) } \frac{dR}{dt} = -k[S][R] \qquad \text{d) } \frac{dP}{dt} = k[S][R]$$

the concentrations of the three species *S*, *R* and P at time *t* are calculated using the three differential equations b), c) and d); the system can be entirely described using as input parameters the initial concentration of *S* and *P* and the reaction rate constant *k*. Considering more complex biochemical reaction systems, an ODE (Ordinary Differential Equations) system comprising all the coupled differential equation for the network reactions should provide a deterministic unique solution for the calculation of the species amount over time, after the definition of parameters, as the initial species concentrations and the reaction rate constants.

The description of a biochemical network model that uses only differential equations implies that the systems evolves deterministically and continuously. A continuous description of the system evolution can be adequate when considering a huge number of molecules in large sizes (as a test tube or a bulk solution), but does not take in consideration that molecules react individually as integer entities; moreover, molecules do not react in a deterministic manner, but *stochastically*. Chemical reactions take place when reactant molecules randomly hit each other, and only a small percentage of the collision occur with the right orientation and with a sufficient "activation" energy to produce a structural rearrangement and the creation of new chemical bonds.

This random scenario (also known as the "*collision theory*") describing (bio)chemical processes has not a great consequence in the analysis of large systems, where a deterministic kinetic description is a good and numerically cheap approximation, but becomes fundamental at small scales, where random fluctuation in the low molecule numbers becomes fundamental describing the the overall system behavior.

The most correct way to describe the evolution of chemical system is represented by Molecular Dynamics simulations, that allow to track the position and the linear momenta of all the particles in solution, simulating all the trajectory and collision between molecules. Unfortunately, the computational power required for molecular dynamics simulations is extremely high, being unfeasible for the analysis of complex biochemical networks.

Anyway, considering that only a very few number of collisions result in a chemical reaction it is possible to ignore the vast majority of molecular impacts; "non-reactive" collision, even when not giving rise to chemical reaction, affect the state of motion of the particles, and neglecting their effect implies a random description of the position and velocity of molecules. This approximation result in a system representation where molecules are uniformly distributed in space (the "*well-stirred*" condition), and the system evolution is described by the molecule number during time.

As said before, the positions and thus the collisions between molecules are randomized, hence chemical reactions can be fully described as stochastic processes.

The transition from a kinetic continuous equation to a stochastic formulation implies the definition of a state vector X comprising all the species concentrations, and a chemical reactions R_j is defined as the transition event between the different states.

R_j is characterized by two parameters: a state-change vector v_j which defines the stoichiometry of the reaction, and the propensity $a_j(x)dt$ that defines the probability that one R_j reaction will occur in the next time interval dt. In a well-stirred system, substrate molecules hit each other whit a rate proportional to their concentrations, hence the propensities are calculated according to the mass action law for the different orders of reactions (Table 2).

Table 2: Conversion between reaction rate constant *k* and propensities for different orders of reactions, from Klipp, 2009 [54]

Reaction order	Formula	Propensity	Scaling
0	NULL -> ...	$a_j=c_j$	$c_j=k_j V$
1	A -> ...	$a_j=c_j x_A$	$c_j= k_j$
2	A+B -> ...	$a_j=c_j x_A x_B$	$c_j= k_j /V$
2	2A -> ...	$a_j=½\, c_j x_A (x_A-1)$	$c_j= k_j /V$

However, this stochastic formalization implies the presence of irreversible reaction up to the second order: thus, equilibria or higher order reactions has to be formulated by splitting them in elementary one- or bimolecular unidirectional reactions.

The probability for a certain state X (which comprises the molecular concentrations of all the chemical species x_i) to change in the time interval dt is defined by the *Chemical Master Equation* (CME):

Eq. 8:

$$\frac{\delta P(X,t\,|X_0,t_0)}{\delta t} = \sum_{j=1}^{M}[(a_j(X-v_j)\,P(X-v_j,t\,|X_0,t_0) - a_j(X)P(X,t\,|X_0,t_0)]$$

In the left positive term are enumerated the possibilities to enter the state X w with a reaction v_j, while the negative term, to the right, collects all the realizations that exit from the state X. Every possible state X gives rise to a differential equation as the Eq. 8 and obtaining analytical solution for the CME for every state X is a very difficult task; for large systems, we can ignore fluctuations in molecule number, and the CME is reduced to the Reaction Rate Equation(Eq. 9), here written in terms of propensities and transition vectors:

Eq. 9:

$$\frac{dX(t)}{dt} = \sum_{j=1}^{M} v_j a_j(X(t))$$

Anyway, as the molecule number becomes small, fluctuations cannot be ignored, and a deterministic description becomes uncorrect, thus the use of the RRE is unjustified. There are different methods that can describe the evolution of a system using the same formalization of the CME to define chemical reactions, but without the need to find analytical solutions or approximating to a purely deterministic behavior.

In 1976 Daniel Gillespie proposed a stochastic simulation algorithm [13] for chemical reactions with continuous time and treating molecular species as integer, discrete particles.

Gillespie's Stochastic Simulation Algorithm

Individual random realizations of the system are calculated according to their probabilities, which are defined drawing from the same probability density function defined in the CME; properly distributed random numbers are generated to determine the time course of the different reactions: a first random number sets the time *dt* to the next reaction *R,* and another random value determines the index *j* of that reaction, thus defining the state-change vector v_j and the evolution of the system to the next state:

The variable τ defines the time to the next reaction, and is randomly distributed with a mean that is the inverse of the total sum of propensities a_0, as follows:

Eq. 10:

$$\tau = \frac{-\ln(r_1)}{a_0}$$

The first random number r_1 thus defines the time evolution, while a second random number r_2 is generated to determine the index *j* for the next reaction *R* that will occur in the previously chosen time interval τ:

Eq. 11:

$$j = \mathit{the\ smaller\ integer\ such\ that}\colon \sum_{k=1}^{j} a_k > r_2 a_0$$

Once that the time interval and the reaction have been chosen, the reactions fires as an instantaneous process and the system is updated and the next transition is chosen this process is repeated iteratively until the end of the simulation, yielding a stochastic representation of the entire set of (bio)chemical reactions according to their probabilities; the overall time course of state transitions can be described in following steps:

1) System initialization (t=t_O); reactions, species and their amount are declared (X=X_O)
2) Propensities for all reactions are calculated
3) Random variables τ and *j* are calculated using the propensities as probability weights
4) The reaction instantaneously occurs and the systems is updated (t=t_O+ τ; X=X+υ_j)
5) The algorithm outputs *X* and *t*
6) Re-iterate from step 2) until the end of simulation.

The step succession depicted above describes the *Direct Method* (DM) [55] implementation of the Gillespie's algorithm, while similar formulations have been proposed, as the *First Reaction Method* (FRM) [56] or the *Next Reaction Method* (NRM) [57], which have a differ succession of the single computation steps, but are proven to be equally exact realization of CME, being probabilistically correct and theoretically founded.

DM first computes the total propensites, and then it defines separately the variables τ and j, while FRM computes one time τ for each reaction j, and then makes a selection for the smallest time with the corresponding reaction. NRM optimizes the FRM re-computations. The choice between DM and FRM depends on the characteristic of the simulated system, but generally DM performs better than FRM [58] .

In general, Gillespie's SSA is simple to be implemented, but it requires a relatively high computational power and exhibits some problems (as ODE systems also do) when dealing with highly heterogeneous systems: since all the reactions take place as instantaneous events, there is no difference between very fast or slow reactions involving the time required for the reaction to take place, producing a common problem known as *stiffness*: time steps are dependent by the sum a_0 of all propensities (Eq. 10); systems with very fast reactions result in very small time steps; only fast reaction are very likely to occur when the time step is small (Eq. 11). The combined effect of these phenomena cause the computation to take excessively smaller step sizes than what is necessary for the accuracy requirement to easily follow the system dynamics; trying to overcome these general drawbacks, different solutions were proposed, as the firing of reactions in a delayed fashion [59].

Another method, called the *tau-leaping* method [60, 61], sacrifices the *exactness* of the algorithm by simulating not individual events, but it considers a fixed time interval τ and compute an ensemble of multiple reactions that will occur in that time interval; this simulation approach is faster to compute than the standard SSA, but it represents an approximated method, because propensities are assumed to remain constants in the fixed time interval τ. Intuitively, as τ becomes large the approximation error increases, and the assumption of constant propensities can lose its justification; thus, the tau-leaping method is most used when working with large numbers of molecules. As the number of molecules increases, the SSA suffer the high computational cost, and other method can be used to easily compute the system dynamics.

Some methods uses mixed approaches, with differential equations and Gillespie's SSA used for different processes in the same system; as the particle numbers increases, other approaches can be used to describe system dynamics, as the

Chemical Langevin Equation (CLE) [62], a *Stochastic Differential Equation* which adds a random stochastic term of noise normally distributed (*Gaussian white noise*) to the deterministic formulation of chemical reactions, yielding a continuous, but still stochastic, description of the system.

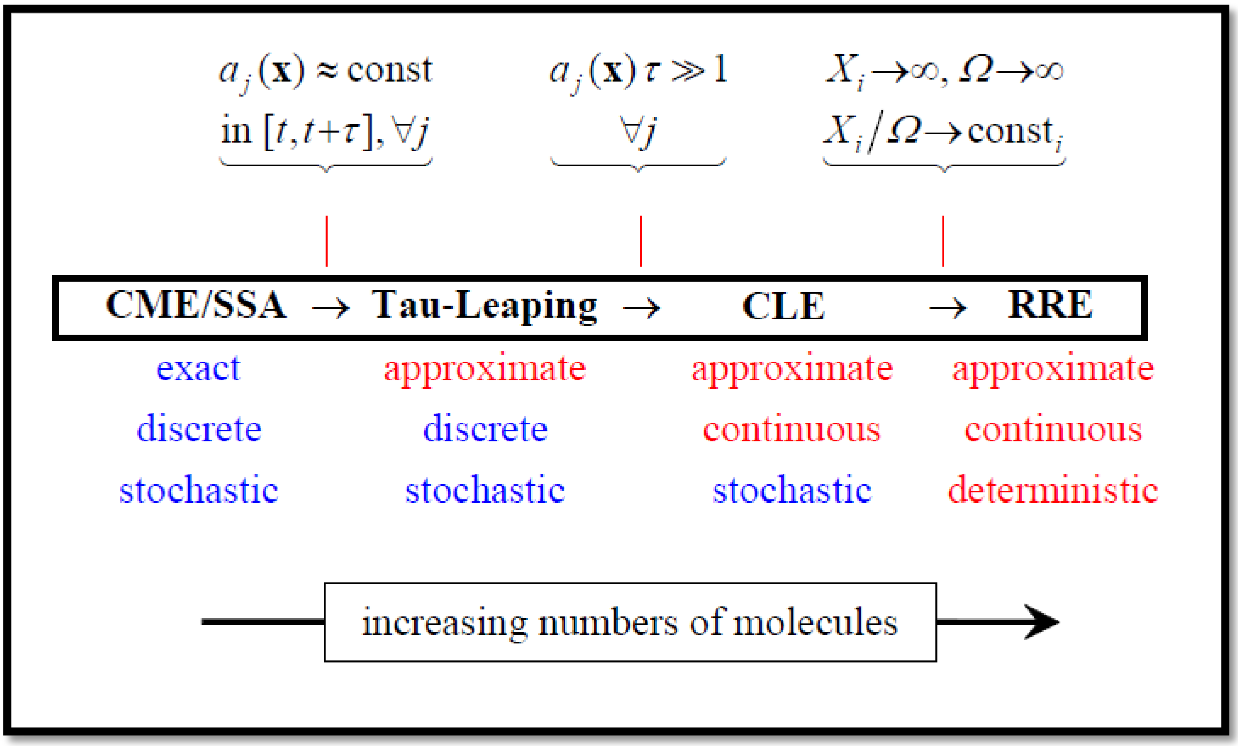

$\underbrace{a_j(\mathbf{x}) \approx \text{const in } [t, t+\tau], \forall j}$
$\underbrace{a_j(\mathbf{x})\tau \gg 1 \; \forall j}$
$\underbrace{X_i \to \infty, \Omega \to \infty \; X_i/\Omega \to \text{const}_i}$
CME/SSA → Tau-Leaping → CLE → RRE
exact
discrete
stochastic
approximate
discrete
stochastic
approximate
continuous
stochastic
approximate
continuous
deterministic
increasing numbers of molecules

Fig. 6: Different mathematical formulations of chemical systems, with their assumptions and general features.

These different mathematical formulations allow correct descriptions for many chemical systems, from small environments to large complex systems for a wide range of volumes, providing different solutions under different general assumptions (**Fig. 6**). Of course, the choice for a correct mathematical formulation has to be made purposefully: a biological problem can be realistically described considering its real characteristics, but the adequateness of a mathematical model principally depends on the particular aspects we are trying to elucidate, which determine the level of detail we are currently focusing on.

STATE OF THE ART AND THESIS AIM

GFP-expressing nanoscale liposomes represent the smallest biological objects capable of genetic expression, opening the way for different consideration for the definition of minimal living organism. Different experimental restrictions are currently limiting the investigation on such synthetic systems at the nanoscale range, where stochastic simulation proved to be a valuable tool, providing a correct description of small chemical environments, where randomness and uncertainty cannot be described simply as an additive noise factor, but they represents the fundamental forces which drive the overall behavior of the system.

A model for the transcription/translation module (PURESYSTEM) was previously published by our research team [63], comprising over 100 biochemical reactions with their kinetic coefficients; although this model is very complex, it was conceived to give a qualitative description of protein synthesis in different sized compartments, according to different entrapment models. As every first attempt, it contained some simplifications:

1) the presence of multiple elongating ribosomes on the same RNA molecule was not described;
2) protein production was conceived without taking into account aminoacids consumption;
3) the transcription process was not included;

These three problems are deeply intertwined, because the transcription and the translation processes are dynamically *coupled*: ribosomes bind to the Shine-Dalgarno sequence (the ribosomal binding site, or RBS) as soon as it is available, even if the complete RNA sequence has not been entirely produced yet. This means that the translation process begins as the RNA polymerase is still transcribing the RNA molecule, and the two processes (transcription and translation) cannot be treated separately.

Moreover, different ribosomes were detected as progressing along the same RNA molecule synthesizing the same protein, forming a cluster of ribosomes named *polysome* or *polyribosome*; the same phenomenon was observed for transcription: multiple RNA polymerases can simultaneously transcribe the same DNA molecule.

The aim of my thesis was to create a suitable model for the PURESYSTEM reaction network that can describe correctly all these interdependent processes, to give subsequently, using stochastic simulation experiments, a quantitative description of the system behavior in different experimental conditions.

Gillespie's SSA describes a chemical environment were all the reactions compete with each other and are randomly chosen with a frequency which oscillates around a value defined by their propensity. This formulation depicts a concurrent-only environment, and a stochastic formalization of a system where sequential processes occur with a defined order of succession represents indeed a difficult task.

This new improved *in silico* formalization of the PURESYSTEM has to account for a detailed description of the single molecular reactions; aminoacids and nucleotides must be consumed in the right quantities to ultimately produce proteins and RNA, permitting a correct evaluation of the system dependency by the initial amount of the different species, consumables included.

Stochastic simulation experiments will determine how the different composition of species affects the overall GFP production kinetics in small volumes.

METHODS

In silico experiments were performed using QDC (Quick Direct-method Controlled, http://gillespie-qdc.sourceforge.net/), a simulation software previously built in our lab to perform stochastic simulation of biochemical systems. QDC uses the Gillespie's Stochastic Simulation Algorithm with the Direct Method implementation, yielding a correct description of small biochemical system as the one we are investigating, where molecules act as integer numbers and randomness represents an important force driving the biochemical reactions.

QDC (Quick Direct-Method Controlled)

Structure and Outputs

QDC has a core written in the C++ language, represented by a single C++ source file, *parser.cpp*; when this file is compiled it generates an executable file (*parser*) which can accept the input file, creating another source file, named *engine.cpp*; the compilation of this additional source file creates the last executable file *engine*, that ultimately runs the simulation (**Fig. 7**).

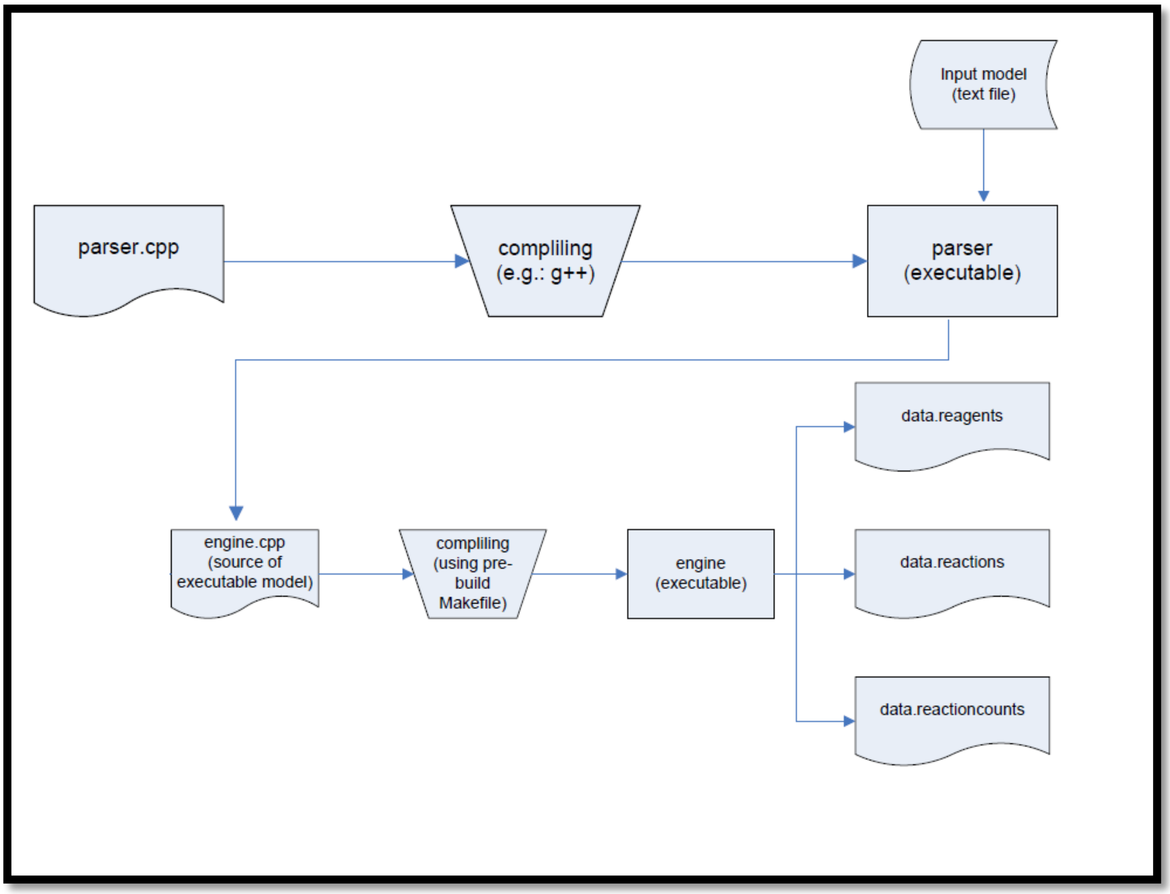


Fig. 7: The architecture of QDC software (http://sourceforge.net/projects/gillespie-qdc/)

The simulation is launched by specifying the total duration of the simulation and the sampling frequency by which the simulator outputs the system status, in three .csv (Comma-separated values) files, that can easily be imported in a spreadsheet, using familiar softwares such as *OpenOffice Calc* or *Microsoft Excel*™:

<input_filename>_reagents.csv
<input_filename>_reactioncounts.csv
<input_filename>_reactions.csv

where *<input_filename>* is the name of the input file used.

The *reagents.csv* file contains the time course of the species during the simulation, reporting in columns the declared species and in rows their respective amount, written as integer numbers, according to the sample frequencies: a simulation of 1000 simulated seconds with a sample frequency of 1 second outputs a *reagents* file with 1000 rows, where the n^{th} row contains the molecule amount after n seconds of simulations.

The *reactioncounts.csv* file enumerates the sum of all the occurrence of each reactions during the simulation, while the *reactions.csv* file comprises the time course of the propensities of the different reactions. It is worth to specify that propensities and species amount are instantaneous values: for example, a value for the molecule $a = 0$ in a time interval $t_1 - t_2$ does not imply that the reaction a -> b has never been fired in that time interval, because the reaction may have occurred with a velocity greater than the sample frequency; also the propensity for that reaction will be, of course, zero, and that is a typical situation where the *reactioncounts* file comes into help, reporting the cumulative number of the fired reactions, and thus avoiding possible misinterpretation about the system dynamics.

SYNTAX AND USEFUL FEATURES

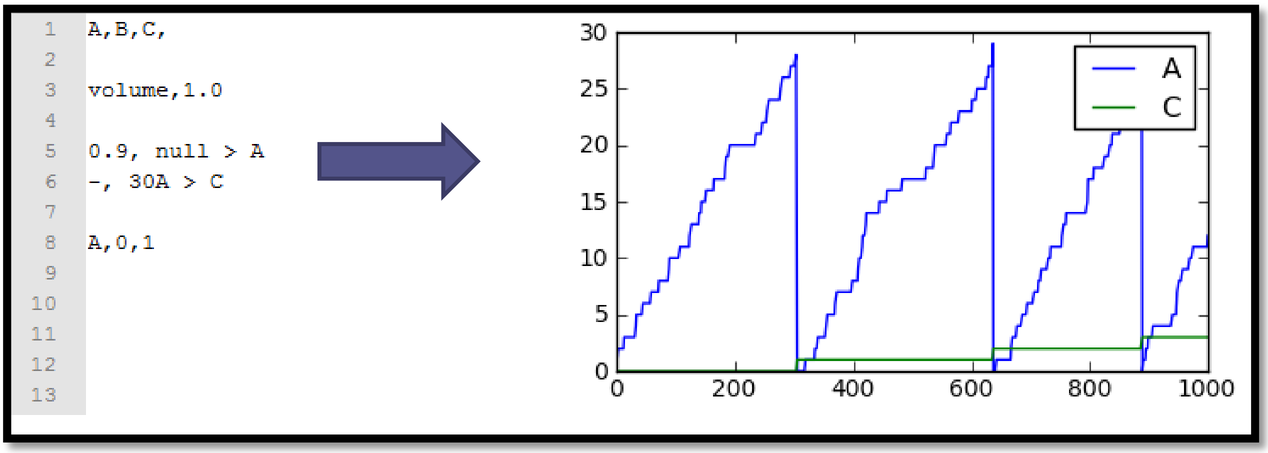
QDC accepts as input file a normal ASCII text file, that can be easily written using a standard text editor; the file has to be structured in different blocks separated by a blank line; in the first block all the biochemical species of the system are declared, separated by a comma.

The second block specifies the total volume of our system, measured in liters, which is subsequently used for the calculation of propensities (see Table 2).

The third area contains all the biochemical reactions of the system: in each line, the first value represent the kinetic rate constant of the reaction, followed by the reagents species and their respective products, separated by the reaction arrow (the ">" symbol); it is possible to simulate the exchange of molecule with the environment, by using the "*null*" term in substitution of the reagents (uptake) or products (excretion); as specified in the previous chapters, only reaction up to the 2nd order are allowed, but QDC allows to write reactions with complex stoichiometry in form of *"immediate"* reactions, defined by the presence of the hyphen symbol ("-") in substitution of the rate constant, but with reagents and products normally written as described before;

immediate reactions lack a real rate constant, and they are *instantaneously* fired once the reagents have reached the correct stoichiometric conditions; thus, *immediate* reactions are not real biochemical events that occur according to their probabilities, but they represents *logical statements*, allowing for description of complex events which cannot be easily included in form of standard biochemical processes (**Fig. 8**).

The fourth block contains the number of molecules supplied to the system together with the time they are supplied (when time=0 the species are present from the beginning of the simulation); the last two blocks are optional, and concern the eventual presence of control variables: it is possible, for example, to simulate the addition of an effector molecule by changing the reaction rate of a desired equation at a specified time.



```
A,B,C,

volume,1.0

0.9, null > A
-, 30A > C

A,0,1
```

Fig. 8: an example of immediate reactions with the resulting *reagents.csv* output file; the A molecule is continuously added by the uptake reaction: null->A; the *immediate* reaction produces one C molecule every 30 A molecule; notice the resulting oscillatory behavior of the A specie.

QDC proved to be an efficient stochastic simulation software when compared to other widely used biochemical simulator, and presents some very useful features, which allowed the construction of a realistic *in silico* model of the PURESYSTEM reactions network.

In Silico PURESYSTEM model

The simultaneous elongation events present in the transcription/translation system depict a scenario with multiple sequential interdependent processes; **Fig. 8** shows a representative case trying to clarify better this complex situation.

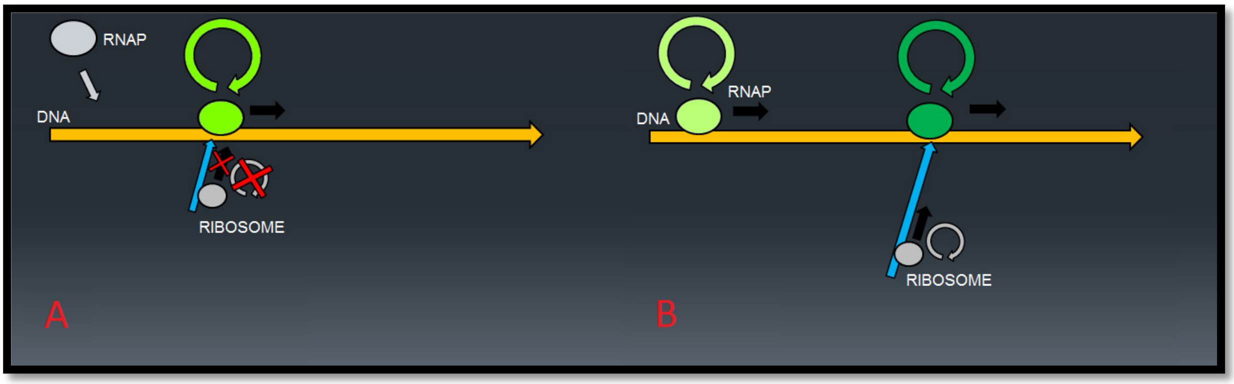


Fig. 9: Multiple elongation events in transcription and translation processes, see text for details.

As said before, a ribosome starts the translation process binding to the RBS, which is a RNA sequence produced in the initial stage of the transcription process; (A) the ribosome cannot incorporate aminoacids and subsequently move forward if the polymerase has not yet produced a sufficiently long RNA sequence; (B) as long as transcription continues, new nucleotides are incorporated, the RNA molecule is elongated, and the ribosome can continue the translation process and progress along the RNA strand, while a new polymerase bind to DNA and begins to produce a new RNA molecule, by keeping a certain distance from the other elongating polymerase due to steric repulsions.

The correct spacing between polymerases and between ribosomes, and, subsequently, a consistent occupancy of RNA and DNA sites, can be fulfilled by splitting all the species involved in the elongation events (DNA, RNA, polymerases and ribosomes) into different entities representing molecular *states*, and describing their evolution coherently (**Fig. 10**).

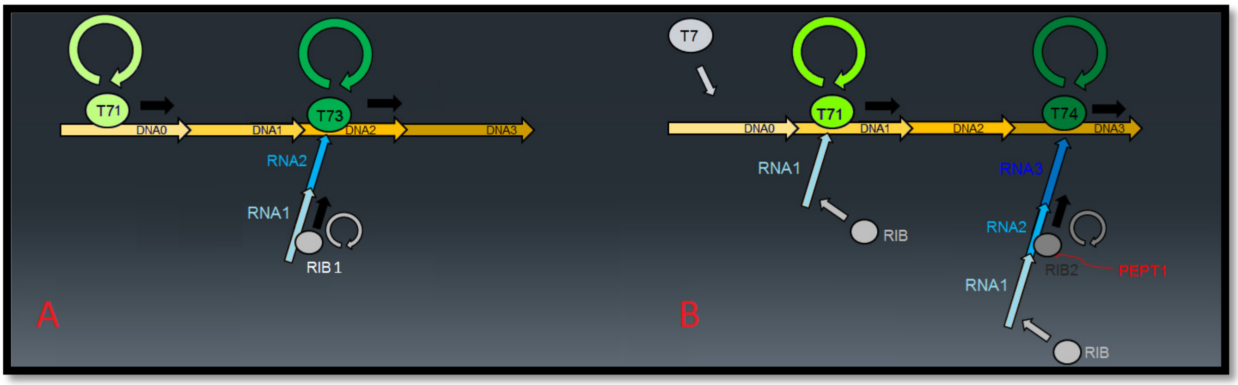


Fig. 10: Transcription/translation events using multiple molecular states, see text for explanations.

The progression from a state to another is thus easily defined: (A) a newly-bond polymerase (T7) can advance to the next DNA sequence (DNA1) which is free, and the ribosome (RIB1) can begin to incorporate aminoacids and move forward, because the RNA sequence has been extended and the adjacent site (RNA2) is free; in the next stage (B), polymerases and ribosomes have advanced, consequently releasing the site they were previously occupying; accordingly, a new polymerase (T7) and a new ribosome (RIB) can bind to their respective target molecules, and start the elongation processes.

The presence of multiple states allows a coherent description of the different steps, but the formalization of the events that describe the forward motion of ribosomes and polymerases it is not an easy task. As explained before, Gillespie's SSA depicts a concurrent environment where all reactions are randomly chosen according to their propensities during the simulation. Transcription and translation are two processes which act in a sequential fashion, and the correct progression of the elongating molecules cannot be described using the standard notation for biochemical reaction.

The solution to this problem was assessed using logical statements written in form of reactions present in the QDC syntax, the aforementioned *immediate* reactions, which were massively used in this new model to regulate the forward motion of polymerases and ribosomes, accounting for correct spacing and coupling.

The DNA sequence encoding for GFP was divided, according its length, into different multiple species, each representing a 80 bp sequence; polymerases bind to the first

DNA sequence and start the transcription process, according to their efficiency; the polymerization process is divided into different reactions to describe a second-order reaction for nucleotide binding, and a first-order reaction for nucleotides incorporation which returns the polymerase molecule (which can bind to another nucleotide, see the blue arrows) and a “dummy” product, that allows to track the number of nucleotides incorporated in the RNA molecule (the term dummy comes from computer science language, where dummy variables are arbitrary chosen variable employed for temporary purposes); the *immediate* reactions determines the transition to the next step, ensuring the following conditions: a) an adjacent DNA site is available, b) a correct number of nucleotides has been added to the RNA sequence, c) the corresponding RNA sequence is produced, d) the previously occupied DNA site is released. Here it is an example for transcription process at the fourth step (GTP and ATP are considered):

1000000, **T7EL4** + GTP > T7pregEL4

28, T7pregEL4 > **T7EL4** + Pi + **gtr4**

1000000, **T7EL4** + ATP > T7preaEL4

28, T7preaEL4 > **T7EL4** + Pi + **atr4**

-, 20 **gtr4** + 20 **atr4** + T7EL4 + *DNA5* > T7EL5 + RNA4 + *DNA3*

This reaction “box” was duplicated different times ensuring the correct succession of molecular states; when only one DNA molecule is available, the polymerases advance one by one, separated by at least one DNA site between each other (elongating RNA polymerases are separated each other by at least 80 bp). The transcription process thus developed was able to produce RNA molecules with an average transcription rate of 28-40 nucleotides per second (depending on the DNA concentration), which is in good agreement with data present in literature [64, 65].

The same strategy was used to describe the translation reactions (note, here the dissociation reactions are included due to their relevance for the overall translation process):

5000000, **eR2** + EFaRGTP > EX2

0.2, EX2 > eR2 + EFaRGTP

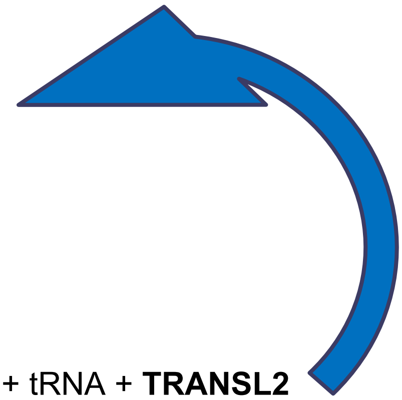
30, EX2 > eRa2 + EFtuGDP + Pi

5000000, eRa2 + EFgGTP > EXb2

0.3, EXb2 > eRa2 + EFgGTP

30, EXb2 > EFg + GDP + Pi + **eR2** + tRNA + **TRANSL2**

-, 27 **TRANSL2** + RNA4 + *PEPT1* + **eR2** > eR3 + *PEPT2* + RNA2



An elongating ribosome (eR2) binds the complex which carry the aminoacid (EFaRGTP), after which moves to the next codon, aided by the elongation factor EFg charged with GTP (EFgGTP); this translocation reaction yield ad additional product which is used to regulate the progression to the next state.

After a fixed number of translocation steps (the minimal space between two elongating ribosomes is, curiously as for polymerases, 80 nt ≈ 27 codons) an *immediate* reaction occurs in a similar fashion as seen for transcription: 1) the correct amount of aminoacids are incorporated, and thus consumed; 2) the next free RNA site is occupied and 3) the previous one is therefore liberated; 4) an entity named PEPT is also produced, allowing to calculate the length of the peptide sequence produced so far: for example, if 4 species named PEPT3 are present in a certain time of the simulation, this means that there are 4 peptides, still bound to the ribosomes, with a length spanning from 27 x 3= 81 to 27 x 4= 108 aminoacids.

The portrayed model includes the single biochemical reactions comprised in the transcription\translation interaction network, avoiding the use of simplified average macroscopic measures for the formalization of initiation, elongation or termination events. Moreover, this formalization accounted for the presence of ordered sequential events: all the elongation processes take into account the steric repulsions between molecules and the correct sites occupancies.

The introduction of new reactions during the simulation can depict the presence of multiple elongation events, as reported in other stochastic models [66].

Anyway, the addition of new events at fixed times during the simulation can jeopardize the stochastic nature of biochemical processes, while in our formulation all these sequential dependencies were included inside the model using the immediate reactions. Here, the dynamical coupling between the multiple elongation events is accounted in the model formulation itself, and the system evolves, according to its stochastic nature, depending solely on its internal biochemical composition.

As pointed out in the introduction, POPC vesicle do not fuse and their membrane is impermeable to big molecules and small charged particles, and therefore all the species are confined into the internal volume; thus the model for the encapsulated PURESYSTEM does not need to include any uptake or excretion reaction.
The overall system behavior is of course influenced by its initial conditions, and all the species influence the overall kinetics of protein productions in different ways.

Stochastic simulation experiments

Stochastic simulations were performed to test how protein production kinetics are affected by the internal composition of the system; different input files were used, using the concentrations of the PURESYSTEM species as reported in Table 1, but lowering the initial quantities of DNA, total enzymes or consumables, to 2/3 or 1/3 of their original value, for a total of 3^3=27 combinations; each combination is defined by a series of 3 digits, accounting for DNA, Enzymes or Consumables; each digit contains a number which is 0,1 or 2, meaning respectively 1/3, 2/3 or 3/3 of their normal concentration.
For example, “201” means 3/3 of DNA, 1/3 of enzymes and 2/3 of consumables (see Table 3). The first set of simulation was performed simulating the PURESYSTEM in a compartment of 10^{-14} liters, representing a vesicle with ≈ 2µm of diameter, to study the general behavior of the system in presence of large number of molecules.

Table 3: Different combination of DNA, Enzymes of Consumables of the PURESYSTEM with the respective concentrations

SAMPLE NAME	DNA	ENZYMES			CONSUMABLES		
		RNApol	aaRS	ribosome	tRNA	AAs	NTPs
	µM	µM	µM	µM	µM	µM	µM
222	0,333	0,100	0,200	1,200	1,91	300	2000
221	0,333	0,100	0,200	1,200	1,273	200	1333
220	0,333	0,100	0,200	1,200	0,637	100	667
212	0,333	0,067	0,130	0,800	1,91	300	2000
211	0,333	0,067	0,130	0,800	1,273	200	1333
210	0,333	0,067	0,130	0,800	0,637	100	667
202	0,333	0,033	0,067	0,400	1,91	300	2000
201	0,333	0,033	0,067	0,400	1,273	200	1333
200	0,333	0,033	0,067	0,400	0,637	100	667
122	0,214	0,100	0,200	1,200	1,91	300	2000
121	0,214	0,100	0,200	1,200	1,273	200	1333
120	0,214	0,100	0,200	1,200	0,637	100	667
112	0,214	0,067	0,130	0,800	1,91	300	2000
111	0,214	0,067	0,130	0,800	1,273	200	1333
110	0,214	0,067	0,130	0,800	0,637	100	667
102	0,214	0,033	0,067	0,400	1,91	300	2000
101	0,214	0,033	0,067	0,400	1,273	200	1333
100	0,214	0,033	0,067	0,400	0,637	100	667
022	0,109	0,100	0,200	1,200	1,91	300	2000
021	0,109	0,100	0,200	1,200	1,273	200	1333
020	0,109	0,100	0,200	1,200	0,637	100	667
012	0,109	0,067	0,130	0,800	1,91	300	2000
011	0,109	0,067	0,130	0,800	1,273	200	1333
010	0,109	0,067	0,130	0,800	0,637	100	667
002	0,109	0,033	0,067	0,400	1,91	300	2000
001	0,109	0,033	0,067	0,400	1,273	200	1333
000	0,109	0,033	0,067	0,400	0,637	100	667

Stochastic simulation experiments did not require high computational power: short simulation with low volumes and few particle numbers were performed using a standard dual-core CPU laptop, while calculations involving higher volumes and thousands of species were performed via remote access with a 8-core machine with medium computational power.

The output files for each combination were imported in SigmaPlot™ (SyStat Software), and data for protein production over time was fitted by a 3-parameter Sigmoid Curve (Eq. 12) using the Nonlinear Regression tool:

Eq. 12:

$$y = \frac{a}{1 + e^{-(\frac{x - x_0}{b})}}$$

The resulting estimated parameters represents:

a = maximum value of protein produced

x_0= time value for y=$a/2$

b = maximum slope of the curve (a low value for b indicates an high steepness)

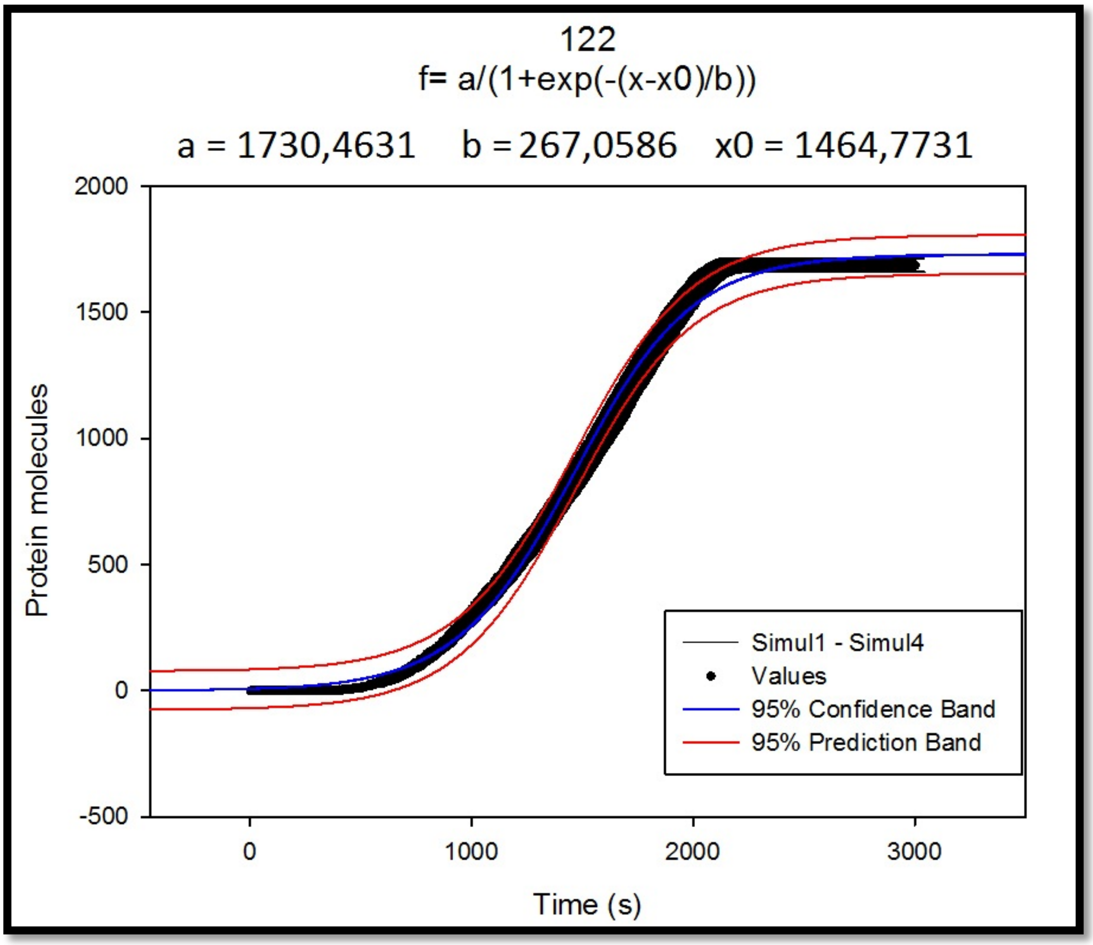


Fig. 11: Protein production curve using the 122 input file: data comes from 4 simulation replicates.

Simulated data was nicely fitted (R^2 > 0.98) by the sigmoid curves (**Fig. 11**), except for cases in which the simulation yielded a low number of proteins (>10), where differences within replicates become significant.

The parameter comparison for different inputs revealed how changes in the initial amount of species severely affects the overall protein production (Table 4).

Table 4: Parameter estimation for different protein production time-courses; in green-scale very low protein productions, in red-scale high protein yield. Simulation were performed in 4 replicates for each input file (volume = 10^{-14}).

Data Source: 000		Data Source: 001		Data Source: 002	
a	1,5014	a	40,6271	a	144,6786
b	140,3791	b	603,8102	b	1059,195
x0	2439,411	x0	3890,285	x0	5801,185
Data Source: 010		Data Source: 011		Data Source: 012	
a	9,7658	a	269,8977	a	868,675
b	113,8108	b	293,4997	b	528,8943
x0	1125,974	x0	1986,326	x0	2774,823
Data Source: 020		Data Source: 021		Data Source: 022	
a	39,3244	a	756,1448	a	2386,73
b	75,1161	b	203,958	b	362,4964
x0	760,8157	x0	1276,52	x0	1804,794
Data Source: 100		Data Source: 101		Data Source: 102	
a	0,25	a	30,8911	a	105,6872
b	0,0603	b	493,2288	b	799,1214
x0	1392,439	x0	3267,887	x0	4683,551
Data Source: 110		Data Source: 111		Data Source: 112	
a	2,9981	a	183,8399	a	658,9507
b	53,143	b	229,806	b	394,3518
x0	955,1248	x0	1645,187	x0	2297,454
Data Source: 120		Data Source: 121		Data Source: 122	
a	16,7787	a	519,7821	a	1730,463
b	55,6302	b	158,144	b	267,0586
x0	640,5336	x0	1060,858	x0	1464,773
Data Source: 200		Data Source: 201		Data Source: 202	
a	0,25	a	23,338	a	87,9321
b	0,0521	b	444,6386	b	752,4419
x0	1743,425	x0	2954,022	x0	4498,174
Data Source: 210		Data Source: 211		Data Source: 212	
a	2,0019	a	156,3254	a	597,2295
b	83,6926	b	203,526	b	361,6856
x0	906,5084	x0	1543,768	x0	2148,991
Data Source: 220		Data Source: 221		Data Source: 222	
a	9,2609	a	433,5804	a	1573,636
b	44,0768	b	140,5483	b	240,864
x0	594,6098	x0	999,6463	x0	1360,217

Not all the combinations led to an efficient protein yield, and parameters comparisons considering their outputs are obviously biased by the low numbers of proteins molecules produced.

An high DNA concentration slightly accelerates the overall protein production (parameter *b* decreases); however, the overall protein yield (parameter *a*) diminishes as the DNA amount increases; in fact, although being a more rapid process, protein production with high DNA concentrations stops at lower time values (see parameter x_o).

Simulations carried with a lower amount of enzymes concentrations resulted in a strong decrease of protein yield: a change from 3/3 to 2/3 in enzymes concentration determined a reduction in protein yield to circa 38% of the total; an additional reduction in enzymes concentration to a 1/3 of the original value led to a 5,5% of produced protein if compared to the 3/3 of enzymes condition. Very slow kinetics (high values for *b* and x_o) are observed as the enzymes concentration decrease.

The amount of consumables is absolutely the determining factor for the overall production: input files with 1/3 of the concentrations of consumables yielded a maximum of 39 protein molecules (internal concentration ≈ 6nM), in presence of high enzyme and low DNA amounts.

The highest protein production is afforded when DNA is low and enzyme and consumables are present in maximum quantity (sample "022"), reaching a total of approximately 2380 proteins (final concentration = 0.39 μM)

The different protein yields for the different PURESYSTEM combinations were tested also in small volumes (10^{-16} liters, corresponding to a vesicle with 570 nm of diameter); parameters were extracted after data fitting using the same procedure discussed before.

Table 5: Parameter estimation for different protein production time-courses; in green-scale very low protein productions, in red-scale high protein yield. Simulation were performed in 4 replicates for each input file (volume = 10^{-16}).

Data Source: 000		Data Source: 001		Data Source: 002	
a	0	a	0,767	a	2,0714
b	1	b	650,6759	b	1647,123
x0	2	x0	3898,484	x0	5879,842
Data Source: 010		Data Source: 011		Data Source: 012	
a	0,25	a	2,2798	a	9,5395
b	0,0529	b	374,4029	b	425,8866
x0	1266,554	x0	1902,351	x0	2381,602
Data Source: 020		Data Source: 021		Data Source: 022	
a	0,25	a	6,7965	a	26,5048
b	0,0555	b	199,6762	b	376,7406
x0	783,5081	x0	1252,539	x0	1808,302
Data Source: 100		Data Source: 101		Data Source: 102	
a	0	a	0,25	a	0,9845
b	1	b	0,0597	b	866,2825
x0	2	x0	2934,46	x0	4245,203
Data Source: 110		Data Source: 111		Data Source: 112	
a	0	a	2,006	a	6,811
b	1	b	204,4793	b	399,1548
x0	2	x0	1318,435	x0	2236,633
Data Source: 120		Data Source: 121		Data Source: 122	
a	0	a	3,7441	a	16,2867
b	1	b	152,321	b	240,3424
x0	2	x0	1014,37	x0	1461,434
Data Source: 200		Data Source: 201		Data Source: 202	
a	0	a	0	a	1,0079
b	1	b	1	b	455,6501
x0	2	x0	2	x0	5001,928
Data Source: 210		Data Source: 211		Data Source: 212	
a	0,25	a	1,7573	a	6,8644
b	0,056	b	177,7419	b	466,6417
x0	893,5071	x0	1553,995	x0	2141,824
Data Source: 220		Data Source: 221		Data Source: 222	
a	0	a	2,7726	a	15,1471
b	1	b	135,219	b	258,7586
x0	2	x0	1044,23	x0	1388,208

Results show as only some combinations for system internal composition can efficiently afford protein synthesis (Table 5), reaching a maximum of approximately 26 proteins (final protein concentration ≈ 0.44 μM) for the "022" sample, as seen also for simulations in higher volumes. The general trend for protein yield and its dependencies by DNA, enzymes and consumables concentrations are similar as described for higher volumes.

Additional stochastic simulation were performed to evaluate the translation and transcription kinetics with lower DNA concentrations. Data for protein production, transcription reaction and GTP depletion rates was fitted as described before, and the extracted parameters were compared (**Fig. 12**).

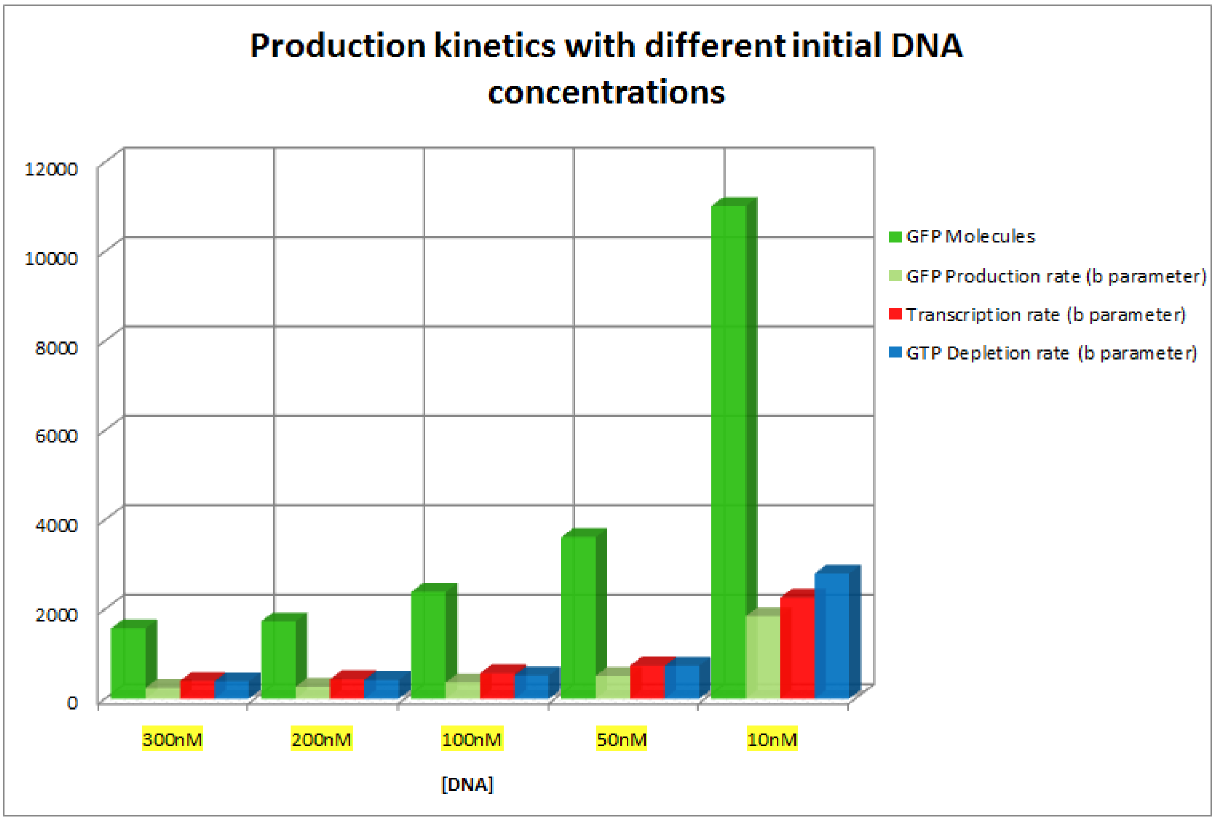


Fig. 12: Parameter comparison for simulated data using lower DNA initial amounts. Note that parameter *b* increases as the reaction kinetics become slower (see text for details). Volume 10^{-14} liters.

As DNA concentration decreases, the transcription process slows down, resulting in a minor GTP consumption (*b* parameters increases for both transcription and GTP depletion rates). GFP production kinetics are slower but the total protein yield is increased to a total of over 11.000 GFP molecules (≈ 1.8 μM).

These preliminary results about the general behavior for different composition of the PURESYSTEM showed how the determining factor which guides the protein production efficiency is the energy availability.

When DNA concentration is high, transcription produces immediately many ribosomal binding sites and the translation initiation (which is one of the rate-limiting process of protein production) is more likely to occur, resulting in a more rapid protein production kinetic compared with inputs containing lower DNA concentrations; at the same time, the transcription process consumes a large amount of nucleotides during the RNA elongation process.

Translation factors (IF1, Ef-Tu etc…) use GTP as energy donor, thus the lack of GTP molecules causes the translation process, and subsequently protein production, to stop. Nucleotide di-phosphate kinase can provide additional GTP molecule from the ATP pool, but the additional GTP is consumed by transcription or in the intermediate translation steps, unlikely resulting in the formation of a significant number of new complete proteins.

When using lower initial DNA concentrations (>100 nM) this competition effect between the transcription and translation processes for energy resources becomes a clearer phenomenon: in fact, results from simulation performed using initial DNA concentrations of 50 and 10 nM showed how the lower transcription rate resulted in a minor GTP consumption, which allowed the translation process to continue.

The total protein yield reached a value up to 1.8 μM considering a compartment of 10^{-14} liters of internal volume, which is in very good agreement with experimental measures in GFP-expressing giant lipid vesicles of equal volume [36].

Although maximizing the protein yield, low DNA concentrations determine a slow kinetic of protein production: simulation performed using even lower DNA amounts resulted in unfeasibly slow GFP productions. Furthermore, protein production using very low DNA amounts encompasses several hours of time, and self-inactivating phenomena, which probably involve ribosomes inactivation, were experimentally observed after approximately 3 hours from the beginning of the experiment [67].

Anyway, these phenomena are still not well described for the PURESYSTEM, and seems to have a significant effect only after different hours of protein production.

For what concerns the other molecular participants, aminoacids concentration never drops to zero, even for low consumables combinations, exerting in this preliminary data a role of secondary importance in the determination of protein production kinetics.

A small initial amount of enzymes resulted in slow kinetics with very low protein production, even in presence of low DNA concentration and medium energy availability; of course not all the enzymes do have the same effect of the system behavior: RNA polymerase is one of the key factor which determines the rate of GTP depletion from the system, exerting the same effect of DNA on the overall protein yield.

Obtained results suggests how a high protein production yield can be afforded when DNA concentrations are low and enzymes are present in high quantity, while many other combinations resulted in very low protein productivity.

These observations, which must still be experimentally validated, point out how vesicles do not need to encapsulate high amount of DNA to efficiently afford protein production, but that an high concentration of enzymes and nucleotides can determine a substantial rate of GFP production. More data about the entrapment phenomenon in liposomes has to be analyzed and discussed, with the aim to compare the behavior of the PURESYSTEM in small systems of different nature, such as water-in-oil droplets, which represent another interesting system for *in lipo* protein-expression studies [68].

CONCLUSIONS

For the first time, a detailed stochastic description of the transcription/translation process was given. All the real topological constraints for the elongation reactions were included inside the model, allowing for the presence of sequential processes, and then yielding a more realistic picture of the overall phenomenon.

Results obtained by the stochastic simulation experiments can surely aid studies about the entrapment phenomenon, by highlighting the presence of some "key" molecules which have a primal role in driving protein production, which can be a possible target for preferential encapsulation hypotheses.

The creation of suitable models can be of great help in assisting experimentation, allowing to impose *in silico* perturbations on the system and make well-founded and testable predictions, where most information about the internal system dynamics are not accessible by experimental approaches.

REFERENCES

1. Deamer, D.W. and R.M. Pashley, *Amphiphilic components of the murchison carbonaceous chondrite: Surface properties and membrane formation.* Origins of Life and Evolution of Biospheres, 1989. **19**(1): p. 21-38.
2. Tereza Pereira de, S., et al., *Spontaneous Crowding of Ribosomes and Proteins inside Vesicles: A Possible Mechanism for the Origin of Cell Metabolism.* ChemBioChem, 2011. **12**(2c6f5d5a-f6b1-7d40-4922-239c9858e542).
3. Stano, P., *Minimal cells: relevance and interplay of physical and biochemical factors.* Biotechnol J, 2011. **6**(e962ae53-9e07-daf0-4d04-239c9854ffc1): p. 850-859.
4. Oparin, A.I., *[The origin of life].* Nord Med, 1961. **65**: p. 693-7.
5. Klyce, B. *Panspermia asks new questions.* in *The Search for Extraterrestrial Intelligence (SETI) in the Optical Spectrum III, Stuart A. Kingsley; Ragbir Bhathal; Eds.* 2001.
6. Martins, Z., et al., *Extraterrestrial nucleobases in the Murchison meteorite.* Earth and Planetary ... 2008(7d959906-6bbe-60fd-0e31-bfb7cec027d2).
7. Luisi, P., *Autopoiesis: a review and a reappraisal.* Naturwissenschaften, 2003. **90**(40356c38-7a78-7547-8a88-bfb7cbf9443e): p. 49-108.
8. Luisi, P., F. Ferri, and P. Stano, *Approaches to semi-synthetic minimal cells: a review.* Naturwissenschaften, 2006. **93**(e975186c-6d35-e5a8-7236-239c98c5ea6f): p. 1-14.
9. Bachmann, P.A., P.L. Luisi, and J. Lang, *Autocatalytic self-replicating micelles as models for prebiotic structures.* Nature, 1992. **357**(6373): p. 57-59.
10. Angelova, M.I. and D.S. Dimitrov, *Liposome Electroformation.* Faraday Discussions, 1986. **81**: p. 303-+.
11. Walde, P., et al., *Giant vesicles: preparations and applications.* Chembiochem, 2010. **11**(7): p. 848-65.
12. Chang, T.M.S., *Artificial cells : biotechnology, nanomedicine, regenerative medicine, blood substitutes, bioencapsulation, cell/stem cell therapy.* Regenerative medicine, artificial cells and nanomedicine2007, Hackensack, N. J.: World Scientific. xxvi, 455 p.
13. Glass, J.I., et al., *Essential genes of a minimal bacterium.* Proc Natl Acad Sci U S A, 2006. **103**(2): p. 425-30.
14. Chiarugi, D., P. Degano, and R. Marangoni, *A computational approach to the functional screening of genomes.* PLoS Comput Biol, 2007. **3**(9): p. 1801-6.
15. Moore, P.B., *How Small is Small? The Minimal Cell*, P.L. Luisi and P. Stano, Editors. 2011, Springer Netherlands. p. 65-71.
16. Szostak, J.W., D.P. Bartel, and P.L. Luisi, *Synthesizing life.* Nature, 2001. **409**(6818): p. 387-90.
17. Mavelli, F., *Stochastic simulations of minimal cells: the Ribocell model.* BMC Bioinformatics, 2012. **13 Suppl 4**: p. S10.
18. Ghosh, I. and J. Chmielewski, *Peptide self-assembly as a model of proteins in the pre-genomic world.* Curr Opin Chem Biol, 2004. **8**(6): p. 640-4.

19. Walde, P., et al., *Oparin's Reactions Revisited: Enzymic Synthesis of Poly(adenylic acid) in Micelles and Self-Reproducing Vesicles*. Journal of the American Chemical Society, 1994. **116**(17): p. 7541-7547.
20. Oberholzer, T., M. Albrizio, and P.L. Luisi, *Polymerase chain reaction in liposomes*. Chem Biol, 1995. **2**(10): p. 677-82.
21. Oberholzer, T., et al., *Enzymatic RNA replication in self-reproducing vesicles: an approach to a minimal cell*. Biochem Biophys Res Commun, 1995. **207**(1): p. 250-7.
22. Kurihara, K., et al., *Self-reproduction of supramolecular giant vesicles combined with the amplification of encapsulated DNA*. Nat Chem, 2011. **3**(10): p. 775-81.
23. Luisi, P.L. and P. Stano, *Synthetic Biology Minimal Cell Mimicry*. Nat Chem, 2011. **3**(10): p. 755-756.
24. Yu, W.E.I., et al., *Synthesis of Functional Protein in Liposome*. Journal of Bioscience and Bioengineering, 2001. **92**(6): p. 590-593.
25. Stano, P., et al., *Semi-synthetic minimal cells as a tool for biochemical ICT*. Biosystems, 2012. **109**(1): p. 24-34.
26. Harris, D.C. and M.C. Jewett, *Cell-free biology: Exploiting the interface between synthetic biology and synthetic chemistry*. Curr Opin Biotechnol, 2012.
27. Yoshihiro, S., K. Takashi, and U. Takuya, *Protein synthesis by pure translation systems*. Methods, 2005. **36**(d40739d9-3635-04de-3a97-239c98cd4d3b).
28. Shimizu, Y., et al., *Cell-free translation reconstituted with purified components*. Nature ... 2001(28f9cb18-d3a3-328d-ae54-239c98cdc21d).
29. Murtas, G., et al., *Protein synthesis in liposomes with a minimal set of enzymes*. Biochemical and biophysical research communications, 2007. **363**(fbad069c-dfeb-d429-435f-bfb7cbf9940c): p. 12-19.
30. Sunami, T., et al., *Femtoliter compartment in liposomes for in vitro selection of proteins*. Anal Biochem, 2006. **357**(1): p. 128-36.
31. Tereza Pereira de, S., S. Pasquale, and L. Pier Luigi, *The Minimal Size of Liposome-Based Model Cells Brings about a Remarkably Enhanced Entrapment and Protein Synthesis*. ChemBioChem, 2009. **10**(aaece635-437c-a62d-75d9-239c9855d981).
32. Kita, H., et al., *Replication of genetic information with self-encoded replicase in liposomes*. Chembiochem, 2008. **9**(15): p. 2403-10.
33. Ichihashi, N., et al., *Constructing Partial Models of Cells*. Cold Spring Harbor Perspectives in Biology, 2010. **2**(545801e0-2890-7dc0-79b3-239c9c5b1a5f).
34. Yutetsu, K., et al., *A synthetic biology approach to the construction of membrane proteins in semi-synthetic minimal cells*. Biochimica et Biophysica Acta (BBA) - Biomembranes, 2009. **1788**(7e058bce-3e92-8697-11fb-94e985abf6d1).
35. Gil, R., et al., *Determination of the core of a minimal bacterial gene set*. Microbiol Mol Biol Rev, 2004. **68**(3): p. 518-37.
36. Sunami, T., et al., *Cellular compartment model for exploring the effect of the lipidic membrane on the kinetics of encapsulated biochemical reactions*. Langmuir : the ACS journal of surfaces and colloids, 2010. **26**(a3f9f92a-b3c3-7067-e1aa-239c9c5d8d41): p. 8544-8595.
37. Boukobza, E., A. Sonnenfeld, and G. Haran, *Immobilization in surface-tethered lipid vesicles as a new tool for single biomolecule spectroscopy*. The

Journal of Physical Chemistry B, 2001. **105**(3b4af02e-7be1-0076-c34f-239c984a9f1d): p. 12165-24335.

38. Heider, E.C., et al., *Quantitative fluorescence microscopy to determine molecular occupancy of phospholipid vesicles*. Anal Chem, 2011. **83**(13): p. 5128-36.
39. Berclaz, N., et al., *Growth and transformation of vesicles studied by ferritin labeling and cryotransmission electron microscopy*. Journal of Physical Chemistry B, 2001. **105**(5): p. 1056-1064.
40. Luisi, P., et al., *Spontaneous protein crowding in liposomes: a new vista for the origin of cellular metabolism*. Chembiochem : a European journal of chemical biology, 2010. **11**(c79f048f-358f-39cb-316a-239c9a1ba949): p. 1989-2081.
41. Kulkarni, S.B., G.V. Betageri, and M. Singh, *Factors affecting microencapsulation of drugs in liposomes*. Journal of Microencapsulation, 1995. **12**(3): p. 229-246.
42. Dominak, L., et al., *Polymeric crowding agents improve passive biomacromolecule encapsulation in lipid vesicles*. Langmuir : the ACS journal of surfaces and colloids, 2010. **26**(49c22ebc-42d2-84b5-b8ff-239c9a1da180): p. 13195-13395.
43. Mitzenmacher, M., *A brief history of generative models for power law and lognormal distributions*.
44. Newman, M., *Power laws, Pareto distributions and Zipf's law*. Contemporary physics, 2005(c47dc421-100e-e02b-a389-239c9850b817).
45. Sornette, D., *Critical phenomena in natural sciences : chaos, fractals, selforganization, and disorder : concepts and tools*. 2nd ed. Springer series in synergetics,2006, Berlin ; New York: Springer. xxii, 528 p.
46. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks*. Science, 1999. **286**(5439): p. 509-12.
47. Drossel, B. and F. Schwabl, *Self-organized critical forest-fire model*. Physical Review Letters, 1992. **69**(11): p. 1629-1632.
48. Yanagida, T., et al., *Brownian motion, fluctuation and life*. Biosystems, 2007. **88**(3): p. 228-42.
49. Takano, M., T.P. Terada, and M. Sasai, *Unidirectional Brownian motion observed in an in silico single molecule experiment of an actomyosin motor*. Proc Natl Acad Sci U S A, 2010. **107**(17): p. 7769-74.
50. Mitzenmacher, M., *A brief history of generative models for power law and lognormal distributions*. Internet mathematics, 2004(9f31ccc5-c3ee-be24-1d66-239c9a1dcd22).
51. Hao, Y.-M. and K.a. Li, *Entrapment and release difference resulting from hydrogen bonding interactions in niosome*. International journal of pharmaceutics, 2011. **403**(17335335-12c2-bf93-3bd9-239c9a1cab1c): p. 245-298.
52. Lohse, B. and P. Bolinger, *Encapsulation efficiency measured on single small unilamellar vesicles*. Journal of the American ... 2008(753c46ce-1ab6-111e-e9bf-239c9a1b079c).
53. Sun, B. and D. Chiu, *Determination of the encapsulation efficiency of individual vesicles using single-vesicle photolysis and confocal single-molecule detection*. Analytical chemistry, 2005. **77**(0a3ad347-8bdb-7f29-2e6a-239c9a2046a6): p. 2770-2776.

54. Klipp, E., *Systems Biology: a Textbook*2009: Wiley-Blackwell.

55. Gillespie, D.T., *Exact Stochastic Simulation of Coupled Chemical-Reactions*. Journal of Physical Chemistry, 1977. **81**(25): p. 2340-2361.

56. Gillespie, D.T., *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*. Journal of Computational Physics, 1976. **22**(4): p. 403-434.

57. Gibson, M.A. and J. Bruck, *Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels*. The Journal of Physical Chemistry A, 2000. **104**(9): p. 1876-1889.

58. Cangelosi, D., *On Improving Stochastic Simulation for Systems Biology*, 2010, University of Pisa.

59. Roussel, M.R. and R. Zhu, *Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression*. Physical Biology, 2006. **3**(4): p. 274-284.

60. Gillespie, D.T., *Approximate accelerated stochastic simulation of chemically reacting systems*. Journal of Chemical Physics, 2001. **115**(4): p. 1716-1733.

61. Cao, Y., D.T. Gillespie, and L.R. Petzold, *Efficient step size selection for the tau-leaping simulation method*. Journal of Chemical Physics, 2006. **124**(4).

62. Gillespie, D.T., *Chemical Langevin equation*. Journal of Chemical Physics, 2000. **113**(1): p. 297-306.

63. Lazzerini-Ospri, L., et al., *Characterization of the emergent properties of a synthetic quasi-cellular system*. BMC Bioinformatics, 2012. **13 Suppl 4**: p. S9.

64. Kim, J.H. and R.G. Larson, *Single-molecule analysis of 1D diffusion and transcription elongation of T7 RNA polymerase along individual stretched DNA molecules*. Nucleic Acids Res, 2007. **35**(11): p. 3848-58.

65. Nichola, D.R., *"Determining the Rate of Transcription of T7 RNA Polymerase Using Single Molecule Fluorescence Imaging"*, 2010, Marshall University.

66. Mäkelä, J., et al., *Stochastic sequence-level model of coupled transcription and translation in prokaryotes*. BMC Bioinformatics, 2011. **12**(79563a14-1272-feb7-481d-85b1f894f09e): p. 121.

67. Tobias, S., et al., *Experiment and mathematical modeling of gene expression dynamics in a cell-free system*. Integrative Biology, 2012. **4**(b372eb93-1286-890c-a0d2-94e985aa0a47).

68. Kato, A., et al., *Cell-Sized confinement in microspheres accelerates the reaction of gene expression*. Sci Rep, 2012. **2**: p. 283.

APPENDIX

PURESYSTEM model with its original composition, volume 10^{-16} liters (≈ 570 nm vesicle):

```
DNA0, DNA1, DNA2, DNA3,
DNA4, DNA5, DNA6, DNA7,
DNA8, T7preEL, T7pregELa,
T7pregELb, T7pregELc, T7, T7IN,
T7EL, T7EL2, T7EL3, T7EL4,
T7EL5, T7EL6, T7EL7, T7EL8,
T7EL9, T7preaEL,  T7preaELa,
T7preaELb,  T7preaELc,
T7preaEL2, T7preaEL3,
T7preaEL4, T7preaEL5,
T7preaEL6, T7preaEL7,
T7preaEL8, T7preaEL9, T7pregEL,
T7pregELa, T7pregELb, T7pregELc,
T7pregEL2, T7pregEL3,
T7pregEL4, T7pregEL5,
T7pregEL6, T7pregEL7,
T7pregEL8, T7pregEL9, T7ELa,
T7ELb, T7ELc, gtr1, gtr1a, gtr1b,
gtr1c, gtr2, gtr3, gtr4, gtr5, gtr6,
gtr7, gtr8, gtr9, atr1, atr1a, atr1b,
atr1c, atr2, atr3, atr4, atr5, atr6,
atr7, atr8, atr9, TRANSL1,
TRANSL2, TRANSL3, TRANSL4,
TRANSL5, TRANSL6, TRANSL7,
TRANSL8, TRANSL9, RNATOT,
PEPT0, PEPT1, PEPT2, PEPT3,
PEPT4, PEPT5, PEPT6, PEPT7,
PEPT8, RNA1, RNA2, RNA3,
RNA4, RNA5, RNA6, RNA7, RNA8,
RNA9, DNA, BS, R, RBS, RBS1,
RBS2, RBS3, ATP, GTP, ADP, GDP,
AMP, PPi, Pi, IF1, IF2, IF3,
IF2GTP, IX, EFts, EFtu, EFtuGTP,
EFtuGDP, EFg, EFgGTP,
EFaRGTP, EFX, EX0, EX1, EX2,
EX3, EX4, EX5, EX6, EX7, EX8,
EX9, RF, RFGTP, TX, PROT,
TRANSL0, RNAEND, preeR1, eR0,
eR1, eR2, eR3, eR4, eR5, eR6, eR7,
eR8, eR9, eRSTOP, eRa0, eRa1,
eRa2, eRa3, eRa4, eRa5, eRa6,
eRa7, eRa8, eRa9, tRNA, aa, Syn,
aaSyn, aaAMPSyn, aatRNA, AX,
AX2, PPase, AK, AKP, NDK, NDKP,
CK, CKP, CrP, Cr, Q, W, Z, Y, D, H,
C, N, J, K, L, M, Met, MSyn,
aaMSyn, aaAMPMSyn, MtRNA,
MetRNA, AMX, AMX2, MTF,
MetMTF, FTHF, MetFMTF,
fMetRNA, THF, EXb0, EXb1,
EXb2, EXb3, EXb4, EXb5, EXb6,
EXb7, EXb8, EXb9, tuts, EFY

volume, 0.0000000000000001

10000000, DNA0 + T7 > T7IN
2.9, T7IN > DNA0 + T7
-, DNA > DNA0 + DNA1 + DNA2 +
DNA3 + DNA4 + DNA5 + DNA6 +
DNA7 + DNA8
0.36, T7IN > T7preEL
-, T7preEL + DNA1 > T7EL
1000000, T7EL + GTP > T7pregEL
28, T7pregEL > T7EL + Pi + gtr1
1000000, T7EL + ATP > T7preaEL
28, T7preaEL > T7EL + Pi + atr1
-, 5 gtr1 + 5 atr1 + T7EL > T7ELa
1000000, T7ELa + GTP >
T7pregELa
28, T7pregELa > T7ELa + Pi + gtr1a
1000000, T7ELa + ATP >
T7preaELa
28, T7preaELa > T7ELa + Pi + atr1a
-, 5 gtr1a + 5 atr1a + T7ELa >
T7ELb
1000000, T7ELb + GTP >
T7pregELb
28, T7pregELb > T7ELb + Pi +
gtr1b
1000000, T7ELb + ATP >
T7preaELb
28, T7preaELb > T7ELb + Pi +
atr1b
-, 5 gtr1b + 5 atr1b + T7ELb >
T7ELc
1000000, T7ELc + GTP >
T7pregELc
28, T7pregELc > T7ELc + Pi + gtr1c
1000000, T7ELc + ATP >
T7preaELc
28, T7preaELc > T7ELc + Pi + atr1c
-, 5 gtr1c + 5 atr1c + T7ELc + DNA2
> T7EL2 + DNA0 + BS + RNA1
1000000, T7EL2 + GTP >
T7pregEL2
28, T7pregEL2 > T7EL2 + Pi + gtr2
1000000, T7EL2 + ATP >
T7preaEL2
28, T7preaEL2 > T7EL2 + Pi + atr2
-, 20 gtr2 + 20 atr2 + T7EL2 +
DNA3 > T7EL3 + RNA2 + DNA1
1000000, T7EL3 + GTP >
T7pregEL3
28, T7pregEL3 > T7EL3 + Pi + gtr3
1000000, T7EL3 + ATP >
T7preaEL3
28, T7preaEL3 > T7EL3 + Pi + atr3
-, 20 gtr3 + 20 atr3 + T7EL3 +
DNA4 > T7EL4 + RNA3 + DNA2
1000000, T7EL4 + GTP >
T7pregEL4
28, T7pregEL4 > T7EL4 + Pi + gtr4
1000000, T7EL4 + ATP >
T7preaEL4
28, T7preaEL4 > T7EL4 + Pi + atr4
-, 20 gtr4 + 20 atr4 + T7EL4 +
DNA5 > T7EL5 + RNA4 + DNA3
1000000, T7EL5 + GTP >
T7pregEL5
```

```
28, T7pregEL5 > T7EL5 + Pi + gtr5
1000000, T7EL5 + ATP > T7preaEL5
28, T7preaEL5 > T7EL5 + Pi + atr5
-, 20 gtr5 + 20 atr5 + T7EL5 + DNA6 > T7EL6 + RNA5 + DNA4
1000000, T7EL6 + GTP > T7pregEL6
28, T7pregEL6 > T7EL6 + Pi + gtr6
1000000, T7EL6 + ATP > T7preaEL6
28, T7preaEL6 > T7EL6 + Pi + atr6
-, 20 gtr6 + 20 atr6 + T7EL6 + DNA7 > T7EL7 + RNA6 + DNA5
1000000, T7EL7 + GTP > T7pregEL7
28, T7pregEL7 > T7EL7 + Pi + gtr7
1000000, T7EL7 + ATP > T7preaEL7
28, T7preaEL7 > T7EL7 + Pi + atr7
-, 20 gtr7 + 20 atr7 + T7EL7 + DNA8 > T7EL8 + RNA7 + DNA6
1000000, T7EL8 + GTP > T7pregEL8
28, T7pregEL8 > T7EL8 + Pi + gtr8
1000000, T7EL8 + ATP > T7preaEL8
28, T7preaEL8 > T7EL8 + Pi + atr8
-, 20 gtr8 + 20 atr8 + T7EL8 > T7EL9 + RNA8 + DNA7
1000000, T7EL9 + GTP > T7pregEL9
28, T7pregEL9 > T7EL9 + Pi + gtr9
1000000, T7EL9 + ATP > T7preaEL9
28, T7preaEL9 > T7EL9 + Pi + atr9
-, 20 gtr9 + 20 atr9 + T7EL9 > T7 + RNA9 + RNATOT + RNAEND + DNA8
1000000, aa + Syn > aaSyn
1, aaSyn > aa + Syn
1000000, aaSyn + ATP > AX
0.1, AX > aaSyn + ATP
100, AX > aaAMPSyn + AMP + PPi
1000000, aaAMPSyn + tRNA > AX2
0.01, AX2 > aaAMPSyn + tRNA
0.84, AX2 > aatRNA + Syn
1000000, aatRNA + EFtuGTP > EFaRGTP
1, EFaRGTP > aatRNA + EFtuGTP
5000000, eR0 + EFaRGTP > EX0
0.2, EX0 > eR0 + EFaRGTP
30, EX0 > eRa0 + EFtuGDP + Pi
5000000, eRa0 + EFgGTP > EXb0
0.3, EXb0 > eRa0 + EFgGTP
30, EXb0 > EFg + GDP + Pi + preeR1 + MtRNA + PEPT0 + TRANSL0
-, TRANSL0 + RNA2 + preeR1 > eR1
5000000, eR1 + EFaRGTP > EX1
0.2, EX1 > eR1 + EFaRGTP
30, EX1 > eRa1 + EFtuGDP + Pi
5000000, eRa1 + EFgGTP > EXb1
0.3, EXb1 > eRa1 + EFgGTP
30, EXb1 > EFg + GDP + Pi + eR1 + tRNA + TRANSL1
-, 27 TRANSL1 + PEPT0 + RNA3 + eR1 > eR2 + PEPT1 + BS + RNA1
5000000, eR2 + EFaRGTP > EX2
0.2, EX2 > eR2 + EFaRGTP
30, EX2 > eRa2 + EFtuGDP + Pi
5000000, eRa2 + EFgGTP > EXb2
0.3, EXb2 > eRa2 + EFgGTP
30, EXb2 > EFg + GDP + Pi + eR2 + tRNA + TRANSL2
-, -, 27 TRANSL2 + PEPT1 + RNA4 + eR2 > eR3 + PEPT2 + RNA2
5000000, eR3 + EFaRGTP > EX3
0.2, EX3 > eR3 + EFaRGTP
30, EX3 > eRa3 + EFtuGDP + Pi
5000000, eRa3 + EFgGTP > EXb3
0.3, EXb3 > eRa3 + EFgGTP
30, EXb3 > EFg + GDP + Pi + eR3 + tRNA + TRANSL3
-, 27 TRANSL3 + PEPT2 + RNA5 + eR3 > eR4 + PEPT3 + RNA3
5000000, eR4 + EFaRGTP > EX4
0.2, EX4 > eR4 + EFaRGTP
30, EX4 > eRa4 + EFtuGDP + Pi
5000000, eRa4 + EFgGTP > EXb4
0.3, EXb4 > eRa4 + EFgGTP
30, EXb4 > EFg + GDP + Pi + eR4 + tRNA + TRANSL4
-, 27 TRANSL4 + RNA6 + PEPT3 + eR4 > eR5 + PEPT4 + RNA3
5000000, eR5 + EFaRGTP > EX5
0.2, EX5 > eR5 + EFaRGTP
30, EX5 > eRa5 + EFtuGDP + Pi
5000000, eRa5 + EFgGTP > EXb5
0.3, EXb5 > eRa5 + EFgGTP
30, EXb5 > EFg + GDP + Pi + eR5 + tRNA + TRANSL5
-, 27 TRANSL5 + RNA7 + PEPT4 + eR5 > eR6 + PEPT5 + RNA5
5000000, eR6 + EFaRGTP > EX6
0.2, EX6 > eR6 + EFaRGTP
30, EX6 > eRa6 + EFtuGDP + Pi
5000000, eRa6 + EFgGTP > EXb6
0.3, EXb6 > eRa6 + EFgGTP
30, EXb6 > EFg + GDP + Pi + eR6 + tRNA + TRANSL6
-, 27 TRANSL6 + RNA8 + PEPT5 + eR6 > eR7 + PEPT6 + RNA6
5000000, eR7 + EFaRGTP > EX7
0.2, EX7 > eR7 + EFaRGTP
30, EX7 > eRa7 + EFtuGDP + Pi
5000000, eRa7 + EFgGTP > EXb7
0.3, EXb7 > eRa7 + EFgGTP
30, EXb7 > EFg + GDP + Pi + eR7 + tRNA + TRANSL7
-, 27 TRANSL7 + RNA9 + PEPT6 + eR7 > eR8 + PEPT7 + RNA7
5000000, eR8 + EFaRGTP > EX8
0.2, EX8 > eR8 + EFaRGTP
30, EX8 > eRa8 + EFtuGDP + Pi
5000000, eRa8 + EFgGTP > EXb8
0.3, EXb8 > eRa8 + EFgGTP
30, EXb8 > EFg + GDP + Pi + eR8 + tRNA + TRANSL8
-, 27 TRANSL8 + RNA9 + PEPT7 + RNAEND +eR8 > eR9 + PEPT8 + RNA7
5000000, eR9 + EFaRGTP > EX9
0.2, EX9 > eR9 + EFaRGTP
30, EX9 > eRa9 + EFtuGDP + Pi
5000000, eRa9 + EFgGTP > EXb9
0.3, EXb9 > eRa9 + EFgGTP
30, EXb9 > EFg + GDP + Pi + eR9 + tRNA + TRANSL9
-, 27 TRANSL9 + eR9 + PEPT8 > eRSTOP + RNA8 + RNA9
0.18, IX > eR0 + IF1 + IF2 + IF3 + GDP + Pi
100000, eRSTOP + RFGTP > TX
1, TX > eRSTOP + RFGTP
10, TX > R + RF + GDP + Pi + PROT + tRNA + RNAEND
1000000, RBS + fMetRNA > IX
```

```
0.23, IX > RBS + fMetRNA
100000, IF2 + GTP > IF2GTP
1.8, IF2GTP > IF2 + GTP
100000, IF1 + R > RBS1
1, RBS1 > IF1 + R
100000, RBS1 + IF3 > RBS3
1, RBS3 > IF3 + RBS1
100000, RBS3 + IF2GTP > RBS2
1, RBS2 > IF2GTP + RBS3
100000, RBS2 + BS > RBS
1, RBS > RBS2 + BS
50000, EFtu + GTP > EFtuGTP
200000, EFtu + GDP > EFtuGDP
1000000, EFtu + EFts > tuts
0.002, EFtuGDP > EFtu + GDP
10000000, EFtuGDP + EFts > EFX
400, EFX > EFtuGDP + EFts
175, EFX > tuts + GDP
0.01, tuts > EFtu + EFts
1000000, tuts + GDP > EFX
400000, tuts + GTP > EFY
90, EFY > tuts + GTP
60, EFY > EFtuGTP + EFts
10000000, EFtuGTP + EFts > EFY
0.01, EFtuGTP > EFtu + GTP
10000, EFg + GTP > EFgGTP
1.8, EFgGTP > EFg + GTP
100000, RF + GTP > RFGTP
1.8, RFGTP > RF + GTP
100, PPase + PPi > PPase + Pi + Pi
100000, AK + ATP > Y
40, Y > AK + ATP
120, Y > AKP + ADP
100000, AK + ADP > Z
90, Z > AK + ADP
120, Z > AKP + AMP
100000, AKP + AMP > W
40, W > AKP + AMP
120, W > AK + ADP
100000, AKP + ADP > Q
90, Q > AKP + ADP
120, Q > AK + ATP
100000, NDK + ATP > D
60, D > NDK + ATP
60, D > NDKP + ADP
100000, NDK + GTP > H
60, H > NDK + GTP
60, H > NDKP + GDP
100000, NDKP + ADP > C
60, C > NDKP + ADP
60, C > NDK + ATP
100000, NDKP + GDP > N
100, N > NDKP + GDP
60, N > NDK + GTP
100000, CK + ATP > K
1, K > CK + ATP
120, K > CKP + ADP
100000, CKP + ADP > J
0.1, J > CKP + ADP
480, J > CK + ATP
100000, CK + CrP > L
20, L > CK + CrP
480, L > CKP + Cr
100000, Cr + CKP > M
20, M > Cr + CKP
150, M > CK + CrP
100000, Met + MSyn > aaMSyn
0.9, aaMSyn > Met + MSyn
100000, aaMSyn + ATP > AMX
0.1, AMX > aaMSyn + ATP
100, AMX > aaAMPMSyn + AMP + PPi
100000, aaAMPMSyn + MtRNA > AMX2
0.01, AMX2 > aaAMPMSyn + MtRNA
2.5, AMX2 > MetRNA + MSyn
100000, MetRNA + MTF > MetMTF
0.01, MetMTF > MetRNA + MTF
100000, MetMTF + FTHF > MetFMTF
1, MetFMTF > MetMTF + FTHF
1.3, MetFMTF > fMetRNA + THF + MTF

DNA0, 0, 0
DNA1, 0, 0
DNA2, 0, 0
DNA3, 0, 0
DNA4, 0, 0
DNA5, 0, 0
DNA6, 0, 0
DNA7, 0, 0
DNA8, 0, 0
T7pregELa, 0, 0
T7pregELb, 0, 0
T7pregELc, 0, 0
T7preaELa, 0, 0
T7preaELb, 0, 0
T7preaELc, 0, 0
T7ELa, 0, 0
T7ELb, 0, 0
T7ELc, 0, 0
gtr1, 0, 0
atr1, 0, 0
gtr1a, 0, 0
atr1a, 0, 0
gtr1b, 0, 0
atr1b, 0, 0
gtr1c, 0, 0
atr1c, 0, 0
gtr2, 0, 0
atr2, 0, 0
gtr3, 0, 0
atr3, 0, 0
gtr4, 0, 0
atr4, 0, 0
gtr5, 0, 0
atr5, 0, 0
gtr6, 0, 0
atr6, 0, 0
gtr7, 0, 0
atr7, 0, 0
gtr8, 0, 0
atr8, 0, 0
gtr9, 0, 0
atr9, 0, 0
TRANSL0, 0, 0
TRANSL1, 0, 0
TRANSL2, 0, 0
TRANSL3, 0, 0
TRANSL4, 0, 0
TRANSL5, 0, 0
TRANSL6, 0, 0
TRANSL7, 0, 0
TRANSL8, 0, 0
TRANSL9, 0, 0
preeR1, 0, 0
eR0, 0, 0
eR1, 0, 0
```

eR2, 0, 0
eR3, 0, 0
eR4, 0, 0
eR5, 0, 0
eR6, 0, 0
eR7, 0, 0
eR8, 0, 0
eR9, 0, 0
eRa0, 0, 0
eRa1, 0, 0
eRa2, 0, 0
eRa3, 0, 0
eRa4, 0, 0
eRa5, 0, 0
eRa6, 0, 0
eRa7, 0, 0
eRa8, 0, 0
eRa9, 0, 0
EX0, 0, 0
EX1, 0, 0
EX2, 0, 0
EX3, 0, 0
EX4, 0, 0
EX5, 0, 0
EX6, 0, 0
EX7, 0, 0
EX8, 0, 0
EX9, 0, 0
EXb0, 0, 0
EXb1, 0, 0
EXb2, 0, 0
EXb3, 0, 0
EXb4, 0, 0
EXb5, 0, 0
EXb6, 0, 0
EXb7, 0, 0
EXb8, 0, 0
EXb9, 0, 0
RNATOT, 0, 0
T7EL, 0, 0
T7ELa, 0, 0
T7ELb, 0, 0
T7ELc, 0, 0
T7EL2, 0, 0
T7EL3, 0, 0
T7EL4, 0, 0
T7EL5, 0, 0
T7EL6, 0, 0
T7EL7, 0, 0
T7EL8, 0, 0
T7EL9, 0, 0
T7pregEL, 0, 0
T7pregELa, 0, 0
T7pregELb, 0, 0
T7pregELc, 0, 0
T7pregEL2, 0, 0
T7pregEL3, 0, 0
T7pregEL4, 0, 0
T7pregEL5, 0, 0
T7pregEL6, 0, 0
T7pregEL7, 0, 0
T7pregEL8, 0, 0
T7pregEL9, 0, 0
T7preaEL, 0, 0
T7preaELa, 0, 0
T7preaELb, 0, 0
T7preaELc, 0, 0
T7preaEL2, 0, 0
T7preaEL3, 0, 0
T7preaEL4, 0, 0
T7preaEL5, 0, 0
T7preaEL6, 0, 0
T7preaEL7, 0, 0
T7preaEL8, 0, 0
T7preaEL9, 0, 0
PEPT0, 0, 0
PEPT1, 0, 0
PEPT2, 0, 0
PEPT3, 0, 0
PEPT4, 0, 0
PEPT5, 0, 0
PEPT6, 0, 0
PEPT7, 0, 0
PEPT8, 0, 0
RNA1, 0, 0
RNA2, 0, 0
RNA3, 0, 0
RNA4, 0, 0
RNA5, 0, 0
RNA6, 0, 0
RNA7, 0, 0
RNA8, 0, 0
RNA9, 0, 0
RNAEND, 0, 0
DNA, 0, 20
T7, 0, 6
T7IN, 0, 0
BS, 0, 0
R, 0, 72
RBS, 0, 0
RBS1, 0, 0
RBS2, 0, 0
RBS3, 0, 0
Cr, 0, 0
ADP, 0, 0
GDP, 0, 0
AMP, 0, 0
PPi, 0, 0
Pi, 0, 0
IF1, 0, 163
IF2, 0, 24
IF3, 0, 90
IF2GTP, 0, 0
IX, 0, 0
EFts, 0, 39
EFtu, 0, 55
EFtuGTP, 0, 0
EFtuGDP, 0, 0
EFg, 0, 15
EFgGTP, 0, 0
EFaRGTP, 0, 0
EFX, 0, 0
RF, 0, 17
RFGTP, 0, 0
TX, 0, 0
PROT, 0, 0
Syn, 0, 54
aaSyn, 0, 0
aaAMPSyn, 0, 0
aatRNA, 0, 0
AX, 0, 0
AX2, 0, 0
PPase, 0, 5
AK, 0, 6
AKP, 0, 0
NDK, 0, 1
NDKP, 0, 0
CK, 0, 3
MSyn, 0, 3
aaMSyn, 0, 0
MTF, 0, 4
MetMTF, 0, 0
CKP, 0, 0
eRSTOP, 0, 0
ATP, 0, 120440
GTP, 0, 120440
CrP, 0, 1204400
tRNA, 0, 5395
aa, 0, 343254
Met, 0, 18066

aaAMPMSyn, 0, 0
MtRNA, 0, 135
MetRNA, 0, 0
AMX, 0, 0
AMX2, 0, 0
FTHF, 0, 1271966
Q, 0, 0
W, 0, 0
Z, 0, 0
Y, 0, 0
D, 0, 0
H, 0, 0
C, 0, 0
N, 0, 0
J, 0, 0
K, 0, 0
L, 0, 0
M, 0, 0
MetFMTF, 0, 0
fMetRNA, 0, 0
THF, 0, 0
tuts, 0, 0
EFY, 0, 0

PURESYSTEM reagents list, with a quick description of their biological counterparts:

DNA --> DNA sequence
T7 --> T7 RNA Polymerase
T7IN --> T7 RNA Polymerase bound with DNA
T7pre(g/a)EL --> T7 RNA Polymerase with bound ATP or GTP
T7EL --> Elongating T7 RNA Polymerase
(g/a)tr --> Fictitious species used to regulate transition steps during transcription
TRANSL --> Fictitious species used to regulate transition steps during translation
eR,, eRa, EX ,EXb --> Elongating ribosome at different stages (binding aminoacids, elongation factors etc..)
RNA --> RNA sequence (RNATOT --> full length RNA sequence)
PEPT --> Fictitious species used to track peptide elongation stages
IF1, IF2, IF3 --> Translation Initiation Factors
R --> Free ribosome
RBS, RBS1, RBS2, RBS3 --> Pre-initiation complex
BS --> Ribosome Binding Site (on RNA)
aatRNA --> Free aminoacyl-tRNA
IX --> Initiation Complex
EFtu, EFts, EFg --> Translation Elongation Factors
EFX --> Guanosine exchange reaction intermediate
EFaRGTP --> Ternary complex (aminoacyl-tRNA + EF-Tu + GTP)
tRNA --> free tRNA
eRSTOP --> Ribosome with Stop codon in A-site
TX --> Translation termination complex
RF --> Ribosome Release Factors
PROT --> Complete protein
aa --> Free aminoacids
Syn --> Aminoacyl-tRNA synthetases
aaSyn --> Charging intermediate (aminoacid bound to aminoacyl-tRNA synthetase)
AX --> Charging intermediate (ATP bound to aminoacyl-tRNA synthetase)
aaAMPSyn --> Aminoacylation intermediate (AMP , aminoacid bound to aminoacyl-tRNA synthetase)
AX2 --> Aminoacylation intermediate (tRNA bound to aminoacyl-tRNA synthetase)
PPase --> Pyrophosphatase
AK(P) --> Adenylate kinase (phosphorylated)

NDK(P) --> Nucleoside-diphosphate kinase (phosphorylated)

CK(P) --> Creatine kinase (phosphorylated)

Cr(P) --> Creatine ((phosphorylated)

Met, --> Free methionine

MSyn, --> Methionyl-tRNA synthetase

aaMSyn, --> Charging intermediate

aaAMPMSyn, --> Aminoacylation intermediate

MtRNA, --> Free Met-anticodon tRNA

MetRNA, --> Free Met-acylated tRNA

AMX, --> Charging intermediate

AMX2, --> Aminoacylation intermediate

MTF, --> Methionyl-tRNA formyltransferase

MetMTF, --> Formylation intermediate

FTHF, --> 10-formyltetrahydrofolate

MetFMTF, --> Formylation intermediate

fMetRNA, --> Free formylmethionyl-tRNA

ATP, ADP, AMP --> Adenosine tri-, di-, mono- phosphate

GTP, GDP --> Guanosine tri-, di- phosphate

THF, --> Tetrahydrofolate

Pi, PPi --> Inorganic phosphate, pyrophosphate

Q, W, Z, Y, D, H, C, N, J, K, L, M   --> Energy recycling intermediates