Université Sorbonne Paris Cité

Thèse de doctorat de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole Doctorale Frontières du Vivant (ED 474)

Laboratoire de Biochimie, UMR 8231 Chimie Biologie Innovation, ESPCI Paris

Pre-evolutionary dynamics in autocatalytic RNA networks

Par Simon Arsène

Thèse de doctorat de biologie des systèmes

Dirigée par Philippe Nghe et Andrew Griffiths

Présentée et soutenue publiquement à Paris le 12 octobre 2018

Président du jury : Pascal Hersen, Directeur de Recherche, Université Paris Diderot

Rapporteurs : Andres Jäschke, Professeur, Universität Heidelberg

Michael Ryckelynck, Maître de Conférences, Université de Strasbourg

Examinateurs : Catherine Isel-Griffiths, Chargé de Recherche, Institut Pasteur

Hervé Isambert, Directeur de Recherche, Institut Curie

Directeurs de thèse : Philippe Nghe, Maître de Conférences, ESPCI Paris

Andrew Griffiths, Professeur, ESPCI Paris

Content

Résumé

Titre : Dynamique pré-évolutive des réseaux ARN autocatalytiques

Les réseaux de molécules interdépendantes sont depuis quelque temps considérés comme de potentiels candidats pour avoir amorcé la transition de la biologie à la chimie. Bien qu'ils aient été intensivement examinés en théorie, il n'existe toujours aucune preuve expérimentale pour confirmer ou infirmer leur supposé rôle crucial dans les origines de la vie. En particulier, il nous manque encore une démonstration empirique des trois ingrédients habituellement présentés comme requis pour l'évolution darwinienne: l'hérédité, la variation et la sélection. Un système qui posséderait les trois tout en étant couplé à un processus de réplication en compartiments serait théoriquement capable d'évoluer au sens darwinien du terme. Par exemple, cela a été montré théoriquement pour les Ensembles Collectivement Autocatalytiques (CAS pour Collectively Aucatalytic Sets en anglais) où chaque molécule de l'ensemble est formée catalytiquement par un autre membre de l'ensemble. Ici, nous utilisons le système de ribozyme *Azoarcus*, qui catalysent des réactions de recombinaisons, pour former expérimentalement des CASs structurellement divers afin d'explorer leurs propriétés évolutives. Dans ce système, les ribozymes peuvent catalyser la formation d'autres ribozymes à partir de fragments plus petits, présents dans l'environnement. Nous utilisons un dispositif de microfluidique en gouttes associé au séquençage haut-débit pour mener une étude à grande échelle sur des milliers de CASs *Azoarcus*. Nous développons une approche perturbative pour identifier les paramètres topologiques importants contrôlant les variations observées dans les CAS à la suite de perturbations de l'environnement, ici l'ajout d'une nouvelle espèce. Nous déterminons ensuite l'ensemble restreint de caractéristiques du réseau régissant la mémoire des conditions initiales dans les CASs *Azoarcus*, un prérequis pour l'hérédité, en utilisant un modèle théorique validé par des données expérimentales. Enfin, nous démontrons qu'il existe dans les CASs *Azoarcus* des processus cataboliques qui les rendent robustes aux perturbations des fragments qui composent leur substrat et donc plus pertinent d'un point de vue prébiotique. Ces résultats démontrent le rôle crucial des CASs à base d'ARN dans les origines de la vie et illustrent comment la structure de leur réseau peut être adaptée pour obtenir des CASs avec des propriétés intéressantes d'un point de vue évolutif, ouvrant la voie à une démonstration expérimentale de l'évolution darwinienne avec système purement moléculaire.

Mots clés : origines de la vie, monde ARN, autocatalyse, ribozyme, réseaux prébiotiques

Abstract

Title: Pre-evolutionary dynamics in autocatalytic RNA networks

Networks of interdependent molecules are considered plausible candidates for initiating the transition from biology to chemistry. Though they have been intensively scrutinized theoretically, there is still no experimental evidence for confirming or denying their supposed crucial role in the origins of life. In particular, we are still lacking experimental proofs of any of the three ingredients usually presented as required for Darwinian evolution: heredity, variation and selection. A system that would possess the three while being coupled to some sort of encapsulated replication process would theoretically be able to undergo Darwinian evolution. As a matter of fact, this has been shown theoretically for Collectively Autocatalytic Sets (CAS) where each molecule of the set is catalytically formed by another member of the ensemble. Here we use the *Azoarcus* recombination ribozyme system to experimentally form structurally diverse CASs to explore their evolutionary properties. In this system, the ribozymes can catalyze the assembly of other ribozymes from smaller fragments, present in the food set. We first use a droplet microfluidics set-up coupled with next-generation sequencing to conduct a large scale study on thousands of *Azoarcus* CASs. We develop a perturbative approach to identify the important topological parameters that control variations in CASs as a result of environmental perturbations, here the addition of a new species. We then determine the small set of network features governing memory of the initial conditions in *Azoarcus* CAS, a pre-requisite for heredity, by using a computational model validated by experimental data. Finally, we demonstrate that *Azoarcus* CAS possess catabolic processes which make them robust to perturbations in the food set and thus more prebiotic relevant. These results provide evidence for the crucial role of RNA CASs in the origins of life and illustrate how the network structure can be tailored to obtain CASs with properties interesting from an evolutionary point of view, paving the way to an experimental demonstration of Darwinian evolution with a purely molecular system.

Keywords: origins of life, RNA world, autocatalysis, ribozyme, prebiotic networks

Remerciements

Je souhaite remercier en premier lieu mes deux directeurs de thèse : Philippe Nghe et Andrew Griffiths pour m'avoir donné la possibilité d'effectuer ma thèse dans leur laboratoire. Je les remercie pour avoir dirigé les projets dans lesquels j'ai été impliqué avec brio. Ils ont tous les deux non seulement fait preuve d'une excellence scientifique que j'admire, mais ils ont aussi toujours su nous faire persévérer au bon moment. Merci de m'avoir accordé votre confiance et votre temps.

Je remercie Andres Jäschke et Michael Rychelynck d'avoir accepté la responsabilité de rapporteur dans mon jury de thèse ainsi que Pascal Hersen, Catherine Isel-Griffiths et Hervé Isambert pour avoir accepté d'y figurer en tant qu'examinateurs. Je souhaite également remercier mes deux tuteurs, Andre Estevez-Torres et Hervé Isambert, pour m'avoir suivi tout au long de ma thèse grâce à nos entretiens annuels qui furent pour moi très bénéfiques.

Je tiens bien évidemment à remercier grandement Sandeep Ameta, collègue et co-auteur, avec qui j'ai collaboré étroitement ces trois dernières années. J'ai énormément appris à son contact et sa motivation sans faille a été d'une grande aide. J'apprécie beaucoup le travail d'équipe et je pense que nous avons eu beaucoup de chance de bien fonctionner ensemble. Malgré un humour piquant mais créatif, il voit toujours le bon côté des choses. Cette collaboration fut un plaisir et je lui souhaite de tout cœur le meilleur pour la suite de retour en Inde.

Je tiens également à remercier chaudement Niles Lehman pour les discussions de grande qualité que nous avons eues lors de ses visites qui ont également beaucoup apporté au projet. Je souhaite également le remercier ainsi que sa compagne pour leur accueil chaleureux en Californie.

Je souhaite adresser un grand merci à tous les membres présents et passés du LBC et de ses voisins pour avoir participé à créer cet environnement de travail très agréable dans lequel j'ai passé trois années épanouissantes dont je garderai un excellent souvenir. Merci à Sophie surtout

d’avoir tant perdu au chifoumi. Merci à Marina/Shakira de m’avoir si bien supporté en tant qu’unique voisin de bureau. Merci à Dany le hippie, autre voisin, plus volatile mais non moins distrayant. Merci à Amandine et Baptiste, ex-membres du B228, dont la bonne-humeur et les manies bizarres nous manquent. Merci à Kévin grâce auquel je me suis senti moins seul face à Sophie. Merci enfin à Stéphanie la yogi, Antoine le compatriote, Raphaël, Stéphane, Marco, Pablo, Roberta, Matt, Andréa et tous les autres. Merci aux Hifibiens avec qui j’ai passé de très bons moments : Véra, Yannick, Adeline, Sami et tous les autres. Enfin, merci à Isa et Hélène pour leur aide indispensable et avec qui j’ai aussi partagé de très bons moments.

Je tiens à remercier enfin et surtout ma famille et mes proches pour leur support tout du long et pour m’avoir permis de partager mes états d’âmes et autres humeurs passagères inhérentes au métier. Merci pour votre soutien donc, et pour tout ce que vous m’apportez au quotidien. Je me considère comme vraiment très chanceux d’être si bien entouré.

1. Introduction

1.1. From the prebiotic synthesis of small polymers to the emergence of collectively replicating networks

The emergence of life on our planet is still an active on-going investigation despite the strong scientific interest that it sparked a long time ago. We probably know more about when life appeared than how: it was sometimes between the formation of Earth, 4.5 billion years ago (Plaxco & Gross 2006), and the oldest evidence that we have of living organisms, estimated between 4.1 and 3.5 billion years old (Hickman-Lewis et al. 2018; Dodd et al. 2017; Westall et al. 2001). During this period, chemical evolution took place transforming simple organic molecules into biological monomers and then into simple biopolymers before finally giving rise to functional biomolecules which were complex enough to be capable of replication and variation as a form of primordial evolution.

1.1.1. Synthesis of small biopolymers by prebiotic chemistry

How simple biological building blocks can be synthetized from basic organic compounds in prebiotic conditions is an advanced area of research (Orgel 2004). It was demonstrated as early as 1861 that formaldehyde can form an autocatalytic reaction, the formose reaction that leads to the formation of sugars (Butlerow 1861). A little less than a century later, Miller and Urey showed with their pioneer experiment the synthesis of amino acids from methane, water and ammonia in conditions chosen so that they would resemble Earth's prebiotic environment (Miller 1953). A few years later, pyrimidine nucleobases were synthetized from hydrogen cyanide (Oró 1961) and more recently, activated pyrimidine ribonucleotides have been formed in prebiotic conditions (Powner et al. 2009). The aforementioned experiments have several weaknesses: they usually require chemically activated reactants in high concentrations and they often yield low amount of

multi-component mixtures whereas contemporary biology utilizes highly specific precursors. Nevertheless, research in this field is advanced enough to now have an atlas (Figure 1) of interconnected synthesis routes producing all classes of major biological buildings blocks in prebiotic conditions (Patel et al. 2015).

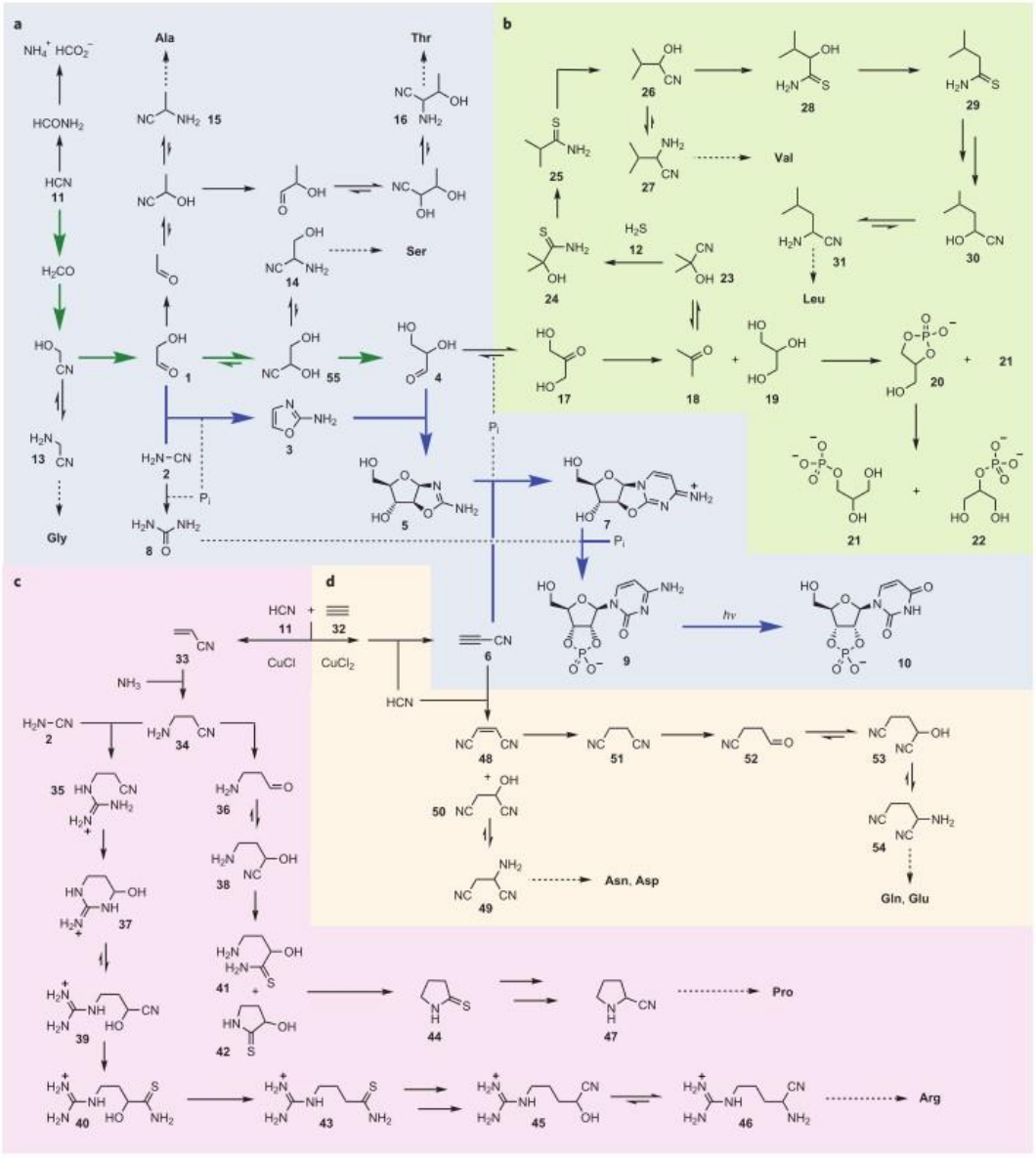


Figure 1. Interconnected network of chemical reactions that transform simple organic compounds thought to be present in early Earth's early conditions such as hydrogen cyanide or acetylene into precursors for the buildings blocks of life biopolymers RNA, proteins and lipids. Reproduced from Patel et al. 2015.

Although figuring out synthesis pathways of the different monomers of life was not trivial in Earth's early conditions, their oligomerization into functional biopolymers is also challenging. Significant progress have been made quite recently, and RNA polymers have been synthesized from energy rich nucleotides using mineral surfaces (Ferris et al. 1996) or phase polymerization (Monnard et al. 2003), from 3'-5' cyclic nucleotides in water (Costanzo et al. 2009), using template-directed non-enzymatic synthesis (Kozlov & Orgel 2000) or with the help of lipids (Rajamani et al. 2008). It should be noted that no polymer longer than 50-mer was obtained with these methods. Although this may be long enough for some miniature ribozyme to be functional (Ferré-D'Amaré & Scott 2010), we are still a long way from the minimal length usually presented as required for a self-replicating and evolvable biopolymer (Kun et al. 2005).

Along these lines, research for a generalist RNA-dependent RNA-polymerase that could eventually self-replicate and thus undergo Darwinian evolution is underway. There has been a lot of progress in this field of research. Starting from RNA ligase and using in-vitro selection (Bartel & Szostak 1993), some ribozymes have been evolved to elongate primers using RNA templates (Zaher & Unrau 2007). The number of nucleotides that could be incorporated have been progressively increased to about 100 (Wochner et al. 2011) and the most advanced version of the ribozyme can even incorporate up to 206 nucleotides (Attwater et al. 2013), which is longer than the ribozyme itself. However, there is a major drawback which is that the ribozyme does not generate its own sequence and thus is still incapable of self-replication. Even though a lot of efforts are dedicated to searching for such a molecule, an alternative view has recently gained a lot of interest.

1.1.2. Theoretical investigations of prebiotic networks as potential pre-life systems

For this alternative line of research, networks of mutually dependent molecular species may constitute more promising and plausible candidates for the appearance of life. Sets of cooperative molecules would represent the fundamental units of evolution rather than single self-replicating entities (Nghe et al. 2015). This was proposed in several contexts and as early as in the 1920's

with Oparin's "coacervates" or bag of interacting molecules. A famous model that falls into this line is Eigen's hypercycle (Figure 2. A) where hyperbolic growth of the members is ensured by mutual beneficial interactions (Eigen & Schuster 1978; Eigen & Schuster 1977). Gánti proposed the chemotron where, encapsulated in a membrane, the replication and metabolic functions form an autocatalytic network (Gánti 2003). Kauffman introduced the notion of collective autocatalytic sets (CAS) where the formation of each molecule is catalyzed by another molecule in the set such that as a whole the system can be considered as autocatalytic (Kauffman 1971; Kauffman 1992). Hordijk and Steel formalized and extended this concept to reflexively autocatalytic and food-generated set (RAF, Figure 2. B) with the benefit of formally including the building blocks, i.e. the food, assumed previously to be provided by the environment (Hordijk 2013; Hordijk et al. 2018).

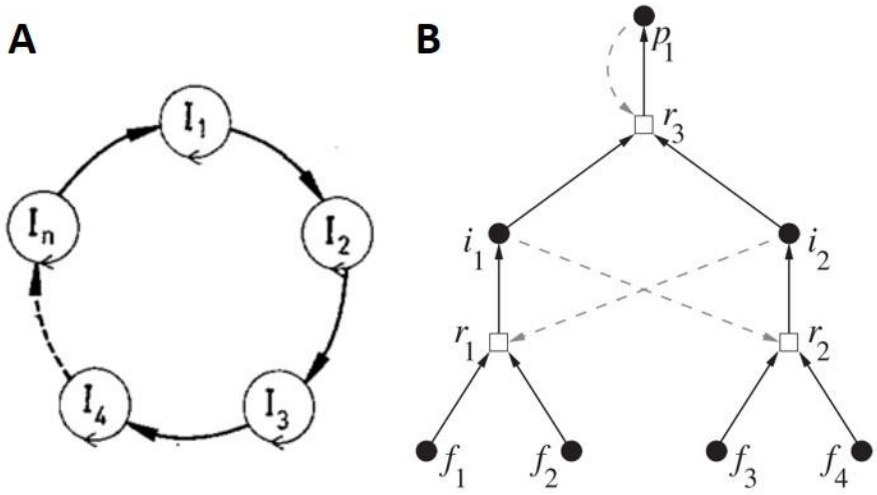


Figure 2. (**A**) An example of a hypercycle with n components I_1 to I_n where each component or node is both an autocatalyst and thus can catalyze its own formation from a food source and a cross-catalyst assisting by catalysis the formation of the next member of the cycle. Reproduced from Eigen & Schuster 1977. (**B**) A minimal example of a reflexively autocatalytic annd food-generated set (RAF) as defined by Hordijk and Steel. White square boxes indicate chemical reactions. The set of incoming nodes to a box represent the set of reactants and the set of out-going nodes is the set of reaction products. In this example, the first two reactions r_1 and r_2 produce products i_1 and i_2 from food molecules $f_{1\text{-}4}$. The products i_1 and i_2 are consumed in the third reaction to produce p_1. This example is a RAF because all the reactions are catalyzed (grey dotted arrows) by products of the reactions. Reproduced from Hordijk et al. 2018.

Cooperative networks have been extensively studied theoretically because of several advantages they have over “selfish” replicators. Because the function is distributed amongst its members (Kauffman 1986), a network is expected to be more robust against potential change in the environment that could result in the disappearance of one of its nodes (Hinshelwood 1952). The distribution of function is also evolutionary favored as single molecules possessing very complex and advanced functions are very unlikely to have appeared by random polymerization (Boza et al. 2014). Finally, a growing body of research suggests that cooperation between molecular species may have been advantageous over single self-replicating molecules (Higgs & Lehman 2015).

Many studies have proposed interesting models of autocatalytic sets. Several of them are based on the graded autocatalysis replication domain (GARD) model. Using this model, they demonstrated the formation of catalytic networks with lipids where information is coded in the relative fraction of their species (Segré et al. 1998; Segré et al. 2000). In a drawn parallel to the genome, Lancet and colleagues named them “composomes”. They also showed with this system that mutual cross-catalysis, i.e. cooperation, was evolutionary favorable compared to selfish replicators (Markovitch & Lancet 2012), yet another argument for the superiority of networks in prebiotic settings.

In a different model, Jain and Krishna demonstrated that the appearance of a small autocatalytic set that would grow to incorporate more and more nodes could not be prevented (Jain & Krishna 2001). Their model is composed of interacting molecular species that can catalyze other species’ formation from a set of food molecules with random probabilities. This network evolves over time in their simulations as the least populated specie gets removed and replaced by a new. Interestingly they also observed random catastrophic events leading close to extinction the dominant autocatalytic set (Jain & Krishna 2002). They concluded that some species, the “keystone” species, are critical for the network to be maintained but are also present in small fractions and thus at risk of being eliminated.

More recently, Vasas et al. studied with a mathematical model and simulations the requirements for autocatalytic sets to evolve (Vasas et al. 2012). They suggested that the coexistence of multiple

viable cores, i.e. self-sustaining subsets of the network, is a necessary condition for evolvability. Evolution in this setting would be transiting from one dominant core to another by the means of a selection pressure. This notion of viable cores was formalized by Hordijk and colleagues as irreducible RAF or irrRAF (Hordijk et al. 2012; Hordijk et al. 2014). They demonstrated that smaller but still autocatalytic subsets can be found in RAF (Hordijk et al. 2015). The smallest of such subsets are the irrRAF since they cannot be further decomposed. They found many irrRAF in a simple polymer model (Hordijk & Steel 2013).

1.1.3. Prebiotically relevant experimental models of networks of replicators

Testing these hypotheses with experimental realizations of autocatalytic sets is another challenge that is critically needed to anchor the study of prebiotic networks in physico-chemical realities. It is only then that it may be possible to demonstrate how autocatalytic sets can really evolve and what has been their role in the path to modern life. In the past 30 years, an increasing number of experimental chemical systems possessing the capacity to self-replicate has been developed (Kosikova & Philp 2017). Even though some of these models are composed of single molecules building on such minimal self-replicators, more complex networks of interacting molecules have also been studied in the lab. Some of these systems are not constituted of what one would consider as biologically relevant molecules but they are nonetheless interesting models for the study of generic properties of self-replicating systems. These systems can be divided into three classes based on which type of molecules they are made of: small organic molecules, peptides or oligonucleotides. Here we review only the experimental systems which involve peptides and oligonucleotides. The interested reader can find more information on replicating networks of small organic molecules in this recent review (Kosikova & Philp 2017).

Self-replication mechanisms almost always invoke the formation of supramolecular assemblies through non-covalent bonds. With oligonucleotides, the sequence of bases determines the hydrogen bonds network that leads to the hybridization of DNA or RNA strands. Peptides on the other hand, with 20 more chemically diverse building blocks, have a much richer structure

repertoire and as a result the molecular recognition mechanisms between peptides do not depend only on the primary sequence of amino acids but also on tertiary interactions.

Ghadiri and colleagues reported in 1996 the first experimental self-replicating peptides (Lee et al. 1996). In this model, a 32-residue alpha-helical peptide acts as a template for the formation of an amine bond between 17-residue N-terminal electrophilic fragment E and 15-residue C-terminal nucleophilic fragment N (Figure 3. A). This reaction goes through a template-directed pathway, mediated by the interaction between the hydrophobic zones of the peptides and is facilitated by their electrophilic outer surface. Addition of an increasing starting amount of full-length template T causes a proportional increase in the initial reaction rate, thus demonstrating the autocatalytic nature of the system. The observed parabolic growth profile is an indication of the product inhibition limiting the reaction rate, which is a common drawback of such systems involving templates.

This initial example of a self-replicating peptide was followed by other demonstrations with similar approaches (Severin et al. 1997; Yao et al. 1997; Yao et al. 1998). For example, a shortened (26 residues only) peptide replicator was shown to reach almost exponential replication (Issac & Chmielewski 2002) and Ashkenazy and co-workers have designed a system in which light exposure controls the amount of initial template and thereby determines the replication rate (Dadon et al. 2010). These initial minimal designs consist of only a single replicator but they were quickly expanded to more and more complex networks of replicators, exhibiting a strong level of cross-catalysis.

Ghadiri and colleagues extended their original replicator design into a network of two mutually replicating peptides (Figure 3. B) (Lee et al. 1997) which they further adapted to allow for error-correction (Severin et al. 1998). To simulate the occurrence of a mutation in biological systems the authors introduced a single mutated residue in each of the two building blocks. The native template was shown to be preferentially formed by the system due to the inactivity of the double mutant template and to the dominant cross-catalytic activity of the two single mutants. They also designed, building on their self-replicator design, a network of replicating peptides showing stereospecificity (Saghatelian et al. 2001). Finally, Ghadiri and co-workers reported the design of

a network composed of 81 peptides of identical length and of similar structure (Figure 3. C). They showed that some pathway in this network could be selectively favored by choosing the initial full-length template concentrations (Ashkenasy et al. 2004). They demonstrated that a sub-system of this network was able to perform basic Boolean logic operations (Ashkenasy & Ghadiri 2004). With a different goal in mind, Chmielewsky and colleagues designed a network of four replicators from four fragments. This network is sensitive to the conditions of the reaction environment and certain pathways leading to the formation of particular replicators can be selected by modulating the pH or the ionic strength (Yao et al. 1998).

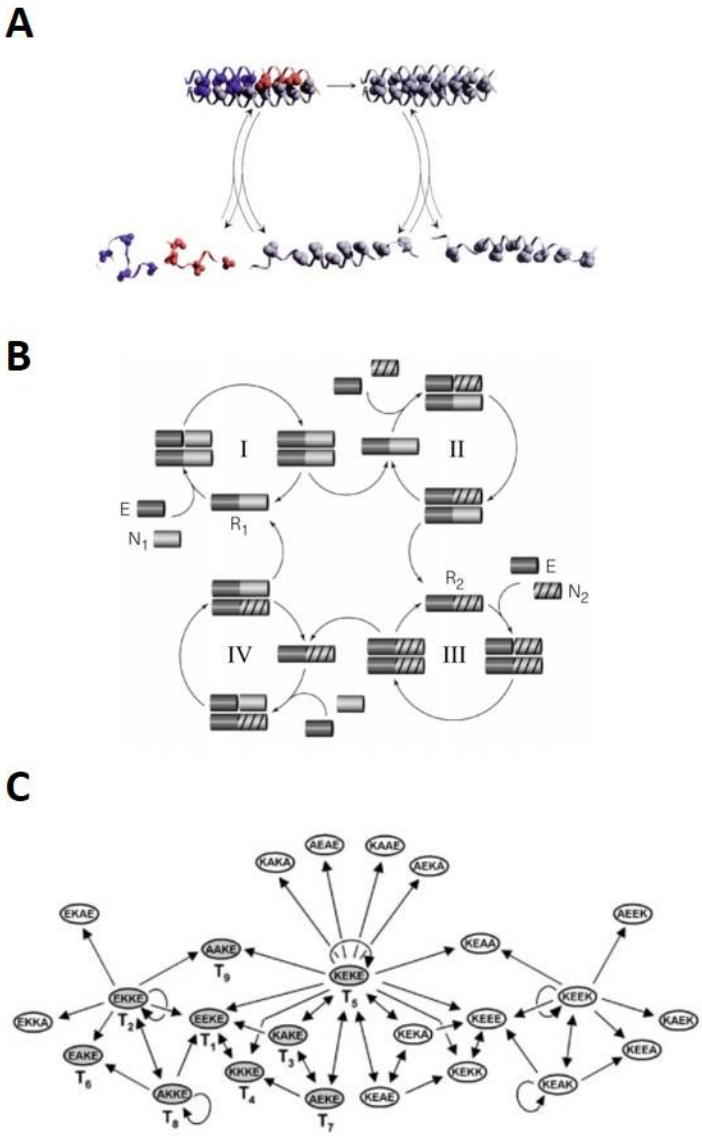


Figure 3. Evolution of the self-replicating peptide system from Ghadiri and colleagues. (**A**) First example of a self-replicating peptide of 32 residues, in white, from two building blocks E and N of 17 and 16 residues respectively depicted in blue and red. The process is autocatalytic since the full-length template self-assembles with the two building blocks and this favors their ligation to produce the full-length template. Reproduced from Lee et al. 1996. (**B**) Extended replicating peptide system with two possible full-length templates E•N_1 and E•N_2 made from three building blocks E, N_1 and N_2. Both full-length templates are not only autocatalysts but also cross-catalysts as they can assemble with a copy of themselves or of the other template. Reproduced from Lee et al. 1997. (**C**) Further extension of the same system with 81 peptides resulting here in a predicted catalysis graph structure composed of 25 nodes corresponding to full-length templates and 53 directed edges corresponding to the template-assisted ligation of another template. Reproduced from Ashkenasy et al. 2004.

The first example of non-enzymatic self-replication of an oligonucleotide was reported in 1986 by von Kriedowski and colleagues (von Kiedrowski 1986). In this experimental model, a hexanucleotide palindromic DNA strand acts as a template for its own formation through Watson-Crick base-pairing facilitating the formation of a phosphodiester bond between its two trinucleotide halves (Figure 4. A). The system was further improved (von Kiedrowski et al. 1991) to show parabolic and sigmoidal growth profile, the key signature of autocatalytic systems. The autocatalytic efficiency as defined by (von Kiedrowski 1993) was measured to be about 420. The system was later adapted to feed from an extended fragment pool of three building blocks (Achilles & von Kiedrowski 1993) and to use template fixed on solid support in order to bypass product inhibition to improve the replication efficiency (Luther et al. 1998). So far, all the versions of the system, if represented as a network, only consist of a single node with a self-loop.

The von Kriedowski group enlarged their minimal self-replicator system to a four-node network (Figure 4. B) in which two nodes are self-replicators and the two others are cross-catalyzing each other (Sievers & von Kiedrowski 1994; Sievers & Von Kiedrowski 1998). To achieve this, the system is composed of two trinucleotide A and B that come into two versions nA, Ap, nB and Bp depending on the location of a reactive group (either an electrophilic phosphate p at the 3'-end or a nucleophilic amine n at the 5'-end). Four different templates can be formed: the two self-complementary self-replicating ApnB and BpnA and the two cross-complementary cross-replicating ApnA and BpnB. Doping in preformed templates resulted in the increase of the formation rate of the target complementary template in all cases.

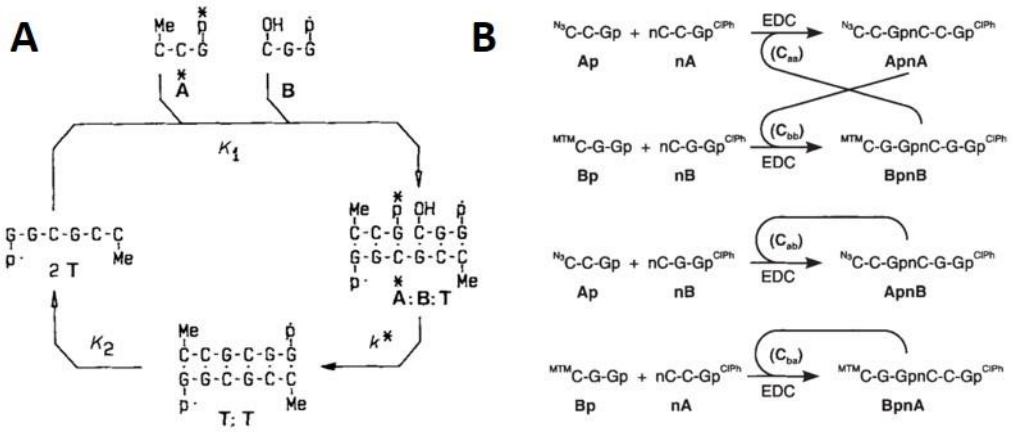


Figure 4. Evolution of the oligonucleotide-based self-replicated system of the von Kiedrowski's team. (**A**) Two palindromic trinucleotides A and B with an activated and a protected end can bind to a full-length template T. This facilitates their ligation into another full-length template T, making the complete process autocatalytic. Reproduced from von Kiedrowski 1986. (**B**) Extension of the system by introducing two versions of the two building blocks A and B where either of the two ends is activated while the other is protected. This makes the self-ligation of A and self-ligation of B possible. As a result, four templates can be obtained: AA, BB, AB and BA. Because AB and BA can self-template, they are autocatalysts while AA or BB needs the complementary template to form AA and BB and thus are cross-catalysts. Reproduced from Sievers & von Kiedrowski 1994.

The first RNA-based self-replicating experimental system was reporter by Paul and Joyce in 2002 (Paul & Joyce 2002). In this model, a RNA ligase ribozyme, the R3C, is modified so that it can ligate two RNA fragments A and B to form itself (Figure 5. A). Again, doping in preformed ligase would result in a proportional increase of the reaction rate demonstrating the autocatalytic behavior of the system. Likewise, this system was taken from a single self-loop bearing node to a two-node cross-catalytic network (Kim & Joyce 2004). In this expanded version, two RNA ligases E and E' can template each other's formation from the corresponding fragments A, B, A' and B' (Figure 5. B). It was demonstrated that by providing preformed template both E and E' can catalyze the formation of the other ligase. E and E' were further improved by in vitro selection (Lincoln & Joyce 2009) to be capable of exponential amplification. They also created 12 pairs of mutual replicators by mutating the original E and E'. They then made them compete against each other in an impressive 20-generation serial transfer experiment. Interestingly, the analysis revealed that the most abundant product was a recombinant template.

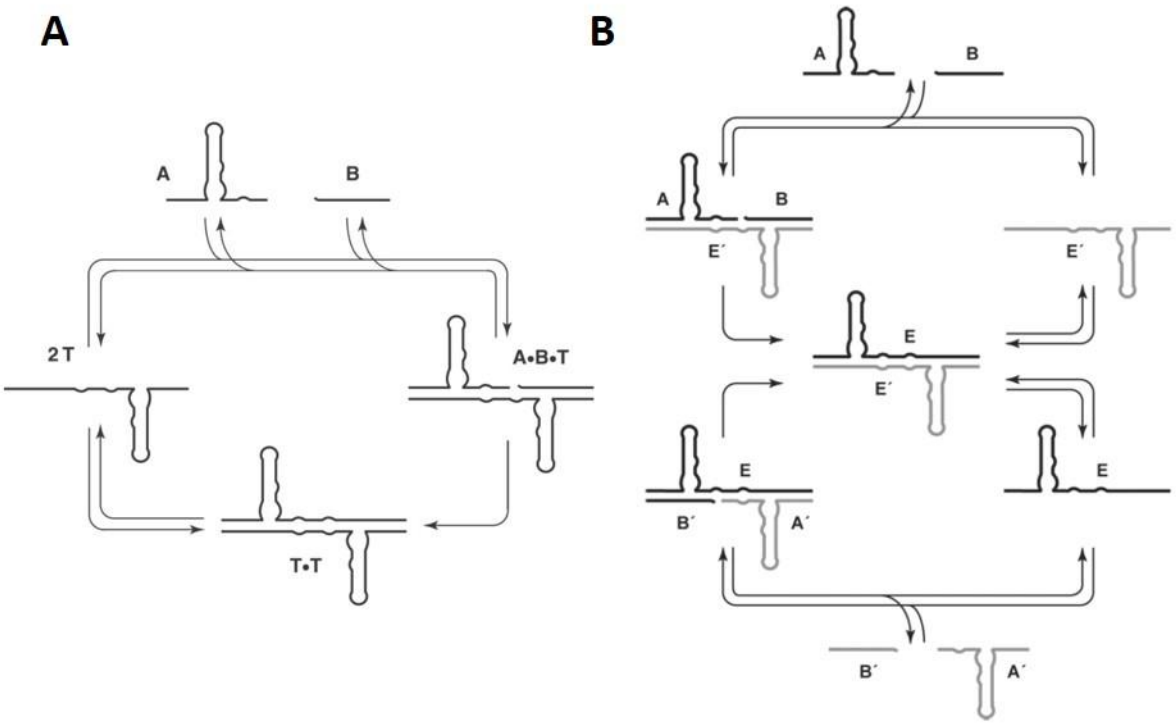


Figure 5. Evolution of the RNA-based self-replicating system from Joyce and colleagues. (**A**) The R3C ligase ribozyme T can self-replicate by templating the ligation of the two RNA fragments A and B resulting in another copy of itself and thus rendering the process autocatalytic. Reproduced from Paul & Joyce 2002. (**B**) Extension of the system to two ligases E and E' that can cross-catalyze the formation of each other. Ligase E (resp. E) can hybridize with fragments A' and B' (resp. A and B) to facilitate their ligation into the ligase E' (resp. E). Reproduced from Kim & Joyce 2004.

RNA-based self-replicating systems are of particular interest in origin of life research because of the central and versatile role that RNA has in contemporary biology and which is discussed in the next section along with the experimental system studied in the present work: the *Azoarcus* ribozyme self-replicating autocatalytic system.

1.2. The RNA world and the *Azoarcus* ribozyme experimental system

Modern life is based on the interdependence of three types of biopolymers: protein, DNA and RNA. While DNA and RNA are required to synthetize proteins, the latter are necessary for synthetizing both DNA and RNA. Hence the question of whether life started simultaneously with all three biopolymers or with only one arises. The first option would represent a famous "chicken and egg" dilemma. An alternative hypothesis that would have the benefit of bypassing this problem is the 'RNA world' hypothesis which proposes that life before we knew it, was mainly based on RNA which would then have played an even more central role than it does today as both an information carrier and a catalyst (Crick 1968; Gilbert 1986; Joyce 2002; Orgel 1968; Pressman et al. 2015).

1.2.1. Advanced catalytic properties of RNA

The involvement of RNA in the most fundamental cellular processes in contemporary organisms provide great support for its supposed crucial role in origin of life (Atkins 2010). Probably the most compelling evidence of this is that RNA is at the heart of the translation of mRNA into proteins as it constitutes the catalytic core of the ribosome (Ban et al. 2000; Chen et al. 2007). RNA is indeed not only a good solution for storing information but can also fold into a wide range of 3D structures with functional activities. For example, RNA aptamers can bind to ligands (Famulok & Mayer 2014; Pfeiffer & Mayer 2016; Sullenger & Nair 2016) and riboswitches can control gene expression by a sensing mechanism that involves binding to metabolites (Breaker 2012; Peselis & Serganov 2014; Serganov & Nudler 2013). Finally and more importantly for this manuscript, some RNA strands, called ribozymes, possess catalytic activity. Many such molecules have been reported during the past 30 years since the discovery of the SSI of Tetrahymena (Kruger et al. 1982) and since it became known that RNAse P is partly made of RNA (Guerrier-Takada et al. 1983). The catalytic properties of ribozymes are based on specific structures involving not only hydrogen bonds through base pairing, but also tertiary interactions. The folding of RNA, whose

timescale range from picoseconds up to seconds (Mustoe et al. 2014) is usually facilitated by the presence of metal ions (Denesyuk & Thirumalai 2015).

We now have precise structural information on many classes of ribozymes. A first class of ribozymes consists of small ribozymes able to cut nucleic acids. Amongst these small nucleolytic ribozymes, a Hammerhead (HHR) variant crystal structure was obtained first (Scott et al. 1995) followed by the Hepatitis delta (HDV) (Ferré-D'Amaré et al. 1998), the hairpin (HP) (Rupert & Ferré-D'Amaré 2001), the glmS ribozyme (Klein & Ferré-D'Amaré 2006) and more recently the Twister ribozyme (Liu et al. 2014), the VS ribozyme (Suslov et al. 2015) and the TS ribozyme (Liu et al. 2017). The second class of ribozymes is composed of more complex ribozymes, the group I intron and the group II intron for which crystal structures are also known at high resolution (Adams et al 2004; Tor et al 2008). Finally for the last class that represents the ribonucleoproteins (RNP, ribozyme-proteins complexes), crystal structures are also known for: RNAse P (Kazantsev et al. 2005), the ribosome (Ban et al. 2000) and the spliceosome (Yan et al. 2015). The catalysis performed by small nucleolytic ribozymes is based on generic acid-base catalysis while it is based on two metal ions catalysis for group I and group II introns, RNAse P and the spliceosome. All ribozymes (except for the ribosome) found in nature catalyze phosphoryl transfer reactions: the mechanism of the reaction proceeds with the attack on the nucleophilic phosphate by the adjacent 2'-OH for the small nucleolytic ribozymes, for the group I intron, it is the 3'-OH of an external guanosine while it is the 2'-OH of an internal adenosine for group II introns and the spliceosome (Lilley & Eckstein 2007).

RNA catalyzed peptidyl-transfer in the ribosome, tRNA maturation by RNase P and RNA splicing by the spliceosome are ubiquitous in contemporary biology, whereas the small nucleolytic ribozymes (HHR, HDV and HP) are quite uncommon and their function is well understood only in viruses. Paradoxically, their sequences have been identified in many organisms across all taxa and some have thus proposed that they are relics from an ancient 'RNA world' (Cech 2012). An alternative explanation is that because of the relative simplicity of the structure and the relaxed sequence requirements, this type of motifs would easily be found everywhere. The HHR motif is indeed the most frequent motif found by *in vitro* selection experiments (Salehi-Ashtiani & Szostak 2001) and in living organisms (Hammann et al. 2012). This would suggest that catalytic RNA

sequences are not that uncommon in sequence space. As a result, in a repertoire of RNA polymers randomly produced by prebiotic processes and of limited size, it should not be improbable to find sequences with catalytic activity (Wachowius et al. 2017). In the context of origin of life research, ribozymes are thus very good model systems to be studied both theoretically and more importantly empirically because they were likely important pre-life actors.

1.2.2. The *Azoarcus* ribozyme, a recombinase ribozyme derived from a group I intron

Taking the above into consideration, group I intron ribozymes are appealing for origin of life research because they have the potential to catalyze recombination reactions. In modern organisms, recombination is used to shuffle genetic information across the genome to increase genetic diversity. In a prebiotic setting, RNA recombination would have been well suited to both build long RNA polymers from a pool of short RNA strands and explore efficiently the sequence space in an energy-wise manner (Lehman 2003; Lehman et al. 2011). Group I intron self-splice from mRNA, tRNA and rRNA precursors by two sequential transesterification reactions in many different types of organisms. The reaction is enthalpically neutral and it was suggested that group I intron may be as ancient as 3.5 billion years old (Kuhsel et al. 1990). Several recombinase ribozymes have been designed from group I introns with similar features (Zaug & Cech 1986). There is one of particular interest for us coming from the $tRNA^{Ile}$ of the purple bacterium *Azoarcus* (Figure 6. A) (Riley & Lehman 2003). In this construct, the exons have been removed and the 5' end has been shortened so that it starts with the internal guide sequence (IGS) that allows the ribozyme to catalyze recombination reactions involving any substrate that has a target sequence complementary to its IGS (Figure 6. B).

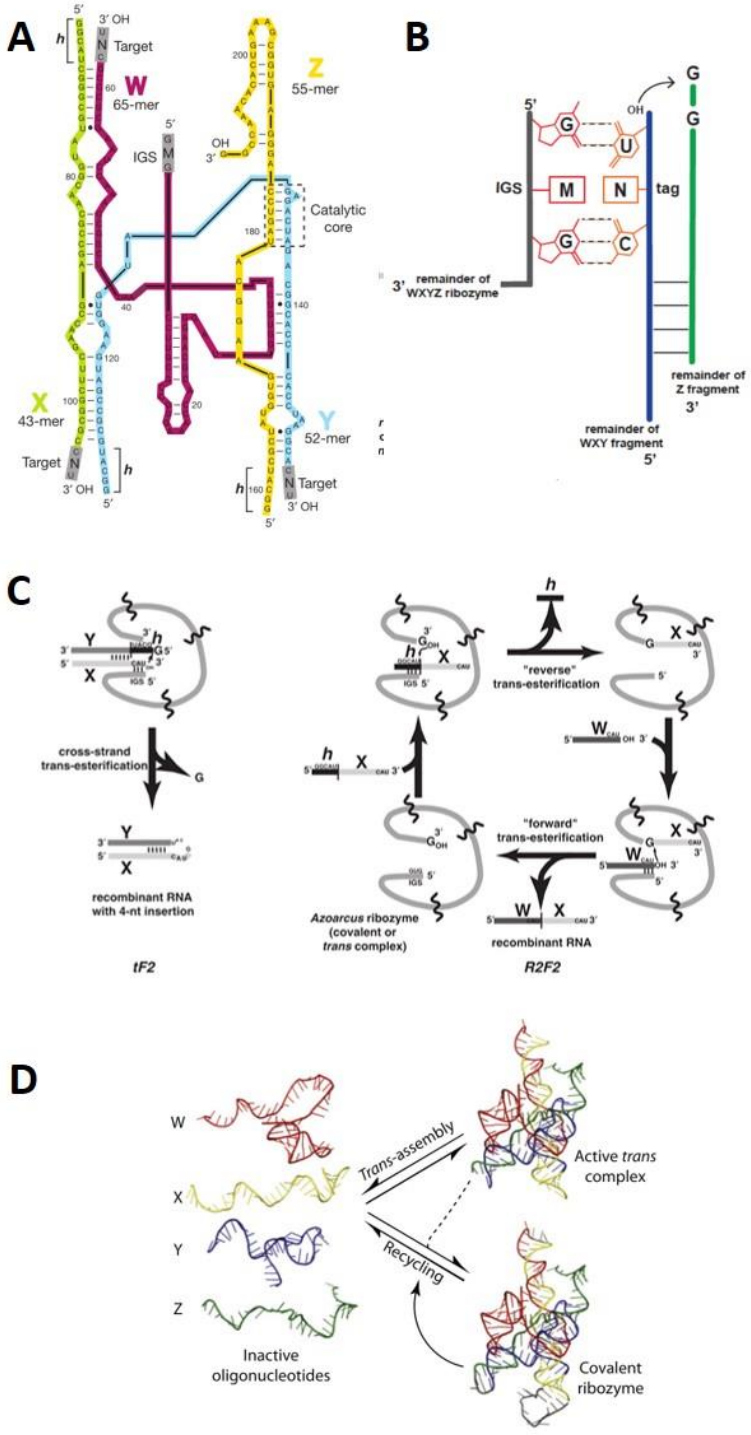


Figure 6. The Azoarcus ribozyme system. (**A**) The engineered version of the Azoarcus ribozyme WXYZ used in this manuscript. Four fragments W, X, Y and Z were created by introducing tag sequences 'CNU' at loop regions. The Internal Guide Sequence (IGS) 'GMG' is located at the 5' of fragment W. Reproduced from Vaidya et al. 2012. (**B**) Mechanism of IGS-tag recognition where IGS of a WXYZ ribozyme binds to the tag of a WXY fragment bound to a Z fragment. The ribozyme facilitates the attack of the 3'-OH of the WXY

fragment on the Z fragment. Reproduced from Satterwhite et al. 2016. (**C**) The two self-assembly recombination mechanisms of Azoarcus ribozyme. The tF2 mechanism goes through a cross-strand transesterification while the R2F2 mechanism is a two-step mechanism where one of the substrate is transiently ligated to the ribozyme itself. Reproduced from Draper et al. 2008. (**D**) Self-assembly of the Azoarcus ribozyme from four fragments. A non-covalent trans complex is first formed by the fragments through hydrogen bonds and which have a residual recombinase activity leading to the formation of the covalent ribozyme WXYZ. This process is autocatalytic since WXYZ can further promote its own formation. Reproduced from Vaidya et al. 2013.

The general reaction that the *Azoarcus* ribozyme can catalyze is the following: A•B + C•D -> A•D + C•B where A, B, C and D are RNA strands with A and C bearing a 'CAU' target sequence, located at their 3' end matching the 'GUG' canonical IGS of the ribozyme. The mechanism for this general scheme is the following (Figure 6. C, right) (Hayden et al. 2005): the first step is the covalent ligation of the 3' tail B of the first substrate A•B to the 3' end of the ribozyme by a transesterification reaction following the base pairing of the IGS and the 'tag' on A•B while the 5' head A is released. During the second step, after the IGS binds to the 'tag' of C•D or of a free C, the 3' end of this substrate attacks the bond formed during the first step between B and the ribozyme resulting in C•B which is one of the final recombined molecules. Exploiting this, *Azoarcus* ribozyme has been shown to recombine structurally complex and catalytically active RNA molecules (Hayden et al. 2005).

1.2.3. *Azoarcus* ribozymes can self-assemble and form diverse networks of cross-catalysis

A very interesting feature of this ribozyme, crucial for the work presented in this manuscript, is that it can recombine RNA fragments in order to form itself (Hayden & Lehman 2006). It was first shown that starting from four fragments named W, X, Y and Z, the full-length ribozyme WXYZ could be reformed (Figure 6. D). These four fragments whose length range from 39 to 63 nucleotides were derived by fragmenting the complete molecule at the loop regions. The full-length ribozyme can be obtained even when only the fragments are present at the start of the reaction. The fragments can indeed form a non-covalent complex, called a *trans* complex, that is structurally similar enough to WXYZ that allows a residual recombinase activity. Because all four

fragments bear tag sequences, three recombination reactions lead to the formation of the covalent ribozyme. While the general mechanism described earlier is observed during this self-assembly process, another mechanism termed 'tF2' has also been described (Draper et al. 2008). In this mechanism, the ribozyme binds a duplex formed by two RNA fragments and catalyzes the cross-stand attack on the 5'-end of one substrate by the 3'-end of the other substrate, resulting in the recombined molecule (Figure 6. C, left). Finally and more importantly, this process was shown to be autocatalytic (Hayden et al. 2008) though with relatively low catalytic efficiency. Recently, it was demonstrated that the *Azoarcus* ribozyme could self-assemble from five fragments when Z was further decomposed into two smaller sub-fragments Z1 and Z2 (Jayathilaka & Lehman 2018).

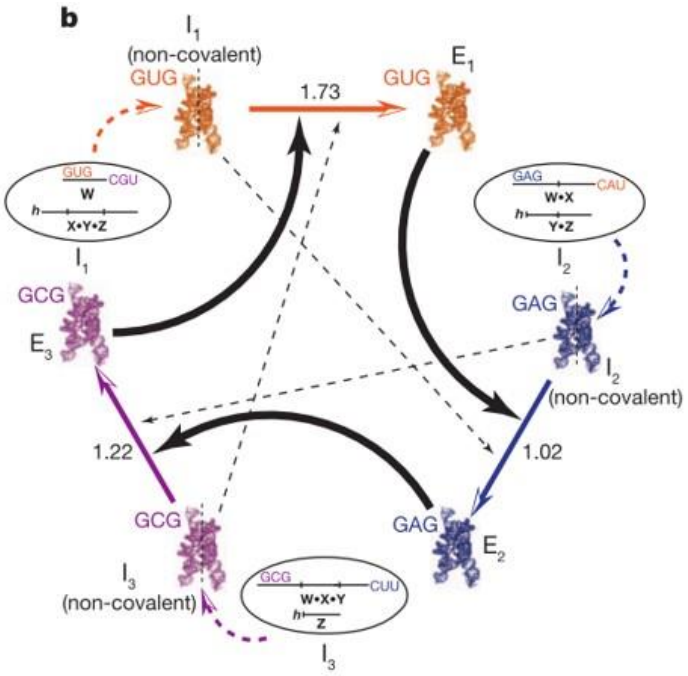


Figure 7. Example of a three-membered Azoarcus network with a closed cycle topology where each member of the network catalyse the formation of the next member. This network was shown to out-compete selfish replicators. Reproduced from Vaidya et al. 2012.

The system can be engineered to another level by varying the middle nucleotide of the IGS and of the tag. This gives many versions of the ribozyme, some of which can no longer efficiently catalyze their own formation from single fragments but that can cross-catalyze the formation of other ribozymes (Satterwhite et al. 2016). Using this, Lehman and colleagues (Vaidya et al. 2012) constructed a three-membered cooperative cycle and showed that it would dominate when competing against selfish single-node replicators (Figure 7). They also reported an analysis of the biggest possible network with this system consisting of 48 nodes. Dynamics of some small networks (up to four nodes) have later been studied in a game theory framework (Yeates et al. 2016) or in a more classical framework (Yeates et al. 2017). Yet these studies could not embrace the full diversity of the *Azoarcus* network space because of the high number of possible networks, up to 2^{48}, for which a high-throughput approach is needed to study at large scale these networks. Finally, recombination is prebiotically relevant also because it allows for potential easy recycling of food fragments. It has been demonstrated with *Azoarcus* ribozyme that mis-recombined products can be recycled into usable fragments (Vaidya et al. 2013); however the question of whether or not such a system could use these recycling properties for expanding its repertoire of usable substrates have not been precisely addressed.

1.3. Droplet microfluidics as an advanced tool for origin of life research

Microfluidics is the manipulation and the study of fluids at the micrometric scale (Denesyuk & Thirumalai 2015; Atencia & Beebe 2005). It allows to miniaturize experiments that would normally require benchtop equipment, to a chip device, which is centimeters wide ("lab on a chip"). Over the last two decades, many applications of microfluidics have been designed, mainly in biotechnology and analytical chemistry. Microfluidics offers indeed the benefit of using very low amounts of reagents by reducing the working volumes to a few picoliters, compared to the few microliters usually required. Because microfluidics allows performing many experiments in parallel, it generates very rich datasets usually with many replicates. This is enhanced by the fact that it gives a very precise tuning of spatial and temporal parameters. Today it is widely used for various applications such as high-throughput combinatorial drug screening, biosensing, single-cell transcriptomics, proteomics and genomics, novel biomaterials synthesis or combinatorial synthetic biology (Melin & Quake 2007; Mark et al. 2010; Auroux et al. 2004).

1.3.1. Historical development of microfluidics

Historically, the earliest microfluidic chips were developed along with the first microfabrication techniques for chip based electrophoresis (Harrison et al. 1992), a natural follow-up of capillary electrophoresis, a technique which was already manipulating fluids at the micrometer scale (Wu et al. 2008). This sparked the development of high-throughput DNA sequencing based on chip-based electrophoretic separation (Kan et al. 2004) or on flow chambers with microfluidic channels (Margulies et al. 2005) and initiated the complete transition of the sequencing platforms to chip-based format.

Microfluidics benefited a lot from the development of soft-lithography techniques involving poly-dimethylsiloxane (PDMS), a polymeric silicon with interesting properties (Tang & Whitesides 2010; McDonald et al. 2000). This led to the elaboration of methods for producing microfluidic chips. Most methods begin with the photolithographic fabrication on a silicon wafer of a design

of the desired microfluidic channels and chambers. PDMS is then poured over the wafer. Once it has polymerized, it is removed and the channels and other details are imprinted in it. A glass slide is then covalently bound to the PDMS to seal the microfluidic channels. After a final treatment of the channels (hydrophobic or hydrophilic), the microfluidic chip is ready to use.

In our case, we are mostly interested in a branch of microfluidics that has developed over the past 10 years: droplet microfluidics which concerns the manipulation of droplets (often water in oil droplets) in micrometer scale channels (Song et al. 2006; Theberge et al. 2010). Droplets are obtained when immiscible oil and water are injected in the channels of a microfluidic chip which has the geometry required for the shearing of the aqueous phase by the oil phase, thus generating droplets at high frequency (Anna et al. 2003). Such droplets, often of the picoliter scale, require a stabilizing molecule not to coalesce when they are in contact. To this end, surfactants can be added to the oil phase. Surfactants are amphiphiles that are concentrated at the water-oil interface and thus stabilize droplets by reducing the surface tension (Nowak et al. 2016; Bibette et al. 1999; Baret 2012).

1.3.2. Droplet microfluidics, a high throughput tool for massive parallelization

Generally, droplet microfluidics allows the compartmentalization of reactions and their precise control in time and space. Droplet generation can be made highly monodisperse and many types of microfluidic devices have been developed to allow for the precise manipulation of droplets (Figure 8). The droplet volume can thus be tuned from the femtoliter (Leman et al. 2015; Li et al. 2015; Arayanarakool et al. 2013) to the nanoliter (Dewan et al. 2012) by carefully designing the chip that generates the droplets (Stan et al. 2009; Dollet et al. 2008; Umbanhowar et al. 2000; Baroud et al. 2010; Abate et al. 2009; Anna et al. 2003). Droplets containing reactants can be incubated off-chip in termocyclers (Mazutis, Araghi, et al. 2009; Zonta et al. 2016) or on-chip, where time of incubation can be chosen by introducing delay lines or by trapping the droplets in chambers (Courtois et al. 2008; Frenz et al. 2009). Droplets can be split into smaller droplets by dividing in two the channels where the droplets are flowing (Abate & Weitz 2011) and by introducing split channels departing from the main channel (Figure 8. C). With a technique known

as pico-injection, reagents can be added to a train of droplets (Abate et al. 2010) and the content of the droplets can then be mixed on-chip usually by employing a wiggle geometry (Courtois et al. 2008). A technique, which is also of particular interest for us, allows for the possibly on demand fusion of different populations of droplets by electro-coalescence, often by using an electric field at a chosen ratio (Mazutis, Baret, et al. 2009; Niu et al. 2009; Zagnoni & Cooper 2009; Zagnoni et al. 2009). Fluorescence emitted by the droplet can be measured on-chip and this way one can measure the droplet size or the concentration of a given chemical in the droplet if a fluorogenic compound is used. Droplets can be sorted based on the fluorescent signal using dielectric forces (Sciambi & Abate 2015; Ahn et al. 2006) or acoustic waves (Franke et al. 2010). This technique is sometimes called Fluorescence Activated Droplet Sorting (FADS) (Baret et al. 2009) in an analogy to the classical Fluorescence Activated Cell Sorting technique (FACS). Such techniques have a limitation when they are used for certain applications, e.g. an enzymatic assay: they only offer one time point measurement and do not allow the derivation of a kinetic profile. This can be circumvented at the expense of losing some throughput by trapping the droplets in 2D chambers so that multiple measurements of the same set of droplets is possible (Eyer et al. 2017).

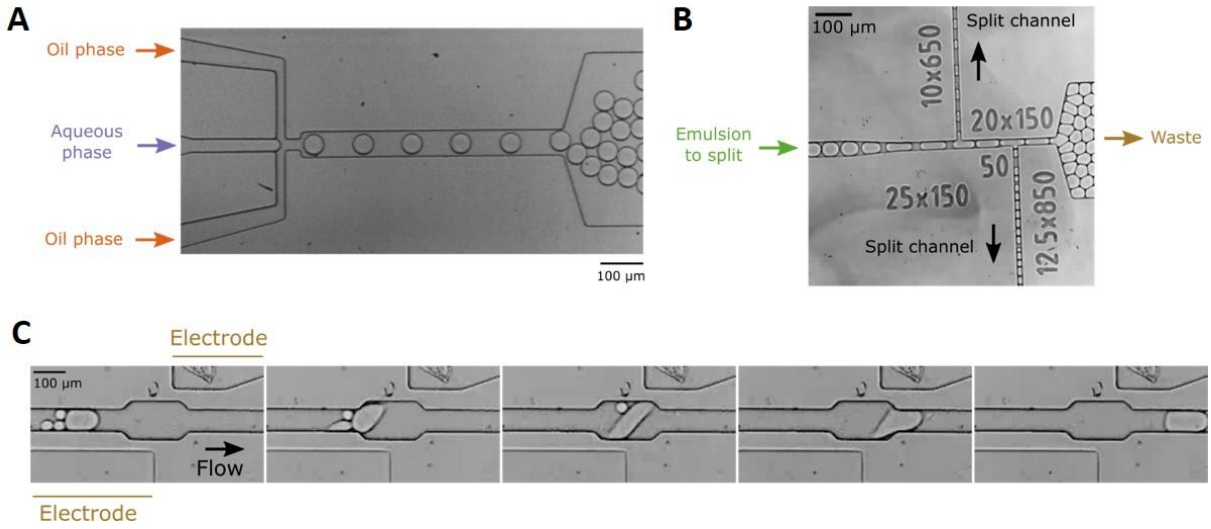


Figure 8. Selection of droplet microfluidics operations relevant for the work presend here. (**A**) Droplet generation device by flow-focusing an aqueous phase by an oil phase. This device can produce droplets about 50 pL, although the volume can be tuned by adjusting the flow rates of the two phases. (**B**) Droplet splitting device. A previously prepared emulsion is re-injected in a main channel from which two split

channels are introduced at 90° and in which a small portion of the droplets is extracted. (**C**) Droplet fusion time-lapse device. A 50 pL droplet is fused with three droplets of 5 pL each by electro-coalescence induced by the electric field between the two electrodes. The droplets are paired in another part of the device so that the desired ratio is obtained.

These operations on droplets can be integrated as part of complex workflows. They can be done sequentially and the resulting emulsion after a given step can be re-injected as the starting emulsion for the next step. However, the emulsions are sensitive and consequently they often become less and less monodisperse as the steps accumulate. Coalescence can be provoked by debris in the tubing, the microfluidic channels or the collection devices. Evaporation can also be problematic if the droplets are incubated for a long time. An alternative option is to integrate many devices performing a given operation into a single microfluidic chip. Nevertheless the number of steps that can be integrated in a single chip is limited by the increase of hydrodynamic resistance as the length of the channels increase (Tabeling 2005). Additionally, the number of input flows is usually limited in a classical microfluidic set-up and as the number of steps increase, the number of inputs also usually increase. Finally it can be difficult to control many steps at the same time and if the chip is loo large it will also be difficult to monitor it entirely with a microscope.

1.3.3. Classical applications of droplet microfluidics particularly relevant for origin of life studies

Two classical applications of droplet microfluidics are of particular interest for us. The first one is single-cell RNA sequencing in droplets (Macosko et al. 2015; Klein et al. 2015). The object of interest in these studies is the RNA content of cells of interest at the single-cell level. For this, one can choose to target either all the messenger RNAs (mRNA) or a set of specific mRNAs corresponding to the set of genes of interest. The first step is the encapsulation of single cells in droplets by diluting the cells enough so that the proportion of double-cell encapsulation is adequately low. Since the encapsulation process follows a Poisson distribution it is straightforward to determine beforehand at which concentration to operate. The next key element is the co-encapsulation of a single cell with a hydrogel bead bearing a clonal population

of barcoded primers. The barcodes are bead-specific and thus cell-specific. The mRNAs bind to the primers and a reverse-transcription reaction allows for the coupling of the mRNA sequence and the cell barcode. All constructs are then pooled by breaking the emulsion and amplified to be subjected to next-generation sequencing. This method is very high-throughput and is currently only limited by the sequencing maximum depth. In the context of origin of life research, such methods are interesting because they couple the RNA content of a reaction vessel to a unique barcode and thus allow a massive parallel read-out by sequencing. One could thus imagine thus to experimentally subject an RNA system relevant to the origin of life to many different conditions and analyze them at high throughput in a microfluidic device.

The other application which is of interest for us is the enzymatic screening, accompanied sometimes with directed evolution, of enzymes using droplet microfluidics. In these *in vivo* experiments, a population of host organism cells bearing mutated versions of the gene which is encoding for the enzyme of interest, is encapsulated at the single-cell level and incubated so that the cells grow and secrete the enzyme (Obexer et al. 2017; Beneyton et al. 2017). *In vitro*, a cell-free expression system is used to produce the enzyme from the encoding gene (Ryckelynck et al. 2015; Fallah-Araghi et al. 2012). A fluorogenic substrate is then usually added either by fusion or pico-injection and the droplets are then sorted for activity based on the fluorescent signal. Starting from the selected mutants, a new library of mutants can be created and subjected to another evolution round for directed evolution. After several rounds, evolved versions of the enzyme of interest are obtained. For example, using such an approach, genes encoding for an active ß-galactosidase were selected *in vitro* against genes coding for an inactive mutant (Fallah-Araghi et al. 2012). Other examples are the *in vitro* evolution of a ribozyme, the X-motif, with the aim of increasing the rate of product release (Ryckelynck et al. 2015), the optimization of an artificial aldolase by directed evolution in *E-coli* to almost challenge natural enzymes (Obexer et al. 2017) and the screening for the activities of several heterologous enzymes by exploiting the secreting capacities of the yeast *Yarrowia lipolytica* (Beneyton et al. 2017).

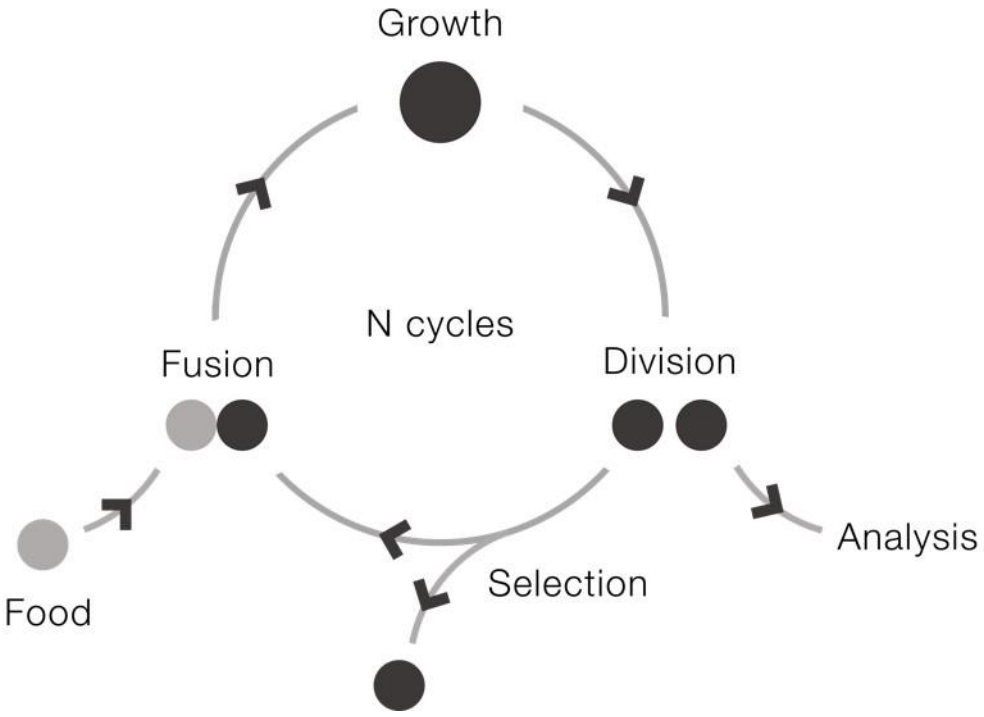


Figure 9. Schematic illustration of an evolution work-flow that could be performed with droplet microfluidics. In this set-up, prebiotic relevant replicating entities are encapsulated in droplets and they go through cycles of growth, division, selection. Fresh food molecules on which the system depends for its growth could be easily introduced at some point of the cycle by fusion. Optionally, if the chosen experimental system does not already possess that, a step introducing variation could be added. Finally, there could also be a step that allows for the analysis of the population of replicating entities at each cycle.

For origin of life studies, these types of work-flows could easily be adapted to prebiotically relevant systems (Figure 9). In such an experiment, the studied system would first be allowed to replicate inside droplets. The replication outcome could be linked to a fluorescent reporter, for example added by picoinjection, that could be used to select droplet using FADS (Baret et al. 2009) or any other droplet sorting method based on fluorescence. This step would effectively put a selection pressure of the prebiotic systems of interest. The selected droplets would then be used for another round of replication/selection. Variation could be intrinsically implemented in the experimental model or could be introduced by an additional step in the work-flow, though its precise nature would depend on the studied system. If such an experiment is achieved, one could potentially observe Darwinian processes *in vitro* demonstrating a form of primordial evolution

with a non-living system. There are many problems to overcome before being able to perform such an experiment. First, the choice of the prebiotic system of interest is itself a challenge as we have seen that there are not so many self-replicating biochemical systems and it is not obvious which would be the easiest to adapt for such a work-flow. Second, the replication mechanism which will be system-dependent has to be implemented in such a way that not only the most successful versions are selected but also that the selected versions give offspring for the next generation. Here, a split and fuse strategy would most likely be part of the solution. Finally it is not guaranteed that variations can be easily introduced in the chosen system which is problematic as variation is a required ingredient for Darwinian evolution.

1.4. Thesis overview

The question of the origins of life despite years of continuous research effort remains unsolved and many pieces of the puzzle are still missing. It is believed that RNA played a crucial role in the process because of its central place in modern biology. As a consequence, a search is currently underway for a generalist RNA molecule that could self-replicate by template-directed polymerization from RNA monomers. This popular hypothesis that life started out with such a molecule is backed up by the developments of prebiotic chemistry that worked out synthesis routes for many biological precursors in young Earth conditions. This scenario as it is has however some limitations since it requires the spontaneous appearance of a RNA molecule long enough for it to be highly accurate and not to disappear because of polymerization errors. These limitations have fueled an alternative view that is not totally incompatible since this alternative scenario could have taken place just before and could have led to the advent of a generalist RNA autoreplicase. According to this other line of research, there was a stage in the origins of life where the evolving entities were collections of interdependent molecules. This concept has been well formalized theoretically, notably with the notion of collective autocatalytic sets (CAS) where the formation of each molecule in the set is catalyzed by another member. Despite the strong theoretical interest, only few experimental systems of CAS exist. They are based on peptides, small organic molecules or oligonucleotides. Among these, the *Azoarcus* recombination ribozyme system allows to form experimentally many diverse CASs. In this system, ribozymes can catalyze the formation of other ribozymes from a pool of common fragments. Three ingredients required for Darwinian evolution are often evoked: heredity, variation and selection.

Providing experimental demonstrations for some of them would represent strong evidence for arguing for the central role of CASs in the origins of life. In this doctoral thesis, we report progress towards this goal with the *Azoarcus* system. In chapter 2, we focus on how variations in CASs could arise from environmental perturbations such as the appearance of novelty in the food set. By exploring a large diversity of RNA CAS, we identified the important structural parameters controlling the response of a RNA network to the addition of a new species. In particular, we

found that the network link density is an efficient buffer for attenuating the response and that pre-existing catalysts similar to the new species would have a significant diminishing effect while the effect of the number of targets of this new species would only be subtle. To achieve this, we used a set-up composed of droplet microfluidics coupled with high-throughput sequencing to perform a large scale study of thousands of *Azoarcus* CASs. In chapter 3, we characterized in *Azoarcus* networks the memory of initial conditions, a pre-requisite of heredity. We found that a small set of network features are governing memory by computationally studying the evolution of networks when one node is seeded at start. Besides the network link density, both the nodes uniqueness in terms of catalysis and the number of downstream connections have particularly significant effect. We further validated our analysis and our model with a diverse experimental dataset. Finally, in chapter 4, we provide evidence for the existence of catabolic processes that allow for an expanded repertoire of fragments to be used in the food set. We showed by providing the *Azoarcus* with modified fragments that the system could transform these into useable substrate for the canonical anabolic assembly process. Although still a lot of progress is required, we are convinced that with these combined results, we are closer than before of an experimental demonstration of Darwinian evolution with a purely molecular system.

2. Mechanism of environmentally induced variations and propagation in RNA collective autocatalytic sets

Note: This chapter is adapted from S. Ameta*, S. Arsène* et al., in preparation (* = equal contribution). Co-authors contributions: Sandeep Ameta designed the droplet microfluidics devices. Baptiste Saudemont and Sophie Foulon designed the hydrogel beads. Sophie Foulon prepared the hydrogel beads. Sandeep Ameta and I performed the experiments. Finally, I carried out the data analysis and theoretical work.

2.1. Abstract

A self-sustaining replicating system is extremely unlikely to have appeared spontaneously on the prebiotic Earth. In this regard, collectively autocatalytic sets (CASs), where in an ensemble molecules can replicate each other, can be seen as a probable scenario for prebiotic evolution. Though CASs have been extensively investigated theoretically, experimental studies to explore a diversity of them are still scarce. Most importantly we are still lacking the understanding of CAS structural features which dictate their response to environmental perturbations, their propagation, and evolution. Here we empirically investigate keys features of CASs by employing a high-throughput experimental set-up combining droplet-based microfluidics and single droplet-level sequencing. We use RNA replicators from group I intron ribozyme of *Azoarcus* bacterium which catalyzes the assembly of other ribozymes from smaller fragments (the food set) in an autocatalytic manner and form diverse RNA networks. With this novel approach, we analyzed the salient features of more than 20,000 RNA CASs with more than 1,800 unique compositions. The results demonstrate general trends relating the network connectivity and species fraction within a CAS. Analyzing the changes in species fraction upon addition or removal of a member in CAS as environmental perturbations, we have identified the set of parameters controlling the network

response to the perturbations. In general, such perturbations shuffles the ranking of other members as a function of connectivity and CASs structure. In particular, we have concretely established that the densely connected networks are less sensitive against perturbations than sparser ones, and we derived simple rules for tailoring network robustness against such changes. Finally, we experimentally illustrated a prebiotic relevant scenario where variations in autocatalytic sets could arise from environmental perturbations and dictate the fate of the next generation upon propagation.

2.2. Introduction

Life probably did not appear on Earth with all the required ingredients to propagate at once but was certainly built through chemical, then biological evolution, starting from elementary interactions between simple molecules. A first milestone in the origins of life may have been the emergence of a self-replicating system (Szathmary 2006) ideally thriving on diverse building blocks. Though significant progress have been made in the search for a template-based RNA self-replicator (Horning & Joyce 2016; Kim & Joyce 2004; Wochner et al. 2011), the process which led to the appearance of self-sustaining biomolecules is still elusive. Even though the synthesis routes for small RNA fragments have been identified in prebiotic conditions (Bowler et al. 2013; Adamala & Szostak 2013; Deamer et al. 2006; Orgel 2004; Huang & Ferris 2006; Huang & Ferris 2003), a path to an efficient self-replicating system is yet to be paved. However, irrespective of its chemical nature, in order to persist such a system has not only to be self-sustaining but also robust against many challenges such as information decay by environmental changes or mutations, substrate limitation, and competition against replicating parasites (Kun et al. 2005; Matsumura et al. 2016). A non-template-based replication system where in an ensemble molecules can replicate each other could overcome several of these limitations and could have preceded template-based replication (Szathmary 2006; Kun et al. 2005; Eigen 1971; Eigen & Schuster 1977; Higgs & Lehman 2015; Levy & Ellington 2001). In this regard, collectively autocatalytic sets (CASs) (Eigen & Schuster 1977; Kauffman 1986) could have been critical for the origin of life. These have been well characterized from a theoretical standpoint (Hordijk & Steel 2015; Hordijk & Steel 2004; Lee et al.

1997; Segré et al. 2000), and yet only a few experimental examples exist of CASs experimental systems based on small-molecules (Butlerow 1861), peptides or oligonucleotides. The range of CAS network structures explored so far is narrow either due to inadequacy in CAS diversity or lack of large-scale study to draw quantitative interpretations (Hayden et al. 2008; Lee et al. 1997; Sievers & von Kiedrowski 1994). As a result, even though some scenarios for their emergence and propagation have been studied theoretically (Hordijk et al. 2012; Jain & Krishna 2001; Nghe et al. 2015; Vasas et al. 2012), CASs have been too scarcely explored experimentally to understand the ingredients required to test such scenarios. One well-studied CAS system is composed of RNA fragments from group I intron of *Azoarcus* bacterium (Reinhold-Hurek & Shub 1992), capable of collective replication using recombination mechanism with some degree of specificity (Vaidya et al. 2012). The RNA networks formed with this system have been explored for demonstrating cooperativity (Vaidya et al. 2012), game-theoric like dynamics (Yeates et al. 2016), or for studying the dynamics of three-membered cores (Yeates et al. 2017). However, in order to propose a plausible scenario for the origin of life and demonstrate evolvability with such system, a quantitative approach is required as a first step and where a large number of diverse CASs can be formed and analyzed at a high-resolution. Such an approach could allow to draw general rules which would relate the distribution of the relative importance of the species involved in a CAS and the interactions between them. In particular, it is critical to determine the important structural parameters that dictate the distribution of species fraction when perturbed by the addition of a new node, in order to empirically demonstrate evolutionary scenarios where such environmental variations can lead to heritable changes in CAS composition.

2.3. Results and discussion

2.3.1 Droplet microfluidics set-up to construct and study a large library of RNA CASs

To put in place such an approach, we have developed a droplet microfluidic-based experimental system where a large diverse library of RNA CASs can be created and composition of each can be analyzed by next-generation sequencing at an unprecedented resolution (Figure 10. A). Using

RNA fragments (WXY, Z) from *Azoarcus* ribozyme system (Figure 10. B), we have constructed RNA CASs where the full-length ribozyme (WXYZ) is assembled in an autocatalytic process using specific 3 nt interactions between internal guide sequence (IGS) and target sequence (tag) at the 5' and 3' end, respectively (Hayden et al. 2008; Draper et al. 2008). In these catalytic networks (Vaidya et al. 2012) 'nodes' are the ribozymes (WXYZ) and the interactions between the nodes are weighted directed edges where the upstream node is a catalyst for the downstream node (Figure 10. C).

We constructed a large set of RNA CASs in water-in-oil emulsions using our droplet microfluidics experimental set-up (Figure 10. A). In this set-up, 24 different 5 pL emulsions ('initial emulsion', each containing a unique combinations of WXY RNA fragments, Table 1) are produced and fused together (at ~100 Hz) with a 50 pL emulsion (containing reaction buffer and Z fragment) in a combinatorial fashion where 2-3 droplets are merged together by electro-coalescence (Figure 11. A). This results in an emulsion where CASs containing 1-16 nodes and diverse compositions are formed after incubation. To determine the composition of CAS in each droplet we developed a droplet barcoding strategy based on single-cell transcriptomics protocols (Macosko et al. 2015; Klein et al. 2015). Here barcoded hydrogel beads are produced by a split-and-pool method, mixed with all the reagents necessary for reverse transcription, singly encapsulated in 50 pL droplets and fused with the CAS containing droplets (Figure 11. B). Release of barcodes from beads (using a restriction endonuclease) and reverse transcription is performed inside the droplet in order to barcode each RNA with a unique barcode per droplet. Droplets are then broken down, barcoded cDNAs are amplified to append sequencing adaptors and subjected to high-throughput sequencing.

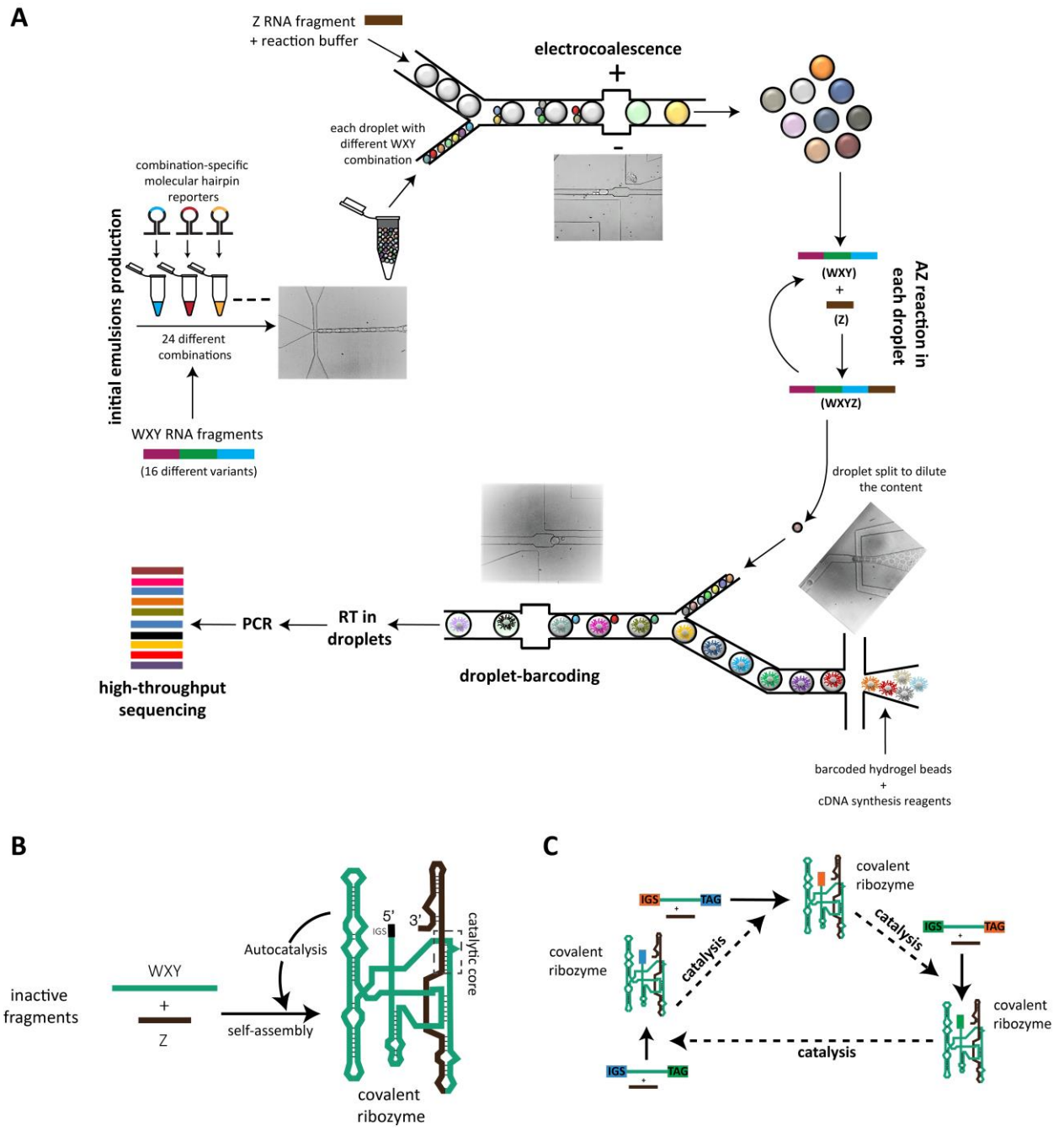


Figure 10. Droplet microfluidic experimental set-up used to analyze thousands of Azoarcus RNA networks. (**A**) The different steps involved in the droplet microfluidic set-up are shown here some of which are detailed in the following figures. At first 24 different initial emulsions (5 pL) were prepared each containing a unique combination of WXY fragments (Table 1) and a unique molecular RNA hairpin reporter. These emulsions are collected in a single tube (1.5 mL), mixed thoroughly and fused with 50 pL droplets containing Z fragment in excess and reaction buffer. The fusion is done in an electrocoalescence device where the pairing of 5 pL and 50 pL droplets are controlled (using air-pressure controlled pumps) such that 2-4 5 pL droplets fused with one 50 pL droplet. The fusion process is usually typically carried out for 3-4 h.

After fusion the resulting emulsion (~65 pL) is incubated at 48°C for an hour and splitted into 5 pL droplets in order to dilute the reaction droplet content before fusion with droplets containting reverse-transcription (RT) reactants. For RT in droplet, barcoded hydrogel beads mixed with all the reagents required for RT are singly encapsulated in ~50 pL droplets and fused with splitted RNA droplets using electrocoalescence. The fusion frequency is kept low such that only one 5 pL RNA droplet is fused with one 50 pL barcoded hydrogel bead droplet. After fusion, RT droplets are incubated at 60°C for 1 h before being broken down to recover the barcoded cDNA which are then PCR amplified to append sequencing adaptors and subjected to high-throughput sequencing. (**B**) Autocatalytic self-assembly of *Azoarcus* ribozyme. WXY and Z fragment self-assemble to form WXYZ *Azoarcus* ribozyme. (**C**) Three membered cooperative network (CAS) formation by *Azoarcus* ribozymes where each node (WXYZ ribozyme) catalyze the synthesis of another node by specific interaction between IGS and tag.

Table 1. Initial emulsions compositions. The different combinations of WXY fragment used to prepare the set of intial 5 pL emulsions. Here all possible 16 $_{gag}$WXY$_{cnu}$ RNA fragments are used where M is the middle nucleotide of IGS and N is the middle nucleotide of TAG.

Initial emulsion no.	$_{gmg}$**WXY**$_{cnu}$ **fragments (MzN) composition**
1	UzA
2	CzA
3	GzA
4	AzG
5	CzU
6	GzC
7	UzU
8	CzU, UzG
9	AzC
10	CzA, GzA, CzG, UzA, AzU
11	GzG, GzC
12	AzU
13	CzA, UzC, GzA, CzC
14	AzC, AzG
15	UzA, GzU, GzG, CzG, UzG, AzG, UzU, UzC
16	GzG
17	AzU, CzA, AzA, AzG
18	GzC
19	GzU
20	CzC, GzU, CzA
21	GzG, GzU, CzC
22	UzU, CzU
23	CzU, CzA
24	UzC

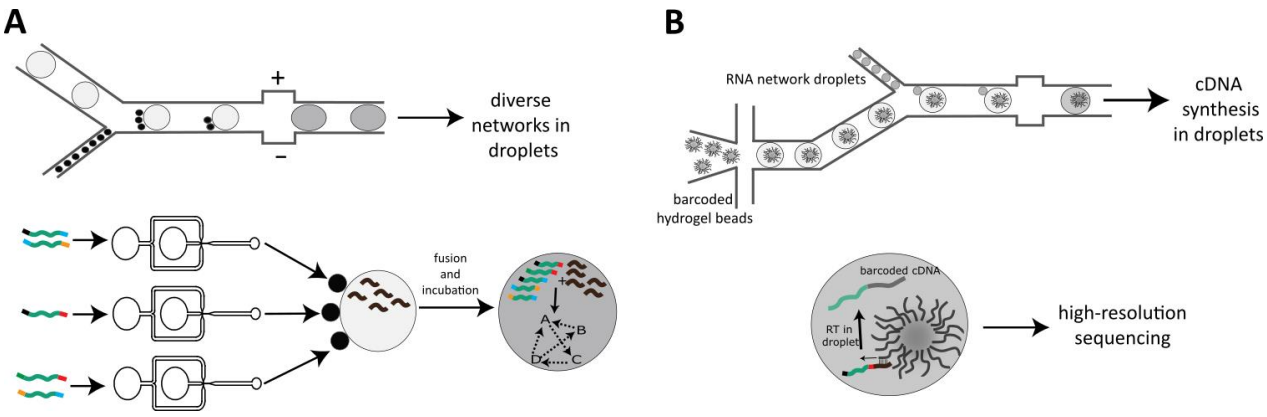


Figure 11. Detailed description of the two main steps of the droplet microfluidic set-up (**A**) Combinatorial droplet-fusion strategy to create a diverse set of CAS. Here at first different sets of IGS/tag combinations of WXY fragments are encapsulated in 5 pL droplets ('initial emulsion', Table 1) and then fused with 50 pL droplets containing reaction buffer and excess of Z fragment in a combinatorial way (two-three 5 pL droplets per 50 pL droplet) using electro-coalescence (Sciambi & Abate 2015). After fusion, emulsions are incubated at 48°C for an hour. (**D**) Droplet-barcoding strategy for sequencing RNA composition of each droplet at a high-resolution. The hydrogel-beads, each containing a clonal population of unique barcode, are prepared by split-pool method, singly encapsulated with all the reagents required for reverse-transcription (RT) and then fused with RNA CAS containing droplets. RT is performed inside the droplets to generate barcoded cDNA and then PCR amplified after breaking the emulsion.

2.3.2 Validation of the droplet microfluidics set-up

With our experimental droplet microfluidics set-up, we get as a result, the number of unique reads (Kivioja et al. 2012) for each covalently assembled WXYZ molecule within each network which allows the quantitative measurement of the relative ribozyme content. Note that the original IGS for each RNA within the droplet is identified using an internal mutation (Section 2.4. Material and methods). Additionally, the composition of the network in each droplet is identified using hairpin reporters, sequenced together with WXYZ in each droplet, coding for the 24 initial emulsions of WXY fragments (Figure 12. A, Table 1). The combined information of the ribozyme and the hairpin reporter content in each droplet gives us enough information to validate rigorously the complete experimental set-up.

First, the distribution of the number of different hairpin reporters detected per droplet matches very well the expected distribution which we obtained by monitoring the first fusion step of the set-up by video acquisition (Figure 12. B). Additionally, the resulting distribution of the number

of nodes per network matches also very well the expected distribution also obtained using the data from video acquisition to simulate the combinatorial fusion step which creates the library of *Azoarcus* networks (Figure 12. C). We computed for each droplet barcode a purity score which is equal to 1 if all the WXYZ detected in the droplet are expected given the network structure determined by the set of hairpin reporters identified in this droplet. The distribution of droplet barcode purity score is very skewed to the perfect score of 1 which demonstrates that we are well able to identify the network in each droplet with hairpin reporters (Figure 12. D). Additionally, because a lot of networks are present in several droplets, we compared the number of replicates per network to the expected number obtained as before by simulating the network library making process and we found that both are in very good agreement (Figure 12. E). Finally, we sequenced the left-over mix of initial emulsions for both replicates of the complete experiment and found that proportions of each initial emulsion match very well the proportions of each hairpin reporter detected (Figure 12, F). The fact that this distribution is quite different between the two replicates shows that it is difficult during the very first step of the droplet microfluidic set-up to produce the 24 initial emulsions in equal proportion.

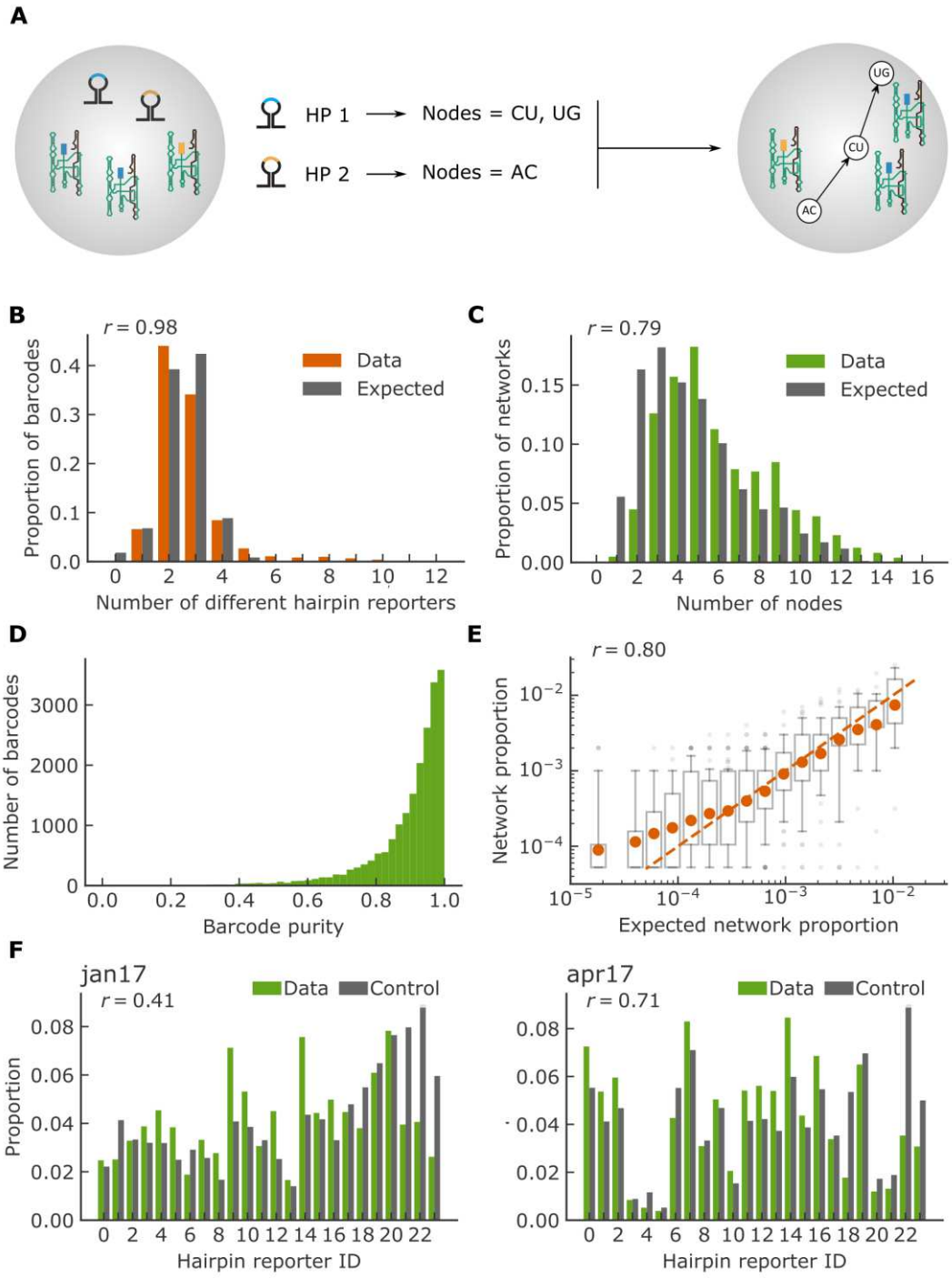


Figure 12. Identification at the droplet level of the network structure by the coding set of hairpin reporters matches the expectations and the controls. (**A**) Schematic illustrating identification of the network structure by the coding set of hairpin reporters. Hairpin reporters with more unique reads than a given percentage of the total number of unique reads for hairpin reporters are considered as part of the coding set. From this the network structure can be derived. Numbers of unique reads for each node of the identified network structure give the nodes fractions. (**B**) In orange, distribution of number of hairpin

reporters that pass the threshold per droplet barcode in the final dataset. In dark grey, distribution of number of 5 pL droplets (containing initial WXY emulsions) fused with a 50 pL droplet (containing reaction buffer and Z) measured by video acquisition. Pearson correlation coefficient is reported. (**C**) In green, distribution of number of nodes in the network per droplet barcode in the final dataset. In dark grey, expectations from simulating the library fusion step with the same number of droplets as in the data and with the fusion frequencies measured by video acquisition. Pearson correlation coefficient is reported. (**D**) Histogram of droplet barcode purity which is the ratio between the total numbers of unique reads associated with nodes that are coded for by the identified coding set of hairpin reporters and the total number of unique reads for *Azoarcus* ribozyme for a droplet barcode. A value of 1 is reached when all the unique reads associated with *Azoarcus* ribozyme are expected given the set of coding hairpin reporters. (**E**) Network proportion is the ratio between the number of replicates of a given network and the total number of replicates for all networks. Here network proportion in the data is plotted against expected network proportion obtained by simulating the library fusion step with 50,000 droplets as in the data and with the fusion frequencies measured by video acquisition. Typically 1/3 of the network in the data do not have any replicate in the simulation and are excluded from the analysis. Data points were binned in 30 $\log_{10}$-spaced bins between 10^{-5} and 1. Bins with less than 10 points are discarded. The box extends from the lower to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5th percentile to the 95th. Flier points are those past the end of the whiskers. Dotted orange line is the identity line. Pearson correlation coefficient is reported. (**F**) In green, proportion of identified coding sets of hairpin reporters that contains a given hairpin reporter for both replicate of the experiment (left and right) in the final dataset. In grey, proportion of unique reads for each hairpin reporter obtained by sequencing the left-over mix of initial emulsions. Pearson correlation coefficient is reported. The green and the grey profile are similar for each replicate, however the profile is different between the two replicates indicating that there is some randomness in the first step during which the 24 initial emulsions are made.

2.3.3 Acquisition of large dataset comprising of more than 20,000 *Azoarcus* networks and estimation of the precision of our measure

Using our droplet microfluidic set-up, we gathered high-quality data from more than 20,000 droplets summing up to >1,800 networks, unique in structure and composition (Figure 13). In this representation of the gathered data, each ray is a network unique in structure and composition but which can be present in several replicates. Each ray is composed of 16 blocks for the 16 possible nodes with a white block indicating that the corresponding node is not present in the network. The links connect two networks that differ only by the addition or the removal of a node. The high density of links indicates that the combinatorial fusion strategy allows us to access to a large portion of the compositional neighborhood of most networks in our data set.

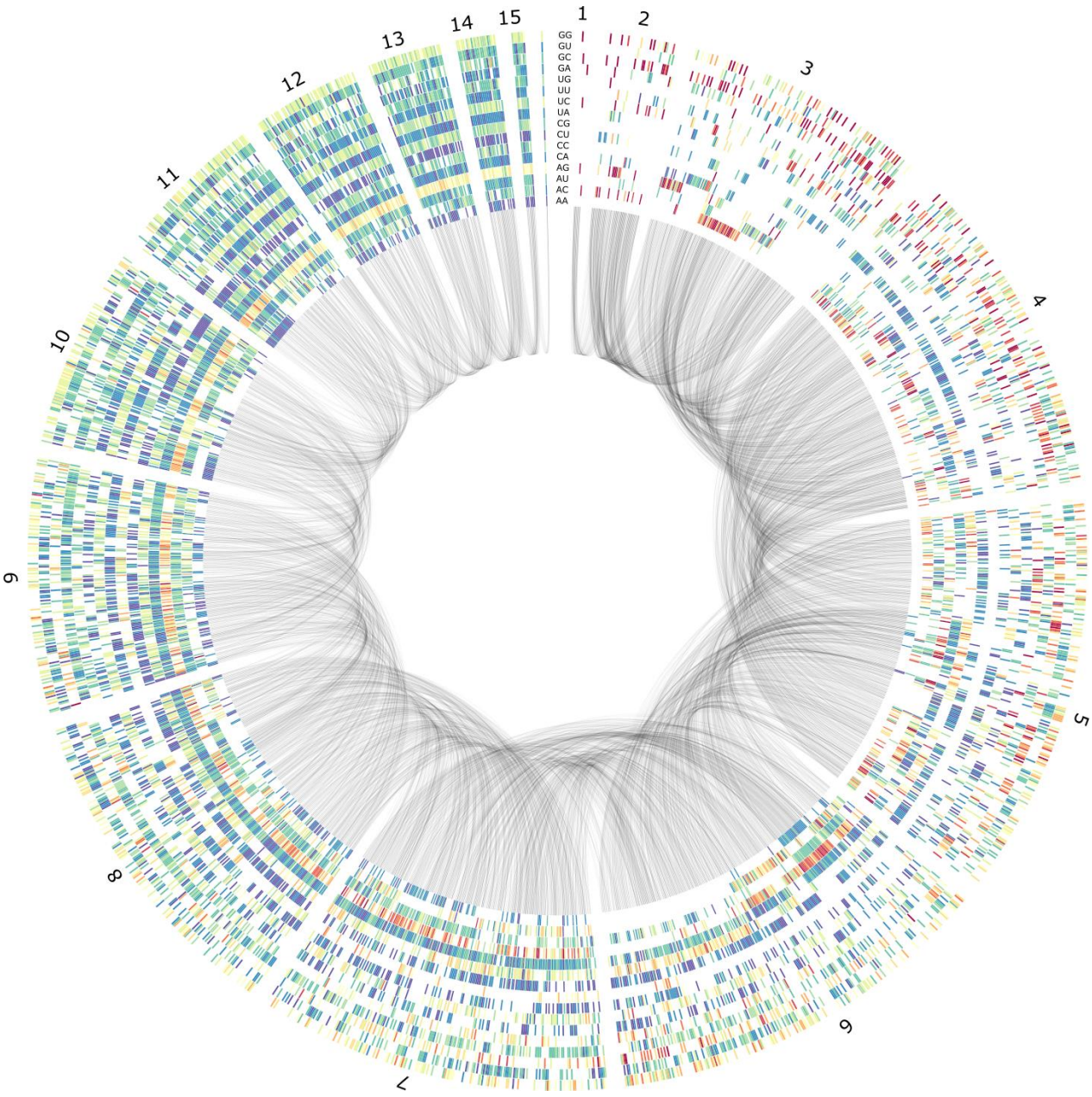


Figure 13. Graphical representation of the all the unique networks data obtained by sequencing the RNA composition of each droplet with our strategy. Here all the networks are arranged circularly (each ray is a network) by number of nodes. Within each block, the networks have the same number of nodes (form 1 to 16 nodes). Within each ray, there are 16 colored box tracks for each of 16 possible nodes. The color represent the relative proportion of this node in the network. A white block indicates that the node is absent in the network. The grey lines between each ray represent the network neighborhood where the networks differ from each other by just one node.

The data is composed of more than 50% networks in duplicates (and ~30% in quintuplicates) (Figure 6.A) allowing low variations on the measured relative proportions of each node in the networks (mean standard deviation of 5 %, Figure 6. B). The standard deviation is maximum for node fraction around 0.5 which is expected as the standard deviation of a fraction y is of the following form: $s.d. = \sqrt{y(1-y)/N}$ where N is the sample size. Additionally, two independent technical replicates of the entire experimental set-up further confirmed the precision of the measure (Figure 6. C). Finally, a control experiment where different ribozymes with known proportions are sequenced in the same way ruled out any quantification biases and cross-talk between droplets introduced by the experimental set-up (Figure 6. D).

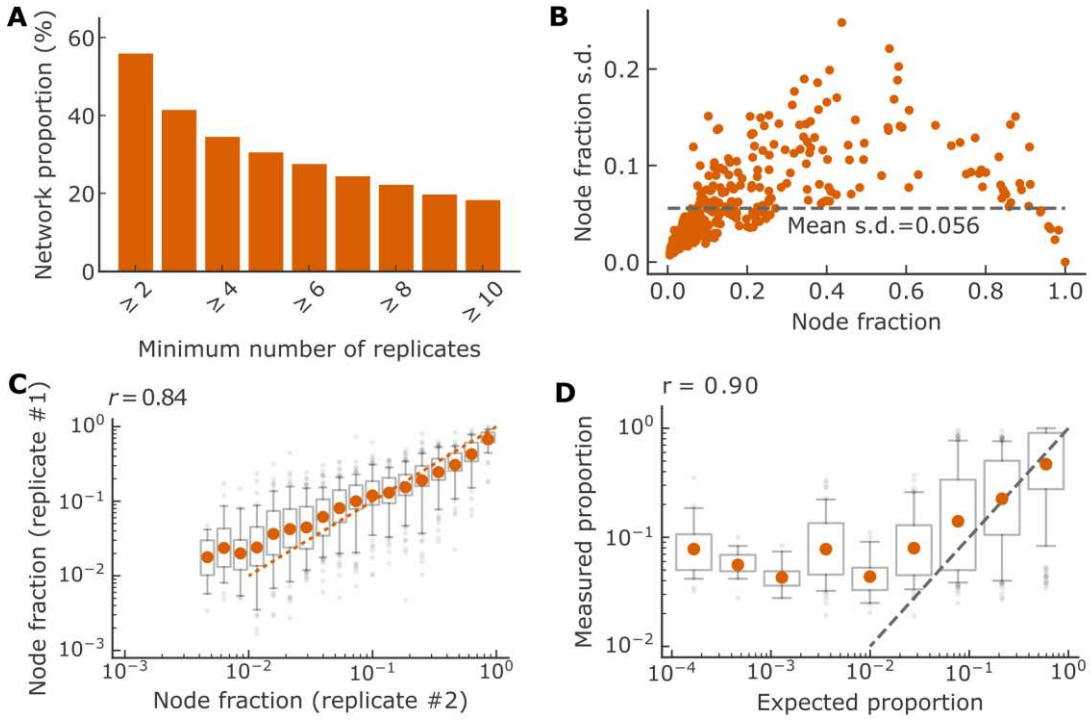


Figure 14. Assessment of the precision and the validity of measurement of nodes fractions with our set-up using replicates within a single experiment, two technical replicates of the complete experiment and an experiment with ribozymes in known proportion. (**A**) Proportion of unique network that have at least a given number of replicates. More than 50% of the networks in the final dataset have at least duplicates and about 20% of them have at least 10 replicates. (**B**) Node fraction standard deviation as a function of node fraction for nodes in network with at least 50 replicates. Dotted grey line is the mean of the distribution of standard deviation. (**C**) Node fractions in two technical replicates. For this data, the mean of node fraction per unique network is taken separately for the two replicates. The data is then restricted

to the common set of unique networks to the two replicates. Data points were binned in 30 $\log_{10}$-spaced bins between 10^{-4} and 1. (**D**) Measured proportion is plotted against expected for ribozymes in droplet with known proportions in a similar experiment using our droplet microfluidic set-up. We prepared for this four populations of 5 pL droplets with 4 different ribozymes each in known concentrations. We mixed these four populations, barcoded and sequenced their content the exact same way as for the droplets containing *Azoarcus* networks. Data points were binned in 10 $\log_{10}$-spaced bins spanning the range of the data. For (**C**) and (**D**), bins with less than 10 points are discarded. The box extends from the lower to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5th percentile to the 95th. Flier points are those past the end of the whiskers. Dotted line is the identity line. Pearson correlation coefficient value is reported.

2.3.4 Detailed description of interactions in *Azoarcus* RNA networks

Ribozymes from the *Azoarcus* system can form directed networks where the interactions ('edges') are weighted ('value over the edge'). An example of such network is shown in Figure 15 (left). The edges' weights quantify how well an upstream node catalyzes the synthesis of its downstream node(s), and are determined by the base-pairing interactions between the 'IGS' of the upstream node and the 'tag' of the downstream node(s). We consider that the rate at which a node catalyzes the formation of another is the sum of two terms. The first term is the product of the catalyst's concentration and of $k^a_{M\to N}$ (min^{-1}) which is the catalytic constant between IGS 'M' and tag 'N'. The second term $k^b_{M\to N}$ ($\mu M.min^{-1}$) quantifies the spontaneous assembly by the non-covalent complexes between WXY and Z fragments. To measure these two sets of rates, the rate of formation of some node from corresponding WXY fragment can be monitored with some full-length covalent catalyst doped in at different concentration. The formation rate linearly depends on the starting concentration of catalysts with the slope of this relationship being $k^a_{M\to N}$ while the intercept is $k^b_{M\to N}$. This was previously measured for all 16 MN pairs (Yeates et al. 2016) but in different conditions than used here (100 mM $MgCl_2$ and 0.5 µM of WXY and Z fragments). We changed the conditions (20 mM $MgCl_2$, 0.1 µM of WXY and 1.6 µM Z fragments) and found that these rates do not compare to each other in the same proportion (Table 2). Notably, the $G \to C$ and $C \to G$ pairs have rates values in our conditions that are 3-5 times higher than the values of the other two pairs. For the 'value over the edge' in our graph description of the system, we chose to simplify by taking the mean for similar pairs (such as $G \to C$ and $C \to G$ or $A \to U$ and $U \to A$).

We also chose to take the $k^b_{M\to N}$ values rather than the $k^a_{M\to N}$ values since near the start of the reaction the first term is negligible compared to the second term.

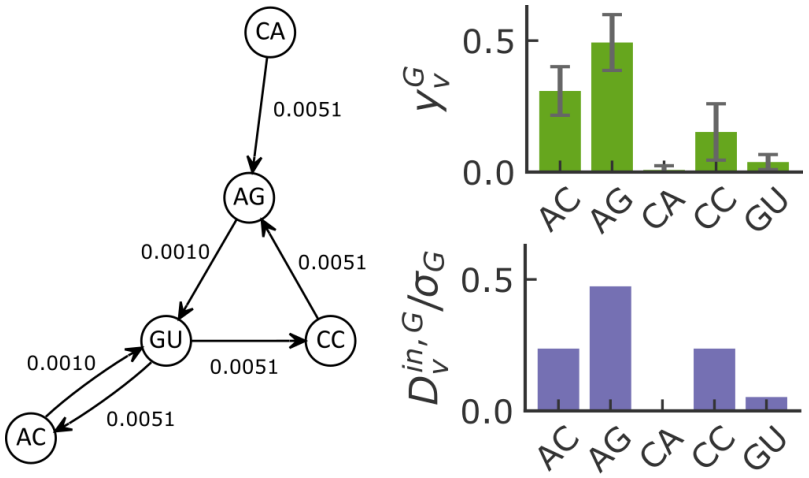


Figure 15. Left: schematic of an *Azoarcus* network G with 5 nodes. Values above the edges represent the strength of the catalytic interaction between the upstream and the downstream nodes in μM.min^{-1}. Right: top, relative fractions of the nodes in G measured in the data with 100 replicates (error bars represent ± 1 standard deviation); bottom, in-degree centrality of the nodes in G where $D_v^{in,G}$ is the in-degree of node v (sum of incoming links to the node v) and σ_G is the sum of all the links in the network.

Table 2. Catalyzed and spontaneous rates of formation. A catalyst with M as IGS catalyzes the formation of a catalyst with N as tag at a rate equal to k^a_{MN} times its concentration. WXY fragments with M as IGS will spontaneous form a complex with Z fragment and contribute to the rate of formation of a catalyst with N as tag at a rate equal to k^b_{MN}. The 20 mM values were measured by Bryce Clifton and Niles Lehman for this study and the 100 mM values were measured by Jessica Yeates and Niles Lehman for another study and are reported here for comparison. The mean of three independent measurements is reported along with the 95% confidence interval when available.

$MgCl_2$	20 mM		100 mM	
Rates	$k^a_{M\to N}$ (min^{-1})	$k^b_{M\to N}$ (μM.min^{-1})	$k^a_{M\to N}$ (min^{-1})	$k^b_{M\to N}$ (μM.min^{-1})
Interaction				
$C \to G$	0.0495 ± 0.0070	0.0047 ± 0.0002	0.0319	0.0183
$G \to C$	0.0301 ± 0.0083	0.0055 ± 0.0013	0.0075	0.0091
$U \to A$	0.0158 ± 0.0024	0.0015 ± 0.0001	0.0073	0.0054
$A \to U$	0.0135 ± 0.0010	0.0004 ± 0.0001	0.0096	0.0040

2.3.5 Influence of the network structure on a node's fraction

The catalytic interactions very likely dictate the relative concentrations of nodes in the CASs and analyzing a diversity of CAS should allow us to determine how the network structure shapes the distribution of relative concentration of nodes. By examining the relative fractions of nodes in the dataset we acquired using our experimental droplet microfluidics set-up, we indeed found that the measured fraction of a node v in a network G, noted as y_v^G, can be well determined by how well it is fed-in by other nodes in the network.

Precisely, it can be measured by its weighted in-degree centrality $D_v^{in,G}/\sigma_G$, where $D_v^{in,G}$ is the in-degree of node v (sum of incoming links to the node v) and σ_G is the sum of all the links in the network (Figure 15, Figure 16). On the contrary, the out-degree centrality $D_v^{out,G}/\sigma_G$ does not show any correlation with the node fraction (Figure 17). Notably, the in-degree centrality is a better determinant of a node's fraction for the smaller networks (Pearson's correlation coefficient ~0.8 for networks with ≤ 6 nodes and ~0.6 for networks ≥ 10 nodes) than the bigger ones. We attributed this to the fact that the differences in fraction tend to be less important in larger networks and thus difficult to measure. As the node fraction can be estimated by its in-degree centrality, it empirically derives a direct and generic relation between the network structure (genotype) and the distribution of node's fractions (phenotype), which holds true for any number of nodes in the network (Figure 16).

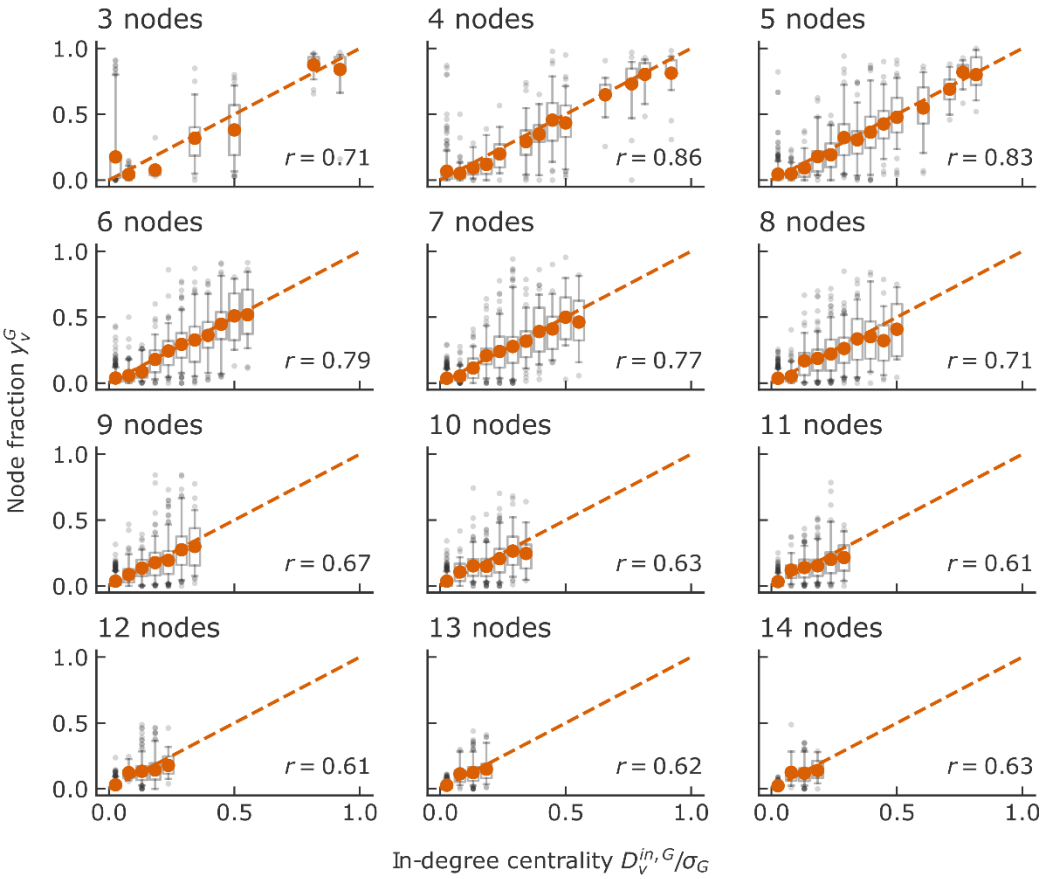


Figure 16. The fraction of a node can be determined by its in-degree centrality whatever the number of nodes in the network. Nodes in network with a given number of nodes are binned in 20 bins linearly spaced between 0 and 1 according to their in-degree centrality. Bins with less than 10 points are discarded. For each bin, a boxplot of the measured node fraction is plotted (mean is taken for the same node in the same network across replicates). The box extends from the lower to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5th percentile to the 95th. Flier points are those past the end of the whiskers. Dotted orange line is the identity line.

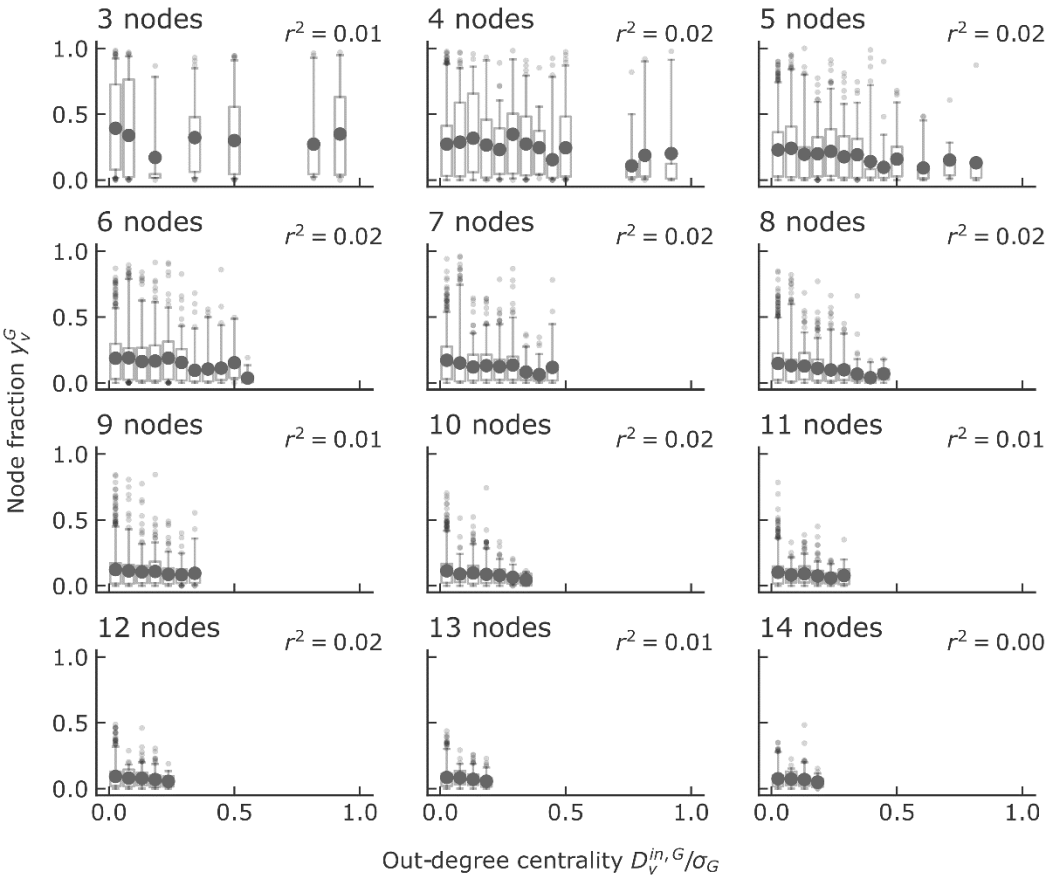


Figure 17. The fraction of a node cannot be determined by its out-degree centrality whatever the number of nodes in the network. Nodes in network with a given number of nodes are binned in 20 bins linearly spaced between 0 and 1 according to their out-degree centrality. Bins with less than 10 points are discarded. For each bin, a boxplot of the measured node fraction is plotted (mean is taken for the same node in the same network across replicates). The box extends from the lower to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5th percentile to the 95th. Flier points are those past the end of the whiskers.

2.3.6 Analytical determination of a small set of network parameters influencing a node's response to a perturbation

Since with our droplet microfluidics set-up we create a library of networks in a combinatorial fashion, we can access most of the compositional neighborhood of our networks. With this, we can analyze how the relative fractions of the nodes of a network are shuffled when the network

is perturbed by the addition of a new node. Focusing on how the relative fraction of a node evolve after such a perturbation, we found that only a restricted set of parameters have a significant effect: the ratio between the network's link density and the catalytic strength of the new node, the number of targets of the new node and the number of pre-existing nodes in the network that are similar to the added node. To demonstrate this, we derive a simple analytical expression quantifying the change in fraction of a node when the network is perturbed and in which these parameters appear.

More precisely, we used a perturbative approach by considering that the addition of node a to an existing network G results in a new network G' where all nodes are common to network G except the added node a (Figure 18. A). The relative change in fraction of common nodes between G and G' is then considered after renormalization. Fraction of node v in G', $y_v^{G'}$ can be taken relative only to the nodes in G (i.e. excluding a) giving $y_v^{G'}|_V = y_v^{G'} / \sum_{u \in V} y_u^{G'}$. Node v response to the perturbation induced by the addition of node a can be measured by $p_v^{G \to G'} = y_v^{G'}|_V - y_v^G$. Using the empirical approximation shown before (fraction of a node in a network is equals to its in degree centrality, i.e. $y_v^G \sim D_v^{in,G} / \sigma_G$) the perturbation $p_v^{G \to G'}$ can be reformulated as:

$$p_v^{G \to G'} = \frac{\frac{D_v^{in,G'}}{\sigma_{G'}}}{\sum_{u \in V} \frac{D_u^{in,G'}}{\sigma_{G'}}} - \frac{D_v^{in,G}}{\sigma_G}$$

$$p_v^{G \to G'} = \frac{D_v^{in,G} + e_{a \to v}}{\sigma_G + D_a^{out,G}} - \frac{D_v^{in,G}}{\sigma_G}$$

$$p_v^{G \to G'} = \frac{e_{a \to v} - \frac{D_v^{in,G} D_a^{out,G}}{\sigma_G}}{\sigma_G + D_a^{out,G}}$$

We can now separate in two cases depending on whether or not the node v is a target for the new node a. If v is target for a, then $e_{a \to v} = e$, and we can introduce the normalized in-degree $m_v = D_v^{in,G} / e$, the normalized sum of all the weights in the network $\sigma_e = \sigma_G / e$ and $n = D_v^{out,G} / e$ which is the number of targets of a in the network (Figure 18. B).

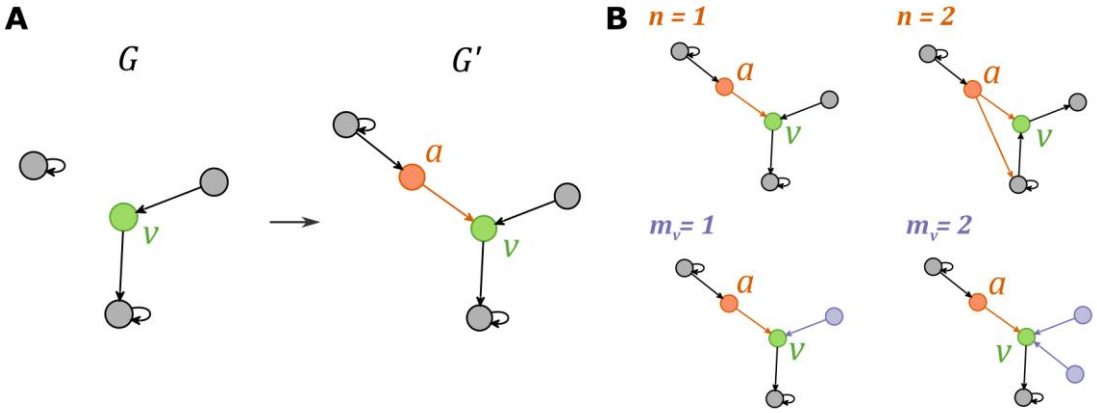


Figure 18. (**A**) Schematic of the addition of node a to network G results in network G'. Here the change in fraction for the node of interest v is measured. (**B**) Schematic of four situations where the parameters n and m_v have different values. Top, left: $n = 1$ as the added node a has only one target. Top, right: $n = 2$ as node a has two targets. Bottom, left: another node despite node a is feeding node v so $m_v = 1$. Bottom, right: tow other nodes are feeding node v is m_v=2.

It follows by dividing by e that:

$$p_v^{G \to G'} = \frac{1 - n \frac{m_v}{\sigma_e}}{\sigma_e + n} \quad (1)$$

Because $\sigma_G \geq n m_v e$, in this case $p_v^{G \to G'} \geq 0$. If v is not a target for a, then $e_{a \to v} = 0$ but we can still divide by e which is the typical weight of an outgoing edge from a. With the same parameters introduced before, it gives in this case:

$$p_v^{G \to G'} = \frac{-n \frac{m_v}{\sigma_e}}{\sigma_e + n} \quad (2)$$

In this case, it is clear that $p_v^{G \to G'} \leq 0$. To summarize, the perturbation ($p_v^{G \to G'}$) which quantifies the change in relative fraction of a node v, caused by the addition of a new node a, can be described by two simple relations depending on whether node v is a target for a or not. Depending on the case, it can be written as following:

$$p_v^{G\to G'} = \begin{cases} \dfrac{1 - n\,\dfrac{m_v}{\sigma_e}}{\sigma_e + n} \geq 0 \; if \; v \; is \; a \; target \; for \; a \; (1) \\ \dfrac{-n\,\dfrac{m_v}{\sigma_e}}{\sigma_e + n} \leq 0 \;\; if \; v \; is \; not \; a \; target \; for \; a \; (2) \end{cases}$$

In the case where v is a target for a, m_v can only be equal to 0, 1, 2 or 3 and it represents then the number of nodes in the network that share the same IGS as a. Using this expression we can quantitatively predict whether the addition of a new species will increase or decrease the relative fraction of the existing nodes (Figure 19). Percentage of correct sign predictions is above random guessing (58% compared to 50%) but more interestingly it increases sharply (to almost reach 100%) when only the perturbation expected to give higher change in fraction are considered. This is expected as we are not able to measure reliably fractions with a precision of 1%. We can as well predict the magnitude of the perturbation within the limit of our measurement precision (Figure 20) and additionally we can observe that $|p_v^{G\to G'}|$ tend to be underestimated by (1) and (2) for larger networks which is consistent with the approximation of a node's fraction by its in-degree centrality becoming less good in networks with many nodes.

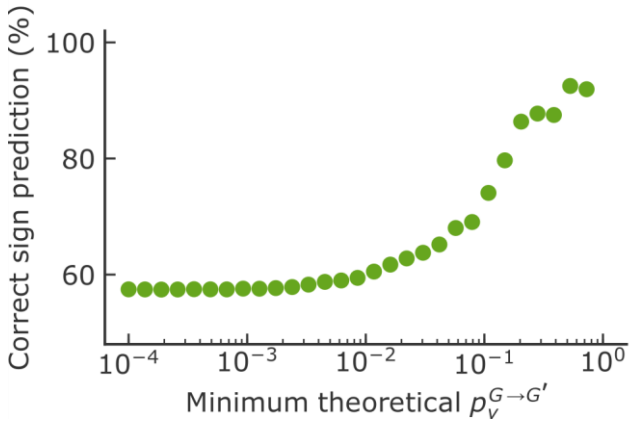


Figure 19. Percentage of correct sign prediction between the measured and the theoretical change of relative fraction of a node upon addition of a new node as a function of the minimum theoretical change taken into consideration. Nodes for which the theoretical change in fraction is below the chosen threshold are discarded and the percentage of equal sign between the measured and the theoretical change in fraction is computed and plotted. The minimums are $\log_{10}$-spaced between 10^{-4} and 1. Here networks with 3 to 11 nodes were considered.

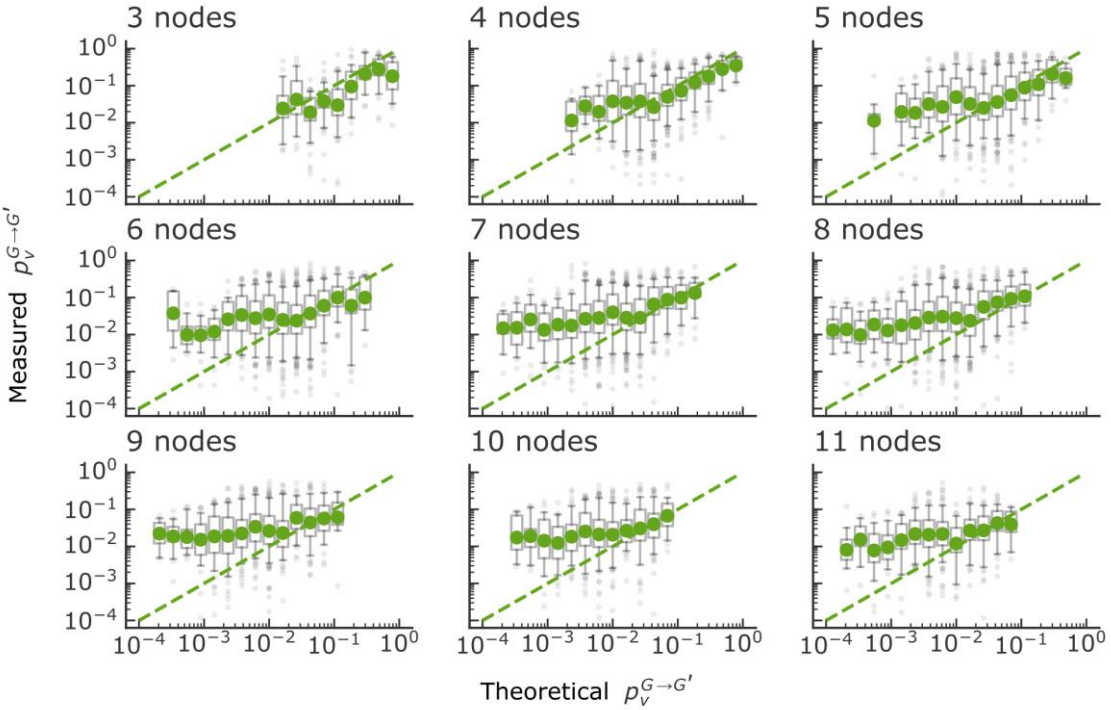


Figure 20. Magnitude of perturbation of a node's fraction upon addition of a new specie can be determined by expression (1) and (2). Measured change in fraction $p_v^{G\to G'}$ in the data plotted against theoretical expectation upon addition of a node for networks with a given number of nodes. Only the data points with measured $p_v^{G\to G'}$ and theoretical expectation of the same sign and not null were considered. Data points were binned in 20 log-spaced bins between 10^{-4} and 10^{0}. Bins with less than 10 points are discarded. The box extends from the lower to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5th percentile to the 95th. Flier points are those past the end of the whiskers. Dotted green line is the identity line.

2.3.7 Differential effects of the identified structural parameters

The analytical expressions (1) and (2) allow us to identify the factors which make some networks more resistant than others against perturbations as with such networks the distribution of their nodes fractions is not very sensitive to the addition of a new node. In particular, we found that the normalized network total connectivity σ_e has a significant buffering effect, decreasing the amplitude of the change in fraction in densely connected networks. We also found that the number of targets of the new node n has only a subtle effect while the normalized in-degree m_v,

which is also the number of pre-existing catalysts similar to the new one, has a strong effect and causes a steep decrease of the response when its value increase.

To analyze expression (1) in more detail we compared two situations where m_v is fixed to 0 and n increase by one unit as a function of σ_e (Figure 21. A). We observed that regardless of the value of n, the relative fraction change is positive and highest for small values of σ_e (it reaches 0.25 for $\sigma_e = 1$ when $n = 1$ and $m_v = 0$, Figure 21. A, orange curve). However, it is quickly attenuated as σ_e increases (attenuating factor; i.e. the ratio of y-axis values; of 1/5 when σ_e increases from 1 to 10 with $n = 1$ and $m_v = 0$, Figure 21. A, orange curve). This indicates that only the nodes in sparse networks are susceptible to high gains in fraction and σ_e act as a buffer, diluting the beneficial effect of the new node. Additionally, we observed than $n = 2$ gives smaller gains in fraction than $n = 1$ (there is a factor of ~0.5 for $\sigma_e = 1.5$, Figure 21. A, top, comparing orange and purple curves), indeed increasing n by one unit is equivalent to adding another target of a besides v, and thus the benefit brought by a is now shared amongst two nodes instead of one, reducing the gain in fraction for v. The decrease in the gain is even more significant when m_v goes from 0 to 1 at constant $n = 1$ (Figure 21. A, bottom, comparing orange and pink curves). While $m_v = 0$ represents a situation when there are no other nodes with the same IGS as a and where v has then no existing in-degree, at $m_v = 1$ there is already a node feeding v. Therefore at $m_v = 0$, a brings a novel catalytic link and the benefit for v can be high provided that σ_e is small enough, i.e. if v is in a sparse network.

Using a similar approach, the expression (2) quantifies the loss of relative fraction when the new added node is not a catalyst for node v. Again, the strong expected attenuating effect of σ_e can be seen as only the nodes in sparse networks (it reaches almost -0.2 for $\sigma_e \sim 1$ when $n = 1$ and $m_v = 1$, Figure 21. B, green curve, top panel) are susceptible to the loss of fraction. In denser networks, σ_e acts again as a buffer and protects the nodes from fraction loss (attenuating factor < 1/10 when σ_e increases from 1 to 10 with $n = 1$ and $m_v = 1$, Figure 3. D, green curve, top panel). The fact that the new node a has more targets in the network, going from $n = 1$ to $n = 2$, has only a slight effect on the relative fraction loss of node v (Figure 3. D, compare green and brown curves). This is expected because the loss in fraction is a result of the collective fraction gain of other nodes in the network and if n increases then a benefits to more nodes but the gain

in fraction is split amongst all the receiving nodes. However, increasing m_v has a more drastic effect on the fraction loss (Figure 21. B, compare green and grey curves, bottom panel) as higher m_v means higher in-degree for v, thus higher fraction before perturbation and therefore more loss after adding a node. In all the cases analyzed here, the data is in very good agreement with the theoretical predictions obtained with expression (1) and (2) (Figure 21. A, B, left panels).

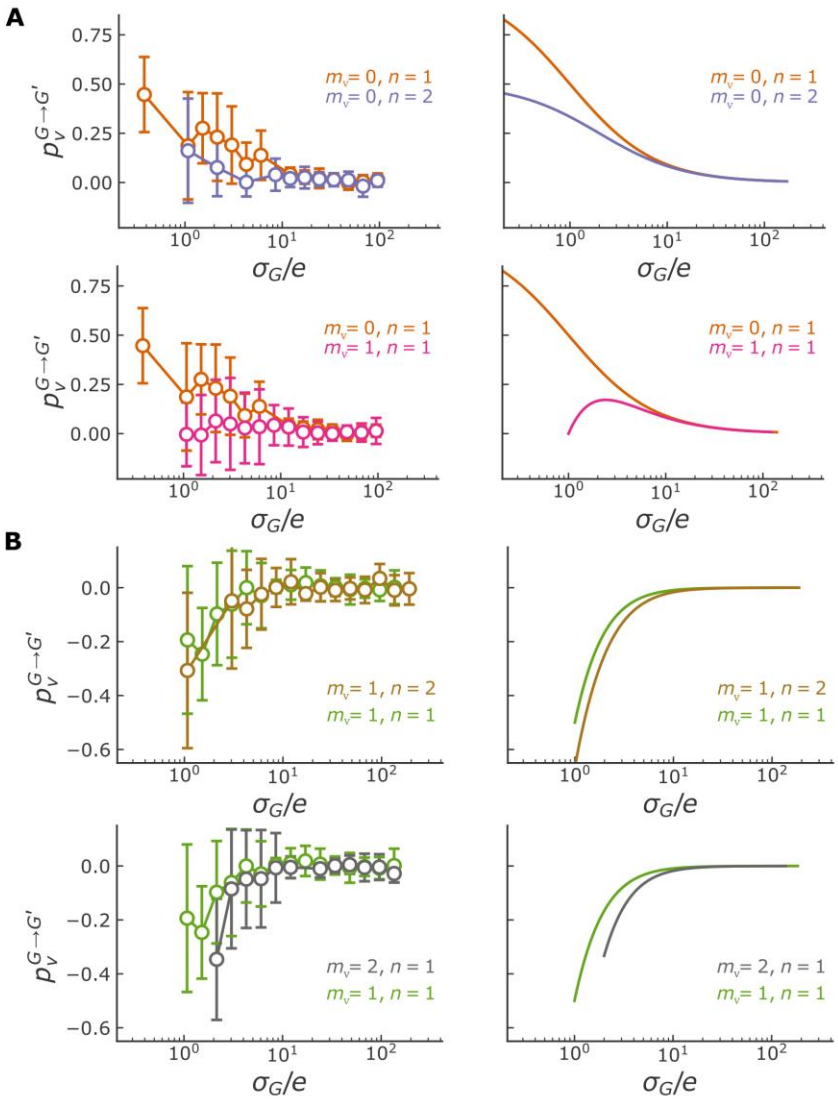


Figure 21. Influence of different parameters on the relative gain or loss of fraction of a node upon addition of a new node to a network. (**A**) $p_v^{G\to G'}$ plotted against σ_e for different values of the parameters m_v and n in the case where v is a target for a and for both the data (left) and the analytical results (right). Data points were binned along x-axis in 20 log-spaced bins between $10^{-0.5}$ and $10^{2.5}$. Bins with less than

3 points are discarded and mean and standard deviation for each bin are plotted. Networks with less than 2 nodes were discarded for the analysis. (**B**) Similar as (A) but in the case where v is a not target for a.

2.3.8 Total network's response to the addition of a new node

While the above analysis focuses on the node-level perturbations, a quantitative evaluation of the network's total response to the perturbation can also be done using $p_{G\to G'} = \sum_{v \in V} |p_v^{G\to G'}|$. This is the sum of the net change in fraction per node and it measures how much the complete distribution of the nodes fractions in G is changed after addition of a new node. By analytically deriving an expression for this, we show that the dependence on network structural parameters is very similar to the previous analysis where we focused on only one node. Notably, here the number of nodes in the network with the same IGS as the new node is now an important parameter while the slight effect of the number of targets is not detected. As before, we found that the magnitude of the response was lower with dense networks, thus with high σ_e, than with sparse ones.

To derive an analytical expression for $p_{G\to G'}$ we consider $V_a = \{v \in V, e_{a\to v} > 0\}$, the set of nodes in G for which a is a catalyst and $V_a^* = \{v \in V, e_{a\to v} = 0\}$, the set of nodes for which it is not. We assume that a is a catalyst for at least one node ($n > 0$) and that there is at least one link in G ($\sigma_G > 0$). $p_{G\to G'}$ can be reformulated as:

$$p_{G\to G'} = \sum_{v \in V_a} \left(\frac{1 - n\,\frac{m_v}{\sigma_e}}{\sigma_e + n}\right) + \sum_{v \in V_a^*} \frac{n\,\frac{m_v}{\sigma_e}}{\sigma_e + n}$$

$$p_{G\to G'} = \frac{n}{\sigma_e + n}\left(1 + \frac{1}{\sigma_e}\left(-\sum_{v \in V_a} m_v + \sum_{v \in V_a^*} m_v\right)\right)$$

Let m_a be the number of nodes in G with the same IGS as a, in other words, these m nodes and a are similar catalysts. It follows that $m_a = \sum_{v \in V_a} m_v/n$ and that $\sum_{v \in V_a^*} m_v = \sigma_e - n m_a$. This gives the following expression:

$$p_{G\to G'} = 2n\,\frac{1 - n\,\frac{m_a}{\sigma_e}}{\sigma_e + n} \quad (3)$$

Just as before, σ_e brings a strong attenuating effect which makes sparser networks more susceptible to high perturbations (attenuating factor of ~1/10 when σ_e increases from 1 to 10 with $n = 1$ and $m_a = 0$, Figure 4. A, orange curve). While increasing n, the number of targets of a, does not have a very significant effect (Figure 22. A), increasing m_a brings a strong decrease to the effect on perturbation (Figure 22. B). The contrast is stark between situations where a has no competitor, thus maximal perturbations are reachable and where a has only one competitor, i.e. a node with the same IGS. These change in relative fraction of nodes can also be visualized as a perturbability rank distribution where networks are ranked based on their perturbablity (Figure 22. C). Here we observe that even when the number of nodes is fixed there is a large diversity in the network responses as a large number of networks tends to be resistant against perturbations whereas some are highly perturbed.

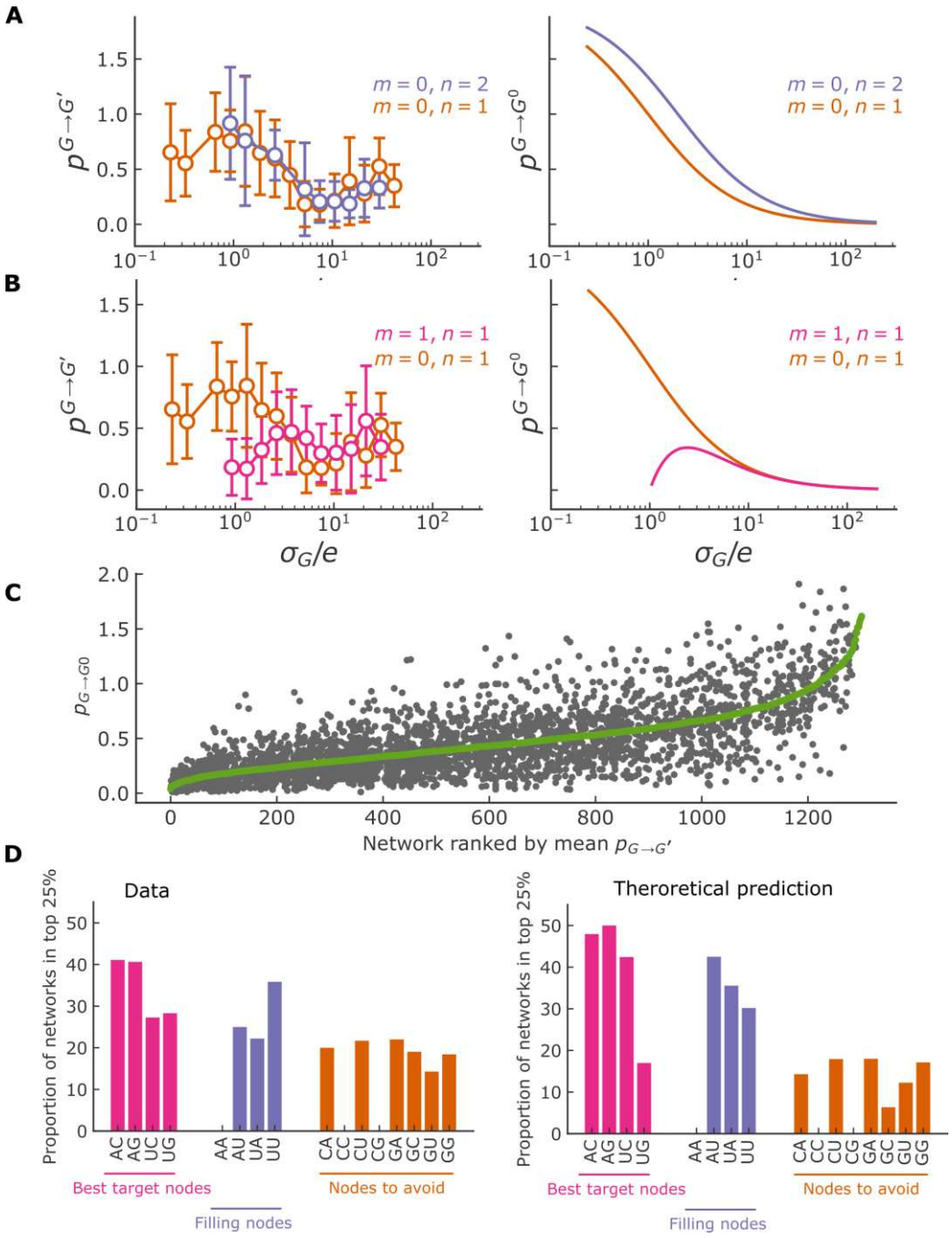


Figure 22. Whole network's response to the perturbation induced by the addition of a node. (**A, B**) Network sum of net change in fraction $p_{G\to G'}$, plotted against σ_e for different values of the parameters m_v and n for both the data (left) and the analytical results (right). Data points were binned along x-axis in 20 log-spaced bins. Bins with less than 3 points are discarded and mean and standard deviation for each bin are plotted. Networks with a number of nodes between 3 and 5 were considered for the analysis. (**C**) Perturbablity distribution of the networks with four nodes. Here networks are ranked (x-axis) based on the mean perturbation (y-axis) with the mean perturbation is shown in green. (**D**) For each node, the proportion of networks containing this node in the top 25% of the most perturbed networks is plotted for

both the data (left) and theoretical predictions (right). Only the networks with 4 node were considered for this analysis.

2.3.9 Deriving simple rules for tailoring network's response to perturbation

Since we now understand better what network parameters are important for controlling how much change is introduced when a node is added to a network, we can draw rules to design CASs, based on *Azoarcus* ribozyme system, with desired sensitivity to perturbations.

In order to construct a RNA CAS which can be perturbed easily, a first design rule is to minimize σ_G compared to e. As the G-C links are about 5 times stronger than A-U links (Table 2), the nodes to be targeted should be placed upstream and should contain 'A' or 'U' as IGS but 'G' or 'C' as tag, such that upon adding a node a strong and new G-C link is created, resulting in a high change in fraction. Therefore the best target nodes are AC, AG, UC and UG and the nodes to avoid are the nodes with a 'G' or 'C' for IGS. The rest of the nodes can act as filling nodes (e.g. AA, AU, UA and UU). Implementing these rules to our network dataset shows that the node categorization is indeed well respected. The proportions of the networks containing a 'best target' or a 'filling' node in the top 25% perturbed network are higher than those of the networks containing a node to avoid (Figure 22. D, right panel) and shows a good correlation with theoretical expectations (Figure 22. D, left panel).

2.3.10 Experimental illustration of the concept of environmentally induced variations in a plausible prebiotic scenario

So far, we have determined the parameters that control how much variation in the nodes proportions is introduced when a node is added to a network. We have done this analysis by studying networks with and without added node separately and thus there was no transfer of material involved. As a result, in what we have studied so far, the species fractions were entirely determined by the food set, i.e. by the environment, whereas in a plausible prebiotic scenario there must exist a process by which some material is passed over to the next generation. Here, to address this, we design an experiment to illustrate a prebiotically relevant scenario where variations in autocatalytic sets arise from environmental perturbations, where part of the

catalysts are transmitted to the next generation and where they have an influence on the species fractions.

More precisely, here, a common network (with nodes AC, AU, UA and UG) in different compartments was exposed to strong environmental perturbations, either CA or GU. As a control, one compartment was kept un-perturbed to assess the relative change of perturbation in all three cases. When perturbing with CA the difference in relative proportions of the two downstream nodes UA and AU changed from 0.4 to 0.6 in first generation, as CA is a strong catalyst for UG which is directly upstream of UA (Figure 23. A, G_0 upper panel, Figure 23. B, G_0 blue curve). Similarly, when GU is added to other compartment, the difference of relative fraction between UA and AU is reduced from 0.4 to 0.08 because GU catalyzes the formation of AC which in turn is a catalyst for AU (Figure 23. A, G_0 lower panel, Figure 23. B, G_0 pink curve). However, no change in the relative proportion is observed when no perturbation is added (Figure 23. A, G_0 middle panel, Figure 23. B, G_0 orange curve). These results shows that there is a direct effect of change in the environment of these networks. To test whether the effect of perturbation can also transmitted to the next-generation, we seeded the fresh food-set ('G_1') of these networks with first generation ('G_0') by serial-transfer. The results clearly demonstrate that the differences of fraction are indeed maintained during the second generation even though no perturbation is introduced here (Figure 23. B, G_1 data points). This illustrates a scenario where autocatalytic set faces environmental perturbations in the food set which causes differences in relative proportion of members and propagates to the next generation.

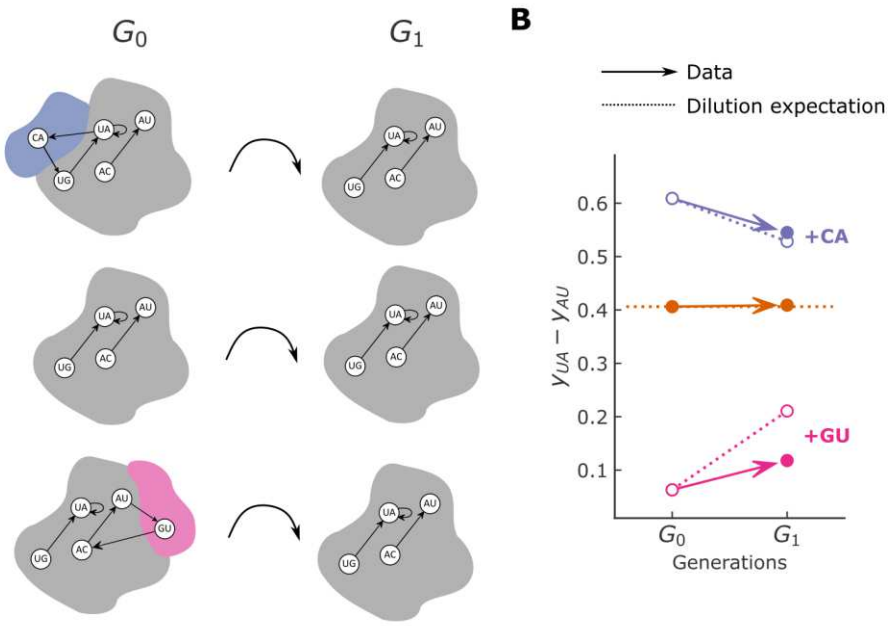


Figure 23. Experimental illustration of environmental perturbations decide the fate of the next generation. (**A**) Serial transfer experiment with a common network (AC, AU, UA and UG) with two generations and three lineages: one unperturbed (middle), one perturbed with CA (top) and one with GU (bottom). The WXY fragment for each node is at 0.1 µM and Z is in excess at 1.6 µM for the first generation G_0. Reactions were incubated for an hour at 48°C and analyzed by next-generation sequencing. 50% is transferred from G_0 to G_1 where fresh WXY fragment for each node is at 0.06 µM and Z at 0.81 µM. Relative fraction of the two downstream nodes, UA and AU is reported. (**B**) Difference of relative fraction between UA and AU is reported for the two generations of the serial transfer experiment (solid arrow) as well as the dilution expectation (dashed line). The dilution expectation is obtained by assuming than what is transferred from G_0 to G_1 is completely inert.

2.3.11 Discussion

On the prebiotic Earth, several factors could have dictated the emergence, stability and propagation of self-reproducing systems. Understanding them can allow us to imagine plausible scenarios for the origin of life and empirically test them. In a plausible RNA world scenario, a self-reproducing systems based on RNA collectively autocatalytic sets could have preceded the first template-based RNA replicases on prebiotic Earth. Though extensive theoretical work have been dedicated to CAS, empirically studying them remained a challenge. The work presented here shows a large-scale empirical studies of thousands of CAS aiming at identifying key properties

important for the CAS propagation on the prebiotic Earth. In a scheme where CAS structure is growing by incorporating new nodes the network structural features are crucial. We analytically explored the role of CAS structure in the response of both complete network and the individual nodes when they are perturbed by adding new nodes. We observe that some nodes are protected, as their relative fraction is unchanged upon perturbation, either because the new node only brings weak non-specific catalysis or because the network is very densely connected and act as a buffer. On the contrary some nodes can significantly gain or lose in fraction depending on whether they are catalyzed by the new node, and the connectivity of the existing network. Such differential response to environmental changes are important to imagine scenarios where some networks persist better than other under the pressure of perturbations. Furthermore, we also demonstrated that these perturbations act as environmental variations and can be propagated to transmit the effect to the next-generation. These observations will pave the path for future CAS selections where understanding such perturbation dynamics can help in deciding appropriate selection pressures. Additionally, the experimental set-up developed here can be easily adapted to perform cycles of 'evolution' with CASs encapsulated in droplets and selecting based on a fluorogenic phenotype coupled to the catalytic activity.

2.4. Materials and methods

2.4.1 General material & methods

For all the experiments, DNAse/RNase free water was used from Thermo Fisher Scientific (UltraPure™ DNAse/RNase Free distilled water, Product No.: 10977035). All the chemicals were purchased from Sigma-Aldrich (unless specified otherwise). 4-(2-Hydroxyethyl)-1-piperazinepropanesulfonic acid (EPPS) was purchased from Alfa Aesar (Product no.: J60511, CAS no.: 16052-06-5). Denaturing polyacrylamide gels were prepared using gel stock solution from Roth and run in 1X TBE (Tris-Borate EDTA, prepared from 10X TBE from Roth). Samples were prepared by mixing with gel-loading buffer containing 70% formamide, 0.01% of each xylene

cyanol and bromophenol blue, and ~2 equivalents of EDTA (ethylenediaminetetraacetic acid). All the gel-based analysis were performed using 12% denaturing polyacrylamide gels run for at least 2-3 h at constant power of 24 W. Gels were stained with 1X GelRedTM (Biotium). Gel analysis and calculation of conversions were carried out with ImageJ software (https://imagej.nih.gov/ij/). All DNA oligonucleotides were obtained from IDT DNA technologies (https://eu.idtdna.com) and available in Table 4. RNA concentrations were measured using Qubit® RNA HS Assay Kit (Thermo Scientific, Product No.: Q32852).

2.4.2. *In vitro* transcription

The different RNA were *in vitro* transcribed using the same protocol as described earlier (Arsène et al. 2018). After transcription, the reaction was subjected to Phenol-Chloroform-Isoamyl alcohol extraction followed by isopropanol precipitation. The template dsDNA was digested using DNaseI treatment. For this, iso-propanol precipitated transcription reaction is re-suspended in water mixed with 1X DNAseI buffer (Thermo Fisher Scientific; 10 mM Tris-HCl pH 7.5, 2.5 mM $MgCl_2$, 0.1 mM $CaCl_2$) and 10 units of DNaseI enzyme (Thermo Fisher Scientific, Product No.: EN0521). The reaction was incubated at 37°C for 40 min and heat-inactivated at 70°C for 10 min. An additional, DNaseI treatment (with 100 units DNase I, Thermo Scientific, Product No.: EN0523) was done after Phenol-Chloroform-Isoamyl alcohol extraction followed by isopropanol precipitation. Then treated samples were mixed with 1X loading dye (70% formamide, 0.005% bromophenol blue, 0.005% xylene cynol) and subjected to purification on 12% denaturing polyacrylamide gels. Band of interest was excised, RNA was eluted in 0.3 M sodium-acetate pH 5.5 overnight at 25°C, and iso-propanol precipitated. The RNA pellets, after 3 times wash with 70% ethanol and vacuum dried, were re-suspended in water, and measured for their concentration using Qubit® RNA HS Assay Kit (Thermo Scientific, Product No.: Q32852).

2.4.5 General *Azoarcus* reaction protocol

For *Azoarcus* reactions, both WXY and Z RNA fragments were folded prior to reaction by heating at 80°C for 3 min and gradually cooling down to 20°C (at a rate of 0.1°C/sec). Then 1X reaction buffer (30 mM of EPPS (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer pH 7.4, 20 mM MgCl2) was added and the reaction was incubated at 48°C. For the droplet experiments WXY and Z were folded separately prior to encapsulation while for bulk experiments they were folded together.

2.4.6 Internal IGS duplication

In order to sequence 'IGS' nucleotide from 5' end of the RNA (without erasing it with forward primer during PCR or to avoid ligation biases), IGS nucleotide was duplicated within the WT *Azoarcus* (used in the study of Vaidya *et al.*) (Vaidya et al. 2012). This was done by creating a mutation at 25th position of WT *Azoarcus* ('A' nucleotides in WT) where it was replaced to either C, U, G with corresponding base-pairing mutation at 7th position. To develop these mutations, site-directed mutagenesis was performed by amplifying WT DNA template (0.25 pg/µL) using respective M25 mutation primer (Oligos 1-4, Table 4) as forward primer, and Oligo 5 (Table 4) as reverse primer. Each of the respective PCR product was cloned in pJET2.1 vector using CloneJET PCR cloning kit (Thermo Scientific, Product No.: K1231) following manufacturer's protocol. The cloned products were transformed in chemical competent *E.coli* (Top10 Chemical competent cells, Thermo Scientific, Product No.: C404010), incubated on LB agar plates supplemented with ampicillin, positive colonies were selected, plasmids were isolated (using NucleoSpin® Plasmid kit from Macherey-Nagel, Product No.: 740588.10), and sequenced by Sanger sequencing (GATC Biotech) to identify the correct clone for each of the M25 mutation. In order to assess the effect of these mutations on total amount of the product (WXYZ) formed, an *Azoarcus* reaction was performed between WXY and Z RNA fragments containing GAG/CUU as IGS/TAG but different M25 mutation (along with corresponding base-pairing M7 mutation). These RNAs were generated by standard *in vitro* transcription protocol using T7-promoter containing dsDNA templates (generated by PCR reactions using respective plasmid, with Oligos 10-13 as forward primer and

Oligo 8 as reverse primer, Table 4). For the *Azoarcus* reaction, 0.5 µM of respective WXY fragment was mixed with 0.5 µM of Z fragment and reacted in 1X *Azoarcus* reaction buffer (30 mM EPPS, 20 mM $MgCl_2$). Samples were taken out at different time intervals, analysed on a 12% denaturing polyacrylamide gels and product band intensities were analysed (Figure 24). As expected, results show that WT (U7, A25) has the maximum activity with a slight decrease in activity for the other mutants. The differences in activity are less important at lower time-points, and an incubation time of around an hour is used for the droplet experiment. Using this M25 nucleotide information, the number of different IGS in each droplet can be identified.

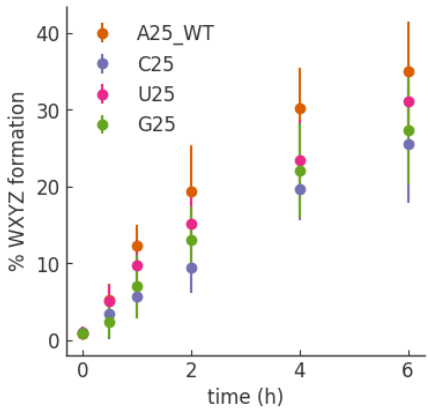


Figure 24. Effect of duplication of IGS internally on the WXYZ ribozyme formation. Time courses showing the effect of duplication of IGS nucleotide at 25th position on the WXYZ formation. Yield (%) of WXYZ formation for all mutants, C25 (U7G, A25C), U25 (U7A, A25U), G25 (U7C, A25G) and WT (U7, A25) is plotted against time (in hours). The self-assembly reactions for each of the mutant and WT are analysed on 12% denaturing polyacrylamide gels and data is extracted from band intensities. All the time-courses are done in triplicates and mean WXYZ product formation is plotted along with standard deviation.

2.4.7 Quality control for the original IGS

In order to correlate the original IGS and the identity of the nucleotide at the 25th position in WXY RNA fragments, the quality of RNAs were checked by subjecting them to next-generation sequencing. For this all 16 WXY RNA fragments were poly-A tailed and ligated with RNA adaptor at their 5' end (Oligo 14, Table 4), converted to cDNA, appended with sequencing adaptors and

subjected to sequencing. For polyA tailing, 2.5 µM of each WXY RNA (in different reaction) was mixed with 1X reaction buffer (50 mM Tris-Hcl pH 7.9, 250 mM NaCl, 10 mM $MgCl_2$), 2 mM ATP, 50 U/µL of *E. coli* poly(A) polymerase (New England Biolabs, Product No.: M0276S) and incubated at 37°C for 40 min. After heat inactivation and isopropanol precipitation, the samples were dephosphorylated using phosphatase enzyme. Precipitated RNA were re-suspended in water, mixed with 1X reaction buffer (10 mM Tris-HCl pH 8.0, 5 mM MgCl2, 100 mM KCl, 0.02% Triton X-100, 0.1 mg/mL BSA), 0.1 U/µL of phosphatase enzyme (FastAP, Thermo Fisher Scientific, Product No.: EF0652) and incubated at 37°C for 1 h. After heat inactivation, these reactions were directly used for kinase reaction to add monophosphate. For phosphorylation, the dephosphorylated RNA reaction was mixed with 1X reaction buffer (50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM spermidine), 2 mM ATP, 0.25 U/µL of T4 Polynucleotide Kinase (Thermo Fisher Scientific, Product No.: EK0031) and incubated at 37°C for 1 h. After heat inactivation, RNAs were purified on AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) and subjected to ligation with RNA adaptor. For ligation, mono-phosphorylated RNA were mixed with 1X reaction buffer (50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 1 mM DTT), 2 mM of ATP, 3% of PEG8000, 1 µM of adaptor RNA (5'-CCUACCAGUACCCUACCA-3'), 2.2 U/µL of T4 RNA ligase 1 (New England Biolabs, Product No.: M0204S) and incubated at 16°C overnight. After heat inactivation RNAs were purified on AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) and subjected to reverse transcription (RT). For RT, purified ligated RNAs were mixed with 1X reaction buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 0.5 mM of each dNTP, 5 mM of DTT, 2.5 µM of RT primer (Oligo 15, Table 4), 10 U/µL of Superscript III enzyme (Thermo Fisher Scientific, Product No.: 18080085) and incubated at 55°C for 1 h. After RT, cDNA were purified using AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) and subjected to PCR to append sequencing adaptors. Final amplicons were subjected to 2*150 bp pair-end microMiSeq at Institute Curie High Throughput Sequencing platform, Paris. To analyse this data, first the sample barcode (the 6 first bp of read 1) was extracted, allowing no mismatch. It codes for the identity of the WXY fragment. For controlling the original IGS, the adaptor sequence was searched and if it was found the following 50 bp were extracted and the Levenshtein distance with the first 50 bp of the canonical WXY sequence was computed. Only the reads with a distance of strictly less than

10 were kept. The original IGS and the two neutral mutations were then extracted by searching for the sequence between the original IGS and the first neutral mutation ('GCCT') and looking at the flanking bases and by searching for the 5 bp before the second mutation and taking the next base if found. Reads with ambiguity in any of the three IGS coding bases was discarded. What is reported in Table 3 is the proportion of these reads where the combinations of theses three bases is correct. For controlling the tag, for each read, the first occurrence of 5 consecutive 'A' was searched and if it was found the 50 bp before were compared against the last 50 bp of the canonical WXY sequence by computing the Levenshtein distance. Again only the reads with a distance of strictly less than 10 were kept. 'tag' was extracted as described above in Sequencing data processing section and what is reported in Table 3 is percentage of correct 'tag' excluding the cases where no 'tag' could be confidently identified.

Table 3. WXY fragment sequencing control. WXY fragments were sequenced using a strategy designed to avoid to erase the original 'IGS' and 'tag' with primer. Here is reporter the percentage of correct 'IGS' (resp. correct 'tag') when the 'IGS' could be confidently identified (resp. the 'tag). N is the total number of sequences where this is the case.

Specie	Correct 'IGS' (%)	N	Correct 'tag' (%)	N
AA	99.2	28820	92.3	188882
AC	99.2	75450	97.8	312848
AU	99.2	97032	81.9	247055
AG	99.3	126108	96.4	214207
CC	99.1	10884	91.4	335472
CA	99.1	42253	97.5	333490
CU	99.1	101719	83.5	295666
CG	99.4	57846	96.6	297463
UA	98.4	34719	97.6	273379
UC	98.1	7413	91.0	295321
UU	94.2	9455	84.1	305542
UG	96.4	16845	95.9	275245
GA	93.1	5958	88.6	249829
GC	95.1	7131	97.6	287838
GU	90.7	6743	81.4	311116
GG	87.9	9788	96.3	316361

2.4.8 General materials and methods for microfluidics

For all the microfluidic experiments per-fluorinated oil HFE 7500 (3M™ Novec™, Product No.: 98-0212-2928-5) and fluoro-surfactant (RAN Biotechnologies, Product No.: 008-FluoroSurfactant) were used. All the devices were designed using AutoCAD software (Student license) and printed at Selba S.A (Route des Fayards, 1290 Versoix, Switzerland). Devices were fabricated by using soft-lithography protocols (Mazutis et al. 2013; Duffy et al. 1998). Device wafers were prepared using MJ Mask Aligner with SU-8 photoresist coating over silicon wafer. Height of the designs on wafer were measured by profilometer (Dektak). All the device were prepared in PDMS polymer (polydimethylsiloxane) over glass slides. For electro-coalescence devices indium tin oxide coated glass slides were used (Delta Technologies, Product No.: CG-90IN-S215). After plasma bonding, channels were treated with 2% 1H,1H,2H,2H-perfluorodecyltrichlorosilane (ABCR, Product No.: AB111155) in HFE 7500. Droplet-based microfluidics experiments were performed using microscope (Nikon eclipse *ti*) station set-up built in the lab as described earlier (Mazutis et al. 2013). Flow in devices were controlled by either syringe pumps (Harvard Apparatus Inc.) or by air-pressure control pumps (MFCS™-EZ, Fluigent SA).

2.4.9 Barcoded hydrogel beads synthesis

Barcoded hydrogel beads were produced in two steps: production of hydrogel beads and generation of barcoded hydrogel beads library. Hydrogel beads were produced by co-encapsulating 10% (w/w) polyethylene glycol diacrylate (PEG-DA) 6000 (Sigma, Product No.: 701963), 1% PEG-DA-700 (Sigma, Product No.: 455008), 400 µM of acrydite-modified double-stranded DNA (Oligo 16 and 17, Table 4), 10µM of FITC-Na, 1% (v/v) photo-initiator (2-hydroxy-2-méthylpropiophenone, Sigma, Product No.: 405655) in buffer (75 mM Trizma HCl pH 7.4, 50 mM of NaCl) along HFE 7500 oil supplemented with 2% surfactant. The encapsulation was done on a dropmaker device (Figure 25) and flows of 150 µL/h for aqueous solution and 500 µL/h for oil with surfactant were maintained using syringe pumps (Harvard Apparatus Inc.). The resulting 9 pL droplets were concurrently passed under UV-light (360 mW) to initiate the polymerization and were collected in 5 mL collector tube. The complete set-up was strictly maintained in dark in order

to avoid any spontaneous polymerization with external light source. The collected droplets containing beads were washed one time with hexane and then 3 times with binding-wash buffer (20 mM Trizma HCl pH 7.4, 50 mM NaCl, 0.1% of Tween 20). Beads were then filtered using Steriflip 20 µM Nylon filters (Millipore, Product No.: SCNY00020) and stored in binding-wash buffer (supplemented with 1 mM EDTA).

The second step is to build barcode over these beads using the split-and-pool method (Macosko et al. 2015). 250 µL of beads (~10 million beads) were washed thoroughly with binding-wash buffer 3 times to remove EDTA. Then they were subjected first to a common adaptor ligation using the overhang on the bead (Oligo 16 and 17, Table 4). For this, the beads were mixed with T7 DNA Ligase buffer (66 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 1 mM ATP, 1 mM DTT, 7.5% PEG 6000), 50 µM common adaptor (Oligo 18 and 19, Table 4), 30 U/µL of T7 DNA Ligase (New England Biolabs, Product No.: M0318L) and incubated at room temperature for 30 min. Beads were then washed 3 times with binding-wash buffer and subjected to first split step. For split, the beads are mixed with T7 DNA Ligase buffer, 30 U/µL of T7 DNA Ligase and divided in 96-well plate (16 µL each) already filled with 4 µM of the 1st barcode index (Oligo 20 and 21, Table 4), unique to each well. Each index is 16 base pair long and there are 96 versions of each. Plate was sealed and incubated in usual bench top centrifuge at 25°C with 600 rpm for 15 min, then at room temperature for 10 min. Reaction was stopped by heat inactivation at 65°C for 10 min and then all the wells were pooled together. Pooling is done by adding 200 µL of cold binding-wash buffer to each well and collecting them in 4 different 5 mL tubes which are merged together after washing and removing the supernatant. The beads were then again subjected to same process to split and pool to ligate 2nd and 3rd barcode (Oligo 22-25, Table 4). After the final pooling step, primer required for the reverse transcription (Oligo 26 and 27, Table 4) was ligated to all the beads (common to all) using the same ligation protocol. The final barcoded hydrogel bead library was washed with binding-wash buffer (supplemented with 1 mM EDTA).

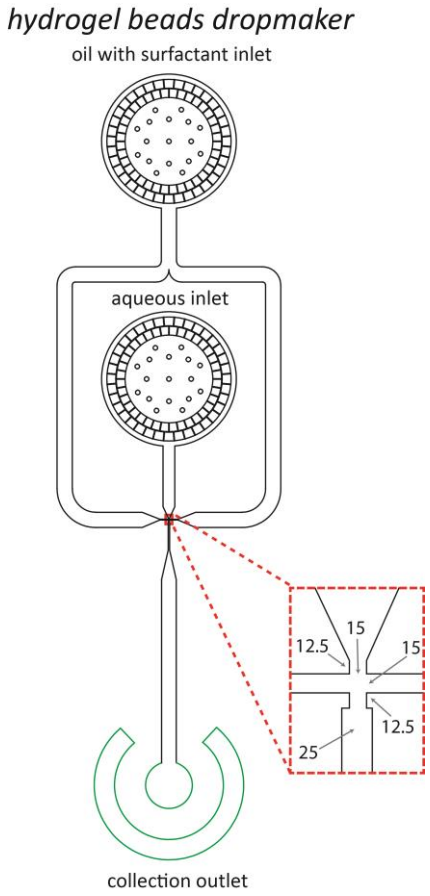


Figure 25. Design for the hydrogel beads production. Dropmaker design used to produce 9 pL hydrogel beads containing droplets. These beads are then used to build barcoded library using split and pool method. The zoomed inset images in each show the design details of important part of the design where dimensions are mentioned as numbers (in µm).

2.4.10 Droplet microfluidic experimental set-up

The complete experimental set-up consisted of five consecutive steps (Figure 10):

1) Initial emulsions production
2) Combinatorial fusion of initial emulsion
3) Incubation of droplets and splitting
4) Droplet barcoding and RT in droplets
5) Sequencing sample preparation

Initial emulsions production

Using all possible 16 different WXY RNA fragments ($_{gmg}WXY_{cnu}$), 24 unique combinations were prepared each containing different WXY fragments each (Table 1). These combinations were chosen to prevent over dominance of nodes, create uniform mixing during electrocoalescence step and covering a vast range of networks, as diverse as possible in topological features. For each of the 24 combinations, 1 µM of respective WXY fragment(s) along with 0.03 µM of molecular RNA reporter mixed in water and encapsulated individually using 5 pL dropmaker device (Figure 26. A). Here concentrations of RNA fragments were kept 10 times higher so that after the fusion by electrocoalescence step (see below) they are diluted to 0.1 µM each for WXY fragments and to 3 nM for hairpin reporters. Flow rates were maintained using syringe pumps at 100-150 µL/h for aqueous phase (containing RNA) and 150-170 µL/h for oil phase (HFE 7500 with 2% surfactant), and each emulsion was collected for 5 min. All the 24 emulsions were collected together in a 1.5 mL collector tube (with PDMS cap) containing oil with 2% surfactant and mixed thoroughly.

In parallel to these 24 emulsions, a 50 pL emulsion containing 1.6 µM of Z fragment (in order to have an excess of Z fragment) and 1X AZ reaction buffer (30 mM EPPS pH 7.4, 20 mM $MgCl_2$) was prepared using 50 pL dropmaker device (Figure 26. B). Flow rates were maintained (using syringe pumps) at 350 µL/h for aqueous phase (containing RNA) and 250 µL/h for oil phase (HFE7500 with 2% surfactant).

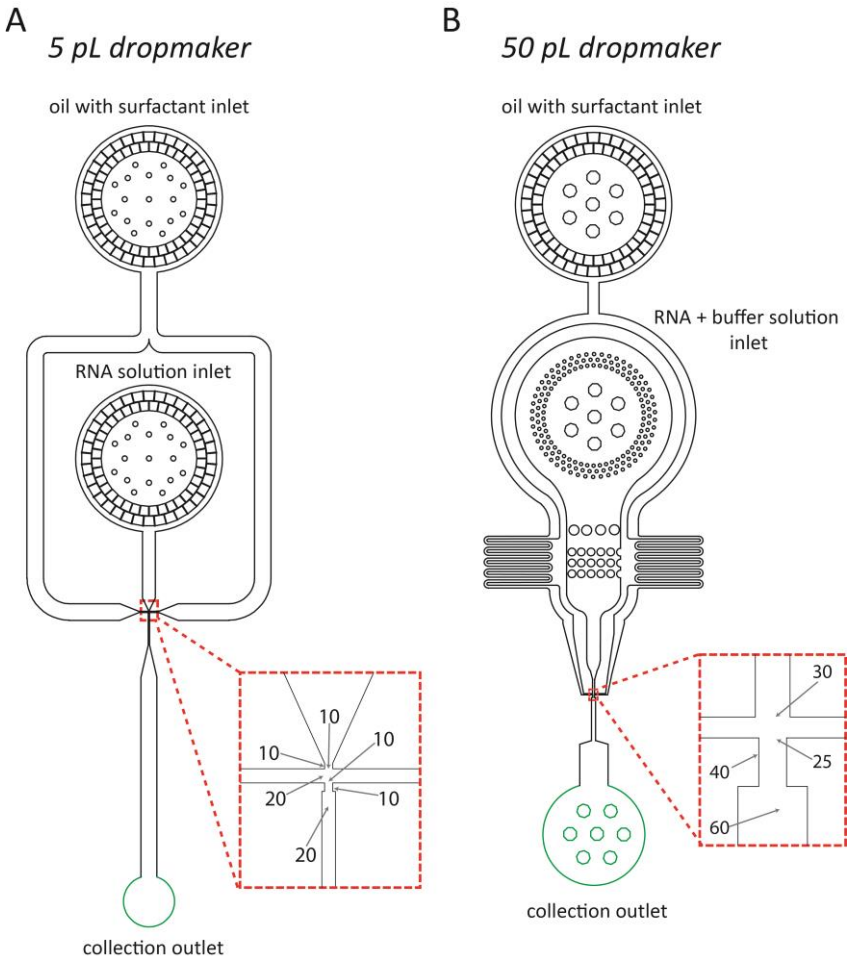


Figure 26. Dropmaker designs for 5 and 50 pL emulsions. (**A**) 5 pL dropmaker design used to produce 24 different initial emulsions containing WXY fragments with a unique combination of IGS/tag in each (Table 1). (**B**) 50 pL dropmaker design used to produce emulsion with Z fragment and the reaction buffer. Both 5 pL (pooled) and 50 pL emulsions were used in the fusion device to create a diverse library of RNA CASs. The zoomed inset images in each show the details of important part of the design where dimensions are mentioned in µm.

Electrocoalescence of initial emulsions

To create RNA CASs with diverse compositions, in this step, combinatorial fusion by electrocoalescence of 5 pL droplets (containing WXY fragments) with 50 pL droplet (containing Z fragment along with reaction buffer) was performed. Here fusion frequency was kept such that on an average two to three 5 pL droplets coalesces with one 50 pL droplet. (Figure 12. B). To achieve such high-throughput combinatorial droplet fusion, a fusion device was developed

(Figure 27) where both 5 pL and 50 pL emulsions were re-injected under a controlled oil flow (with surfactant) and coalesced using liquid electrodes (Sciambi & Abate 2015) modified to contain 3 M NaCl solution and connected through metal wires. Electrocoalescence was achieved using signal (as a sine function from Agilent 33522A waveform generator) of 4 kHz at 400 mV with 50 Ω. Signal was sent to high-voltage amplifier (TREK) to amplify signal 1000 times before sending to microfluidic chip. All the flows were kept under the control of a microfluidic flow control system (MFCS™-EZ, Fluigent S.A), and fusion frequency was achieved by carefully adjusting and coupling the flows of the two emulsions. Pressure values of 380-400, 350-380, 470-520 mbar were adjusted for 50 pL, 5 pL and oil separator channel respectively. In the complete device only oil with 2% surfactant was used. The electrocoalescence and collection was performed for ~3.5 h. All the emulsions were strictly kept on ice during the complete process. Collection was done in 0.2 mL collector tube (with PDMS cap) containing oil with 2% surfactant.

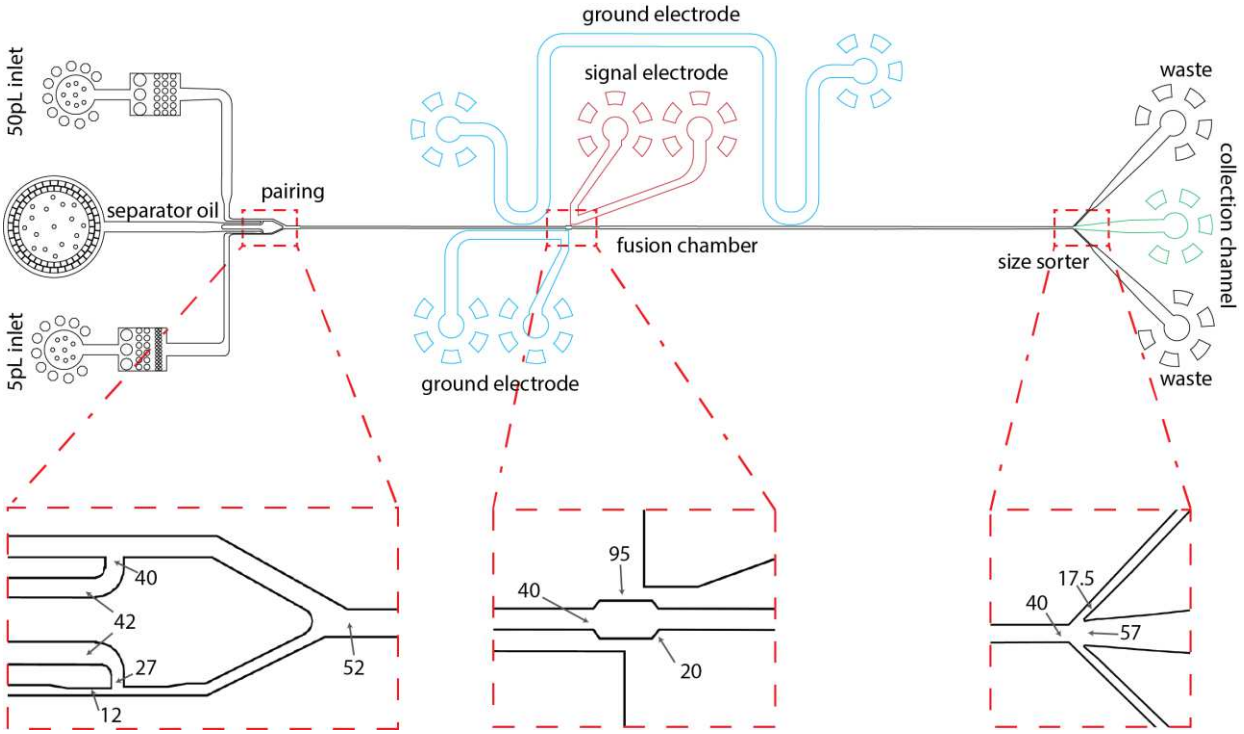


Figure 27. Design for the combinatorial fusion of droplets by electrocoalescence. The design of the microfluidic device used to perform the combinatorial electrocoalescence between 5 pL and 50 pL droplets. Signal and ground electrodes are highlighted in red and blue respectively. To produce potential, the electrode channels were filled with 3 M NaCl solution and connected directly with metal wire to power

generator. Unfused and satellite droplets were sorted out using a size sorter at the end of the design and fused droplets collected in the channel (highlighted in green). The zoomed inset images show the details of important part of the design where dimensions are mentioned in µm.

Incubation of droplets and split

After collection, fused droplets were incubated at 48°C in a thermo-block for an hour. In this next step, droplets (now of a volume of ~60 pL) were splitted in to smaller droplets (of a volume of ~5 pL). These 5 pL droplets were used in the next step of droplet barcoding and RT in droplet. Split design is a T-junction based microfluidic device (Link et al. 2004) with droplet and oil inlet and three split output channels (Figure 28). Here only splitted droplets (~5 pL) were collected and kept for the next step and the other two outlets were connected to waste. Flow rates were maintained using air-pressures controlled pumps (MFCS™-EZ, Fluigent S.A) and pressure values of 800-1000, 150-250 were adjusted for oil separator channel and fused droplets inlet, respectively. Splitted droplets were collected in 0.2 mL collector tube (with PDMS cap) containing oil with 2% surfactant.

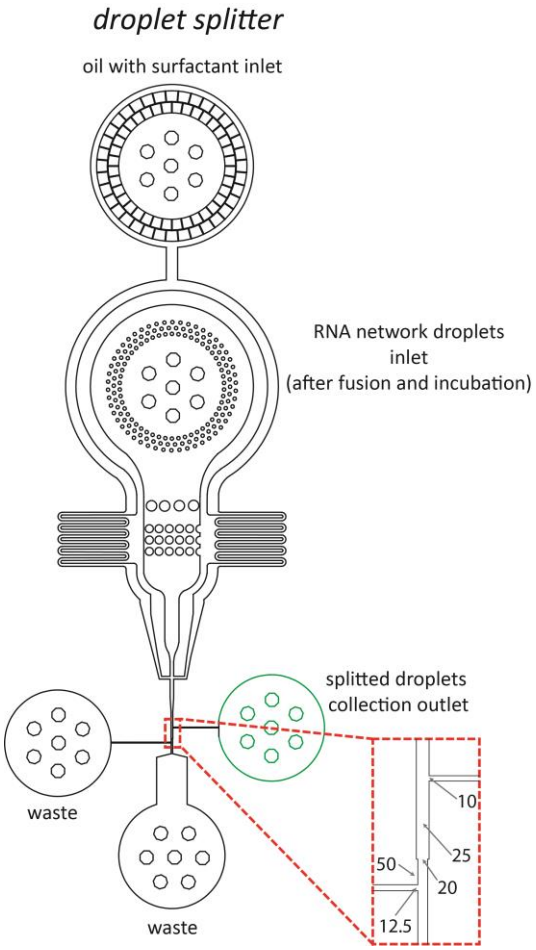


Figure 28. Design for the droplet splitter. Drop splitter design used to split the RNA CASs droplets after incubation. The fused RNA CAS droplets (after electrocoalescence in fusion device, Figure 27) are ~65 pL and splitted here to collect ~5 pL droplets (first collection outlet highlighted as green) to use them for RT in droplets. The zoomed inset images in each show the design details of important part of the design where dimensions are mentioned in µm.

Droplet barcoding and reverse-transcription (RT) in droplet

The strategy that allows the content of each splitted droplet to be sequenced using droplet barcoding and reverse transcription (RT) in droplets was developed from single-cell transcriptomics methods using droplet microfluidics (Macosko et al. 2015; Klein et al. 2015). Here barcoded hydrogel beads were singly encapsulated with all the necessary reagents for RT and fused with RNA containing droplets.

First, the barcoded beads (~50 µL) were washed 5 times with 500 µL of binding-wash buffer by centrifuging them at 13 K rpm for 1 min. The washed beads were mixed with 0.5 mM dNTPs, 1X RT buffer (SSIII RT buffer, 50 mM Tris-HCl pH8.3, 75 mM KCl, 3 mM $MgCl_2$), 5 mM DTT, 0.4% of Tween20, 2U/µL of SUPERase·In™ (SUPERase·In™ RNase inhibitor, Thermo Fisher Scientific, Product No.: AM2694) and incubated at 37°C for 30 min. The beads were spinned down and excess of liquid was carefully removed leaving only 4 µL of it over the beads. Then 500 units of reverse transcriptase (SuperScript III, Thermo Fisher Scientific, Product No.: 18080044) and 25 units of BclI restriction enzyme (New England Biolabs, Product No.: R0160L) were added to the beads, mixed thoroughly and used for encapsulation for the RT in droplets. BclI was added to release the barcode DNA from hydrogel beads inside the droplets when the emulsion is incubated at 60°C during RT (BclI has no activity at 4°C).

Then the hydrogel beads together with cDNA synthesis reagents were singly encapsulated in ~50 pL droplets and fused with RNA droplets on the same microfluidic device (Figure 29). All the flows were controled using air-pressured controlled pumps (MFCS™-EZ, Fluigent SA), and fusion frequency was achieved by carefully adjusting the flows of the two emulsions with the following pressure values around 550, 225, 650, 450 mbar for hydrogel beads, 5 pL droplets, separator oil, and oil for beads encapsulation, respectively. The fusion frequency was kept low such that only one 5 pL RNA droplet is fused with one 50 pL barcoded hydrogel bead droplet. In the complete device only oil with 2 % surfactant was used. Electocoalescence was done using liquid electrodes (Sciambi & Abate 2015) containing 3 M NaCl solution in channels and connected thorough metal wires using signal (as a sine function from Agilent 33522A waveform generator) of 4 kHz at 400 mV with 50 Ω. Signal was sent to high-voltage amplifier (TREK) to amplify it 1000 times before sending it to microfluidic chip. Fused droplets were collected in a 0.2 mL collector tube (with PDMS cap) containing oil with 2% surfactant.

The fused droplets were then incubated at 60°C in a thermo-block for 1 h to perform cDNA synthesis and at the same time DNA barcodes were released from the hydrogels beads using BclI restriction enzyme. As all the produced cDNA are drop-specifically barcoded, post incubation droplets were broken down and all the barcoded cDNA were pooled in a 1.5 m tube. The emulsion

breaking was achieved by using 2 equivalent of 1*H*,1*H*,2*H*,2*H*-Perfluoro-1-octanol and extracting with water. These extracted cDNAs were iso-propanol precipitated to be used in the next step.

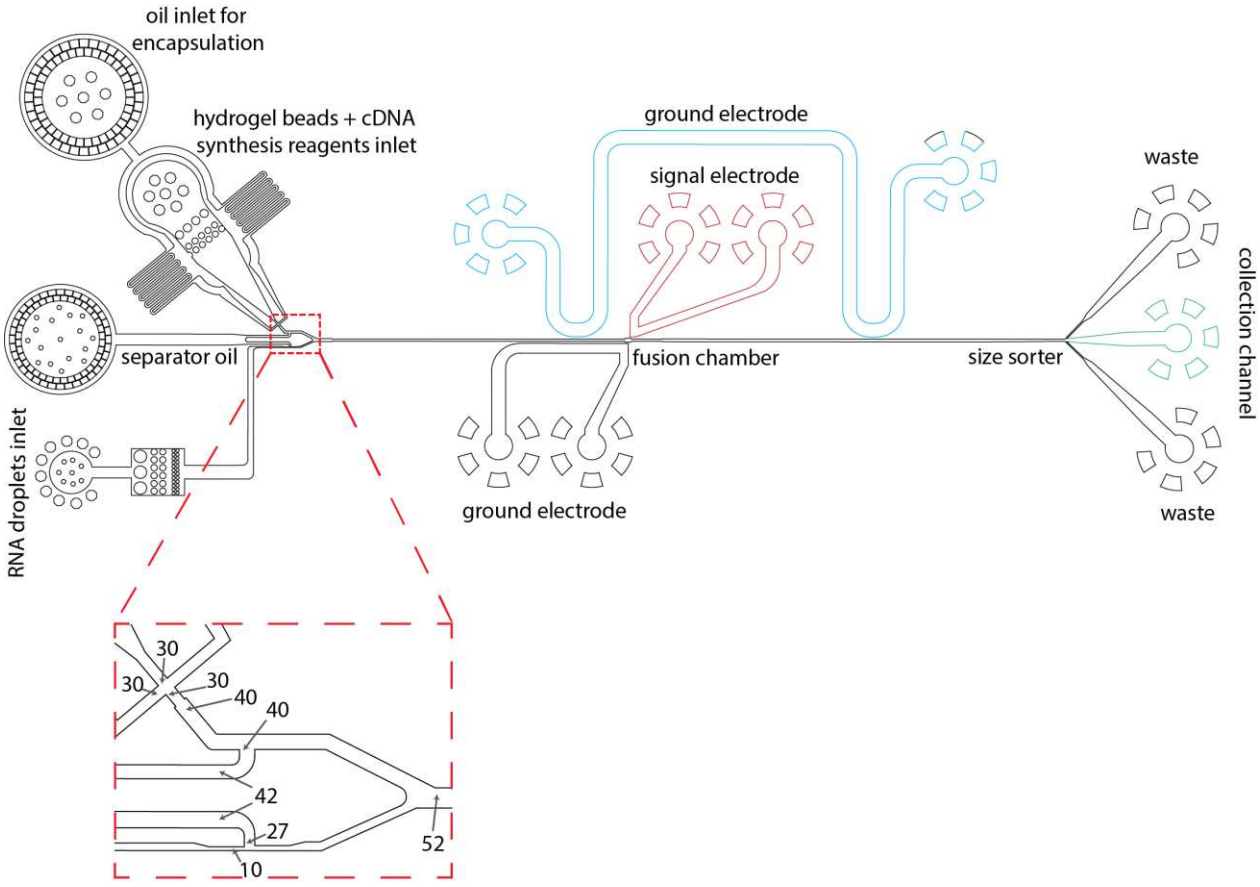


Figure 29. Design for the encapsulation of hydrogel beads and fusion by electrocoalescence with RNA droplets. The design of the microfluidic device used to encapsulate the hydrogel beads (containing cDNA synthesis reactants) and fusion between splitted RNA droplet (~5 pL) and droplets containing hydrogel beads. Signal and ground electrodes are highlighted in red and blue respectively. 3 M NaCl solution filled in the channel connected directly with the metal wire used as electrodes. Unfused and satellite droplets were sorted out using a size sorter at the end of the design and fused droplets collected in the channel (highlighted in green). Fusion chamber and size sorter dimensions are exactly the same as mentioned in the other electrocoalescence device (Figure 27). The zoomed inset images show the design details of important part of the design where dimensions are mentioned in µm.

Sequencing sample preparation

In order to block all the free 3' priming end of the unused barcodes (which could be used as primers during the next PCR steps leading to artefacts), the extracted cDNA is subjected to TdT (terminal transferase) treatment with ddCTP (dideoxy CTP). For that, the cDNA were mixed with 1X TdT reaction buffer (50 mM Potassium acetate, 20 mM Tris-Acetate, 10 mM $MgCl_2$), 0.25 mM $CoCl_2$, 0.4 mM ddCTP (Roche CustomBiotech, Product No.: 12158183103) and 0.4 UµL of TdT enzyme (Terminal Transferase, New England Biolabs, Product No.: M0315L). The reaction was incubated for 30 min at 37°C and then TdT was heat inactivated at 70°C for 10 min. The TdT treated cDNAs were then purified over magnetic beads (AMPure XP, Beckman Coulter, Product No.: A63881) following manufacturer's protocol and subjected to PCR to add sequencing adaptors.

Sequencing adaptors were appended using two sequential PCR steps. PCR was performed using 0.5 µM forward primer, 0.5 µM of reverse primer in 1X PCR buffer (Thermo Scientific: 25 mM TAPS-HCl pH 9.3, 50 mM KCl, 1.5 mM MgCl2), 0.2 mM dNTPs, 0.01 U/µL of polymerase (Thermo Scientific Phusion Hot Start II, Product No.: F459). Amplification was done using following protocol: initial heating 98°C/30sec, cycles of 98°C/10sec, annealed for 30sec, 72°C/30sec, final extension 72°C/3min. The first PCR was done with a high annealing temperature of 72°C (to avoid non-specific amplification) using Oligos 28 and 29 as primers (Table 4) and amplified for 18 cycles. PCR products were purified using AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) following manufacturer's protocol. For the second PCR, Oligos 30 and 31 (Table 4) were used as primers with an annealing temperature of 56°C and 10 cycles of PCR. The final sequencing library was purified on magnetic beads, analysed on TapeStation (Agilent 2200 TapeStation, using high sensitivity D1000 ScreenTape®, Product No.: 5067-5584) and quantified by Qubit® dsDNA HS Assay Kit (Thermo Scientific, Product No.: Q32854). Sequencing libraries were subjected to high-throughput sequencing using NextSeq 550 system for 2*150 High Output mode at Genotyping and Sequencing Core Facility, ICM Paris.

2.4.11 Sequencing data processing

The libraries were sequenced in paired-end mode with 180 bp devoted to read 1 and 120 devoted to read 2. Read 1 was used to determine the sample barcode (used for multiplexing samples for sequencing) and the RNA molecule identity (either an *Azoarcus* ribozyme or a hairpin reporter). Read 2 was used to determine the Unique Molecular Identifier (UMI) (Kivioja et al. 2012) and the droplet barcode. The structure of read 1 and 2 is summarized in Figure 30.

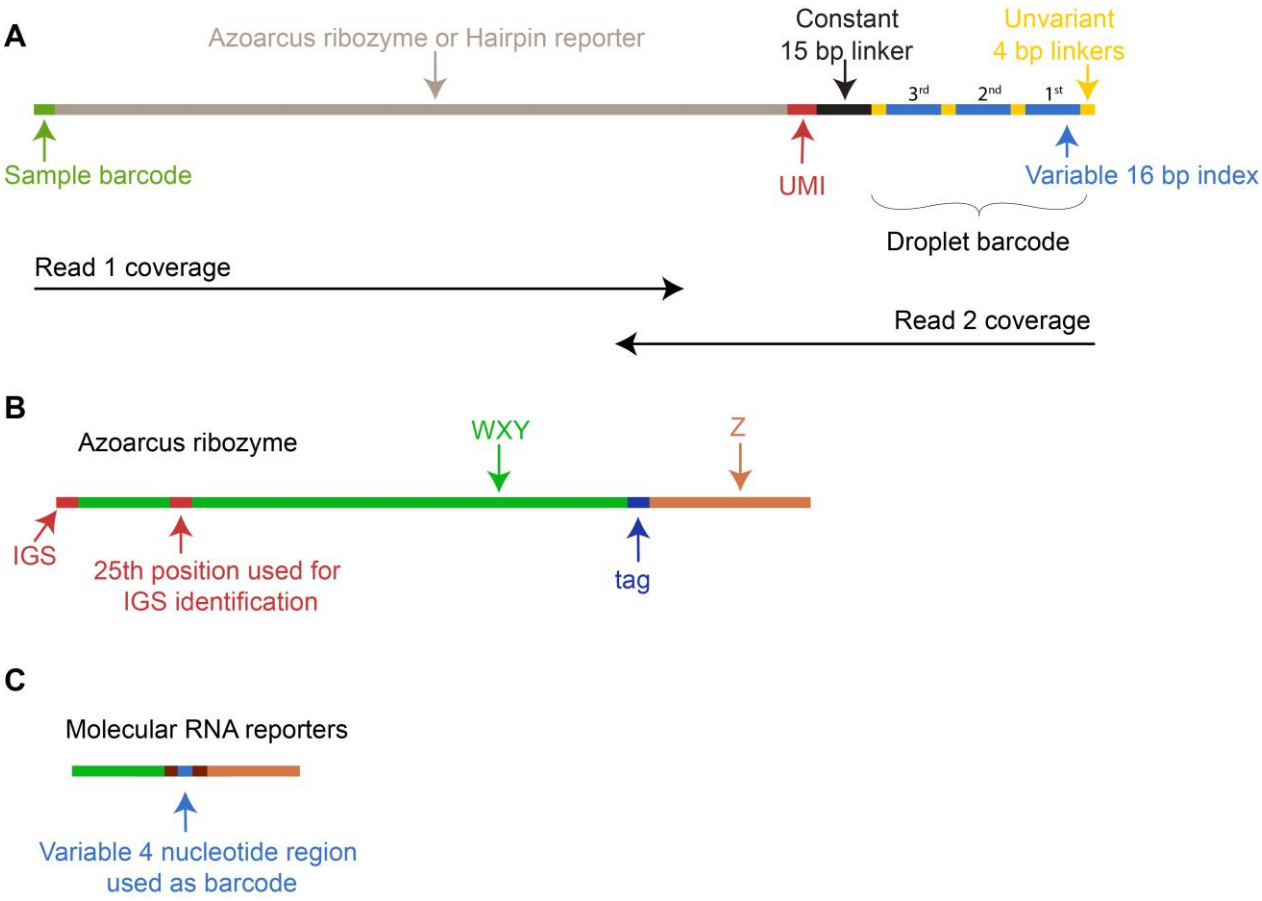


Figure 30. Sequencing read structure. (**A**) Generic paired-end reads structure. (**B**) Detailed structure of the read for *Azoarcus* ribozyme and for hairpin reporter annotated with the region where the specific information is searched for. (**C**) Structure of molecular RNA reporters ('hairpin reporters') where they share the same 5' and 3' part as WXYZ for the amplification but forms hairpin with a loop region of 4 nt serving as barcode.

Droplet barcode identification

The droplet barcode is composed of three 16 bp long variable regions denoted as indexes (Figure 30) separated by 4 bp long invariant regions denoted as linkers. The first step is to find the linkers positions. For this they are searched with a sliding window approach around their expected positions in a window of 5 bp allowing a maximum Hamming distance of 1. The sequence in between two found linkers is extracted and aligned against the 96 possible variants for the corresponding index using Bowtie 2 version 2.2.9 with the "very-sensitive" parameters. Alignments with a mapping quality lower than 20 were discarded. This gives the identity of the 3 indexes composing the droplet barcode for each read.

UMI identification

In read 2, there is a constant 15 bp sequence before the UMI. To identify the UMI, the position of this sequence is searched around its expected position in a window of 20 bp allowing a maximum Hamming distance of 2. If a position is found and if the last 3 bp are exactly matching the end of the constant 15 bp sequence then the following 8 bp are extracted as the UMI for the read.

IGS identification

As explained before the 25^{th} position in the *Azoarcus* ribozyme codes for the IGS. In read 1 this position correspond to the 31^{th} position. To identify the base at this 31^{th} position the constant 7 bp region 2 bp upstream of the 25^{th} position is searched around its expected position in a window of 10 bp allowing a maximum Hamming distance of 2. If a position is found and if the 3 bp located 7 bp after this position are in the following format: "CNA" then N is extracted as the IGS for the read.

tag identification

The tag identification is similar to the IGS identification. The constant 10 bp region upstream of the tag (CNU) in read 1 is searched around its expected position in a window of 10 bp allowing a maximum Hamming distance of 2. If a position is found and if the 3 bp located 10 bp after this position are in the following format: “CNU” then N is extracted as the tag for the read.

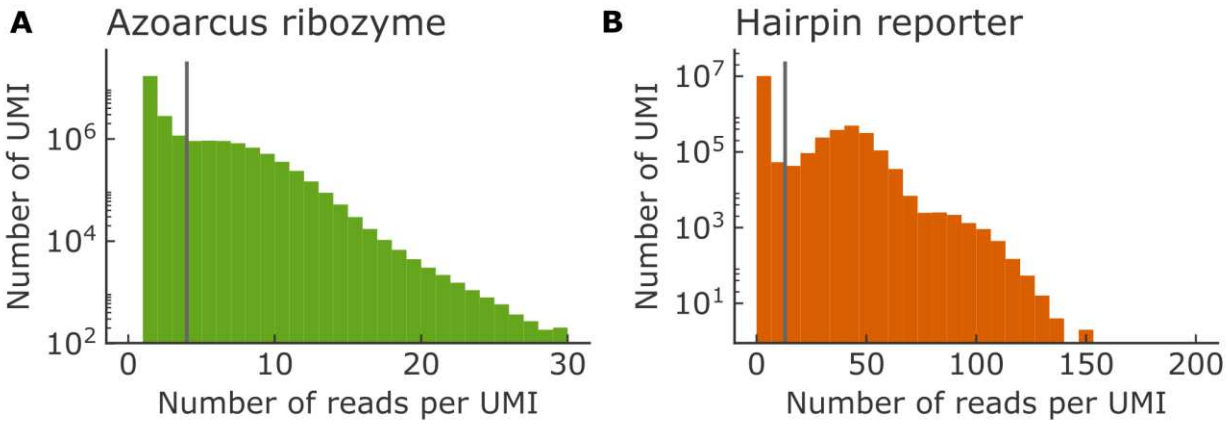
Hairpin reporter identification

The hairpin reporter sequence contains a variable 4 bp region in the middle loop region (Figure 30) which is coding for its identity. To test whether a read corresponds to a hairpin reporter, the constant 10 bp region before and after the variable region are searched around their expected positions in windows of 10 bp and allowing a maximum Hamming distance of 2. If the two positions are found the read is declared to be a hairpin reporter read. If the two positions are 14 bp distant and if the 3 bp just before the supposedly identified variable region exactly match the last 3 bp of the constant region before the variable region then the next 4 bp are taken as the read hairpin variable region. This sequence is then compared to the list of possible variable region and the closest one in terms of Hamming distance (but not more distant than 1) is taken as the read hairpin reporter identity.

UMI normalization

At this stage, four meta fields are associated for each pair of reads: sample barcode, droplet barcode, UMI and RNA molecule identity. Any read where any of the four meta fields is empty is discarded. Typically around 1/3 of the reads are discarded at this stage. The reads with the same meta information are collapsed into a single row with an extra meta fields with the number of reads that were collapsed. This field is used to filter out noise due to sequencing errors for which the value of this field is expected to be low. The thresholds were chosen based on the shape of the distribution of number of reads (Figure 31. A-B) and are summarized in Figure **31**. C. Typically this threshold is different for *Azoarcus* ribozyme rows or for hairpin reporter rows as the hairpin

reporters are smaller, they are more amplified during the PCRs used to add the sequencing primers.



C

Experiment	Sequencing run	Minimum number of reads per UMI	
		Azoarcus ribozyme	*Hairpin reporter*
Replicate 1	NextSeq 1	2	21
Replicate 2	NextSeq 2	51	51
Replicate 2	NextSeq 3	4	13

Figure 31. UMI normalization. (**A**) Distribution of number of UMI per number of reads per UMI for reads associated with *Azoarcus* ribozyme for the replicate 2 and the sequencing run 'NextSeq 3'. Any UMI that is left of the grey bar, the threshold for this replicate and sequencing run, was discarded. (**B**) Same as (A) but for reads associated with hairpin reporter. Note that values on the x axis are bigger because of the higher amplification rate of hairpin reporter during PCR due to their smaller size. (**C**) Table summarizing what threshold values were used for discarding UMIs with not enough reads based on the distribution of number of reads per UMI as shown in (A) and (B) depending on the replicate and on the sequencing run. These values are not always the same because they depend on the sequencing depth chosen.

Final data processing

Filtered UMI-normalized rows are then grouped by droplet barcode to count the number of UMI per RNA molecule type per droplet barcode. For each droplet barcode the coding set of hairpin reporter is determined by taking any hairpin reporter that has more UMIs than 7.5% of the total number of UMI associated with hairpin reporters for this droplet barcode. This threshold value was chosen to match the measured fusion distribution acquired by video during the experiments (Figure 12. B). Once the set of coding hairpin is determined, the *Azoarcus* network is also

determined and any UMI that does not correspond to a coding hairpin reporter or to a ribozyme not coded by the set of coding hairpin reporters is discarded. Finally the droplet barcodes with more than 10 UMIs associated with hairpin reporters and 20 with ribozymes are kept and the fraction of each node in the network is computed. Here again, these thresholds were chosen so that the data would match the most the videos acquired during the experiments (Figure 32). Finally, mean of node fraction is taken across replicates of the same network.

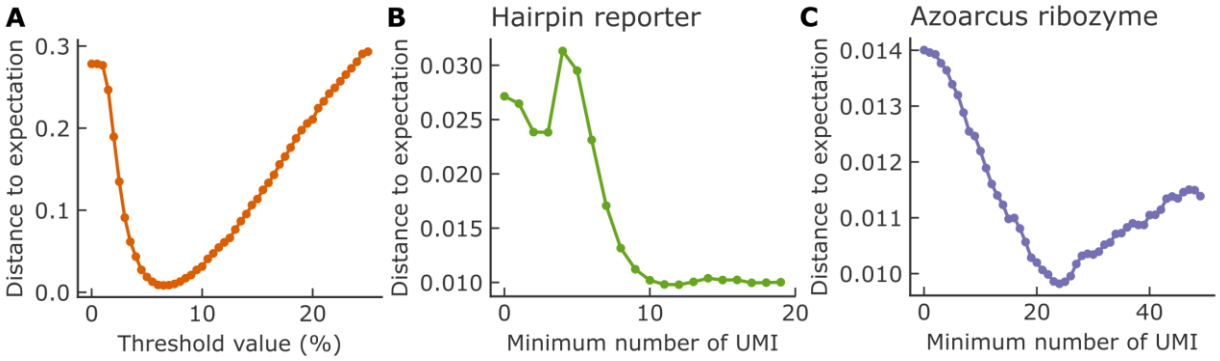


Figure 32. Choosing the thresholds. Here distance to expectation is the Euclidian distance between the distribution of the number of hairpin reporters per droplet barcode once the three thresholds have been applied and the distribution of number of 5 pL droplets containing initial WXY emulsions to a 50 pL containing reaction buffer and an excess of Z during the library fusion step measured by video acquisition.The three thresholds are: (**A**) the minimum percentage of total hairpin reporter UMI for a hairpin reporter to be declared part of the coding set of hairpin reporter for a given droplet barcode (default value = 7.5%), (**B**) the minimum number of UMI of hairpin reporter for a droplet barcode to pass to be considered for the final dataset (default value = 10) and (**C**) the minimum number of UMI of *Azoarcus* ribozyme for a droplet barcode to pass to be considered for the final dataset (default value = 20).

2.4.12 Reducing $MgCl_2$ concentration

Unlike previous study (Vaidya et al. 2012) where the WXYZ ribozyme were purified before subjecting to RT-PCR, in this study as networks are formed in droplets and directly subjected to reverse transcription, thus, there is no possibility to purify the products. While the amplification protocol were optimized to give PCR product only from full-length WXYZ ribozymes (not from substrates), high $MgCl_2$ used earlier in reaction (100 mM) could be problematic in downstream steps as well the *Azoarcus* reaction could still go on during the incubation for RT. Therefore $MgCl_2$

concentration for RT is reduced by two ways; optimizing the reaction at lower $MgCl_2$ concentration and by diluting the RNA network containing droplets by splitting and then fusing with RT droplets. By performing self-assembling reaction at different $MgCl_2$ concentration (10-100 mM), 20 mM $MgCl_2$ was chosen for the experiments as it gave enough WXYZ product at 1 h (Figure 33) and after dilution by split ($1/10^{th}$), the final concentration of $MgCl_2$ from the reaction was just 2 mM.

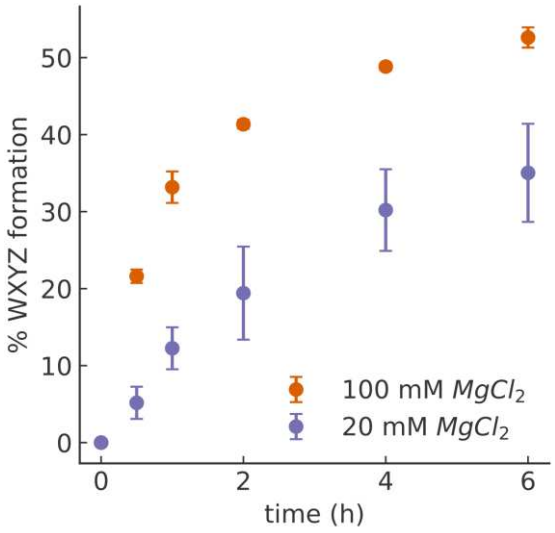


Figure 33. Effect of reducing $MgCl_2$ concentration on the WXYZ ribozyme formation. Time courses showing the effect of decreasing $MgCl_2$ concentration from 100 mM to 20 mM on the WXYZ formation. The self-assembly reaction is analysed on 12% denaturing polyacrylamide gels and data is extracted from band intensities. All the time-courses are done in triplicates and mean WXYZ product formation is plotted along with standard deviation.

2.4.13 Quality control of beads

The quality of the barcoded beads was checked by analyzing the percentage of full-length barcodes in the final library and as well the diversity of barcodes on beads. Percentage full-length barcode was analyzed by subjecting a small portion of the beads to restriction endonuclease treatment (BclI, New England Biolabs, Product No.: R0160L). The barcodes beads have unique BclI site in the first common adaptor which is also used to release barcodes during RT in droplets. The

beads were centrifuged and supernatant was analyzed on TapeStation (Agilent 2200 TapeStation, using high sensitivity D1000 ScreenTape®, Product No.: 5067-5584). On average ~70% of the total oligonucleotide released contained full-length barcodes.

For checking the diversity of barcodes on beads, beads were sorted in 96-well plate using Fluorescence-activated cell sorting (FACS) (Institut Cochin, INSERM U1016, Plateforme de Cytométrie et d'Immunobiologie CYBIO, Paris, France) such that each well contains only one beads in 4 µL of water. RT was performed in each well in a total volume of 10 µL with 0.1 µM of row-specific harpin reporter for the synthesis of cDNA coupling the bead barcode and the row-specific hairpin reporter. The reaction was incubated for an hour at 60°C and 2 µl of each well was transferred to a new plate. Two sequential PCR steps were carried out in each well as described in section below (see Sequencing sample preparation section) though the first PCR was performed using column-specific forward primer. PCR products after the first PCR were purified separately in each well using 1.2 equivalent of AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) following manufacturer's protocol. 71/96 wells showed correct size amplification and 1 µL of PCR products were taken out from the positive wells in a pooled sample that was purified using 1.2 equivalent of AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) following manufacturer's protocol. The final sample was subjected to high-throughput sequencing using HiSeq XXX system for 2*150 High Output mode at BGI, Hong-Kong, China. To analyse this data, the sample barcode (first 6 bp of read 1), the hairpin reporter identity and the bead barcode were extracted for each read (see Sequencing data processing section). A pair sample barcode, hairpin reporter identifies a single well. For each pair of sample barcode, hairpin reporter the percentage of reads of the most abundant barcode is computed and constitutes our measure of bead purity (Figure 34). All beads but three have near perfect purity value with a mean value at 93.6%. Three beads have purity values around 50% and for the three corresponding wells, the second most abundant barcodes also represent around 50% of the reads which indicates that probably two beads instead of a single one were put in these wells.

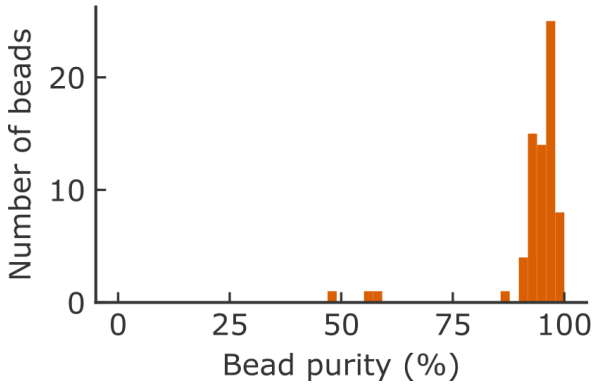


Figure 34. Histogram of bead purity in %. Bead purity is the percentage of the reads associated with the most abundant barcode per well. For this, single hydrogel bead were sorted in a 96-well plate using FACS. Datapoints were binned in 100 bins linearly spaced between 0% and 100%.

Table 4. Oligonucleotides used in the present chapter.

Name	Sequence	Description
Oligo 1	ctgcagaattctaatacgactcactatagagccttgcgccgggaaaccacgcaag ggatgg	Forward primer to have 'GAG' as IGS with 'A' at the 25th and 'U' at the 7th position
Oligo 2	ctgcagaattctaatacgactcactatagcgcctggcgccgggaaaccacgccag ggatgg	Forward primer to have 'GCG' as IGS with 'C' at the 25th and 'G' at the 7th position
Oligo 3	ctgcagaattctaatacgactcactatagtgcctagcgccgggaaaccacgctagg gatgg	Forward primer to have 'GUG' as IGS with 'U' at the 25th and 'A' at the 7th position
Oligo 4	ctgcagaattctaatacgactcactatagggcctcgcgccgggaaaccacgcgag ggatgg	Forward primer to have 'GGG' as IGS with 'G' at the 25th and 'C' at the 7th position
Oligo 5	ccggtttgtgtgactttcgcc	Reverse primer targeting 3' end of Z fragment
Oligo 6	atgtgccttaggtgggtgc	Reverse primer to add 'A' tag in WXY fragment
Oligo 7	aggtgccttaggtgggtgc	Reverse primer to add 'C' tag in WXY fragment
Oligo 8	aagtgccttaggtgggtgc	Reverse primer to add 'U' tag in WXY fragment
Oligo 9	acgtgccttaggtgggtgc	Reverse primer to add 'G' tag in WXY fragment
Oligo 10	ctgcagaattctaatacgactcactatagagccttgcgcc	Forward primer to have 'GAG' as IGS with 'A' at the 25th and 'U' at the 7th position
Oligo 11	ctgcagaattctaatacgactcactatagagcctggcgcc	Forward primer to have 'GAG' as IGS with 'C' at the 25th and 'G' at the 7th position

Oligo 12	ctgcagaattctaatacgactcactatagagcctagcgcc	Forward primer to have 'GAG' as IGS with 'U' at the 25th and 'A' at the 7th position
Oligo 13	ctgcagaattctaatacgactcactatagagcctcgcgcc	Forward primer to have 'GAG' as IGS with 'G' at the 25th and 'C' at the 7th position
Oligo 14	ccuaccaguacccuacca	RNA adaptor to ligate
Oligo 15	cgtgtgctcttccgatctnnnnnnnnttttttttttttttttttvn	RT primer for WXY control
Oligo 16	/5Acryd/tcttcacggaacga	Top strand of the double-stranded DNA linked to the hydrogel beads (/5Acryd/ = acrylic phosphoramidite 5'-modification.)
Oligo 17	cagttcgttccgtgaaga	Bottom strand of the double-stranded DNA linked to the hydrogel beads
Oligo 18	actggactagaatgatcagtgactggagttcagacgtgtgctcttccgatct	Top strand of the common adaptor used to ligate the 1st index to the bead
Oligo 19	cgaaagatcggaagagcacacgtctgaactccagtcactgatcattctagtc	Bottom strand of the common adaptor used to ligate the 1st index to the bead
Oligo 20	/5Phos/ttcg-1st index-	Top strand 1st index (/5Phos/ indicates that the oligo is phosphorylated at the 5' end)
Oligo 21	gtca-1st index-	Bottom strand 1st index (/5Phos/ indicates that the oligo is phosphorylated at the 5' end)
Oligo 22	/5Phos/tgac-2nd index-	Top strand 2nd index (/5Phos/ indicates that the oligo is phosphorylated at the 5' end)
Oligo 23	tggt-2nd index-	Bottom strand 2nd index (/5Phos/ indicates that the oligo is phosphorylated at the 5' end)
Oligo 24	/5Phos/acca-3rd index-	Top strand 3rd index (/5Phos/ indicates that the oligo is phosphorylated at the 5' end)
Oligo 25	gttg-3rd index-	Bottom strand 3rd index (/5Phos/ indicates that the oligo is phosphorylated at the 5' end)
Oligo 26	accatacgctacggaacgannnnnnnnccggtttgtgtgactttcgccactc	Primer for RT ligated after the indexes (top strand)
Oligo 27	tcgttccgtagcgta	Bottom strand used for the ligation of the primer for RT
Oligo 28	ttccctacacgacgctcttccgatct-Sample barcode (6 nt)-gagccttgcgccgggaaacca	Illumina sequencing primer I (forward)
Oligo 29	gtgactggagttcagacgtgtgctcttccg	Illumina sequencing primer I (reverse)
Oligo 30	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttcc	Illumina sequencing primer II (forward, Rd1)

Oligo 31	caagcagaagacggcatacgagat gtgactggagttcagacgtgtgctcttccga tct	Illumina sequencing primer II (reverse, Rd2)
Oligo 32	ttccctacacgacgctcttccgatct-Sample barcode (6 nt)-cctaccagtaccctacc	Illumina sequencing primer I (forward) used for original IGS control

3. Characterization of memory of initial conditions, pre-requisite for heritability, in prebiotic RNA networks

Note: this chapter is adapted from a manuscript in preparation.

3.1. Abstract

Collective Autocatalytic Sets (CASs), where in an ensemble molecules catalyze the formation of each other, are often regarded as the most plausible candidates for the first self-reproducing molecular systems on prebiotic Earth. Though extensive theoretical efforts have been dedicated to their study, Darwinian-like evolution with CAS requires empirical demonstration of basic ingredients; variation, heredity and selection. Exploring heritability involves examination of 'memory' of the initial state seeded in an ensemble, which depends on CAS structure and thus dictates the final state. Possessing memory is a necessary pre-requisite for heredity as a CAS without memory would have only one trivial stable state. Here using RNA CASs derived from *Azoarcus* ribozyme system, we identified the set of structural parameters influencing the memory in CASs and used a simple computational model to quantify the effect of these parameters. We observed that besides the total sum of weights of all the connections in a network, catalyst's uniqueness plays a significant role. We then experimentally studied a set of networks with different initial states to test the validity of the model developed here.

3.2. Introduction

Interest in Collectively Autocatalytic Sets (CAS) (Kauffman 1992) as potential candidates for the origins of life has grown over the past decades as they possess several advantages over the

general template-based replicases (Higgs & Lehman 2015). In context of RNA-world hypothesis, though small biopolymers have been synthetized in prebiotic conditions (Orgel 2004; Patel et al. 2015; Ferris et al. 1996; Rajamani et al. 2008), none have reached the length required to overcome the "error-threshold" that would be needed for a generalist auto-replicase in order to self-replicate without information decay (Kun et al. 2015). RNA-based CASs overcome this problem as each molecule's formation in the set is catalyzed by another member and as they can self-reproduce from a food-set composed of small polymers (Hordijk et al. 2012).

Even though CASs represent attractive models for the origins of life, a formal proof of evolution with CASs is yet to be found. For this, required ingredients of Darwinian evolution: variation, heredity, and selection must be present in CAS systems (Nghe et al. 2015) and an empirical demonstration of any of these ingredients will strengthen the plausible origins of life scenarios with CASs. Though CAS have been extensively investigated theoretically (Eigen & Schuster 1977; Eigen & Schuster 1978; Hordijk 2013; Gánti 2003; Segré et al. 1998; Jain & Krishna 2001; Vasas et al. 2012), experimental studies exploiting these ingredients are largely lacking. Regarding 'variations', there exist empirical examples of peptide-based networks (Lee et al. 1997; Ashkenasy et al. 2004) and RNA-based networks (Kim & Joyce 2004; Lincoln & Joyce 2009) which show that such molecular CASs system possess some diversity, however 'heredity' has not yet been addressed using these systems. Even though the idea of compositional heritance has been theoretical studied by Segré and colleagues (Segré et al. 2000), it was found to be lacking evolvability in this system due to the lack of multiple stable states (Vasas et al. 2010). For a system of CAS to show heredity, it would require having several stable states such that each state can be transmitted to future generations, for example in a serial dilution experiments. In order for evolution to happen in such a generation experiment, it is necessary for the CAS system to have certain memory of the initial conditions which can be propagated, and ideally such memory should be function of CAS structure such that different structures possesses different level of memory. Due to the scarcity of experimental systems (Lee et al. 1997; Ashkenasy & Ghadiri 2004; Sievers & von Kiedrowski 1994; Paul & Joyce 2002; Kim & Joyce 2004), the structural characteristics of CAS important to have memory of initial conditions and the parameters memory depends on have never been investigated systematically, neither experimentally or theoretically.

Here we exploit RNA-based CAS system derived from group I intron of the *Azoarcus* (Cech 1990; Reinhold-Hurek & Shub 1992) bacteria to explore memory of initial conditions in CASs, both computationally and experimentally. In this RNA system, an RNA recombinase WXYZ can catalyze its own formation from the two fragments WXY and Z (Draper et al. 2008; Riley & Lehman 2003) (Figure 35). The self-assembly depends on the interactions between 3-nt of the internal guide sequence (IGS) and of the target sequence (tag) at the 5' end and 3' end of the WXY fragment, respectively (Figure 35. A). The reaction can be catalyzed either by a fully covalent catalyst or less efficiently by the non-covalent complex of the two fragments WXY and Z (Hayden et al. 2008). The middle nucleotide of these two recognition sequences (IGS and the tag) can be varied. As a result, a diversity of catalyst (16 different WXYZ) can be used to generate closed catalytic cycles (Figure 35. B), and multi-node RNA networks (Vaidya et al. 2012; Yeates et al. 2016; Yeates et al. 2017). By studying this system computationally, we found that memory of initial conditions depends only on a limited set of network parameters. Due to differences in the network architecture, some networks possess a lot of memory while some do not, and even the same network can have some memory for some initial conditions while it does not for some other. We developed a simple first-order approximation to analytically quantify the effect of the identified parameters on memory. Finally, we tested the validity of our computational model by experimentally studying different initial conditions with eight RNA networks. The results here are an empirical proof that CAS can retain memory of initial conditions and that it depends on its structure of specific interactions. These observations show that even though rudimentary, these RNA-based CASs can be complex enough to keep memory of the information provided; a process crucial for modern day organisms.

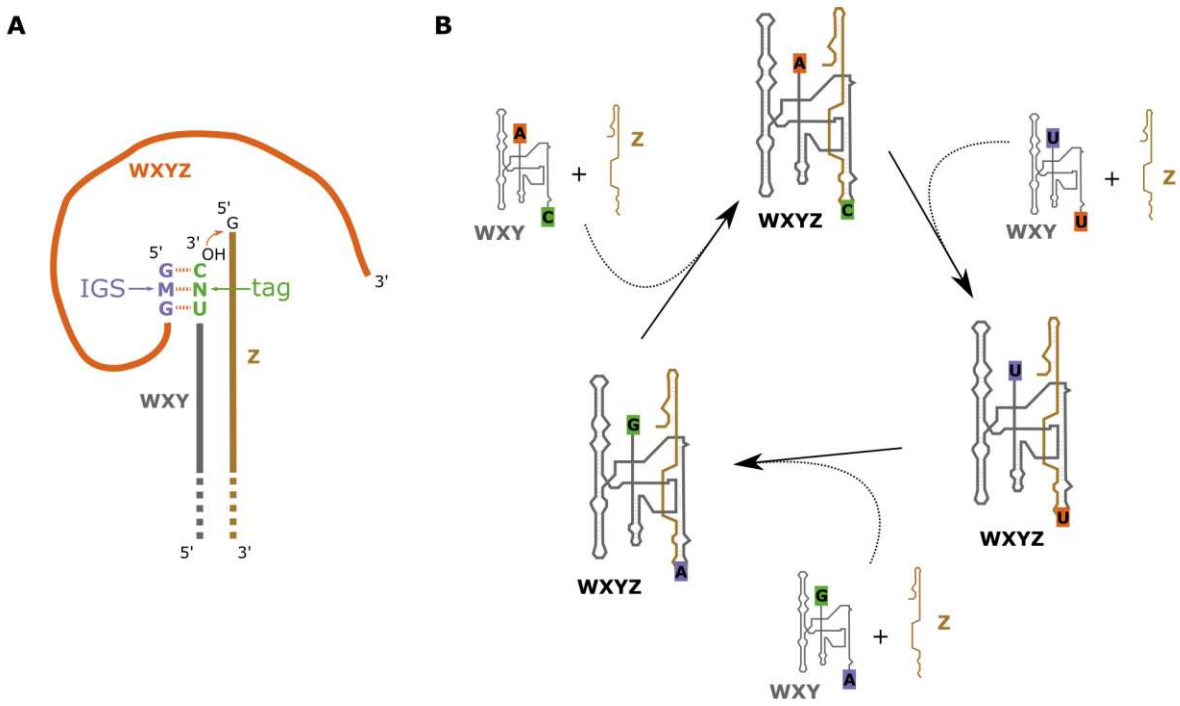


Figure 35. The *Azoarcus* ribozyme system. (**A**) Mechanism of the recombination between fragment WXY and Z catalyzed by WXYZ in an autocatalytic manner. The catalysis proceeds through the binding of the Internal Guide Sequence (IGS) of a catalyst 'GMG' to the target (tag) sequence 'CNU' of WXY fragment. (**B**) Example of a three-member *Azoarcus* RNA network where AC catalyzes (solid arrow) the formation of UU from the corresponding fragments (dashed arrow). Similarly UU catalyzes the formation of GA that catalyzes the formation of AC, forming a closed catalytic cycle.

3.3 Results and discussion

3.3.1 Model of *Azoarcus* networks

In order to computationally study memory of initial conditions in CASs, we use a theoretical model of *Azoarcus* RNA networks which we describe here by first specifying the nomenclature used and then by defining the mathematical expressions controlling the temporal evolution of these networks.

In this system, the interaction between IGS and tag, which are mentionned above, can be defined as 'MN' where M and N are respectively the middle nucleotide of the IGS 'GMG' and of tag 'CNU', thus the catalyst $_{5'GMG}WXY_{CNU}Z_{3'}$ is noted as MN. If M can form a Watson-Crick base pair with N then the MN ribozyme can efficiently catalyze its own formation (Yeates et al. 2016). However if M can base pair with N', the tag of another ribozyme M'N' then the first ribozyme can cross-catalyze the formation of the second ribozyme. This allows the construction of RNA catalytic networks (Vaidya et al. 2012) in which the nodes are the MN ribozymes and the interactions are weighted directed links between two nodes if the upstream node is a catalyst for the downstream node. The weight of this link quantifies how well the upstream node can catalyze the formation of its target. For the computational model, we consider that the variation of the concentration x_i of a node i has two terms, one which is proportional to the concentrations of its upstream nodes and one which is constant which quantifies the rate at which the non-covalent complexes can assemble this node. If $X = (x_i)_i$ is the vector of the concentrations of all the nodes, then its evolution over time is governed by the following system of ordinary differential equations: $dX/dt = AX + B$ where $A = \left(k_{ij}^a\right)_{ij}$ and $B = \left(b_i = \sum_j k_{ij}^b\right)_i$. The values for the k^a (min^{-1}) and for the k^b (µM.min^{-1}) rates were taken from measurements done for another study and are summarized in Table 5.

Table 5. Catalyzed and spontaneous formation rates. A catalyst with M as IGS catalyzes the formation of a catalyst with N as tag at a rate equal to k_{MN}^a multiplied by its concentration. WXY fragments with M as IGS will spontaneous form a complex with Z fragment and contribute to the rate of formation of a catalyst with N as tag at a rate equal to k_{MN}^b if concentration of WXY fragment is at 0.1 µM and if Z fragment is in excess compared to this concentration. These values were experimentally measured for another study (see 2.3. Results and discussion, Table 2) and here we took the mean between similar types of interactions (A/U and C/G interactions).

Interaction $M \rightarrow N$	k_{MN}^a (min^{-1})	k_{MN}^b (µM.min^{-1})
$A \rightarrow U$	0.015	0.0010
$U \rightarrow A$	0.015	0.0010
$C \rightarrow G$	0.040	0.0051

$G \rightarrow C$	0.040	0.0051

3.3.2 Measure of a network's memory of initial conditions using the model

Using the model described in the previous section, we analyzed how the concentrations of the nodes of 5-member networks are changing over time when one particular node, 'the seed', is doped in at the start of the reaction (at concentration x_0 = 50 nM). Comparing the situations between different two different seeds, seed 1 and seed 2, allows us to evaluate memory of initial conditions. We term the measure of this difference d_{12}. If the two situations between seed 1 and seed 2 are very different, i.e. that the two distributions of node relative abundance are not alike and d_{12} is high, then it can be said that the network has good memory of these two initial states (Figure 36. A). On the other hand, if the two distributions are very similar and d_{12} is small, then it can be said that the network 'forgot' the initial states and thus has no memory of them (Figure 36. B).

More precisely, using the model, we let the system evolve until the total concentration of the catalysts in the system has doubled ($2x_0$ = 100 nM) since it is the minimal requirement for a duplicating system to maintain stable concentrations. The distribution of relative concentrations of the remaining nodes is then assessed and compared between two different seeds, seed 1 and seed 2. The relative abundance of a node j can be written as y_1^j or y_2^j when it is seeded with seed 1 or 2, respectively. We then assess the absolute difference $|y_1^j - y_2^j|$ to compare the situations for node j with the two different seeds. By analyzing this difference of relative concentration for all the 10 possible pairs of seeds for 5-membered networks, we observed that for some networks, the difference $|y_1^j - y_2^j|$ is very high, thus they have high memory (Figure 37. A), while others have little memory having $|y_1^j - y_2^j|$ values close to 0 (Figure 37. B).

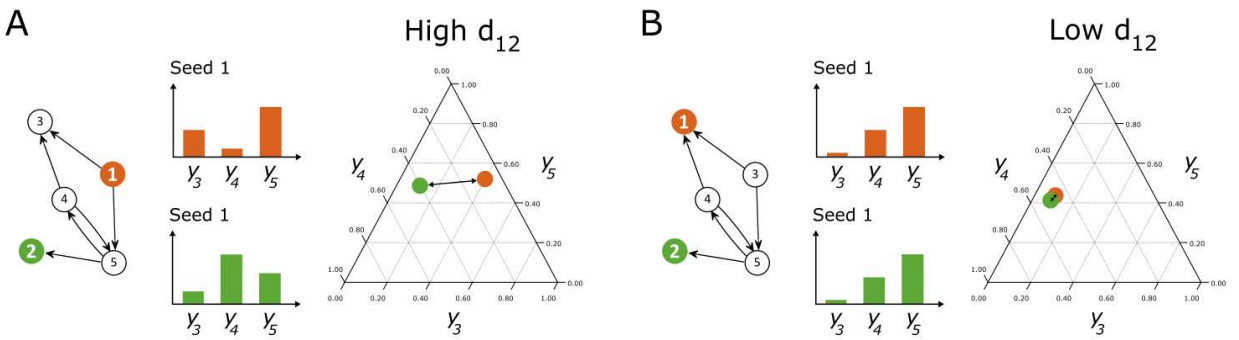


Figure 36. Schematic illustrating the concept of memory of initial conditions and d_{12} our measure of it. Nodes 1 and 2 are the two seeded nodes and we analyze the relative abundance of nodes 3, 4 and 5. Depending on which node is seeded at start, two different distributions of y_i are obtained. Graphically in a ternary plot, if the network retains memory of the two initial states (**A**) then the two points are well separated and d_{12} is high. On the contrary, if the network has no memory of the initial conditions (**B**) and 'forgets' them quickly then the two points are very close to one another.

To compare between two pairs of seed but for all the nodes, the sum of $|y_1^j - y_2^j|$ for all three nodes can be computed and thus can d_{12} be written as $d_{12} = \sum_{j=3}^{5} |y_1^j - y_2^j|$, at the expense of reordering the nodes so that nodes 1 and 2 are the two seeds. Again, this is a measure of what memory the network has of the two different initial conditions with seed 1 or seed 2. In the extreme case where $d_{12} = 0$, the fractions of the nodes are exactly the same whatever the seed and thus given the final state, it would not be possible to determine the initial state. On the other hand, if the fractions of the nodes are very different depending on which seed was used, then d_{12} is high and the network has good memory of these two initial states.

Other options were considered for the measure of memory. Instead of comparing between two different seeds, a first option is to compare the final state with one seed to the initial state. This is more appropriate to study final stable states but here we are interesting in detecting differentiable transient states and so it is needed for us to compare between potentially different final states. A second option would have been to compare the final state with one particular seed to the asymptotic state obtained if the system is left to evolve for a long time thus mimicking a flow reactor where quantity of substrate is unlimited and where there is continuous dilution. Though this asymptotic state can be computationally obtained by computing the first eigenvector, i.e. associated with the maximum eigenvalue, of the network's adjacency matrix, it cannot be

analytically derived and so it would not be possible to analytically determine the important network parameters responsible for memory.

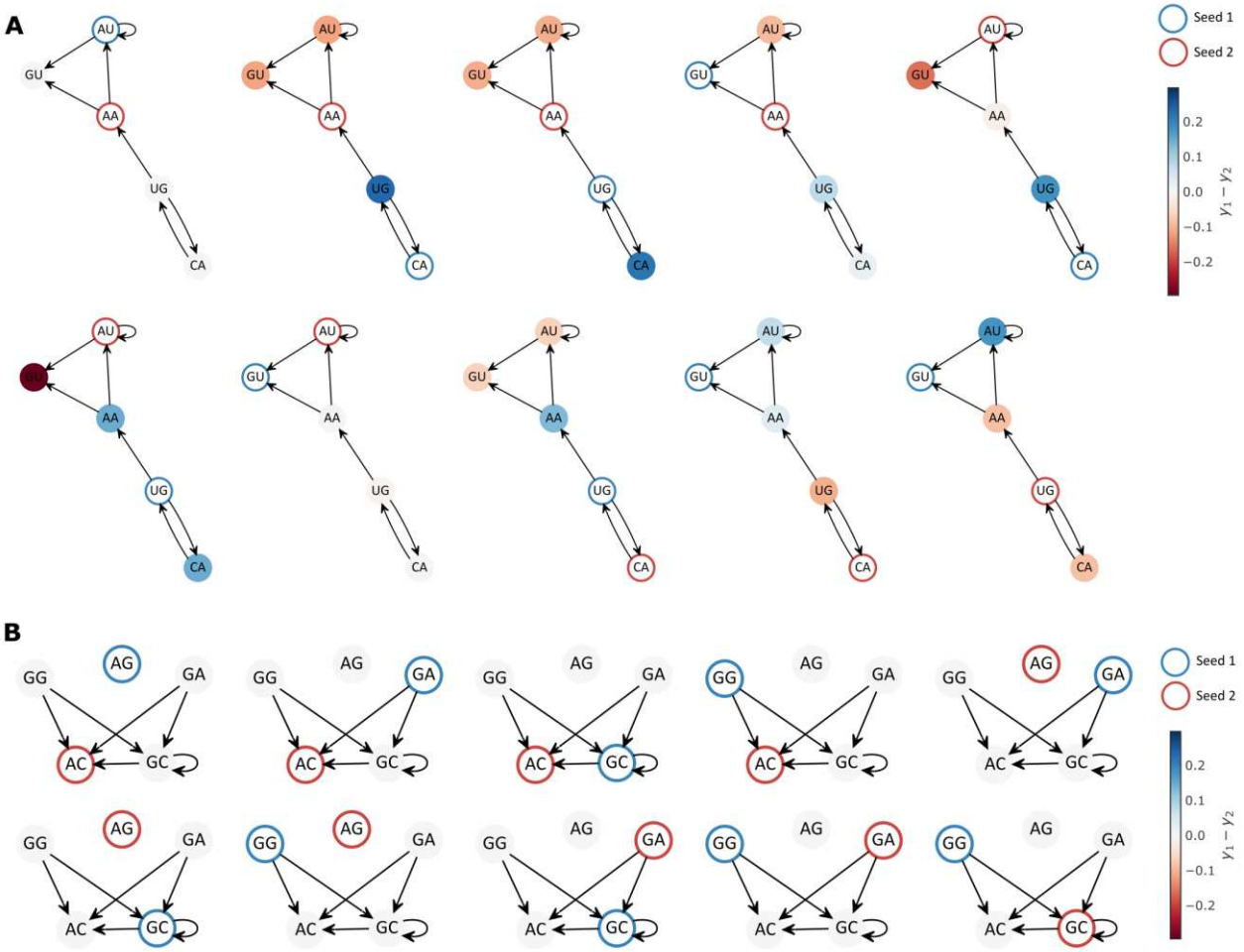


Figure 37. Model results with all pairs of seeds for two networks with high and low memory of initial conditions. Seed 1 is circled in blue while seed 2 is circled in red. For the three remaining nodes, the color of the node codes for the value of $y_1 - y_2$ where y_1 (resp. y_2) is the relative concentration of the node when the seeded node is seed 1 (resp. seed 2) compared to the rest of the network but excluding the two seeds.

3.3.3 Identification of a restricted set of parameters controlling the memory of initial conditions

We develop in what follows a first-order approximation to obtain a simple analytical expression for d_{12} that allows us to identify which network parameters are important for memory. We found

that d_{12} can be parameterized, apart from the initial seed concentration x_0, with only 5 parameters: n_1 and n_2, the number of targets for the two seeds, m_1 and m_2, the number of nodes in the network with the same IGS as the two seeds and the sum of the spontaneous formation rates of the three nodes which are not the seeds. This last parameter accounts for the network link density since a very dense network would have a high value whereas a sparse network would not.

More precisely, to obtain an analytical expression for d_{12}, we proceed as what follows. The two rate matrices mentioned before are $A = \left(k_{ij}^a\right)_{ij}$ and $B = \left(b_i = \sum_j k_{ij}^b\right)_i$ where $k_{ij}^a = k_{M\to N}^a$ and $k_{ij}^b = k_{M\to N}^b$ with M the IGS of node j and N the tag of node i. Let x_i be the concentration of node i whose variations follows the relationship $dx_i/dt = \sum_j k_{ij}^a x_j + b_i$. We make a first-order approximation by considering that dx_i/dt is constant and that $dx_i/dt = k_{ik}^a x_0 + b_i$ where node k is the seeded node, initially at the concentration x_0. The concentration of node i after Δ_k minutes is $x_i(\Delta_k) = (k_{ik}^a x_0 + b_i)\Delta_k$ where Δ_k is the time required for the sum of all concentrations to reach $2x_0$. Let us reorder the nodes so that node 1 and node 2 are the two seeds. Then, we obtain:

$$d_{12} = \sum_{j=3}^{5} \left| \frac{x_i(\Delta_1)}{\sum_{i=3}^{5} x_i(\Delta_1)} - \frac{x_i(\Delta_2)}{\sum_{i=3}^{5} x_i(\Delta_2)} \right| = \sum_{j=3}^{5} \left| \frac{(k_{ik}^a x_0 + b_i)\Delta_1}{\sum_{i=3}^{5}(k_{ik}^a x_0 + b_i)\Delta_1} - \frac{(k_{ik}^a x_0 + b_i)\Delta_2}{\sum_{i=3}^{5}(k_{ik}^a x_0 + b_i)\Delta_2} \right|$$

which can be rearranged by normalizing each term by $x_0 e_2$ as:

$$d_{12} = \sum_{j=3}^{5} \left| \frac{\delta_{j2} + \frac{b_j}{x_0 e_2}}{n_2 + \frac{\beta}{x_0 e_2}} - \frac{\delta_{j1} + \frac{b_j}{x_0 e_2}}{n_1 + \frac{\beta}{x_0 e_2}} \right|$$

where $\delta_{jk} = 1$ if seed k is a catalyst for node j, x_0 is the initial seed concentration, e_k is the catalytic strength of seed k (for example $e_k = k_{AU}^a$ if seed k has A as IGS), n_k is the number of targets for seed k and $\beta = \sum_{j=3}^{5} b_j$ is the sum of the spontaneous formation rates of nodes 3, 4 and 5. The cases where both seeds have the same IGS are excluded as they would have the exact same effect and thus d_{12} would be null. According to the rates (Table 5), the C/G catalysts can be considered as strong catalysts while the A/U as weak catalysts because k^a and k^b values are 3-5 times smaller. Here the first seed (seed 1) is always by convention with A/U as IGS (weak seed)

and the second seed (seed 2) is with C/G as IGS (strong seed). We analyzed d_{12} as a function of the normalized network's background, β/x_0e_2. We get as a result several families of curves depending on the values of n_1 and n_2 (Figure 38, Table 6). In each of these families, the curves depend only on another pair of parameters m_1 and m_2 which represent the number of nodes in the network that have the same IGS as seed 1 or seed 2.

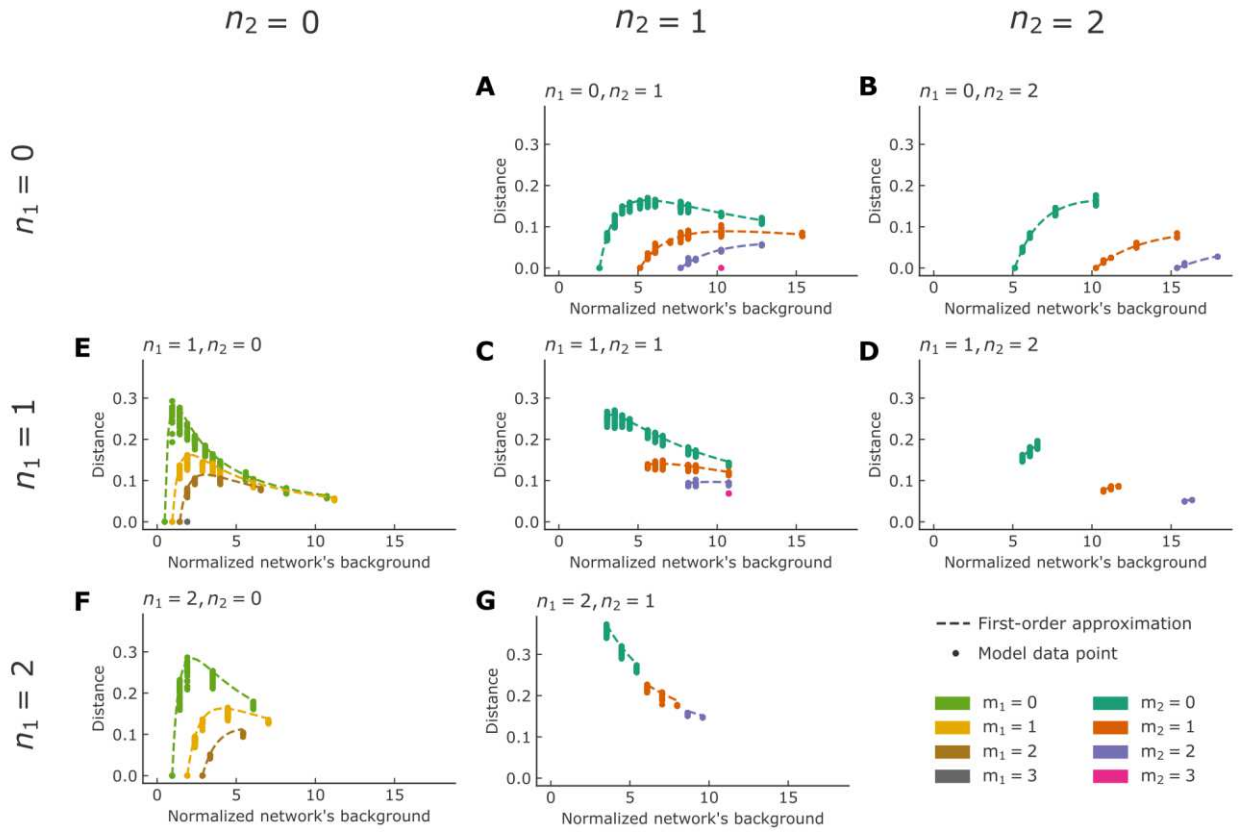


Figure 38. The distance (d_{12}) between the two distributions of relative abundance of the nodes in a 5-member network when two different nodes are doped in as seed. This depends notably of the number of targets for the two seeds (n_1 and n_2), on the number of nodes that share the same IGS with the seeds, i.e. that are similar catalysts (m_1 and m_2) and finally on the normalized network's background β/x_0e_2 which is the sum of the spontaneous formation rates of the 3 nodes (excluding the two seeds) normalized by x_0e_2 where x_0 is the initial seed concentration and e_2 is the catalytic strength of seed 2. (**A-G**) d_{12} is plotted against β/x_0e_2 for different values of n_1, n_2, m_1 and m_2. Here only the cases where seed 1 is a weak seed (with A/U IGS) and seed 2 is a strong seed (with C/G IGS) are considered. Dashed line is the first-order approximation and solid points are the complete model data points.

Table 6. First-order approximation expressions for d_{12}. This first-order approximation presented above can be further parameterized for particular values of n_1 and n_2. Here are reported the expressions used in Figure 38 for plotting the first-order approximation line. So in these particular cases, seed 1 has A/U for IGS and seed 2 has C/G so here $e_2 = k^b_{C\to G}$ and $e_1 = k^b_{A\to U}$. Similar types of expression can be obtained when the two seeds are of the same type.

n_1	n_2	m_1	m_2	d_{12}
0	1	N/A	$\in \{0,1,2,3\}$	$\frac{2}{\beta/x_0e_2\left(1+\beta/x_0e_2\right)}\left(\beta/x_0e_2 - \frac{(m_2+1)k^a_{C\to G}}{x_0e_2}\right)$
0	2	N/A	$\in \{0,1,2\}$	$\frac{4}{\beta/x_0e_2\left(2+\beta/x_0e_2\right)}\left(\beta/x_0e_2 - 2\frac{(m_2+1)k^a_{C\to G}}{x_0e_2}\right)$
1	0	$\in \{0,1,2,3\}$	N/A	$\frac{2}{\beta/x_0e_2\left(1+\frac{e_2}{e_1}.\beta/x_0e_2\right)}\left(\beta/x_0e_2 - \frac{(m_1+1)k^a_{A\to U}}{x_0e_2}\right)$
2	0	$\in \{0,1,2\}$	N/A	$\frac{4}{\beta/x_0e_2\left(2+\frac{e_2}{e_1}.\beta/x_0e_2\right)}\left(\beta/x_0e_2 - 2\frac{(m_1+1)k^a_{A\to U}}{x_0e_2}\right)$
1	1	N/A	$\in \{0,1,2,3\}$	$\frac{1}{\left(1+\beta/x_0e_2\right)\left(1+\frac{e_2}{e_1}.\beta/x_0e_2\right)}\left(\left\|1+\frac{e_2}{e_1}.\beta/x_0e_2+\left(1-\frac{e_2}{e_1}\right)\frac{(m_2+1)k^a_{C\to G}}{x_0e_2}\right\| + \left\|1+\beta/x_0e_2+\left(1-\frac{e_2}{e_1}\right)\frac{(m_1+1)k^a_{A\to U}}{x_0e_2}\right\| + \left\|\left(1-\frac{e_2}{e_1}\right)\left(\beta/x_0e_2 - \frac{(m_2+1)k^a_{C\to G}}{x_0e_2} - \frac{(m_1+1)k^a_{A\to U}}{x_0e_2}\right)\right\|\right)$
1	2	N/A	$\in \{0,1,2\}$	$\frac{4}{\left(2+\beta/x_0e_2\right)\left(1+\frac{e_2}{e_1}.\beta/x_0e_2\right)}\left\|1+\frac{e_2}{e_1}.\beta/x_0e_2+(1-2\frac{e_2}{e_1})\frac{(m_2+1)k^a_{C\to G}}{x_0e_2}\right\|$
2	1	N/A	$\in \{0,1,2\}$	$\frac{1}{\left(1+\beta/x_0e_2\right)\left(2+\frac{e_2}{e_1}.\beta/x_0e_2\right)}\left(\left\|2+\frac{e_2}{e_1}.\beta/x_0e_2+\left(2-\frac{e_2}{e_1}\right)\frac{(m_2+1)k^a_{C\to G}}{x_0e_2}\right\| + 2\left\|1+\frac{1}{2}\frac{e_2}{e_1}\beta/x_0e_2+\left(1-\frac{1}{2}\frac{e_2}{e_1}\right)\frac{(m_2+1)k^a_{C\to G}}{x_0e_2}\right\|\right)$

3.3.4 Attenuating effect of network's background only at high values

We have identified several parameters useful for describing d_{12}, our measure of memory and developed a first-order approximation to obtain a set of analytical expressions depending on the values of the parameters. Remarkably, the first-order approximation (dashed line in Figure 38) conveys very well with the variations of the complete model data (solid dots in Figure 38). Nonetheless, this approximation tends to overestimate d_{12} in some conditions as it does not take into account higher order interactions which are more likely to link two parts of a network

resulting in a smaller d_{12}. Among the set of identified parameters, the normalized network's background β/x_0e_2 has a significant effect. Generally, it can be said that at high values, it has a attenuating effect on d_{12} (Figure 38. A-G) and that quite surprisingly there is often a small regime at low β/x_0e_2 where it has an increasing effect.

For example, when the strong seed has a single target ($n_1 = 0$ and $n_2 = 1$) doubling β/x_0e_2 from 5 to 10 only decreases d_{12} of about 20% at $m_2 = 0$ (Figure 38. A) whereas the decrease reaches 50% for the same increase of β/x_0e_2 when it is the weak seed that has a single target (Figure 38. E). At smaller values, β/x_0e_2 can have an increasing effect on d_{12} (Figure 38. A, B, D, E, F) since the network is directed almost exclusively to the set of targets of the two seeds and so the extra catalysis brought by the seeds does not have a significant effect. However, non-intuitively, when β/x_0e_2 increases more nodes are well catalyzed in the network and as a result d_{12} increases because the nodes which were high in fraction in a less connected network have relatively less fraction now, therefore doping in a seed can have a significant effect.

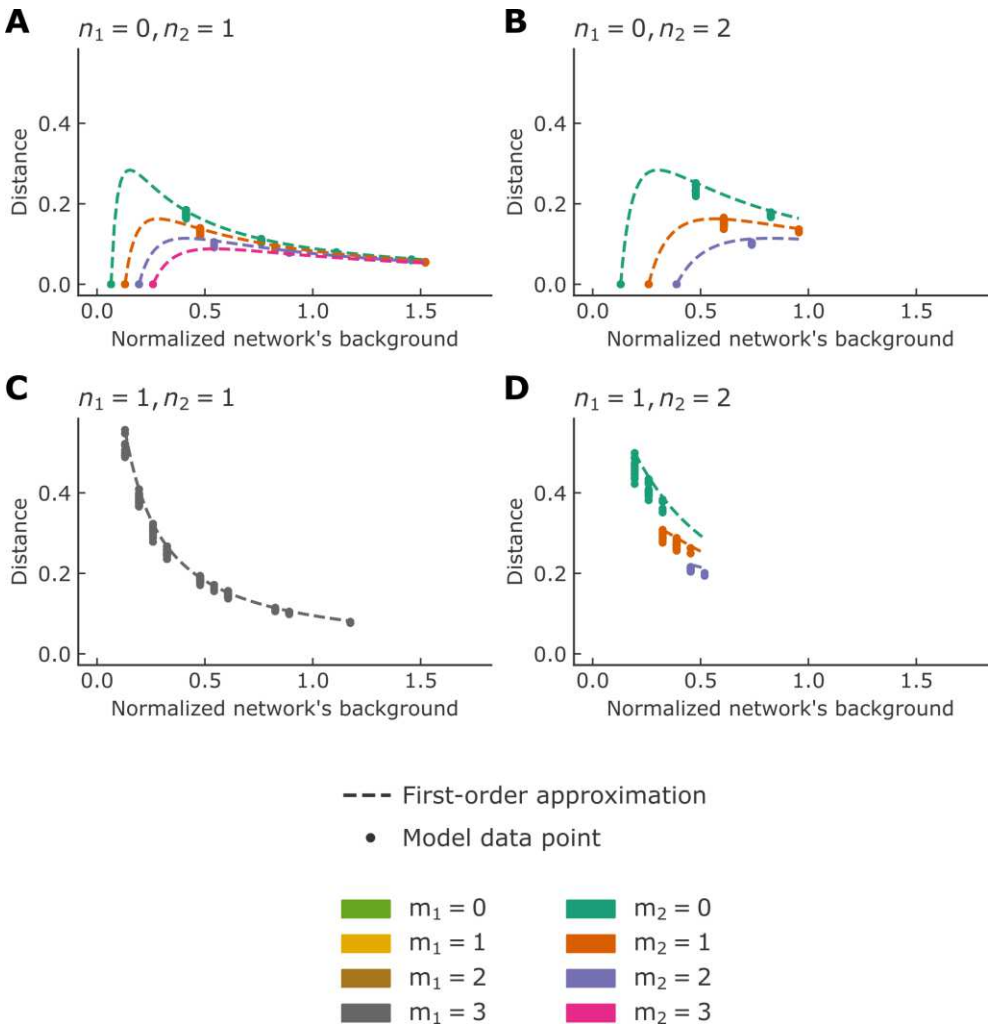


Figure 39. The distance (d_{12}) between two distributions of relative abundance of the nodes in a 5-member network when two different nodes are doped in as seeds at the start of reaction and when the two seeds have A/U as IGS. (**A-D**) d_{12} is plotted against $\beta/x_0 e_2$ for different values of n_1, n_2, m_1 and m_2. Here only the cases where seed 1 and seed 2 are weak seeds (with A/U IGS) are considered. Dashed line is the first-order approximation and solid points are the complete model data points.

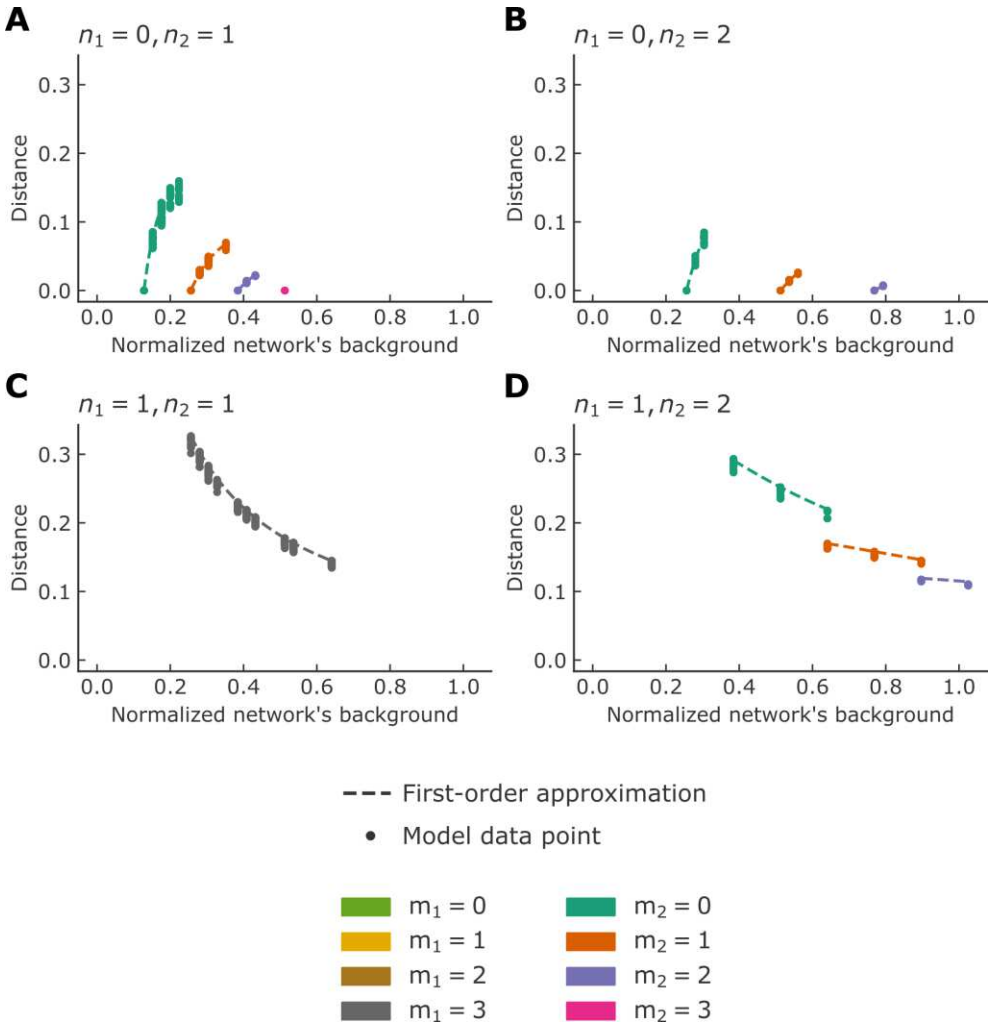


Figure 40. The distance (d_{12}) between two distributions of relative abundance of the nodes in a 5-member network when two different nodes are doped in as seeds at the start of reaction and when the two seeds have C/G as IGS. (**A-D**) d_{12} is plotted against $\beta/x_0 e_2$ for different values of n_1, n_2, m_1 and m_2. Here only the cases where seed 1 and seed 2 are strong seeds (with C/G IGS) are considered. Dashed line is the first-order approximation and solid points are the complete model data points.

3.3.5 Significant effect of catalyst uniqueness

Strikingly, there is a very significant attenuating effect of m_1 or m_2 which the number of nodes in the network with the same IGS as the seeds. They measure the seed's uniqueness because if $m_i = 0$, it means that the seed is catalytically unique since there are no other nodes in the network with the same IGS. Generally, more uniqueness, i.e. smaller values of m_i, are associated with higher memory of the initial conditions.

More precisely, the effect of m_1 or m_2 depends on the values of n_1, n_2 and $\beta/x_0 e_2$ (Figure 38. A-G). For example, the curves are very well separated when $n_1 = 0$ and $n_2 = 1$ (Figure 38. A) whereas they almost coincide when $n_1 = 2$ and $n_2 = 1$ (Figure 38. G), showing that in the latter case it matters less that there are other seed-like catalysts in the network. However, in the first case, the value of m_2 influence strongly d_{12} since the seed effect is less significant when m_2 is high because the effect is diluted by the spontaneous formation by the non-covalent complexes brought by the m_2 other nodes in the network. Thus to maximize d_{12} i.e., memory in a network, the seeds must target nodes that are as upstream as possible and not buried in the network structure. The highest values of d_{12}, i.e. above 0.2, are only reachable when the weak seed has at least one target (Figure 38. C, E, F, G) because the ratio between the catalyzed rate k^a and the spontaneous rate k^b is about two times higher (~1.98) for the weak A/U seeds than for the strong C/G seeds. Therefore, if the network is sparse enough, the effect of an A/U seed will relatively be stronger. Increasing the number of targets of the weak or of the strong seed can also has various effects. For example, changing from $n_2 = 1$ to $n_2 = 2$ while $m_1 = 0$ (Figure 38. A-B) does not increase the maximum d_{12} reached and maintains its distribution. This is intuitive given the symmetry between the two situations: for increasing d_{12}, it is quite similar to better catalyze one node in order to increase its fraction while the other two are losing some or to catalyze only two nodes and with no catalysis for the third one.

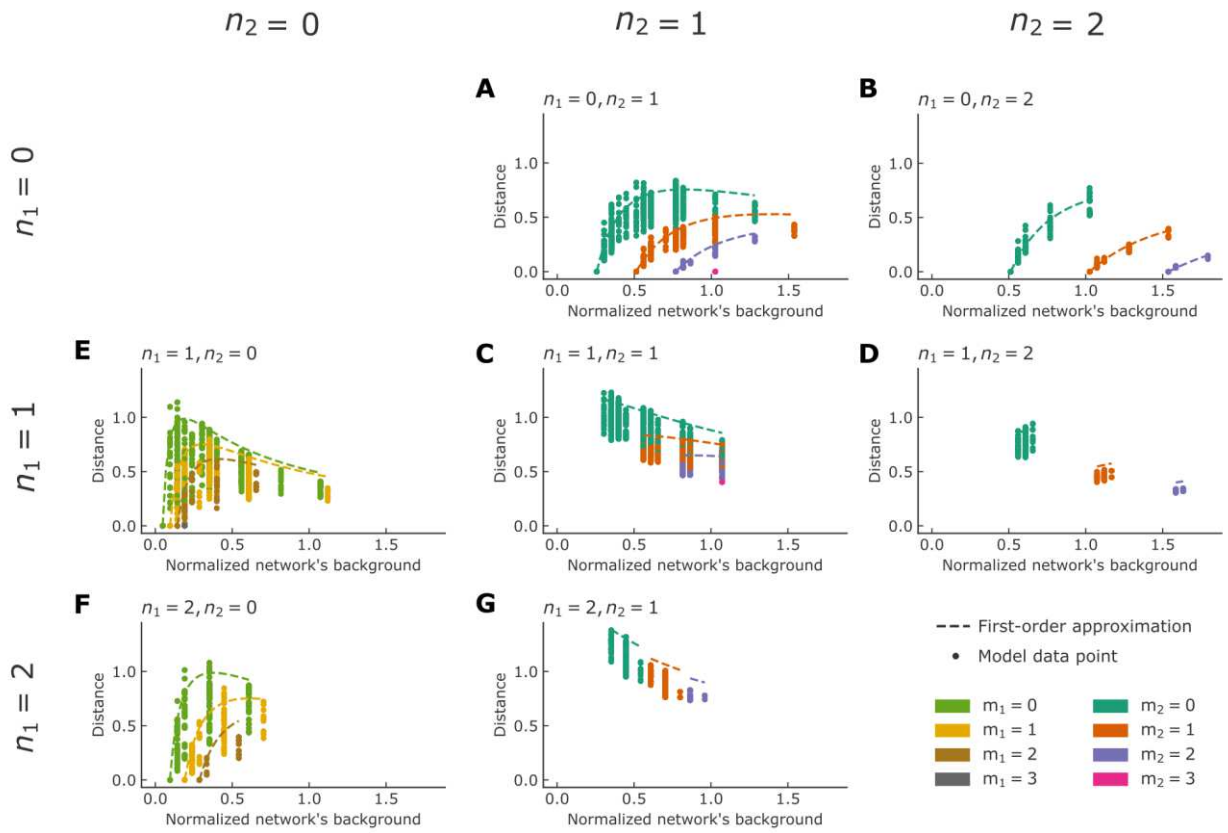


Figure 41. Same as Figure 38 but the results were obtained with with k^b values 10 times smaller.

3.3.6. Similar conclusions with other cases

The analysis presented above allowed us to identify what network features are responsible for the propensity of networks to maintain memory of the initial conditions. First, designing a network with a lot of memory, thus with high d_{12}, can be easily achieved in sparse networks since the spontaneous formation of its members is minimal. Second, to increase d_{12}, it is better to maximize the number of almost-upstream nodes which are is nodes with a single incoming link in order to get smaller values of m_1 and m_2. Though the results above are obtained for the cases where both seeds are of different types, similar observations can be made if the catalyst are of same type, i.e. either both weak (Figure 39) or both strong (Figure 40). Note that here in order to be in the same range as experimental data, the values of k^b used are rather high and at $t = 0$, $k^b \geq x_0 k^a$ for both A/U or C/G catalysts. However, even with high kb, we observed that many network structures have high memory of initial condition. It would certainly be advantageous if

the k^b is lower since it would enhance the difference between two seeds thus making it more pronounced. Indeed when in the model, k^b values are reduced by an order of magnitude, it increases d_{12} values but the shape of its distribution remains similar, implying that the general qualitative results are still valid though the first-order approximation is less good (Figure 41).

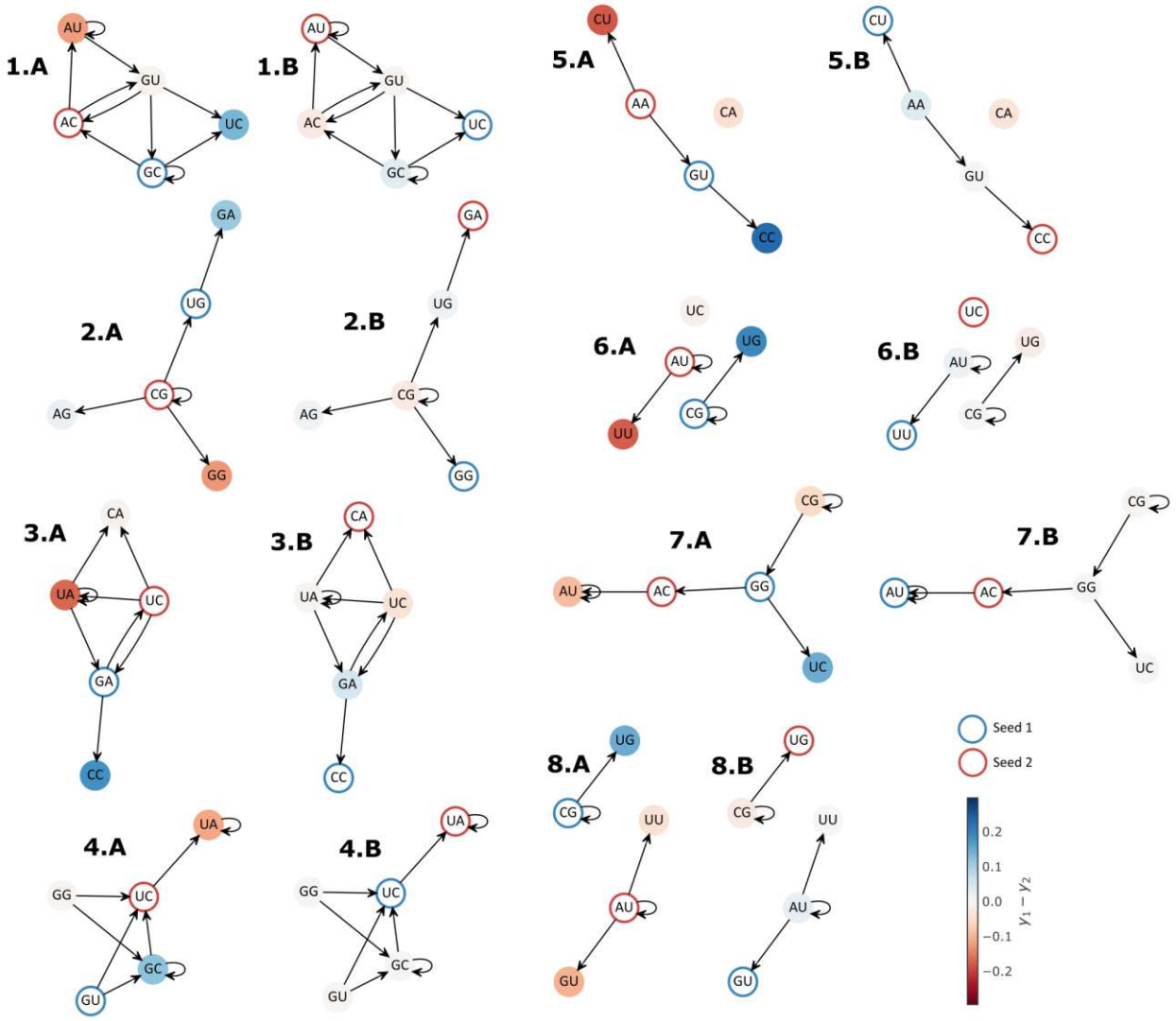


Figure 42. Experimental results for two different pairs of seeds for 8 different networks of five nodes. Seed 1 is circled in blue while seed 2 is circled in red. For the three remaining nodes, the color of the node codes for the value of $y_1 - y_2$ where y_1 (resp. y_2) is the relative concentration of the node when the seeded node is seed 1 (resp. seed 2) compared to the rest of the network, excluding the two seeds. Pair of seeds **A** is expected to cause high values of $|y_1 - y_2|$ while **B** is not.

Table 7. Values of the different parameters for the experimental data in Figure 42.

Network	Pair of seeds	n_1	n_2	m_1	m_2
1	A	1	2	1	1
1	B	0	1	0	1
2	A	1	2	0	0
2	B	0	0	0	0
3	A	1	2	0	1
3	B	0	0	0	0
4	A	1	1	2	1
4	B	0	0	0	0
5	A	1	1	0	0
5	B	0	0	2	2
6	A	1	1	0	0
6	B	0	0	1	1
7	A	1	1	0	1
7	B	0	0	0	0
8	A	1	2	0	0
8	B	0	0	0	1

3.3.7 Validation of the model and its conclusions by experimentally seeding a subset of networks

The conclusions and results discussed above were obtained computationally and analytically with a theoretical model. We thus acquired a non-negligible amount experimental data to further test the validity of this model and found that the experimental data was in good agreement with our model's expectations.

We did this by performing an experiment where 8 different networks are seeded, one node at a time, and the relative abundance of each node is measured. For each pair of seeded nodes, similarly we consider d_{12} as the sum of the net change in relative concentration of the remaining three nodes. Two pairs of seeds for each network are chosen, for high d_{12} (Figure 42, pair of seeds A), and low d_{12} (Figure 42, pair of seeds B). The parameters identified previously are reported for the chosen pairs of seeds in Table 2. Both from the values of these parameters and from intuition by looking at the network structure, one can get an immediate feeling of whether or not d_{12} will be important for a given network and pair of seed. For example, with pairs of seeds B the parameters n_1 and n_2 are almost always null, a case which trivially gives expected null values for

d_{12}. With pairs of seeds A, not only n_1 and n_2 are not null but m_1 and m_2 are small enough (i.e. the seeds are unique catalysts in the network) so that d_{12} can reach high values. This is a first indication that so far the experimental data seems to follow expectations from our model. The rest of the data with all pairs of seeds for every network can be found in 6.1. Annex 1: Seeding eight Azoarcus RNA networks one node at a time, the complete dataset. Finally, we compared the difference in relative abundance for each node between two different seeds to the model predictions and observed a good agreement (Figure 43), indicating that the parameters identified above are applicable to a large set of experimental networks. Notably because the model is independent of the number of nodes in the network, we expect that qualitative results both experimental and theoretical will be the same irrespective of the number of nodes.

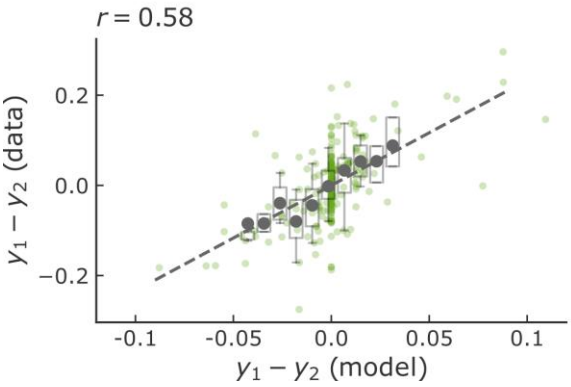


Figure 43. Good agreement between the experimental results and the model results. Experimentally measured versus theoretically predicted $y_1 - y_2$ is plotted for all common three nodes for all pairs of seeds for the 8 networks tested experimentally. Green solid dots are the data points. Data points were also binned in 25 bins linearly spaced according to their x value. Bins with less than 10 points are discarded. For each bin, a boxplot is superimposed. The box extends from the lower to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5th percentile to the 95th. Flier points are those past the end of the whiskers. Dotted grey line is a linear regression line.

3.3.8 Discussion

Even though collective autocatalytic sets are often envisioned as plausible candidates for the first self-sustaining systems in the prebiotic evolution processes, there is no experimental proof of their evolvability even though it is supported by a strong body of theoretical studies. One of the main pre-requisite for a CAS to be evolvable (Vasas et al. 2012) is 'multi-stability', i.e. it must

contain several stable cores which should be heritable, therefore transmit their dominance over to the next generation. Having enough variations in the initial states, several stable final states with selection can lead to Darwinian evolution. Variation can be implemented via several ways, and simplest would involve the environment, e.g. either by changing composition of the food set or the physico-chemical conditions. However, empirical demonstration for the existence of stable heritable cores in CASs system is not trivial. It requires the evaluation of network structural parameters governing the memory of the initial conditions among other things. Here we identified these parameters and quantified their effect computationally in a model experimental RNA-based CAS system derived from *Azoarcus* ribozymes. We showed that only a restricted set of parameters controls memory of initial conditions in a network. We evaluated this by comparing the evolution of the relative concentrations of the nodes in 5-member networks when one of the nodes is spiked in as seed. Notably one important parameter despite the overall density of links in the network is the 'uniqueness' of the seeded node. As a result a network with a high diversity of specific catalysts, i.e. a high diversity of IGS will have a long lasting memory of the initial conditions. Interestingly there are many such networks in the *Azoarcus* system. The model we used was chosen to be as close as possible to the experimental settings and we further demonstrated its validity by experimentally seeding 8 different networks. The differences of relative abundance of the nodes when two different nodes are seeded closely matched the expectations from the model showing that the model on the set of structural parameters influencing network memory are valid. This could pave the way for evolution experiments using 'serial transfer' format where the networks can be specifically designed to retain memory of the initial conditions and carried over to the next generations. This would allow to experimentally demonstrate heredity or at least 'limited heredity' (Jablonka & Szathmáry 1995) in CAS systems.

3.4. Materials and methods

3.4.1 General material and methods

DNase/RNase free water was purchased from Thermo Fisher Scientific (UltraPure™ DNAse/RNase Free distilled water, Product No.: 10977035) and was used for all the experiments. DNA oligonucleotides were obtained from IDT DNA technologies (https://eu.idtdna.com) unless specified otherwise. Table 9 contains the list of all oligonucleotides used in this study. RNA concentrations were measured using Qubit® RNA HS Assay Kit (Thermo Scientific, Product No.: Q32852). DNA concentrations were measured on a NanoDrop-1000 UV-spectrophotometer (Peqlab) unless specified otherwise.

3.4.2 RNA *in vitro* transcription

WXY fragments were transcribed with the same protocol as described earlier (Arsène et al. 2018). Z fragment was purchased from IDT DNA technologies (https://eu.idtdna.com). The DNA templates used to in vitro transcribe WXYZ were prepared by extension overlap PCR. For each full-length ribozyme, a combinations of four primers (for example for AA ribozyme, Oligo 1, 6, 10 and 13 were used), summarized in Table 9, were mixed at a final concentration of 2 µM along with 1X High-Fidelity buffer (Thermo Fischer Scientific, Product No.: F518L), 0.2 mM dNTP (Thermo Fischer Scientific, Product No.: R0192) and 0.02 U/µL of Phusion (Thermo Fischer Scientific, Product No.: F530S) and thermocycled as follows: step 1: 98°C/30 sec, step 2: 98°C/10 sec, step 3: 50°C/30 sec, step 4: 72°C/30 sec with one additional cycle from step 2 to 4 and final extension at 72°C/5 min. 2 µL from each sample were then used as template for another PCR with the same conditions as before except that the forward primer, Oligo 14-17 (Table 9) and the reverse primer Oligo 18 (Table 9), were put at 0.5 µM and that the annealing step was performed at 25°C and that 25 cycles of thermocycling were performed in total. The PCR products were run on a 2% agarose gel (run under standard electrophoresis conditions; 1X TAE, 110 V, 40 min) and the band of interest was excised and purified with NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel, Product No.: 740609.250) following the manufacturer's instructions. Purified

products were then diluted to 25 pg/µL and were then used as template for a PCR with forward primer Oligo 14-17 (Table 9) and reverse primer Oligo 18 (Table 9) with the same conditions to produce the dsDNA template used for *in vitro* transcriptions.

3.4.3 *Azoarcus* seeded reaction protocol

For Azoarcus seeded reactions, the WXY fragment corresponding to each of the five nodes was mixed at a final concentration of 0.1 µM with the Z fragment at a final concentration of 0.5 µM and with the WXYZ full-length catalyst corresponding to the seeded node at a final concentration of 0.05 µM. All RNAs were folded before the reaction by heating at 80°C for 3 min and by cooling down to 20°C at a rate of 0.1°C/sec. 4X Azoarcus reaction buffer (30 mM of EPPS (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer pH 7.4, 20 mM MgCl2) was then added to a final concentration of 1X and the mixture was incubated at 48°C for 30 minutes.

3.4.4 Sample preparation for sequencing

After the *Azoarcus* seeded reaction, 1 µL were taken out to be used as template for a reverse-transcription reaction (RT). RT was performed in a total volume of 10 µL with 0.1 µM of primer Oligo 23 (Table 9) with 1X First-Strand buffer (Thermo Fischer Scientific, Product No.: 18080093), 50 mM of DTT (Thermo Fischer Scientific, Product No.: P2325), 0.2 mM of each dNTP (Thermo Fischer Scientific, Product No.: R0192) and 10 U/µL of Superscript III (Thermo Fischer Scientific, Product No.: 18080093). RT was incubated for an hour at 60°C followed by a heat-inactivation step of 15 min at 75°C. The resulting cDNAs were purified over 1.2 equivalent of magnetic beads (AMPure XP, Beckman Coulter, Product No.: A63881) following manufacturer's protocol and diluted 5,000. 2 µl of diluted cDNAs were used in order to add the sequencing adaptors using two sequential PCR steps whose conditions are described in 2.4. Materials and methods. For the first PCR primers Oligo 24 (Table 9) and Oligo 25 (Table 9) were used while for the second PCR, primers Oligo 26 (Table 9) and Oligo 27 (Table 9) were used. PCR products were purified using AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) following manufacturer's protocol. The

final library was analyzed on TapeStation (Agilent 2200 TapeStation, using high sensitivity D1000 ScreenTape®, Product No.: 5067-5584) and quantified by Qubit® dsDNA HS Assay Kit (Thermo Scientific, Product No.: Q32854). Sequencing libraries were subjected to high-throughput sequencing using MiSeq V2micro-300 2*150 at Institut Curie, Paris.

3.4.5 Sequencing data processing

The final library was sequenced in paired-end mode with 180 bp devoted to read 1 and 120 devoted to read 2 and the methods described in 2.4. Materials and methods were used to associate three meta fields to each read: a sample barcode (used for multiplexing samples), the RNA molecule identity (either an *Azoarcus* ribozyme MN where M is the IGS and N is the tag or a harpin reporter) and a Unique Molecular Identifer (UMI) (Kivioja et al. 2012).

Reads with the same meta information were collapsed into a single row with an extra meta field with the number of reads that were collapsed. As described in 2.4. Materials and methods, this field was used to filter out noise arising from sequencing errors. For this a threshold of number of reads per UMI is fixed and whichever UMI has less reads than the threshold is discarded. The thresholds were chosen based on the shape of the distribution of number of reads (Figure 44) and are summarized in Table 8 for the 40 samples.

Filtered UMI-normalized rows are then grouped by sample barcode to count the number of UMI per RNA molecule type per sample. Counts for expected *Azoarcus* ribozyme MN are normalized to get relative abundance of each node.

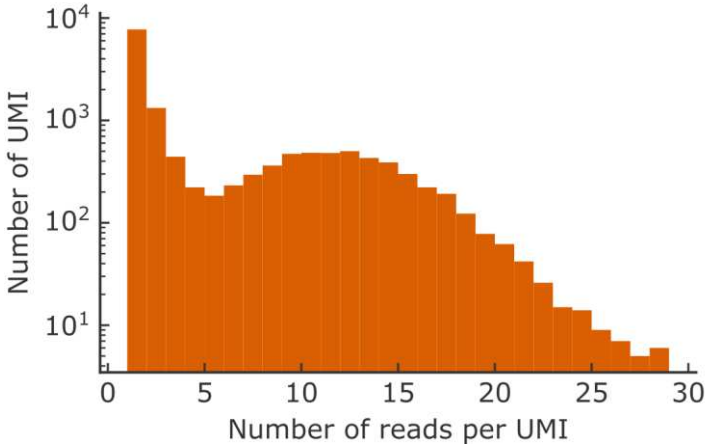


Figure 44. Distribution of number of reads per UMI for the sample with network AC, AU, UC, GC and GU with seed GU. Two overlapping distributions can be observed: the noise distribution with low number of reads per UMI arising from sequencing errors creating reads with new UMIs and the real distribution which is centred around 12 reads per UMI.

Table 8. Minimum number of reads per UMI per sample for a group of collapse reads not to be discarded. This number slightly depends on the sample because the two PCR performed to add the sequencing adaptors were done separately for each sample.

Sample	Network	Seeded node	Threshold value (min number of reads per UMI)
1	AC, AU, UC, GC, GU	AC	1
2	AC, AU, UC, GC, GU	AU	3
3	AC, AU, UC, GC, GU	UC	2
4	AC, AU, UC, GC, GU	GC	1
5	AC, AU, UC, GC, GU	GU	4
6	AG, CG, UG, GA, GG	AG	1
7	AG, CG, UG, GA, GG	CG	3
8	AG, CG, UG, GA, GG	UG	5
9	AG, CG, UG, GA, GG	GA	2
10	AG, CG, UG, GA, GG	GG	4
11	CA, CC, UA, UC, GA	CA	2
12	CA, CC, UA, UC, GA	CC	3
13	CA, CC, UA, UC, GA	UA	3
14	CA, CC, UA, UC, GA	UC	2
15	CA, CC, UA, UC, GA	GA	2
16	UA, UC, GC, GU, GG	UA	2
17	UA, UC, GC, GU, GG	UC	2
18	UA, UC, GC, GU, GG	GC	4

19	UA, UC, GC, GU, GG	GU	4
20	UA, UC, GC, GU, GG	GG	4
21	AA, CA, CC, CU, GU	AA	3
22	AA, CA, CC, CU, GU	CA	5
23	AA, CA, CC, CU, GU	CC	2
24	AA, CA, CC, CU, GU	CU	4
25	AA, CA, CC, CU, GU	GU	3
26	AU, CG, UC, UU, UG	AU	1
27	AU, CG, UC, UU, UG	CG	3
28	AU, CG, UC, UU, UG	UC	2
29	AU, CG, UC, UU, UG	UU	3
30	AU, CG, UC, UU, UG	UG	4
31	AC, AU, CG, UC, GG	AC	1
32	AC, AU, CG, UC, GG	AU	2
33	AC, AU, CG, UC, GG	CG	3
34	AC, AU, CG, UC, GG	UC	2
35	AC, AU, CG, UC, GG	GG	4
36	AU, CG, UU, UG, GU	AU	2
37	AU, CG, UU, UG, GU	CG	3
38	AU, CG, UU, UG, GU	UU	3
39	AU, CG, UU, UG, GU	UG	2
40	AU, CG, UU, UG, GU	GU	2

Table 9. List of oligonucleotides used in this chapter

Name	Sequence	Description
Oligo 1	gagccttgcgccgggaaaccacgcaagggatggtgtcaaattcggc gaaacctaagcgcccgcccgggcgtatggcaacgccgagccaagct tcggcgcctgcgccgatgaaggtgtagagactagacggca	First oligonucleotide adding ‘A’ as IGS with corresponding mutations at the 7th and 25th positions for extension PCR to assemble template for transcription of WXYZ catalyst
Oligo 2	gcgcctggcgccgggaaaccacgccagggatggtgtcaaattcggc gaaacctaagcgcccgcccgggcgtatggcaacgccgagccaagct tcggcgcctgcgccgatgaaggtgtagagactagacggca	First oligonucleotide adding ‘C’ as IGS with corresponding mutations at the 7th and 25th positions for extension PCR to assemble template for transcription of WXYZ catalyst
Oligo 3	gtgcctagcgccgggaaaccacgctagggatggtgtcaaattcggcg aaacctaagcgcccgcccgggcgtatggcaacgccgagccaagctt cggcgcctgcgccgatgaaggtgtagagactagacggca	First oligonucleotide adding ‘U’ as IGS with corresponding mutations at the 7th and 25th positions for extension PCR to assemble template for transcription of WXYZ catalyst
Oligo 4	gggcctcgcgccgggaaaccacgcgagggatggtgtcaaattcggc gaaacctaagcgcccgcccgggcgtatggcaacgccgagccaagct tcggcgcctgcgccgatgaaggtgtagagactagacggca	First oligonucleotide adding ‘G’ as IGS with corresponding mutations at the 7th and 25th positions for extension PCR to assemble template for transcription of WXYZ catalyst

Oligo 5	actatgccttcaccatagcgatgcaagtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘U’ as tag with a G insertion before the tag
Oligo 6	actatgccttcaccatagcgatgcatgtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘A’ as tag with a G insertion before the tag
Oligo 7	actatgccttcaccatagcgatgcaggtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘C’ as tag with a G insertion before the tag
Oligo 8	actatgccttcaccatagcgatgcacgtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘G’ as tag with a G insertion before the tag
Oligo 9	actatgccttcaccatagcgatgaagtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘U’ as tag without insertion
Oligo 10	actatgccttcaccatagcgatgatgtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘A’ as tag without insertion
Oligo 11	actatgccttcaccatagcgatgaggtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘C’ as tag without insertion
Oligo 12	actatgccttcaccatagcgatgacgtgccttaggtgggtgccgtctagtctctacacc	Second oligonucleotide for assembling WXYZ template adding ‘G’ as tag without insertion
Oligo 13	cgctatggtgaaggcatagtccagggagtggcgaaagtcacacaaaccgg	Third oligonucleotide for assembling WXYZ template common to all MN combination
Oligo 14	ctgcagaattctaatacgactcactatagagccttgcgccgggaaaccacgcaagggatgg	Forward primer to have 'GAG' as IGS with 'A' at the 25th and 'U' at the 7th position
Oligo 15	ctgcagaattctaatacgactcactatagcgcctggcgccgggaaaccacgccagggatgg	Forward primer to have 'GCG' as IGS with 'C' at the 25th and 'G' at the 7th position
Oligo 16	ctgcagaattctaatacgactcactatagtgcctagcgccgggaaaccacgctagggatgg	Forward primer to have 'GUG' as IGS with 'U' at the 25th and 'A' at the 7th position
Oligo 17	ctgcagaattctaatacgactcactatagggcctcgcgccgggaaaccacgcgagggatgg	Forward primer to have 'GGG' as IGS with 'G' at the 25th and 'C' at the 7th position

Oligo 18	ccggtttgtgtgactttcgcc	Reverse primer targeting 3’ end of Z fragment
Oligo 19	atgtgccttaggtgggtgc	Reverse primer to add 'A' tag in WXY fragment
Oligo 20	aggtgccttaggtgggtgc	Reverse primer to add 'C' tag in WXY fragment
Oligo 21	aagtgccttaggtgggtgc	Reverse primer to add 'U' tag in WXY fragment
Oligo 22	acgtgccttaggtgggtgc	Reverse primer to add 'G' tag in WXY fragment
Oligo 23	gtgactggagttcagacgtgtgctcttccgatctcagactcactcaca acggcttgacggctcatacggtaactaccacaaatccctcgcatgac aactacgctacggaacgannnnnnnnccggtttgtgtgactttcgc cactc	Primer used for reverse-transcription (RT) introducing a UMI of 8 base pairs N
Oligo 24	ttccctacacgacgctcttccgatct-sample barcode (6 nt)-gagccttgcgccgggaaacca	Illumina sequencing primer I (forward)
Oligo 25	gtgactggagttcagacgtgtgctcttccg	Illumina sequencing primer I (reverse)
Oligo 26	aatgatacggcgaccaccgagatctacactctttccctacacgacgct cttcc	Illumina sequencing primer II (forward, Rd1)
Oligo 27	caagcagaagacggcatacgagat gtgactggagttcagacgtgtg ctcttccgatct	Illumina sequencing primer II (reverse, Rd2)

4. Coupled catabolism and anabolism in autocatalytic RNA sets

Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky598

Coupled catabolism and anabolism in autocatalytic RNA sets

Simon Arsène[1,†], Sandeep Ameta[1,†], Niles Lehman[2], Andrew D. Griffiths[1,*] and Philippe Nghe[1,*]

[1]Laboratoire de Biochimie, École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI Paris), CNRS UMR 8231 Chimie Biologie Innovation, PSL Research University, 10 rue Vauquelin, 75005 Paris, France and [2]Department of Chemistry, Portland State University, P.O. Box 751, Portland, OR 97207, USA

Received March 29, 2018; Revised June 15, 2018; Editorial Decision June 20, 2018; Accepted June 22, 2018

ABSTRACT

The ability to process molecules available in the environment into useable building blocks characterizes catabolism in contemporary cells and was probably critical for the initiation of life. Here we show that a catabolic process in collectively autocatalytic sets of RNAs allows diversified substrates to be assimilated. We modify fragments of the *Azoarcus* group I intron and find that the system is able to restore the original native fragments by a multi-step reaction pathway. This allows in turn the formation of catalysts by an anabolic process, eventually leading to the accumulation of ribozymes. These results demonstrate that rudimentary self-reproducing RNA systems based on recombination possess an inherent capacity to assimilate an expanded repertoire of chemical resources and suggest that coupled catabolism and anabolism could have arisen at a very early stage in primordial living systems.

INTRODUCTION

Collective autocatalytic sets (CASs) (1–3), where an ensemble of molecules can reproduce each other, have been envisaged as a possible scenario for the origin of life (4–10). A fundamental feature of such sets is their ability to self-sustain using substrates available in the environment (the food set) (6,10,11), a property which has reached a very high level of complexity and diversity in contemporary metabolisms. In the context of origin of life and the RNA world, where directly useable substrates were limited (12–17), it would have been advantageous for a self-reproducing system to thrive on a broad range of resources by pre-processing them (Figure 1A).

In this regard, RNAs derived from the group I intron (18) of the *Azoarcus* bacterium (19) are an attractive model, as they can self-reproduce and recycle RNA materials by recombination reactions (17). They are 200 nt long RNA recombinases (WXYZ, Figure 1B) which can catalyze their own assembly from the fragments WXY + Z via transesterification in an autocatalytic process (20). For the assembly, they exploit Watson–Crick interactions between the 3 nt at both extremities of WXY fragments (the internal guide sequence 'IGS' and the target sequence 'tag' at the 5' and 3' ends, respectively). The IGS and tag of the fragments can be engineered to form autocatalytic recombination networks (21). However, the autocatalytic character of this system has so far only been demonstrated in a purely anabolic manner, relying on designed substrates obtained by fragmentation of the ribozyme (22). In a more realistic prebiotic setting, the initial reaction mixtures would likely consist of a much broader range of molecules. This could cause self-reproduction of CASs to be inhibited or stopped for several reasons: (i) the available molecules cannot be used as substrates by the catalysts; (ii) available molecules can be used as substrates, but lead to futile products that are not capable of catalysis; (iii) the available molecules are assembled to form novel catalysts, but these catalysts do not allow formation of an autocatalytic cycle.

Here, we mimic a prebiotic environment containing substrates that cannot be used directly to form autocatalytic ribozymes by fueling the *Azoarcus* ribozyme system with only modified RNA substrates. We observe that unmodified fragments are, nevertheless, reformed, leading to the production of wild-type catalysts (WXYZ). Kinetic and biochemical analyses show that the transformation of the raw material is catalyzed by the reaction products via a multi-step reaction pathway involving a series of specific but unexpected binding interactions between the RNA substrates and the ribozyme. The combination of catabolic and anabolic steps enables collective autocatalysis. We furthermore show that CASs comprising multiple species are maintained in similar conditions. These results highlight a form of rudimentary catabolism, where the catalysts transform the available

*To whom correspondence should be addressed. Philippe Nghe: Tel: +33 140794586; Fax: +33 140794776; Email: philippe.nghe@espci.fr, Andrew D. Griffiths: Tel: +33 140794539; Fax: +33 140794776; Email: andrew.griffiths@espci.fr
†The first two authors should be regarded as joint First Authors.

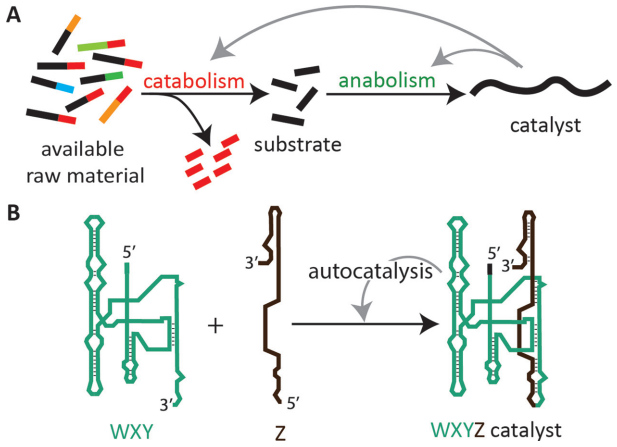


Figure 1. Coupled catabolism and anabolism in self-reproducing systems. (**A**) Schematic showing how catabolism in a self-reproducing system can process unusable raw material into usable substrates. (**B**) Autocatalytic synthesis of covalent RNA catalyst (WXYZ) from inactive RNA substrates (WXY and Z) using an anabolic autocatalytic process in the *Azoarcus* ribozyme system.

resources into building blocks that drive their own formation.

MATERIALS AND METHODS

Materials

All chemicals were purchased from Sigma-Aldrich (unless specified otherwise). 4-(2-Hydroxyethyl)-1-piperazinepropanesulfonic acid, EPPS was purchased from Alfa Aesar (Product no.: J60511, CAS no.: 16052-06-5). For all the reactions water was used from ThermoFisher Scientific (UltraPureTM DNAse/RNase free, Product no.: 10977035) or from the MilliQ water purifier system (Millipore). RNA concentrations were measured on a NanoDrop-1000 UV-spectrophotometer (Peqlab). Denaturing polyacrylamide gels were prepared using gel stock solution from Roth and run in 1× TBE (Tris-Borate ethylenediaminetetraacetic acid (EDTA), prepared from 10× TBE from Roth). All analysis were performed using 12% denaturing polyacrylamide gels containing 8.3 M urea and run for at least 2–3 h at constant power of 24 W. Gels were stained with 1× GelRedTM (Biotium). Gel analysis and calculation of conversions were carried out with ImageJ software (https://imagej.nih.gov/ij/). All DNA oligonucleotides were obtained from IDT DNA technologies (https://eu.idtdna.com) and are described in Supplementary Table S1.

RNA preparation

dsDNA templates for *in vitro* transcription reactions were produced using standard PCR reactions. For PCR, ∼25 pg of plasmid bearing dsWXYZ sequence was mixed with 1× PCR buffer (ThermoFisher Scientific), 0.5 μM of each forward and reverse primer (Supplementary Table S1), 0.2 mM of each dNTP, 0.02 U/μl Hot Start Phusion polymerase (ThermoFisher Scientific, Product no.: F-549L) and thermocycled as follows: step 1: 98°C/30 s, step 2: 98°C/10 s, step 3: 57°C/30 s, step 4: 72°C/30 s with 24 additional cycles from step 2 to 4 and final extension at 72°C/5 min. The purity of the dsDNA was checked on a 2% agarose gel (stained with GelRed™, run under standard electrophoresis conditions; 1× Tris-acetate-EDTA (TAE), 110 V, 40 min). After PCR, amplified dsDNA templates were isopropanol precipitated, pellets were washed with 70% ethanol, dissolved in water and used directly for *in vitro* transcription. *In vitro* transcription reactions were performed at 100 μl scale with dsDNA mixed with 1× transcription buffer (ThermoFisher Scientific), 16 mM additional $MgCl_2$, 4 mM of each NTP and 10 U/μl of T7 RNA polymerase (ThermoFisher Scientific, Product no.: EP0111) and incubated at 37°C for 4 h. The reaction was stopped by addition of gel loading buffer (70% formamide containing 0.1% of each xylene cyanol and bromophenol blue) and purified on 12% denaturing polyacrylamide gels using standard electrophoresis conditions (1× TBE buffer, run at 24 W for 2–3 h). Transcript bands were excised and eluted in 0.3 M Na-Acetate pH 5.5 overnight at room temperature. The eluted solution was isopropanol precipitated, washed with 70% Ethanol, dissolved in water and concentrations were measured.

Autocatalytic trans-esterification reaction

For the autocatalytic trans-esterification reactions RNA substrates were mixed in water to give a final concentration of 0.5 μM for each RNA. For the cooperative network formation each of the three WXY substrates were mixed to give a final concentration of 0.5 μM for each of the three RNAs (1.5 μM total) and the other substrate (Z or Z-mod) was added to a final concentration of 1.5 μM (to give 1:1 stoichiometry of WXY with Z or Z-mod). To fold the RNA, the mixture was heated at 80°C for 3 min and gradually cooled down to 20°C (at a rate of 0.1°C/s). Then 30 mM of EPPS (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer pH 7.4 and 100 mM $MgCl_2$ was added and the reaction was incubated at 48°C. For gel-based kinetic analysis, at each time point 2 μl of reaction was taken out and mixed with gel loading buffer containing 70% formamide, 0.01% of each xylene cyanol and bromophenol blue, and ∼2 equivalents of EDTA (ethylenediaminetetraacetic acid) and analyzed on denaturing polyacrylamide gels.

Sanger sequencing

To confirm the identity of the different products formed with different substrate combinations (Figure 2; c2, c3, c4) the sequences were analyzed by Sanger sequencing. For each reaction, the expected size product band was excised from polyacrylamide gel, RNA were eluted in 0.3 M sodium-acetate pH 5.5 overnight and isopropanol precipitated. The pellets were washed three-times with 70% ethanol and re-suspended in 30 μl water. A total of 5 μl was used for reverse-transcription (RT) using 2.5 μM of primer complementary to the 3' end of the Z fragment (primer 11, Supplementary Table S1). For the sequencing of WXY-mod, RNA was additionally tailed with poly-guanosine (23) by *Escherichia coli* polyA polymerase (New England Biolabs, Product no.: M0276L) using 2 mM guanosine triphosphate (GTP) and the standard protocol.

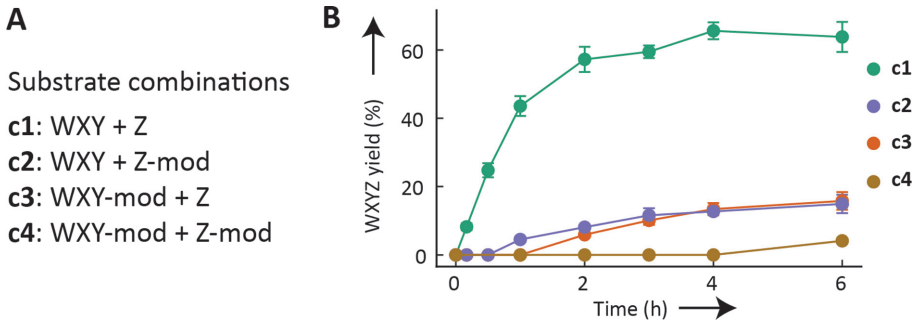


Figure 2. Synthesis of WXYZ catalyst with modified RNA substrates. (**A**) Different substrate combinations used to study the catabolic properties of the *Azoarcus* ribozyme system. (**B**) Kinetics of the synthesis of WXYZ using all four substrate combinations (c1 to c4). The reported yield is the substrate to product conversion in percentage. Error bars represent ±1 standard deviation (triplicates).

For the RT of poly-guanosine tailed samples, primer 24 was used and for PCR amplification primer 25 and primer 15 were used as forward and reverse primers, respectively. The RT reaction (20 μl total volume) was performed in 1× first-strand buffer (250 mM Tris–HCl (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$), 0.2 mM dNTPs, 5 mM DTT and 10 U/μl of enzyme (SuperScript® III Reverse Transcriptase; ThermoFisher Scientific; Product no.: 18080093). The reaction was incubated at 55°C for 1 h and samples were then purified using 1.2 equivalent of AMPure XP magnetic beads (Beckman Coulter, Product no.: A63881) following the manufacturer's protocol. A total of 10 μl of purified RT products were then PCR amplified using the same protocol as for *in vitro* transcription using primer 1 as the forward primer and primer 11 as the reverse primer (Supplementary Table S1). Polymerase chain reaction (PCR) products were purified using AMPure XP magnetic beads. Approximately 20 ng of purified PCR products were ligated into pJET1.2/blunt vector using CloneJET PCR Cloning Kit (ThermoFisher Scientific; Product no.: K1232). A total of 5 μl of ligation mix was mixed with 50 μl of MegaX DH10B™ T1R Electrocomp™ Cells (ThermoFisher Scientific; Product No.: C640003), kept for 2 min on ice, then at 42°C for 1 min 30 s and then put back on ice. A total of 1 ml of LB medium was added and cells were grown for 1 h at 37°C before plating on a LB agar plate supplemented with 100 μg/ml ampicillin and plates were incubated at 37°C overnight. The positive colonies were checked by colony PCR: picked colonies were added into 3 μl water from which half was used for colony PCR and other half to start a 5 ml overnight culture at 37°C in LB medium for positive colonies. Plasmids were extracted from overnight cultures using NucleoSpin® Plasmid kit (Macherey-Nagel; Product no.: 740588.250) and sequenced using the GATC 'Light-run' Sanger sequencing service.

Kinetic modeling

The kinetic model used here was developed using a similar approach as described previously (22). From a small set of assumptions (covalent bond formation is reversible and catalysts can be covalent ribozymes or non-convalent complexes between the two substrate fragments), we derived a set of 14 chemical reactions (Supplementary Table S2) for the complete model. We also constructed reduced models based on additional assumptions and computed a set of model selection criteria (Supplementary Table S3) to investigate whether the number of rate constants could be reduced. Our approach is described in detail in the Supplementary Text. For all models, the reaction rates were fitted using Scipy (24) nonlinear least-squares problem solving routine with 'trf' method (Trust Region Reflective algorithm) and 'cauchy' loss function (results are similar with 'linear' loss function). Reactions and fitted rates are summarized in Supplementary Table S2.

Network genotyping

In order to obtain the genotypic distribution, the reaction products WXYZ ribozymes were sequenced by Illumina high-throughput sequencing. RNA samples after 6 h of reaction were reverse transcribed using a specific primer (Primer 13, see Supplementary Table S1) and PCR amplified following the protocol above (see Sanger sequencing section). PCR was performed in two steps to sequentially add the Illumina sequencing adaptors. For the first step, we used primer 14 and primer 15 and amplified the samples for total of 18 cycles. After the first step, samples were purified using AMPure XP magnetic beads (Beckman Coulter, Product no.: A63881) and subjected to an additional 10 cycles of amplification with primer 16 and primer 17. After this, the final amplicons were purified with AMPure XP magnetic beads (Beckman Coulter, Product no.: A63881), quantified using Quibit (Qubit® 3.0 Fluorometer) and subjected to 2 × 150 paired-end nanoMiseq (Institut Curie High Throughput Sequencing platform, Paris). The relative number of Unique Molecular Identifiers (UMIs) (25) for each genotype is de-convoluted using a custom software pipeline developed in house.

RESULTS

Effect of substrate modification on the formation of *Azoarcus* ribozymes

To investigate the impact of substrate modifications on the *Azoarcus* system, we modified the native substrates WXY and Z into WXY-mod and Z-mod, respectively, by appending a foreign sequence at the 3' end (mod): a poly-adenosine stretch preceded by 'agugcc' (Supplementary Figure S1). These modifications are relevant in the context of prebiotic

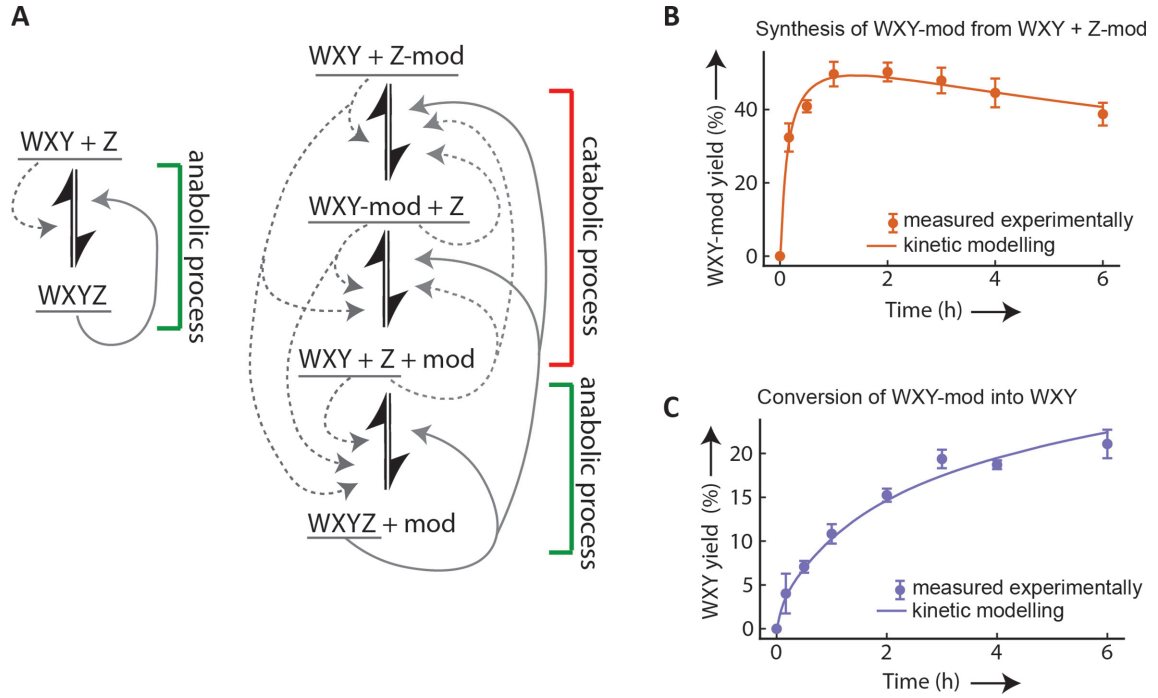


Figure 3. Multi-step reaction pathway to catabolize the starting material. (**A**) Diagrammatic representation of the catabolic steps involved in the synthesis of substrates WXY and Z from the modified RNAs (WXY-mod and Z-mod) to produce WXYZ. Dotted and continuous gray arrow show feed-back by non-covalent and covalent catalyst, respectively. (**B**) Kinetic analysis of the formation of WXY-mod with substrate combination c2. (**C**) Kinetic analysis of the formation of WXY from WXY-mod with substrate combination c3. For both graphs (**B**) and (**C**), the circles represent experimental data obtained by polyacrylamide gel electrophoresis and the lines represents the data obtained from kinetic modeling (Supplementary Text). The reported yield is the substrate to product conversion in percentage. Error bars represent ±1 standard deviation (triplicates).

Earth where any random ligation, recombination or chemical addition could result in the appendage of a stretch of oligonucleotides. Homopolymers are plausible candidates for inert appendages as they cannot form stable folded structures, and spontaneous synthesis of long nucleic acids has so far been shown to be biased toward such homopolymers (26,27). We analyzed the effect of different combinations (Figure 2A) of these modified RNA substrates on the kinetics of self-assembly reactions (Figure 2B). Remarkably, the modifications do not block the reaction completely, as we observe the appearance of measurable amounts of WXYZ catalyst in each case (Figure 2B and Supplementary Figure S2). However, the modification drastically slows down the reaction (Figure 2B; compare c1 with c2, c3 or c4). When only one substrate is modified (c2 and c3), WXYZ synthesis starts rather slowly and reaches ∼15% of initial material after 6 h of reaction as opposed to ∼60% with non-modified substrates (c1). When both substrates are modified the WXYZ production drops to only ∼4% after 6 h (c4).

Additionally, we investigated other modifications of the substrate by changing the tail from poly-adenosine to poly-guanosine or poly-uridine and by changing the sequence preceding the tail (Supplementary Figure S3). In all these cases, we observed the formation of WXYZ catalysts.

These results highlight the inherent capacity of the *Azoarcus* ribozyme system to overcome the modification burden by somehow processing the available raw material. The simplest scenario to purge the modification would involve direct cleavage, as for example proposed in an earlier theoretical study (28), of mod part from the substrates WXY-mod and Z-mod. However, we do not observe direct cleavage of mod from Z-mod, even in the presence of WXYZ (Supplementary Figure S4). Instead mod is transferred from Z-mod to WXY to generate WXY-mod (Supplementary Figure S5). Conversely, transfer of mod from WXY-mod to Z is not detected (Supplementary Figure S6). Ultimately, the modification is slowly cleaved from WXY-mod, catalyzed by WXYZ, generating unmodified WXY (Supplementary Figure S7). Note that while the experiments were routinely performed at 100 mM $MgCl_2$, we observed that these reaction steps lead to measurable amounts of products even down to 10 mM $MgCl_2$ (Supplementary Figure S8).

Multi-step reaction pathway and kinetic modeling

These observations suggest a reaction path (Figure 3A) where substrate composition c2 is converted into c3 by transfer of mod from Z-mod to WXY, followed by conversion into the canonical fragments (substrate combination c1), by cleavage of mod from WXY-mod which then react to generate WXYZ via the autocatalytic anabolic reaction described by Hayden *et al.* (22). The kinetic traces of the proposed intermediates WXY-mod and WXY are consistent with this reaction pathway: we observe the formation of WXY-mod, which peaks at almost 50% conversion within an hour, followed by a slow decrease (Figure 3B). Starting from combination c3 leads to conversion from WXY-mod to WXY which happens on slower timescales, with <10% of

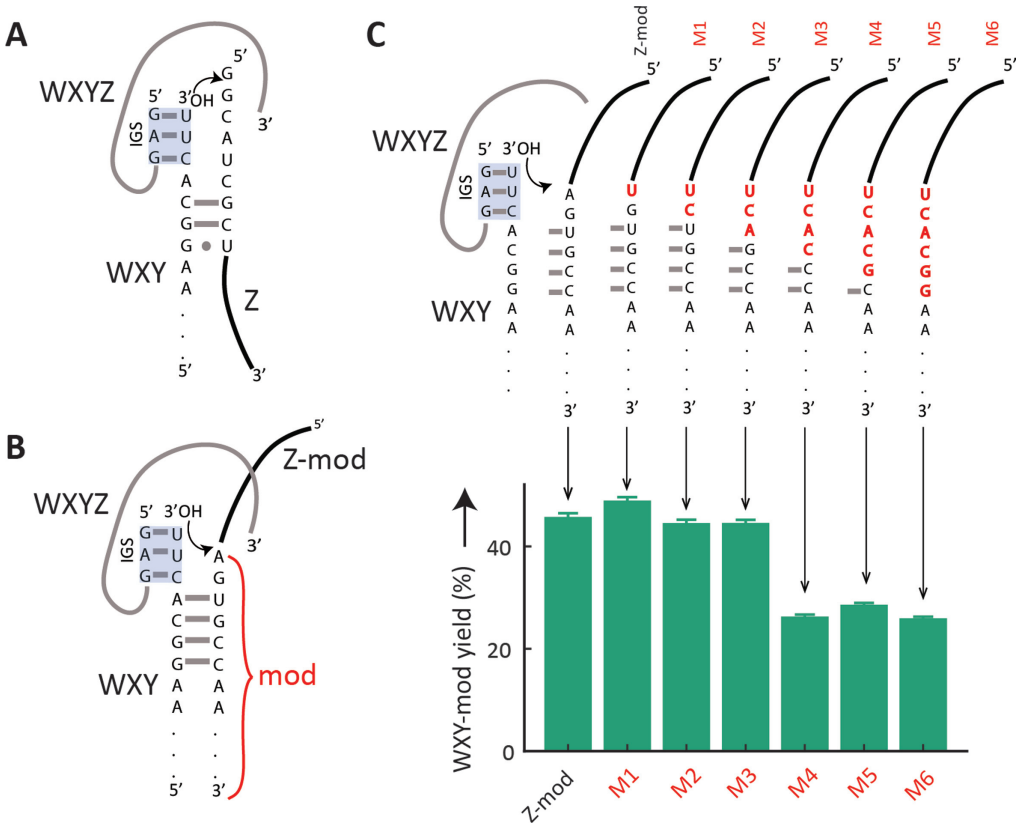


Figure 4. Proposed binding of Z and Z-mod and effect of mutations in Z-mod. (**A**) Diagrammatic representation of the binding of Z to WXY in the case of non-modified substrates (20,29). (**B**) Proposed binding of Z-mod to WXY. (**C**) Sequence of mutated variants (M1-6, substitutions in red) of Z-mod and the yield of WXY-mod after 1 h reaction between WXY and non-mutated or mutated variants of Z-mod obtained by polyacrylamide gel electrophoresis. The reported yield is the substrate to product conversion in percentage. Error bars represent ±1 standard deviation (triplicates).

WXY-mod being converted to WXY within an hour (Figure 3C).

We then built a model to test whether the kinetics of species for different substrate combinations could be described by a single kinetic model of the entire reaction scheme. The model is based on the recombination mechanisms reported earlier for *Azoarcus* ribozymes (22) (Supplementary Table S2): non-covalent complexes of the substrates can catalyze recombination reactions at a slow rate (dashed gray arrows, Figure 3A) and covalent ribozymes at a higher rate (solid gray arrows, Figure 3A). We obtained a single set of parameters from the devised kinetic model by fitting to the experimental data (Figure 3B and C; Supplementary Figure S9). In particular, the model reproduces well the critical steps of the pathway: the rapid formation of WXY-mod and its slow consumption, as well as WXYZ formation. We further investigated the validity of the model by reducing the number of parameters and by computing different model selection criteria (Supplementary Figure S10 and Table S3).

Mechanisms of WXY-mod formation

Based on the previously characterized reaction mechanism (20,29) and the product sequences (Supplementary Figure S2), we hypothesized that the transfer of the modification from Z-mod to WXY is mediated by Watson–Crick base-pairing of Z-mod to the 3' end of WXY (Figure 4A). Whereas normally the 5' end of the unmodified substrate Z base-pairs with the 3' end of WXY, here, the region of Z-mod immediately 5' to the poly-adenosine sequence binds to the 3' end of WXY via four Watson–Crick base-pairs (Figure 4B). We probed the contribution of these nucleotides in the formation of WXY-mod by sequentially substituting them to disrupt the proposed Watson–Crick base-pairing and analyzed the amount of WXY-mod product formed (Figure 4C). Changing nucleotides that are not proposed to base-pair (mutants M1 and M2) as well as that proposed to form the first (A-U) base-pair (mutant M3) has no significant effect on WXY-mod formation. Additionally disrupting the second (G-C) base-pair (mutant M4) reduces the formation of WXY-mod by almost half (only ∼25% of WXY converted to WXY-mod in an hour). However, mutating all the other proposed base-pairing residues (mutants M5 and M6) does not reduce WXY-mod formation further, indicating that other interactions must be involved and could be important in the reaction mechanism.

Cooperative network is not hampered by modified substrates

Azoarcus ribozymes are known to form multi-species autocatalytic networks where the growth of a species is dictated by how well its formation is catalyzed by other species in the network (21). These networks are formed by base-pairing

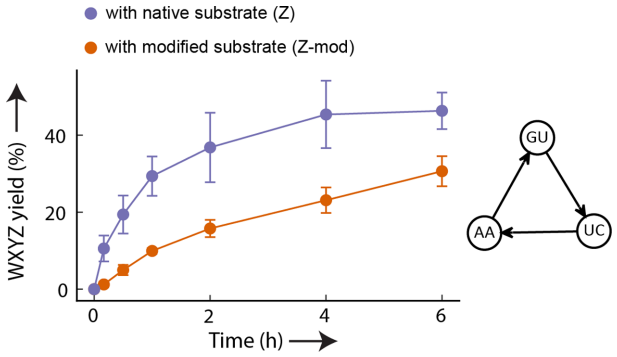


Figure 5. Formation of cooperative RNA network using modified substrates. Reactions were started by adding WXY fragments with different IGS-tag sequences (MN = AA, GU and UC, right panel) as used in previous studies (30,31). The reaction is either provided with native (Z, purple) or modified substrate (Z-mod, orange). Kinetic analysis was performed by analysing the formation of WXYZ at different time points using polyacrylamide gel electrophoresis. The reported yield is the substrate to product conversion in percentage. Error bars represent ±1 standard deviation (triplicates).

interactions between the IGS (GMG) and the tag (CNU) of different fragments, where M and N are variable nucleotides (A, C, U or G) in IGS-tag combination, denoted as MN (20,21). For example, IGS 'A' (GAG) from WXYZ recognizes tag 'U' (CUU) in WXY to catalyze its assembly with Z. Appropriate combinations of such interactions allows the formation of a closed catalytic cycle as shown in Figure 5, right panel. However, it is unclear whether such networks can function when fueled with modified substrates. We thus constructed a three-membered network (30,31) with either modified (Z-mod) or canonical (Z) substrate and analyzed the synthesis of the RNA catalyst (WXYZ) (Figure 5). Even when only modified substrates are used, the *Azoarcus* ribozyme network is still functional and the yield of WXYZ is only slightly lower than with unmodified substrates (~45% after 6 h with Z compared to ~30% with Z-mod). This shows that the catabolic steps do not interfere with the collective dynamics. High-throughput sequencing confirmed that not only the growth (amount of WXYZ formed), but also the relative distribution of the members of the network are maintained (Supplementary Figure S11). In particular, we found the same ordering of the members as reported in an earlier study with the same network of non-modified fragments (31), where the 'UC' ribozyme is at higher concentration than the other two ribozymes ('AA' and 'GU').

DISCUSSION

The work presented here demonstrates how catabolic steps in collectively autocatalytic sets can expand the range of usable substrates, which would have been highly advantageous in a heterogeneous prebiotic milieu. The assimilation of modified substrates here is enabled by the recombination activity of *Azoarcus* ribozymes, and involves a multi-step reaction pathway, which contrasts with other experimental autocatalytic systems (32–34). We have furthermore shown that a cooperative network made of several species and fueled with modified substrates is almost as efficient at synthesizing catalysts as with unmodified substrates and maintains the relative distribution of the members. Although RNA fragments used here are longer than what is usually imagined in prebiotic scenarios, it should be noted that *Azoarcus* ribozymes can be formed from much shorter RNA fragments (<50 nt) (21,35).

Similar catabolic processes may be transposed to smaller molecules CAS (e.g. the formose reaction, where sugars are formed from formaldehyde (36) as soon as the catalysts formed by the autocatalytic reactions are able to reshuffle chemicals. However it remains difficult to elaborate such scenarios at present, given the scarcity of experimental models.

In addition to the basic ingredients of fragment recycling and autocatalysis, the recovery of the unmodified catalytic sequences indicates an inherent form of selection at the molecular level. This selection may be related to the differential protection of functional and parasitic folds from recombination, as envisaged earlier (4). This idea is reminiscent of the concepts of dynamical combinatorial chemistry (37), where self-assembly is replaced by covalent recombination, and differential product stability may arise from thermodynamic as well as from kinetic factors. Interestingly, these ingredients are applied here to an autocatalytic system and suggest a mechanism to limit the extinction of early metabolic cycles by side reactions highlighted by Orgel (38).

Coupled catabolism and anabolism is a universal feature of contemporary living systems: diverse complex molecules are broken down into simpler building blocks that are used to construct new biomolecules. Although the conditions and mechanisms observed in this study differ from those observed in contemporary cellular life, it is striking to witness that coupled catabolism and anabolism could have arisen in a system with a much lower level of complexity, at an early stage in the RNA world.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Estelle Mendes, Nympha Elisa Sia and Bilal Mazhar for their technical assistance in the realization of the experiments.

FUNDING

European Union Seventh Framework Program (FP7/2007-2013) [294332 (EvoEvo)]; PSL Research University (OCAV project); Ecole Polytechnique for PhD fellowship (AMX) to S. Arsène; Ecole Doctorale FdV (Programme Bettencourt). Funding for open access charge: European Union Seventh Framework Program (FP7/2007–2013) [294332 (EvoEvo)].
Conflict of interest statement. None declared.

REFERENCES

1. Eigen,M. (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, **58**, 465–523.
2. Eigen,M. and Schuster,P. (1977) The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften*, **64**, 541–565.

3. Kauffman,S.A. (1986) Autocatalytic sets of proteins. *J. Theor. Biol.*, **119**, 1–24.
4. Higgs,P.G. and Lehman,N. (2015) The RNA World: molecular cooperation at the origins of life. *Nat. Rev. Genet.*, **16**, 7–17.
5. Hordijk,W. and Steel,M. (2017) Chasing the tail: the emergence of autocatalytic networks. *Biosystems*, **152**, 1–10.
6. Hordijk,W., Steel,M. and Kauffman,S.A. (2012) The structure of autocatalytic sets: evolvability, enablement, and emergence. *Acta. Biotheor.*, **60**, 379–392.
7. Jain,S. and Krishna,S. (1998) Autocatalytic sets and the growth of complexity in an evolutionary model. *Phys. Rev. Lett.* **81**, 5684–5687.
8. Jain,S. and Krishna,S. (2001) A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 543–547.
9. Lee,D.H., Severin,K. and Ghadiri,M.R. (1997) Autocatalytic networks: the transition from molecular self-replication to molecular ecosystems. *Curr. Opin. Chem. Biol.*, **1**, 491–496.
10. Vasas,V., Fernando,C., Santos,M., Kauffman,S.A. and Szathmáry,E. (2012) Evolution before genes. *Biol. Direct*, **7**, 1–14.
11. Hordijk,W. and Steel,M. (2004) Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *J. Theor. Biol.*, **227**, 451–461.
12. Chen,X., Li,N. and Ellington,A.D. (2007) Ribozyme catalysis of metabolism in the RNA world. *Chem. Biodivers.*, **4**, 633–655.
13. Copley,S.D., Smith,E. and Morowitz,H.J. (2007) The origin of the RNA world: co-evolution of genes and metabolism. *Bioorg. Chem.*, **35**, 430–443.
14. King,G.A. (1982) Recycling, reproduction, and life's origins. *Biosystems*, **15**, 89–97.
15. Orgel,L.E. (2004) Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.*, **39**, 99–123.
16. Robertson,M.P. and Joyce,G.F. (2012) The origins of the RNA world. *Cold Spring Harb. Perspect. Biol.*, **4**, 1–23.
17. Vaidya,N., Walker,S.I. and Lehman,N. (2013) Recycling of informational units leads to selection of replicators in a prebiotic soup. *Chem. Biol.*, **20**, 241–252.
18. Cech,T.R. (1990) Self-splicing of group I introns. *Annu. Rev. Biochem.*, **59**, 543–568.
19. Reinhold-Hurek,B. and Shub,D.A. (1992) Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature*, **357**, 173–176.
20. Draper,W.E., Hayden,E.J. and Lehman,N. (2008) Mechanisms of covalent self-assembly of the *Azoarcus* ribozyme from four fragment oligonucleotides. *Nucleic Acids Res.*, **36**, 520–531.
21. Vaidya,N., Manapat,M.L., Chen,I.A., Xulvi-Brunet,R., Hayden,E.J. and Lehman,N. (2012) Spontaneous network formation among cooperative RNA replicators. *Nature*, **491**, 72–77.
22. Hayden,E.J., von Kiedrowski,G. and Lehman,N. (2008) Systems chemistry on ribozyme self-construction: evidence for anabolic autocatalysis in a recombination network. *Angew. Chem. Int. Ed. Engl.*, **47**, 8424–8428.
23. Yehudai-Resheff,S. and Schuster,G. (2000) Characterization of the E. coli poly(A) polymerase: nucleotide specificity, RNA-binding affinities and RNA structure dependence. *Nucleic Acids Res.*, **28**, 1139–1144.
24. Oliphant,T.E. (2001) Python for Scientific Computing. *Comput. Sci. Eng.*, **9**, 10–20.
25. Kivioja,T., Vaharautio,A., Karlsson,K., Bonke,M., Enge,M., Linnarsson,S. and Taipale,J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.
26. Huang,W. and Ferris,J.P. (2003) Synthesis of 35–40 mers of RNA oligomers from unblocked monomers. A simple approach to the RNA world. *Chem. Commun.*, **0**, 1458–1459.
27. Huang,W. and Ferris,J.P. (2006) One-step, regioselective synthesis of up to 50-mers of RNA oligomers by montmorillonite catalysis. *J. Am. Chem. Soc.*, **128**, 8914–8919.
28. Hordijk,W. and Steel,M. (2106) Autocatalytic sets in polymer networks with variable catalysis distributions. *J. Math. Chem.*, **54**, 1997–2021.
29. Adams,P.L., Stahley,M.R., Kosek,A.B., Wang,J. and Strobel,S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45–50.
30. Yeates,J.A.M., Hilbe,C., Zwick,M., Nowak,M.A. and Lehman,N. (2016) Dynamics of prebiotic RNA reproduction illuminated by chemical game theory. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 5030–5035.
31. Yeates,J.A.M., Nghe,P. and Lehman,N. (2017) Topological and thermodynamic factors that influence the evolution of small networks of catalytic RNA species. *RNA*, **23**, 1088–1096.
32. Ashkenasy,G., Jagasia,R., Yadav,M. and Ghadiri,M.R. (2004) Design of a directed molecular network. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 10872–10877.
33. Kim,D.E. and Joyce,G.F. (2004) Cross-catalytic replication of an RNA ligase ribozyme. *Chem. Biol.*, **11**, 1505–1512.
34. Sievers,D. and von Kiedrowski,G. (1994) Self-replication of complementary nucleotide-based oligomers. *Nature*, **369**, 221–224.
35. Jayathilaka,T.S. and Lehman,N. (2018) Spontaneous covalent self-assembly of the *Azoarcus* ribozyme from five fragments. *Chembiochem*, **19**, 217–220.
36. Butlerov,A.M. (1861) Einiges über die chemische Structur der Körper. *Zeitschrift für Chem.*, **4**, 549–560.
37. Corbett,P.T., Leclaire,J., Vial,L., West,K.R., Wietor,J.L., Sanders,J.K. and Otto,S. (2006) Dynamic combinatorial chemistry. *Chem. Rev.*, **106**, 3652–3711.
38. Orgel,L.E. (2008) The implausibility of metabolic cycles on the prebiotic Earth. *PLoS Biol.*, **6**, e18.

Supplementary Information

Coupled catabolism and anabolism in autocatalytic RNA sets

Simon Arsène[1,‡], Sandeep Ameta[1,‡], Niles Lehman[2], Andrew D. Griffiths[1,], Philippe Nghe[1,*]*

[‡]These authors contributed equally to this work.

[1]Laboratoire de Biochimie, CNRS UMR8231, Chimie Biologie Innovation, ESPCI Paris, 10 Rue Vauquelin, 75005, Paris, France.

[2]Department of Chemistry, Portland State University, P.O. Box 751, Portland, OR, USA 97207.

E-mail: andrew.griffiths@espci.fr, philippe.nghe@espci.fr

SUPPLEMENTARY TEXT

Kinetic modelling

We derive here a kinetic model for the multi-step reaction pathway presented in Figure 3 using a similar approach as used earlier (22). Our assumptions are the following: (i) covalent bonds formation is reversible and (ii) catalysis can occur either by a covalent ribozyme or by a non-covalent complex. In order to construct our model, similar to the formation of WXY:Z non-covalent complex, we assume that WXY can bind to Z-mod, and WXY-mod can bind to Z to form the supramolecular complex, WXY:Z-mod and WXY-mod:Z respectively. However we assume that the potential complex WXY-mod:Z-mod can only form in negligible amounts because of steric considerations. Contrary to what was done earlier (22), we chose not to fix arbitrarily the rates of association and dissociation of the non-covalent complexes as they are not expected to be equally stable. Formation of the covalent bond between WXY and Z (at the 'YZ' junction) can be catalyzed by either fully covalent WXYZ or by a non-covalent complex (20). Likewise, we assumed that the transfer of the modification part (mod) to WXY from Z-mod can be catalyzed by either fully covalent WXYZ or by a non-covalent complex. Cleavage of mod from WXY-mod can happen by attack of the 3' OH (of the 3' terminal 'G' of Z) from either covalent WXYZ or non-covalent complexes (WXY:Z and WXY-mod:Z). Finally as we do not observe any significant band of higher molecular weight than WXYZ (Supplementary Figure S4, S5) we neglect the possible formation of WXYZ-mod. Slow accumulation of WXYZ (Figure 2B) shows that WXYZ-mod is not formed quickly as an unstable intermediate. Indeed, if this was the case, the yield of WXYZ would increase more quickly than we observe because the cleavage of mod from WXYZ-mod would give WXYZ as a product. These assumptions allow us to limit the number of rate constants to fit. Models with more parameters can also be constructed with, for example, an independent description of the substrate binding and substrate release steps as proposed earlier (22). Our modeling approach resulted in 14 chemical reactions ('a' to 'n', Supplementary Information, Table S2). Reactions 'a-c' account for the formation of non-covalent complexes. Reactions 'd-g' account for the covalent bond formation between WXY and Z (at the 'YZ' junction) while reactions 'h-k' account for transfer of mod from Z-mod to WXY. Finally reactions 'l-n' account for cleavage of mod from WXY-mod.

The model fits well the experimental data with a root mean square error (RMSE) value of 1.2% (Supplementary Figure S9) and thus provide solid support for the multi-step reaction pathway proposed in Figure 3. The fit is slightly less good for WXYZ formation in reaction c2 and c3 for which we have less sensitivity on the measurement because of low intensity of the bands.

We tried to reduce the number of iterable rate constants based on several reasonable hypothesis: 1) all catalysts are assumed to be equally efficient and all complexes are assumed to be equally stable (8 parameters); 2) only WXY:Z-mod and WXY-mod:Z are assumed to be equally stable and all catalysts are assumed to be equally efficient (10 parameters); 3) only WXY:Z-mod and WXY-mod:Z are assumed to be equally stable and all catalysts are presumed to be equally efficient

expect for WXYZ as a catalyst for its own formation (12 parameters); 4) only WXY:Z-mod and WXY-mod:Z are assumed to be equally stable and all complexes are assumed to be equally efficient (16 parameters); 5) only WXY:Z-mod and WXY-mod:Z are assumed to be equally stable and equally efficient (22 parameters). For each of these hypotheses we fitted the rates to the corresponding reduced model (Supplementary Figure S10). Results show qualitatively that there is always at least one specie whose time course of formation is not well predicted by reduced models: either the monotony of the curve is not respected or there exist a systematic model deviation below or above what the spread of data points would allow. As comparing the overall RMSE neither conveys these trends nor allows to justify the number of parameters used, we also computed a set of different model selection criteria described by Turner *et al.* (39). These criteria quantify the trade-off between the number of parameters and the fitting quality, thus avoid overfitting. All these criteria are in favor of choosing the final model (Supplementary Table S3).

SUPPLEMENTARY TABLES

Table S1. Oligonucleotides used in the study.

Oligo name	Sequence	Description
Primer 1	ctgcagaattctaatacgactcactatagagccttgcgccgggaaacc acgcaagggatgg	Forward primer to add 'A' IGS
Primer 2	ctgcagaattctaatacgactcactatagtgcctagcgccgggaaacc acgctagggatgg	Forward primer to add 'U' IGS
Primer 3	ctgcagaattctaatacgactcactatagggcctcgcgccgggaaacc acgcgagggatgg	Forward primer to add 'G' IGS
Primer 4	atgtgccttaggtgggtgc	Reverse primer to add 'A' tag
Primer 5	aggtgccttaggtgggtgc	Reverse primer to add 'C' tag
Primer 6	aagtgccttaggtgggtgc	Reverse primer to add 'U' tag
Primer 7	ttttttttttttttttttttttggcactatgtgccttaggtgggtgc	Reverse primer to generate modified substrate WXY with 'A' tag
Primer 8	ttttttttttttttttttttttggcactaggtgccttaggtgggtgc	Reverse primer to generate modified substrate WXY with 'C' tag
Primer 9	ttttttttttttttttttttttggcactaagtgccttaggtgggtgc	Reverse primer to generate modified substrate WXY with 'U' tag
Primer 10	taatacgactcactataggcatcgctatggtgaaggcatag	Forward primer for Z substrate
Primer 11	ccggtttgtgtgactttcgcc	Reverse primer for Z substrate
Primer 12	ttttttttttttttttttttttggcactccggtttgtgtgactttcgcc	Reverse primer for generate modified Z substrate
Primer 13	cgtgtgctcttccgatctnnnnnnnnttttttttttttttttttttccggttt gtgtgactttcgcc	Reverse primer for reverse transcription
Primer 14	ttccctacacgacgctcttccgatctgatatagagccttgcgccgggaa acca	Illumina sequencing primer I (forward, GSP)
Primer 15	gtgactggagttcagacgtgtgctcttccg	Illumina sequencing primer I (reverse, RT)
Primer 16	aatgatacggcgaccaccgagatctacactctttccctacacgacgctc ttcc	Illumina sequencing primer II (forward, Rd1)
Primer 17	caagcagaagacggcatacgagatgtgactggagttcagacgtgtgc tcttccgatct	Illumina sequencing primer II (reverse, Rd2)
Primer 18	ttttttttttttttttttttggcacaccggtttgtgtgactttcgccactccc	For mutation M1 in Z-mod
Primer 19	ttttttttttttttttttttggcagaccggtttgtgtgactttcgccactccc	For mutation M2 in Z-mod
Primer 20	ttttttttttttttttttttggctgaccggtttgtgtgactttcgccactccc	For mutation M3 in Z-mod
Primer 21	ttttttttttttttttttttgggtgaccggtttgtgtgactttcgccactccc	For mutation M4 in Z-mod
Primer 22	ttttttttttttttttttttgcgtgaccggtttgtgtgactttcgccactccc	For mutation M5 in Z-mod
Primer 23	ttttttttttttttttttttccgtgaccggtttgtgtgactttcgccactccc	For mutation M6 in Z-mod
Primer 24	cgtgtgctcttccgatctnnnnnnnn ttttt ttttt ttttt ttttt gg	Reverse primer for reverse transcription for polyG-tailed samples
Primer 25	ttccctacacgacgctcttccgatct agttcc gag cct tgc gcc ggg aaacca	Forward primer used for poly-G tailed samples amplification

Table S2. Model reactions and fitted rate parameters. The ':' denotes a non-covalent complex between two fragments.

	Reaction	Forward	Reverse
a	WXY + Z -> WXY:Z	1.85E+02 $\mu M^{-1} min^{-1}$	4.44E-02 min^{-1}
b	WXY + Z-mod -> WXY:Z-mod	3.18E+01 μM^{-1} min^{-1}	2.14E+01 min^{-1}
c	WXY-mod + Z -> WXY-mod:Z	8.98E+01 μM^{-1} min^{-1}	2.77E-02 min^{-1}
d	WXYZ + WXY:Z -> WXYZ + WXYZ	8.90E+00 μM^{-1} min^{-1}	9.43E-01 μM^{-1} min^{-1}
e	WXY:Z + WXY:Z -> WXYZ + WXY:Z	1.50E+00 μM^{-1} min^{-1}	7.99E+00 μM^{-1} min^{-1}
f	WXY:Z + WXY-mod:Z -> WXYZ + WXY-mod:Z	1.02E+00 μM^{-1} min^{-1}	1.91E+00 μM^{-1} min^{-1}
g	WXY:Z-mod + WXY:Z -> WXYZ + WXY:Z-mod	5.81E-03 μM^{-1} min^{-1}	8.57E+01 μM^{-1} min^{-1}
h	WXY:Z-mod + WXYZ -> WXY-mod + Z + WXYZ	1.45E-01 μM^{-1} min^{-1}	4.93E+00 μM^{-1} min^{-1}
i	WXY:Z-mod + WXY:Z -> WXY-mod + Z + WXY:Z	9.12E-02 μM^{-1} min^{-1}	2.97E+01 μM^{-1} min^{-1}
j	WXY:Z-mod + WXY-mod:Z -> WXY-mod + Z + WXY-mod:Z	8.85E-01 μM^{-1} min^{-1}	8.37E+00 μM^{-1} min^{-1}
k	WXY:Z-mod + WXY:Z-mod -> WXY-mod + Z + WXY:Z-mod	1.26E+02 μM^{-1} min^{-1}	8.50E+01 μM^{-1} min^{-1}
l	WXY-mod:Z + WXYZ -> WXY:Z + mod + WXYZ	1.70E+01 μM^{-1} min^{-1}	1.48E-04 μM^{-2} min^{-1}
m	WXY-mod:Z + WXY:Z -> WXY:Z + mod + WXY:Z	4.86E-03 μM^{-1} min^{-1}	1.47E-04 μM^{-2} min^{-1}
n	WXY-mod:Z + WXY-mod:Z -> WXY:Z + mod + WXY-mod:Z	3.15E-01 μM^{-1} min^{-1}	7.23E+01 μM^{-2} min^{-1}

Table S3. Different selection criteria for models with reduced number of parameters. p is the number of parameters of the model, n the number of data points, df the degree of freedom of the model, RSS is the sum of squared residuals, AIC is the Akaike Information Criterion, BIC is the Bayesian Information Criterion, F is the F-statistic comparing a model with the chosen model (p = 28) and p-value is the associated p-value with the F-test.

p	n	df	RSS	AIC	BIC	F	p-value
8	120	112	0.1535	-781.38	-756.29	36.39	$p < 0.001$
10	120	110	0.1468	-782.75	-752.09	38.44	$p < 0.001$
12	120	108	0.0765	-857.03	-820.79	19.77	$p < 0.001$
16	120	104	0.0599	-878.36	-830.97	18.98	$p < 0.001$
22	120	98	0.0438	-904.00	-839.89	23.61	$p < 0.001$
28	120	92	0.0172	-1003.86	-923.02	N/A	N/A

SUPPLEMENTARY FIGURES

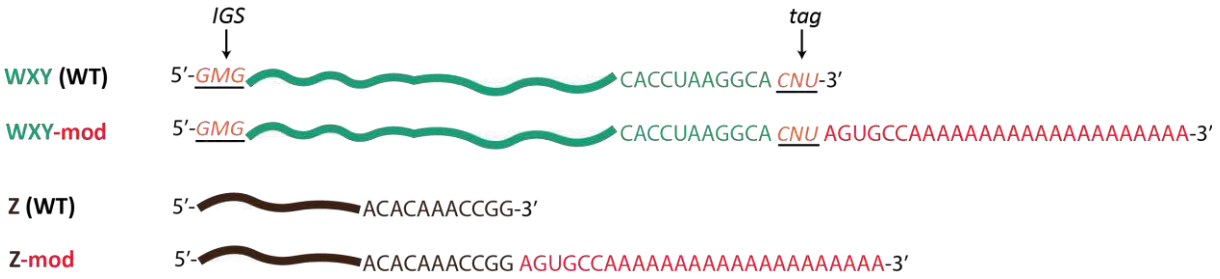


Figure S1. Modification of the substrates. The sequence AGUGCCA followed by a stretch of poly-adenosine (red) is appended to the 3′ end of the substrates. WXY and WXY-mod contains an internal guide sequence (IGS) ("GMG" at the 5′ end, orange and underlined) and a tag ("CNU" at the 3′ end, orange and underlined) required for the transesterification reaction (20, 21) where M and N are variable nucleotides in the IGS and tag (denoted as MN), respectively. While for all the assembly reactions MN is kept as AU (i.e., GAG and CUU as IGS and tag, respectively), for the cooperative network formation AA, GU and UC were used.

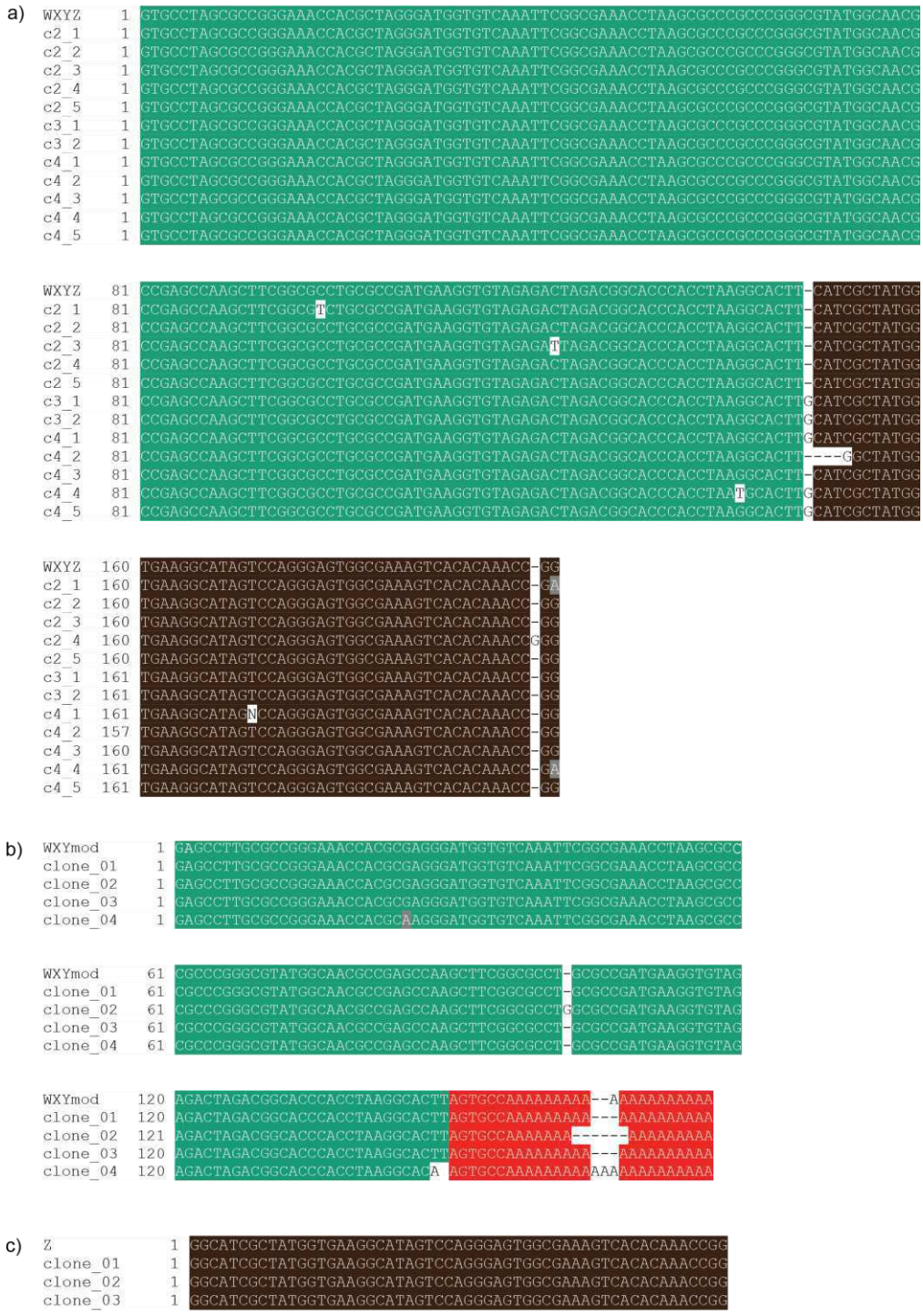


Figure S2. Sanger sequencing results of different reaction products. Sequence alignment of the various products. (**A**) Sequence alignment of WXYZ sequences obtained from the reactions with substrate combination c2, c3 and c4 aligned with the WXYZ reference sequence. (**B**) Sequence alignment of WXY-mod sequences obtained in the reaction with substrate combination c2. The poly-adenosine tail is often poorly sequenced as it contains the same nucleotide repeated several times. (**C**) Sequence alignment of Z sequences obtained in the reaction with substrate combination c2. The alignments were generated using T-Coffee web server (40) with default parameters and BoxShade (https://embnet.vital-it.ch/software/BOX_form.html.) for visualization. Different colors are used to highlight the different parts of the sequence: green for WXY, brown for Z and red for mod.

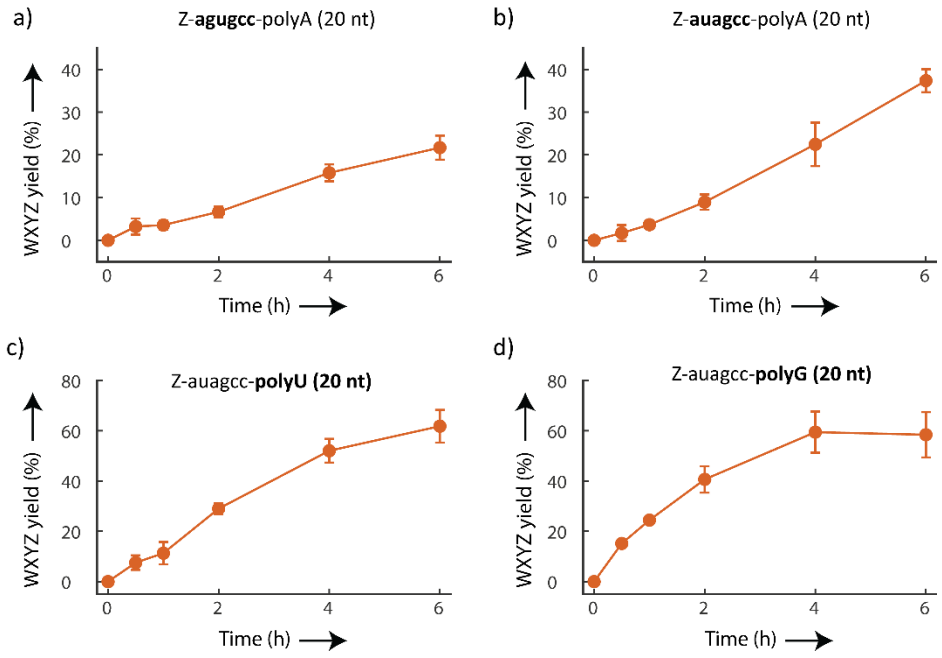


Figure S3. WXYZ formation with diverse modification in the substrate. Kinetic analysis showing that even when different modifications of Z are used the synthesis of catalysts WXYZ is not hampered. (**A**) WXYZ formation with 'agugcc'-polyA modification (used in all other experiments in the study). (**B-D**) WXYZ formation when modification is changed from 'agugcc'-polyA to 'auagcc'-polyA (**B**), 'auagcc'-polyU (**C**) or 'auagcc'-polyG (**D**). Here 0.5 µM of each RNA was incubated in the reaction buffer (see Material and Methods) at 48°C for 6 h. Samples at several time were taken out for gel analysis. The reported yield is the substrate to product conversion in percentage. Error bars represent ±1 standard deviation.

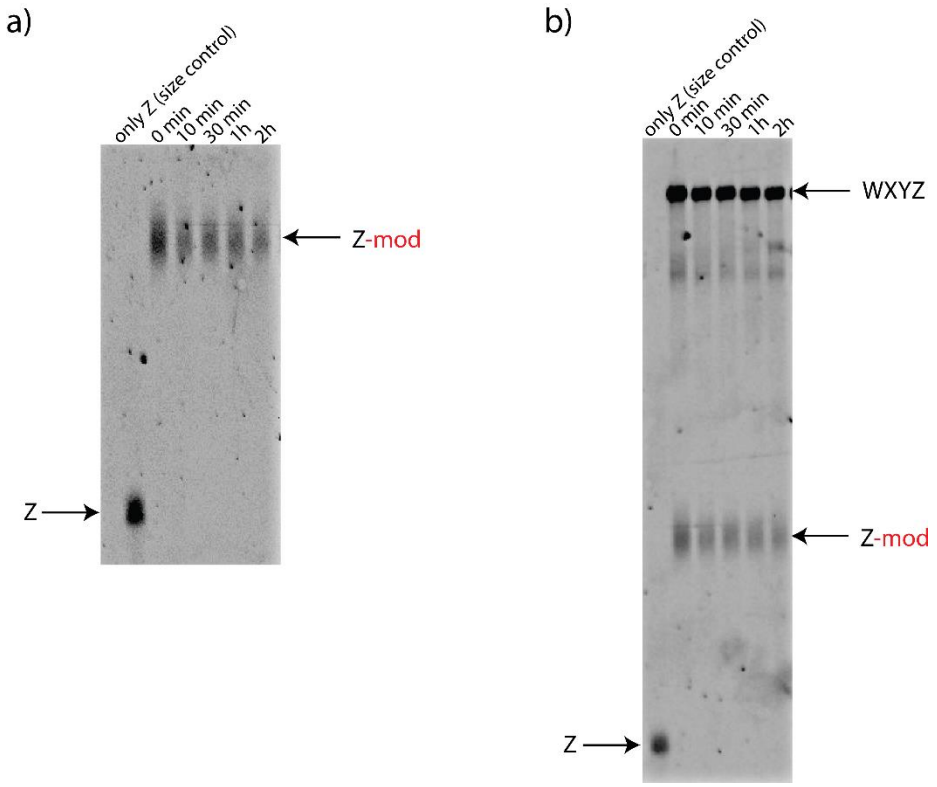


Figure S4. No direct cleavage of modification from the Z-mod substrate. Gel analysis showing that the Z-mod substrate is not hydrolyzed spontaneously. (**A**) Z-mod alone is not hydrolyzed to generate Z in presence of the reaction buffer as no slow migrating band corresponding to Z appears over the course of the reaction. (**B**) Even in presence of catalyst (WXYZ) there is no conversion of Z-mod to Z by cleaving off the modification directly from the substrate. Here 0.5 µM of each RNA was incubated in the reaction buffer at 48°C. Samples at several time points (indicated above the lanes) were taken out for analysis.

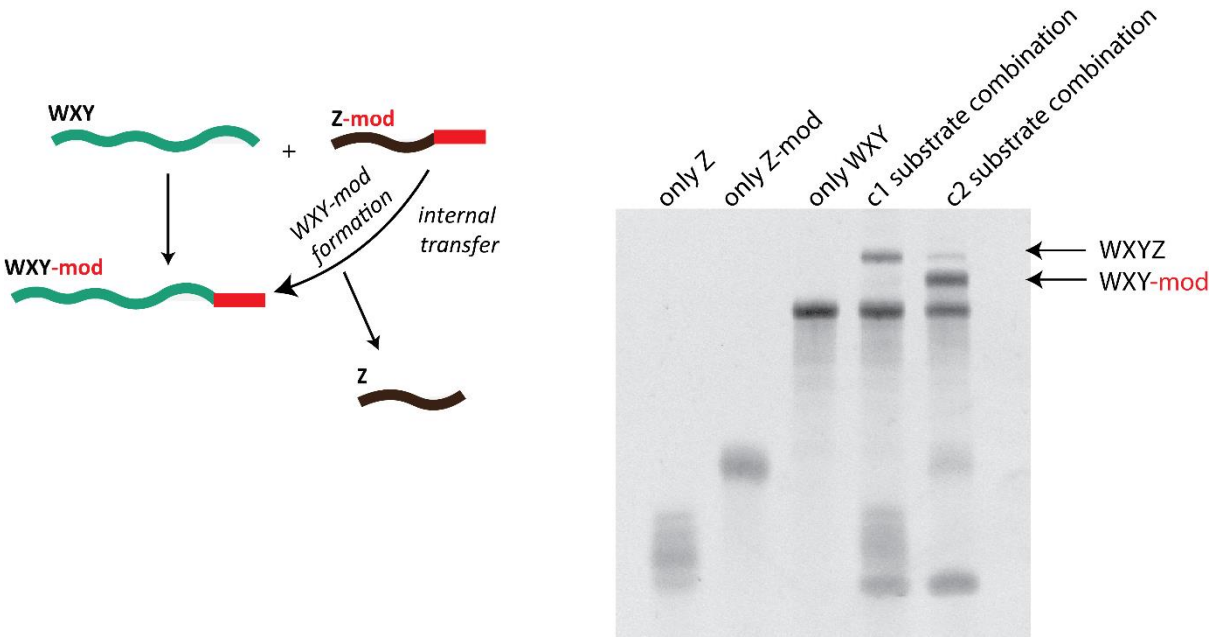


Figure S5. Internal transfer of modification to generate WXY-mod. Schematic showing (left panel) the formation of WXY-mod in the reaction with substrate combination c2 (WXY + Z-mod) by transfer of mod from Z-mod to WXY. Gel analysis (right panel) showing the formation of WXY-mod in the reaction with substrate combination c2. The sequence of WXY-mod was confirmed by sequencing (Supplementary Figure S2). 0.5 µM of each RNA was incubated in the reaction buffer for 1 h at 48°C. As controls, substrate combination c1 (WXY + Z), only Z, only Z-mod and only WXY, were also incubated.

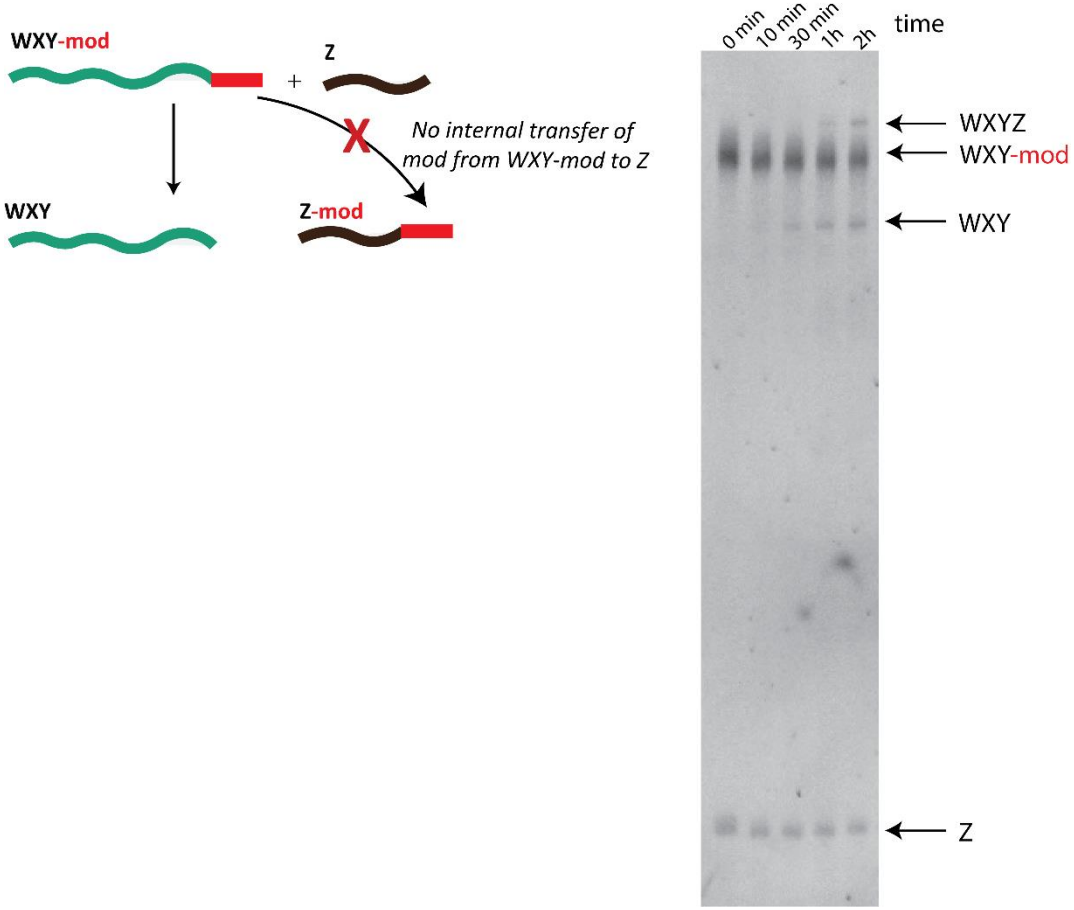


Figure S6. No formation of Z-mod from WXY-mod. Schematic showing (left panel) that there is no formation of Z-mod in the reaction with substrate combination c3 (WXY-mod + Z) by the internal transfer of modification from WXY-mod to Z. Gel analysis (right panel) showing that during the course of the reaction no Z is converted to Z-mod. Here 0.5 µM of each RNA was incubated in the reaction buffer at 48°C. Samples at several time points (indicated above the lanes) were taken out for analysis.

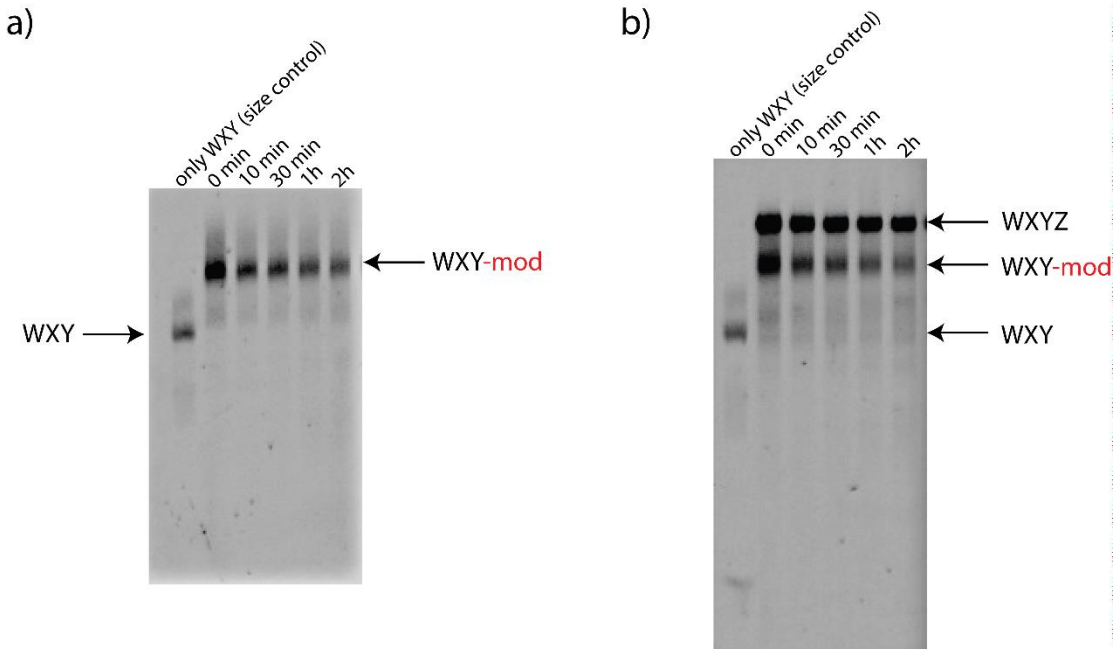


Figure S7. Cleavage of modification from WXY-mod is catalyzed by WXYZ. Gel analysis showing that (**A**) WXY-mod is stable in the reaction buffer as no band corresponding to WXY appears during the course of the reaction, and (**B**) in the presence of the catalyst WXYZ, mod is cleaved off WXY-mod and WXY is generated. Here 0.5 µM of each RNA was incubated in the reaction buffer at 48°C. Samples at several time points (indicated above the lanes) were taken out for analysis.

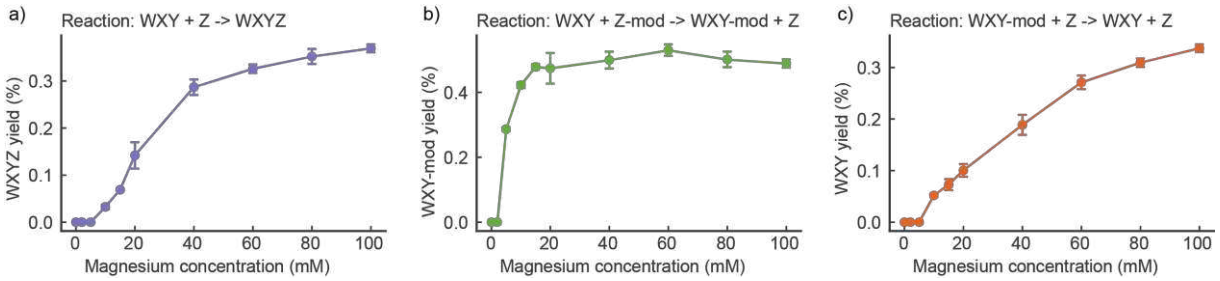


Figure S8. Different steps of the reaction occur even at lower magnesium concentrations. Time course experiment to demonstrate that different steps of the reaction; formation of the catalyst WXYZ from WXY + Z (**A**), formation of WXY-mod (**B**), and processing of WXY-mod to WXY (**C**), can occur at a broad range of $MgCl_2$ concentrations. Here 0.5 µM of each RNA was incubated at 48°C for 1h in the reaction buffer containing different $MgCl_2$ concentration. After 1h, the samples were taken out for the gel analysis. The reported yield is the substrate to product conversion in percentage. Error bars represent ±1 standard deviation.

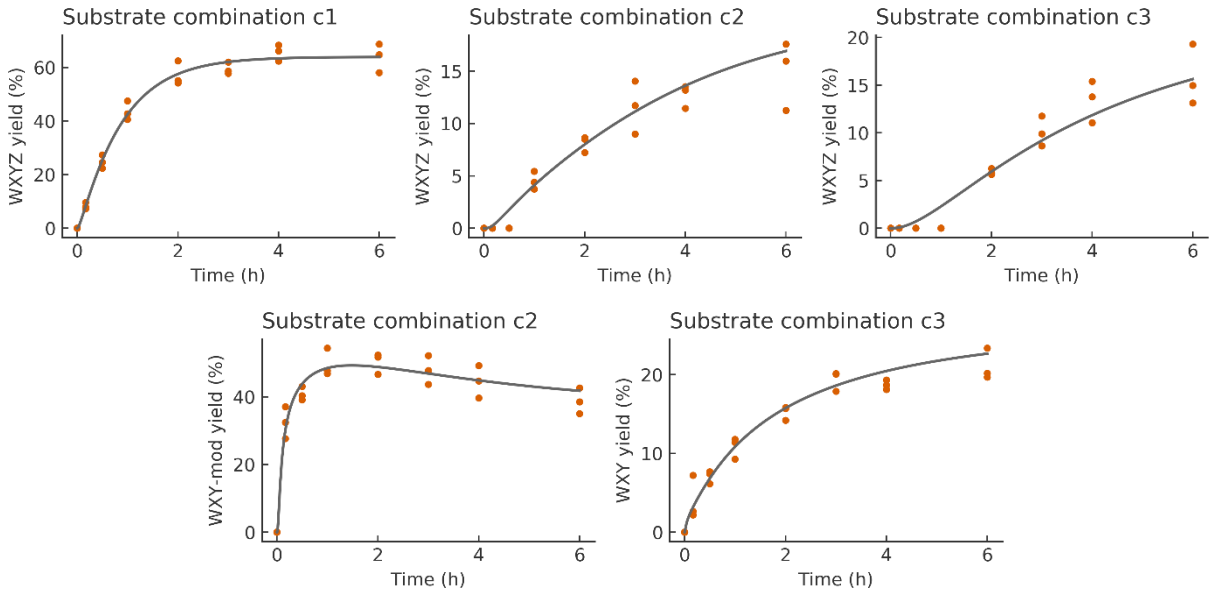


Figure S9. Time course of experimental data (triplicates, orange dots) and kinetic model data (grey line) for the different products formation in different substrate combinations. See main text for description of the reactions. The reported yield is the substrate to product conversion in percentage.

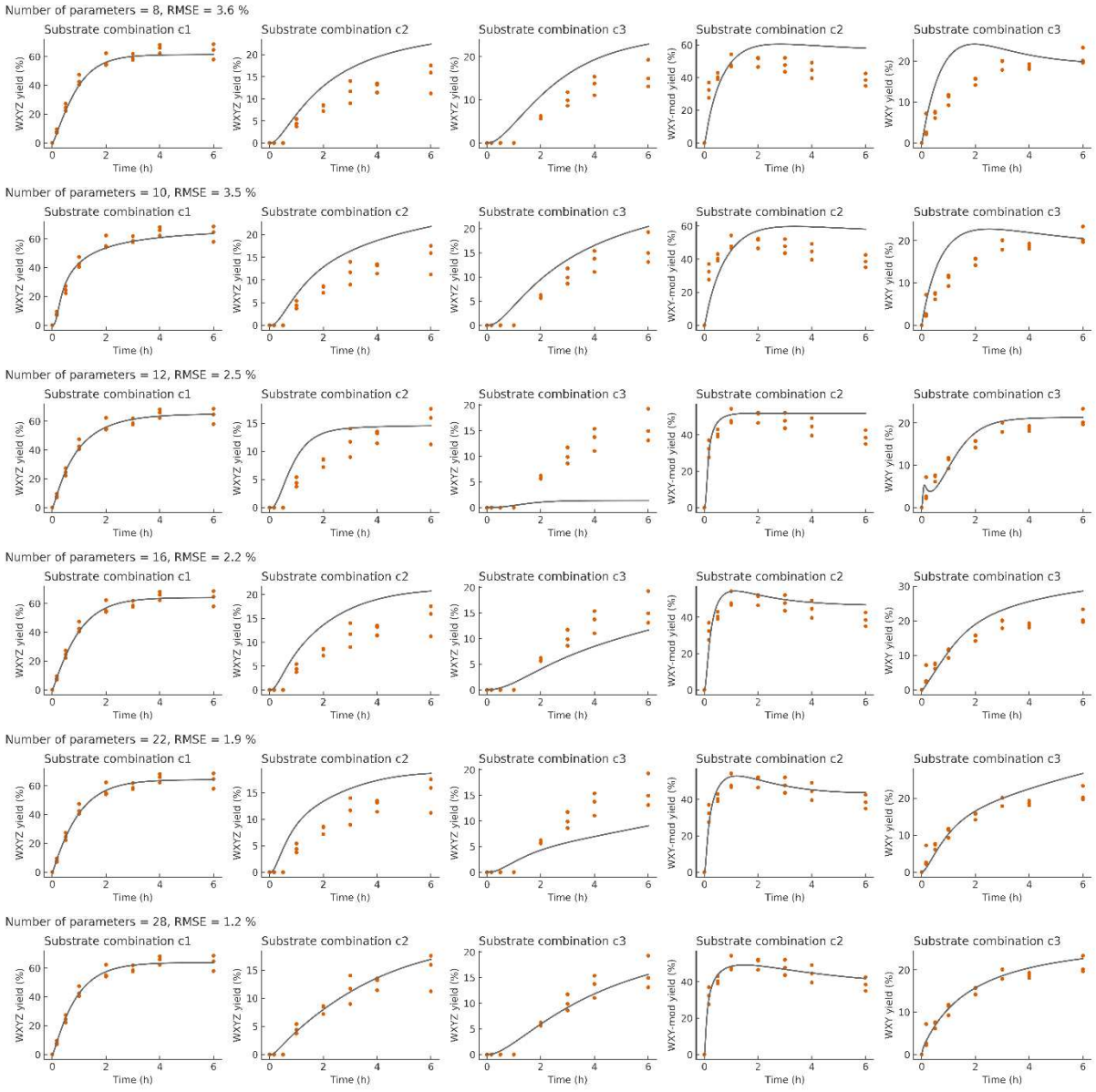


Figure S10. Fitting models with reduced numbers of parameters as described in the Kinetic modelling section. Grey line is fitted model data and orange dots are experimental data points (triplicates). The reported yield is the substrate to product conversion in percentage. For each reduced model the RMSE value is reported.

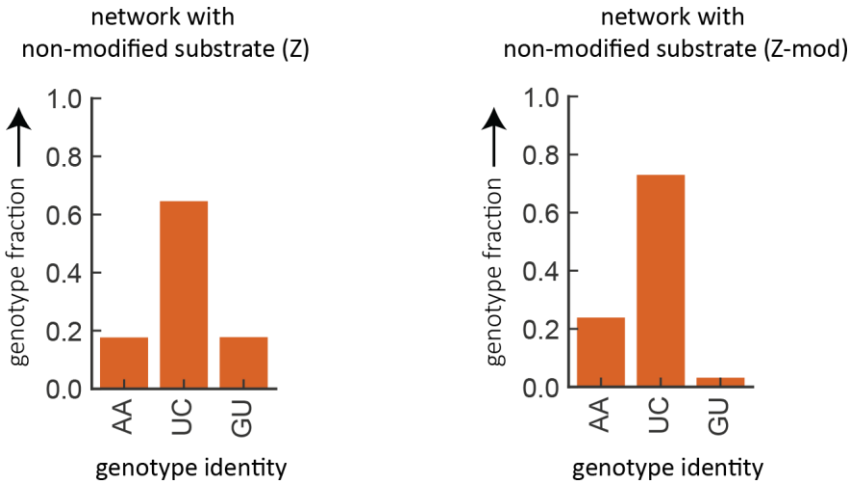


Figure S11. Genotypic distribution of cooperative network formed by modified substrates. Histograms demonstrating that even when the modified Z substrate (Z-mod) is used for cooperative network formation the genotypic distribution of the members in the network is maintained. The Y-axis shows relative genotype fraction derived from number of UMIs (25) obtained for the individual member of the network. The X-axis shows the genotype identity (MN, the sequence of the IGS and tag on each WXY, see Figure 5).

SUPPLEMENTARY REFRENCES

39. Turner,B.D., Henley,B.J., Sleap,S.B. and Sloan,S.W. (2015) Kinetic model selection and the Hill model in geochemistry. *Int. J. Environ. Sci. Technol.*, **12**, 2545-2558.

40. Di Tommaso,P., Moretti,S., Xenarios,I., Orobitg,M., Montanyola,A., Chang,J.M., Taly,J.F. and Notredame,C. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension, *Nucleic Acids Res.*, **39**, W13-17.

5. Conclusion

RNA has a central role in modern biology since not only the information necessary for protein translation is encoded in the sequence of nucleotides of messenger RNAs, but other RNAs possess catalytic properties, which are involved perhaps most notably in the ribosome activity (Ban et al. 2000). This dual nature of RNA, both information carrier and catalyst has fueled the RNA World hypothesis (Gilbert 1986), which is one of the main hypotheses of the origins of life and according to which there was a stage in time where life was mostly based on RNA, before the advent of the trio of contemporary biology, between RNA, DNA and proteins. Such a hypothesis requires as a first step RNA monomers to be synthesized in prebiotic conditions, i.e. with the organic compounds believed to be abundant on Earth before the appearance of life such as cyanamide, glycolaldehyde and inorganic phosphate among other (Powner et al. 2009). A lot of progress has been made towards the precise characterization of synthesis routes yielding, from this starting material, biological precursors such as nucleobases (Patel et al. 2015).

But if life took off with RNA, some process must be found by which RNA monomers polymerize into RNA oligomers long enough to have a chemical activity. Several prebiotically plausible methods of polymerization of ribonucleotides into short RNA strands have been discovered which involve for example use of phase interfaces with lipids or mineral (Ferris et al. 1996; Rajamani et al. 2008). These spontaneous polymerization events result in RNA strands often not long enough to have interesting catalytic activity. Alternatively, RNA enzymes, or ribozymes, derived from modern RNA ligases have been adapted to be able to elongate small RNA primers with very promising results since the most advanced versions can incorporate up to about 200 nucleotides starting from mononucleotides (Attwater et al. 2013). However, despite a lot of effort, none of these molecules able to polymerize its own sequence was yet discovered and this is often an argument used to discredit the RNA world hypothesis. In addition to the fact that no general RNA autoreplicase is known, there are other limitations to the spontaneous appearance of such a hypothesized molecule which would have had to be as long as 200 nucleotides to be able to self-

replicate with enough fidelity, not to decay due to the appearance of parasitic sequences (Kun et al. 2005).

To circumvent these limitations, it has been proposed that life on the early RNA World was based on webs of interdependent molecules, where each member of the set is catalytically formed by other members, rather than on single self-reproducing entities (Higgs & Lehman 2015; Nghe et al. 2015). They would have among other advantages the possibility to self-sustain from oligomers short enough to have been spontaneously synthetized in the environment. This concept has received a lot of attention theoretically and has been well formalized in different forms (Eigen & Schuster 1977; Gánti 2003; Kauffman 1992; Hordijk et al. 2012). A number of fundamental theoretical results have been established. In particular, it has been shown that with Collectively Autocatalytic Sets (CASs), co-existence of multiple stable cores is required for evolvability (Vasas et al. 2012; Jain & Krishna 2001). Evolution in this setting would involve transitioning from one dominant core to another by the combined actions of variation and selection, two other important ingredients for Darwinian evolution. These prebiotic networks have however been less studied in the lab and only few examples exist.

There are three major classes of CAS experimental models, based on peptides, small organic molecules and oligonucleotides (Lee et al. 1997; Sievers & von Kiedrowski 1994; Kim & Joyce 2004). RNA-based CAS models are even scarcer since there are only two main experimental systems purely based on RNA. In the first, RNA ligases were adapted so that they would be complementary to one another and thus be able to cross-catalyze their formation (Kim & Joyce 2004). Though some diversity has been introduced in this system (Lincoln & Joyce 2009), it is hard to disentangle the precise nature of the interactions between the members and thus the detail of the network structure too. In the second system, recombination ribozymes are derived from the *Azoarcus* bacterium group I intron (Riley & Lehman 2003). They utilize specific interactions between two three nucleotide long recognition sequences: the Internal Guide Sequence (IGS) and the target sequence (tag) in order to catalyze their own assembly or the assembly of other versions of the ribozyme where IGS and tag are varied. As a result, a wide diversity of RNA networks can be formed with this system (Vaidya et al. 2012).

There have been a significant number of results that have been established using the Azoarcus system. In particular, the superiority of cooperativity has been demonstrated with a three-membered closed catalytic cycle (Vaidya et al. 2012), game-theoritic dynamics has been applied to networks with two or three nodes (Yeates et al. 2016) and the growth dynamics of three-node cores has been studied in detail (Yeates et al. 2017). However, the order of magnitude of the diversity of possible Azoarcus networks, which reaches 2^{16} possible networks for the simplest version of the system, makes classical studies unable to apprehend this huge network space. To analyze this diversity, droplet microfluidics comes as the perfect tool having both the throughput required and the versatility to form and analyze a large library of *Azoarcus* networks. Droplet microfluidics manipulates emulsions at the micrometer scale using specific devices (Baroud et al. 2010). This allows massive serialization as well as low working volumes, reducing the amount of reactants consumed. It is currently used for many applications ranging from directed evolution of enzymes to single-cell transcriptomics (Klein et al. 2015; Ryckelynck et al. 2015). Droplet microfluidics has a great potential for origins of life research since evolution cycles in droplets can easily be implemented and since selection can be applied at the expense of developing a fluorescent reporter for the system of interest (Baret et al. 2009).

In this thesis, in Chapter 2, we studied the mechanisms of environmentally induced variations in *Azoarcus* CAS system. For this, we used a droplet microfluidics set-up coupled with high-throughput sequencing to create a large library of thousands of *Azoarcus* RNA networks. This allowed us to couple local node connectivity to species relative abundance. We established in particular that in-degree centrality is a good determinant for node's fraction. Because our set-up is combinatorial in nature, we could have access to the compositional neighborhoods of networks. We thus developed a perturbative approach that allowed us to analytically derive the set of network parameters controlling the change of the distribution of species fraction when a new node is added to the network. Notably, this analysis highlighted the importance of the network's link density, of the number of targets for the added node and of the novelty of the added node as catalyst for the network. We could classify the 16 avalaible *Azoarcus* nodes into three groups in order to design networks with the desired level of sensitivity to perturbations such as the addition of a new species. Finally, we illustrated how these results are prebiotically

relevant by experimentally implementing a scenario in which environmental perturbations in the food-set consisting of the addition of a new species resulted in different variations that could be propagated to the next generation by serial transfer.

In chapter 3, we focused on heredity, another important ingredient for evolution, by studying how the network structure can have an impact on its memory of the initial state. One could say that memory of the initial conditions can be considered as a pre-requisite for heredity since a network without memory at all would likely relax quickly to its single trivial stable state. In this chapter, we used a computation model to follow the evolution of networks when one node is seeded at start with some concentration. We could quantify memory of the network between two different seeds and developed a simple first-order approximation to analytically determine the restricted set of parameters controlling amount of memory. In particular, we found that besides network's link density, the uniqueness of its nodes in term of catalysis was an important factor. Finally, we validated our model and results by experimentally seeding one node a time a set of eight *Azoarcus* networks and found that the experimental results were in good agreement with our model's expectations.

In chapter 4, we studied what happened to the *Azoarcus* system when the food-set is chemically perturbed by appending to the native substrates some foreign sequence stretch, thus here the environment is perturbed in a different way than in chapter 2. This is critical to assess since until then, the *Azoarcus* system was built on strongly biased food-set created by fragmenting the final full-length ribozyme. By providing the system with modified substrates, we demonstrated the existence of a multi-step catabolic process, converting the modified substrates into usable building blocks for the classic anabolic process. We constructed a kinetic model describing the system that supports the different steps of the catabolic pathway and showed that a network composed of a closed catalytic cycle is still functional with substrates bearing modifications. To witness catabolism in such a primordial system makes the *Azoarcus* system even more prebiotically relevant and suggests that catabolism could have appeared early in the RNA world.

In conclusion, combined, these results bring substantial material for arguing for the importance of RNA collectively autocatalytic sets in the origins of life and in the RNA world. They represent

progress towards the first experimental demonstration of Darwinian evolution with a purely molecular system. This will be possible after the basic ingredients for evolution are shown with an experimental CAS system; variation, heredity and selection. Concerning the *Azoarcus* system, several leads can be envisioned to make this possible in a near future. Focus should be directed first at reducing the spontaneous assembly by the non-covalent complexes between the RNA fragments as discussed partially in chapter 3. An easy way to improve this point is to pick the junction at which to perform the two-piece assembly that minimize this (Hayden et al. 2008). Additionally, in chapter 4, we found that modifying both fragments resulted in a very strong decrease of the spontaneous assembly. This gives hope that the system could be optimized by appending modifications to prevent the spontaneous assembly since probably it is making the non-covalent complexes more unstable. However, this strategy is dependent on the capacity of the native catalyst to catalyze the recycling of the non-ideal substrate. When this is achieved, it should be quite straightforward to formally prove heredity with very simple networks in a serial transfer setting.

The next step in the road to evolution would involve the development of a fluorescent reporter for the *Azoarcus* system. Options include chemically modifying the fragments with a fluorescent moiety, using an external fluorescent molecular beacon or a fluorescent reporter for pyrophosphate which is released during the reaction. Ideally, one would prefer a method that keeps the activity of the catalysts so that they can be transmitted as seed to the next generation. A complete evolution experiment could then be implemented using droplet microfluidics. A library of networks with different initial state would be encapsulated in droplets and incubated to let the catalysts accumulate. Part of the library would then be selected based on the developed fluorescent reporter. The selected library could then be perturbed by changing the physic-chemical conditions in the environment or by introducing new species before being used as seed for the next round. At each round, some portion of the library, selected or unselected, can be taken out for analysis. With such a set-up, we should be able to witness evolution at play, shaping the distribution of the networks round after round, and this with a molecular RNA-based replicator system.

6. Annexes

6.1. Annex 1: Seeding eight Azoarcus RNA networks one node at a time, the complete dataset

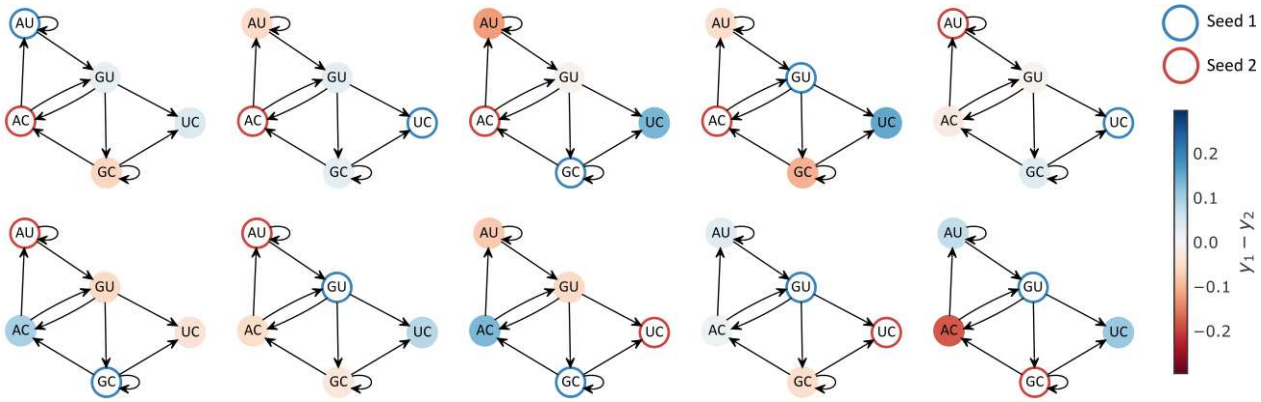


Figure 45. Experimental results for all pairs of seeds for network composed of the nodes AU, AC, UC, GC and GU. Seed 1 is circled in blue while seed 2 is circled in red. For the three remaining nodes, the color of the node codes for the value of $y_1 - y_2$ where y_1 (resp. y_2) is the relative concentration of the node when the seeded node is seed 1 (resp. seed 2) compared to the rest of the network but excluding the two seeds.

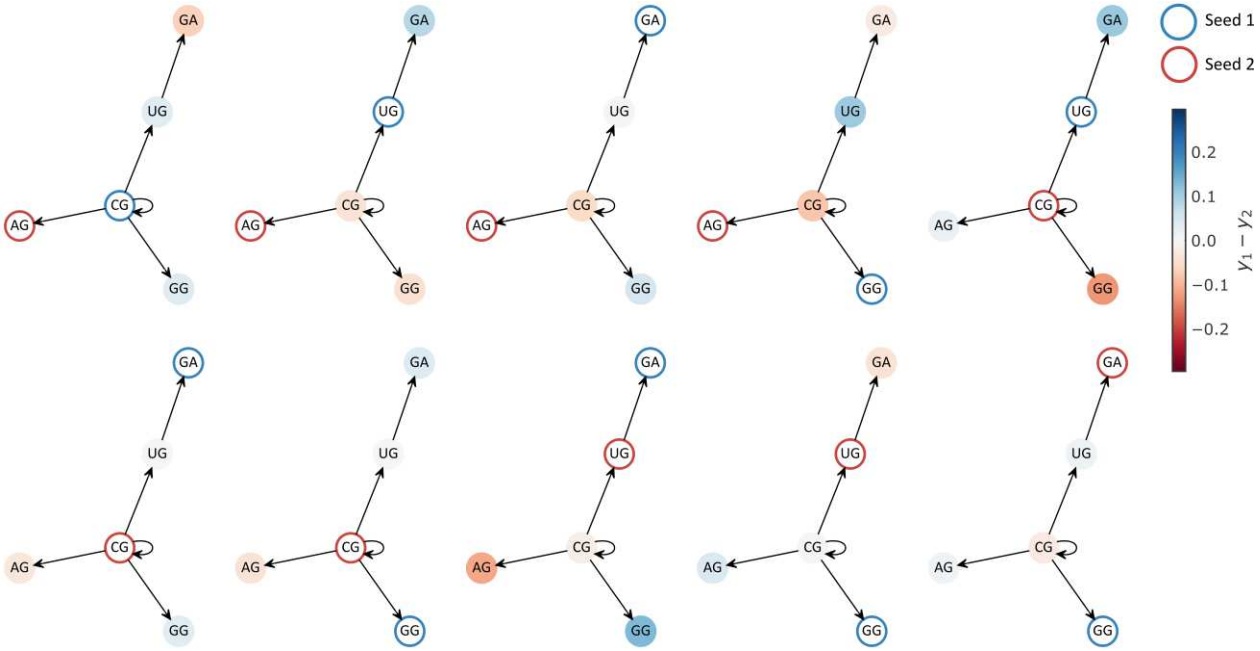


Figure 46. Same representation of experimental results as Figure 45 but for all pairs of seeds for network composed of the nodes AG, CG, UG, GA and GG.

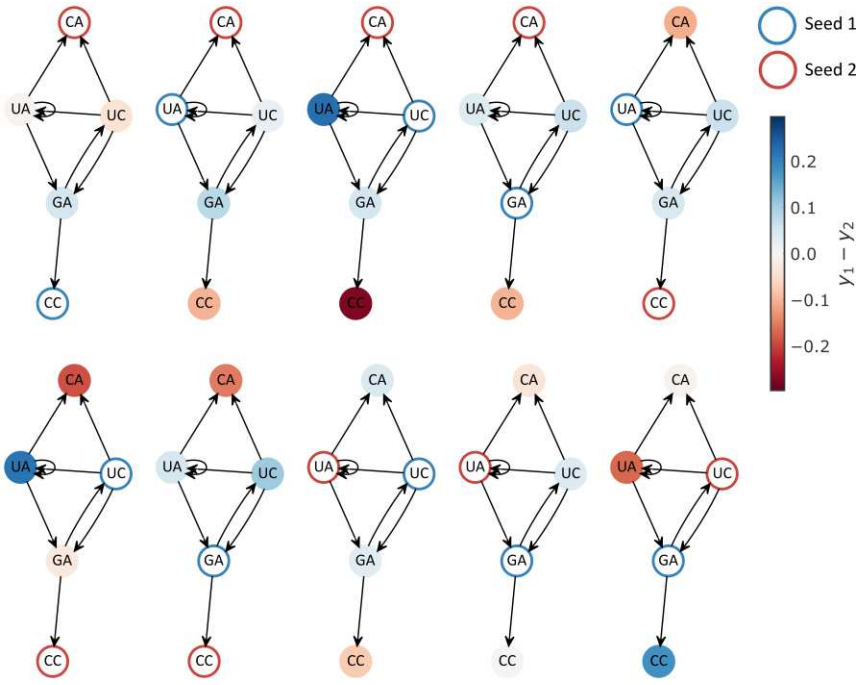


Figure 47. Same representation of experimental results as Figure 45 but for all pairs of seeds for network composed of the nodes AG, CG, UG, GA and GG.

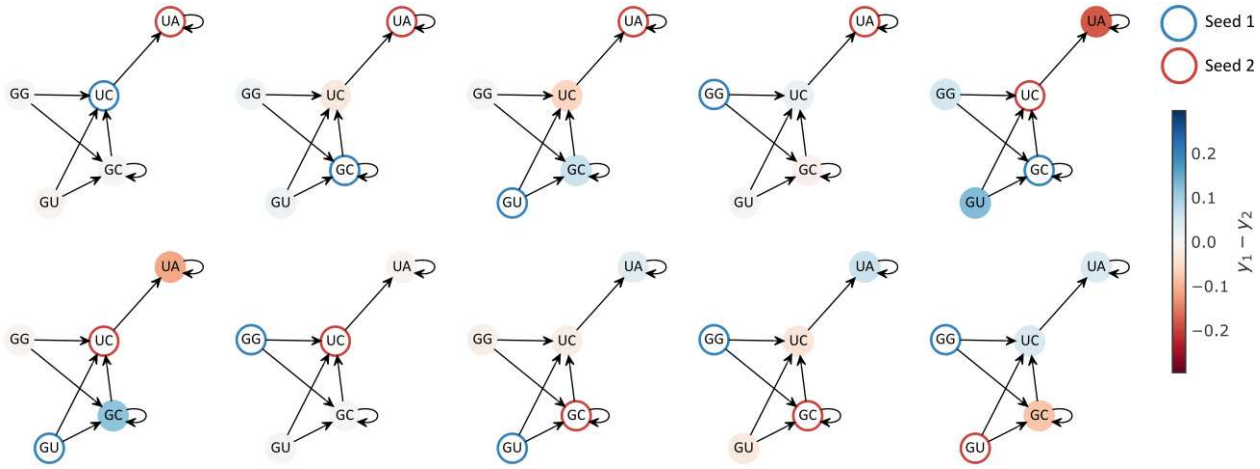


Figure 48. Same representation of experimental results as Figure 45 but for all pairs of seeds for network composed of the nodes AG, CG, UG, GA and GG.

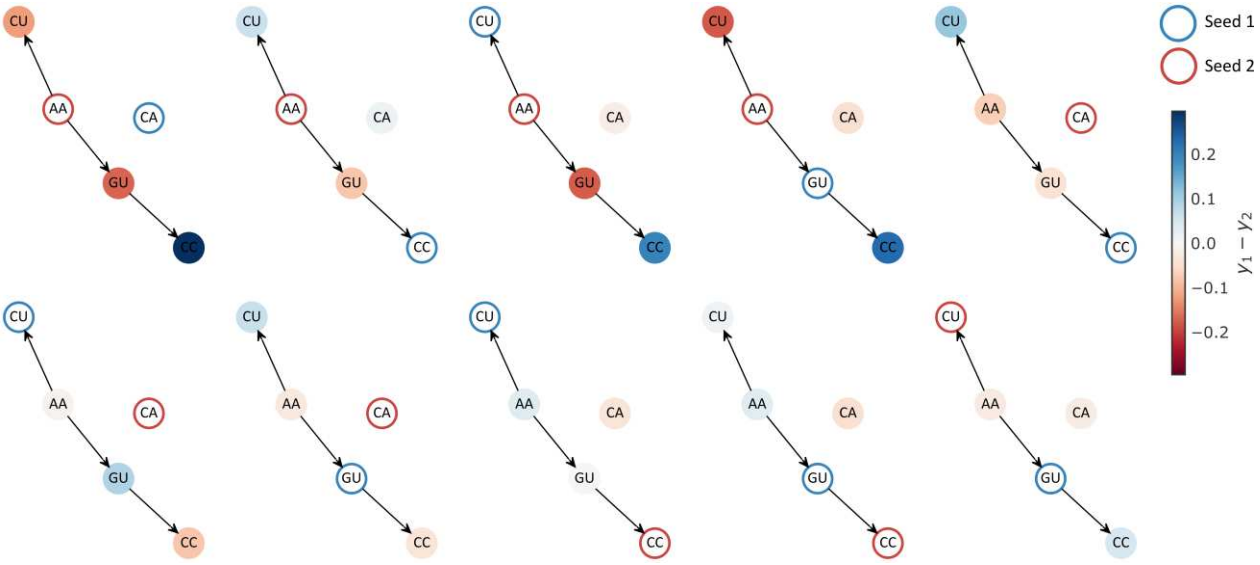


Figure 49. Same representation of experimental results as Figure 45 but for all pairs of seeds for network composed of the nodes AG, CG, UG, GA and GG.

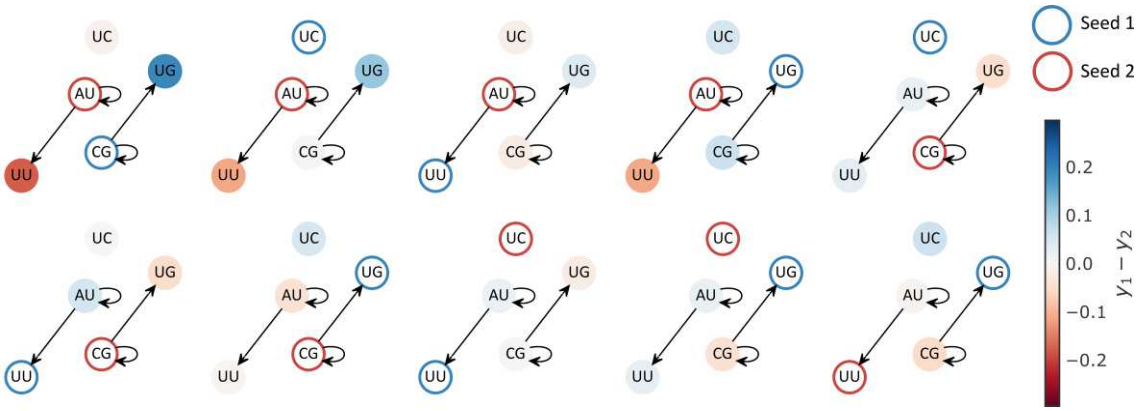


Figure 50. Same representation of experimental results as Figure 45 but for all pairs of seeds for network composed of the nodes AG, CG, UG, GA and GG.

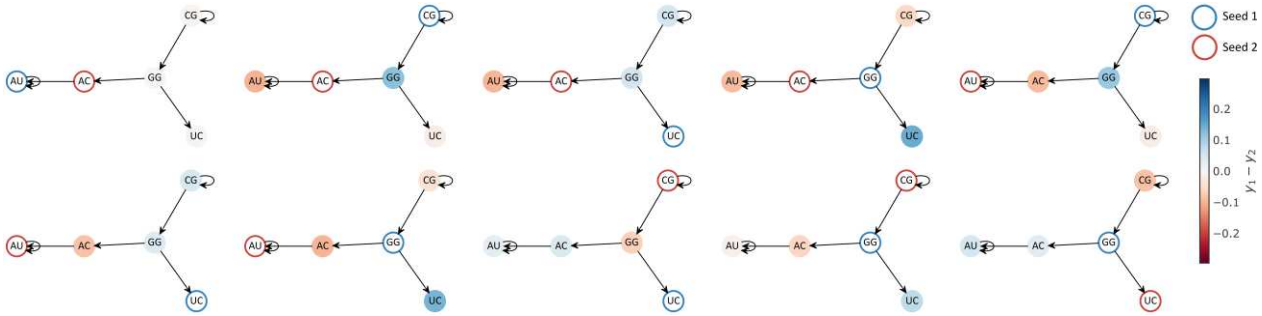


Figure 51. Same representation of experimental results as Figure 45 but for all pairs of seeds for network composed of the nodes AG, CG, UG, GA and GG.

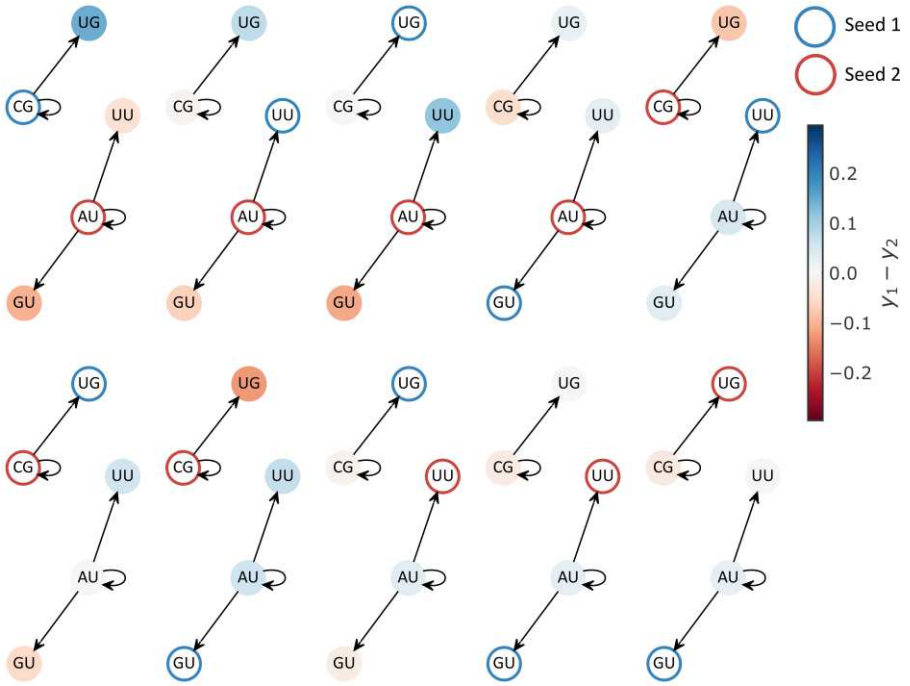


Figure 52. Same representation of experimental results as Figure 45 but for all pairs of seeds for network composed of the nodes AG, CG, UG, GA and GG.

7. References

Abate, A.R. et al., 2010. High-throughput injection with microfluidics using picoinjectors. *Proceedings of the National Academy of Sciences*, 107(45), pp.19163–19166.

Abate, A.R. et al., 2009. Impact of inlet channel geometry on microfluidic drop formation. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(2), pp.1–5.

Abate, A.R. & Weitz, D.A., 2011. Faster multiple emulsification with drop splitting. *Lab on a Chip*, 11(11), p.1911.

Achilles, T. & von Kiedrowski, G., 1993. A Self-Replicating System from Three Starting Materials. *Angewandte Chemie International Edition in English*, 32(8), pp.1198–1201.

Adamala, K. & Szostak, J.W., 2013. Nonenzymatic Template-Directed RNA Synthesis Inside Model Protocells. *Science*, 342(6162), pp.1098–1100.

Ahn, K. et al., 2006. Dielectrophoretic manipulation of drops for high-speed microfluidic sorting devices. *Applied Physics Letters*, 88(2), pp.1–3.

Anna, S.L., Bontoux, N. & Stone, H.A., 2003. Formation of dispersions using “flow focusing” in microchannels. *Applied Physics Letters*, 82(3), pp.364–366.

Arayanarakool, R. et al., 2013. Single-enzyme analysis in a droplet-based micro- and nanofluidic system. *Lab on a Chip*, 13(10), p.1955.

Arsène, S. et al., 2018. Coupled catabolism and anabolism in autocatalytic RNA sets. *Nucleic Acids Research*, (July), pp.1–7.

Ashkenasy, G. et al., 2004. Design of a directed molecular network. *Proceedings of the National Academy of Sciences*, 101(30), pp.10872–10877.

Ashkenasy, G. & Ghadiri, M.R., 2004. Boolean Logic Functions of a Synthetic Peptide Network. *Journal of the American Chemical Society*, 126(36), pp.11140–11141.

Atencia, J. & Beebe, D.J., 2005. Controlled microfluidic interfaces. *Nature*, 437(7059), pp.648–655.

Atkins, J., 2010. *RNA Worlds: From Life's Origins to Diversity in Gene Regulation*, Cold Spring Harbor Laboratory Press.

Attwater, J., Wochner, A. & Holliger, P., 2013. In-ice evolution of RNA polymerase ribozyme activity. *Nature Chemistry*, 5(12), pp.1011–1018.

Auroux, P.-A. et al., 2004. Miniaturised nucleic acid analysis. *Lab on a Chip*, 4(6), p.534.

Ban, N. et al., 2000. The complete atomic structure of the large ribosomal subunit at 2. *Science*, 289(5481), pp.905–920.

Baret, J.-C. et al., 2009. Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab on a Chip*, 9(13), p.1850.

Baret, J.-C., 2012. Surfactants in droplet-based microfluidics. *Lab Chip*, 12(3), pp.422–433.

Baroud, C.N., Gallaire, F. & Dangla, R., 2010. Dynamics of microfluidic droplets. *Lab on a Chip*, 10(16), p.2032.

Bartel, D. & Szostak, J., 1993. Isolation of new ribozymes from a large pool of random sequences [see comment]. *Science*, 261(5127), pp.1411–1418.

Beneyton, T. et al., 2017. Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica. *Microbial Cell Factories*, 16(1), pp.1–14.

Bibette, J., Calderon, F.L. & Poulin, P., 1999. Emulsions: basic principles. *Reports on Progress in Physics*, 62(6), pp.969–1033.

Bowler, F.R. et al., 2013. Prebiotically plausible oligoribonucleotide ligation facilitated by chemoselective acetylation. *Nature Chemistry*, 5(5), pp.383–389.

Boza, G. et al., 2014. Evolution of the Division of Labor between Genes and Enzymes in the RNA World. *PLoS Computational Biology*, 10(12), p.e1003936.

Breaker, R.R., 2012. Riboswitches and the RNA world. *Cold Spring Harbor Perspectives in*

Biology, 4(2), pp.1–15.

Butlerow, A., 1861. Ueber die Aethylmilchsäure. *Annalen der Chemie und Pharmacie*, 118(3), pp.325–330.

Cech, T.R., 1990. Self-Splicing of Group I Introns. *Annual Review of Biochemistry*, 59(1), pp.543–568.

Cech, T.R., 2012. The RNA worlds in context. *Cold Spring Harbor Perspectives in Biology*, 4(7), pp.1–5.

Chen, X., Li, N. & Ellington, A.D., 2007. Ribozyme catalysis of metabolism in the RNA world. *Chemistry and Biodiversity*, 4(4), pp.633–655.

Costanzo, G. et al., 2009. Generation of long RNA chains in water. *Journal of Biological Chemistry*, 284(48), pp.33206–33216.

Courtois, F. et al., 2008. An integrated device for monitoring time-dependent in vitro expression from single genes in picolitre droplets. *ChemBioChem*, 9(3), pp.439–446.

Crick, F.H.C., 1968. The origin of the genetic code. *Journal of Molecular Biology*, 38(3), pp.367–379.

Dadon, Z. et al., 2010. Light-induced peptide replication controls logic operations in small networks. *Chemistry - A European Journal*, 16(40), pp.12096–12099.

Deamer, D. et al., 2006. Self-assembly processes in the prebiotic environment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1474), pp.1809–1818.

Denesyuk, N.A. & Thirumalai, D., 2015. How do metal ions direct ribozyme folding? *Nature Chemistry*, 7(10), pp.793–801.

Dewan, A. et al., 2012. Growth kinetics of microalgae in microfluidic static droplet arrays. *Biotechnology and Bioengineering*, 109(12), pp.2987–2996.

Dodd, M.S. et al., 2017. Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature*, 543(7643), pp.60–64.

Dollet, B. et al., 2008. Role of the channel geometry on the bubble pinch-off in flow-focusing devices. *Physical Review Letters*, 100(3), pp.1–4.

Draper, W.E., Hayden, E.J. & Lehman, N., 2008. Mechanisms of covalent self-assembly of the Azoarcus ribozyme from four fragment oligonucleotides. *Nucleic Acids Research*, 36(2), pp.520–531.

Duffy, D.C. et al., 1998. Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane). *Analytical Chemistry*, 70(23), pp.4974–4984.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58(10), pp.465–523.

Eigen, M. & Schuster, P., 1977. The Hypercycle; A Principle of Natural Self-Organization; Part A: The Emergence of the Hypercycle. *Naturwissenschaften*, 64(11), pp.541–565.

Eigen, M. & Schuster, P., 1978. The Hypercycle; A Principle of Natural Self-Organization; Part B: The Abstract Hypercycle. *Naturwissenschaften*, 65(1), pp.7–41.

Eyer, K. et al., 2017. Single-cell deep phenotyping of IgG-secreting cells for high-resolution immune monitoring. *Nature Biotechnology*, 35(10), pp.977–982.

Fallah-Araghi, A. et al., 2012. A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. *Lab on a Chip*, 12(5), p.882.

Famulok, M. & Mayer, G., 2014. Aptamers and SELEX in chemistry & biology. *Chemistry and Biology*, 21(9), pp.1055–1058.

Ferré-D'Amaré, A.R. & Scott, W.G., 2010. Small self-cleaving ribozymes. *Cold Spring Harbor perspectives in biology*, 2(10).

Ferré-D'Amaré, A.R., Zhou, K. & Doudna, J.A., 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395(6702), pp.567–574.

Ferris, J.P. et al., 1996. Synthesis of long prebiotic oligomers on mineral surfaces. *Nature*,

381(6577), pp.59–61.

Franke, T. et al., 2010. Surface acoustic wave actuated cell sorting (SAWACS). *Lab on a Chip*, 10(6), p.789.

Frenz, L. et al., 2009. Reliable microfluidic on-chip incubation of droplets in delay-lines. *Lab Chip*, 9(10), pp.1344–1348.

Gánti, T., 2003. *The Principles of Life*, Oxford University Press, USA.

Gilbert, W., 1986. The RNA world. *Nature*, 319(February), p.618.

Guerrier-Takada, C. et al., 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 PART 2), pp.849–857.

Hammann, C. et al., 2012. The ubiquitous hammerhead ribozyme. *RNA*, 18(5), pp.871–885.

Harrison, D.J. et al., 1992. Capillary Electrophoresis and Sample Injection Systems Integrated on a Planar Glass Chip. *Analytical Chemistry*, 64(17), pp.1926–1932.

Hayden, E.J. et al., 2005. RNA-directed construction of structurally complex and active ligase ribozymes through recombination. *Rna*, 11(11), pp.1678–1687.

Hayden, E.J., Von Kiedrowski, G. & Lehman, N., 2008. Systems chemistry on ribozyme self-construction: Evidence for anabolic autocatalysis in a recombination network. *Angewandte Chemie - International Edition*, 47(44), pp.8424–8428.

Hayden, E.J. & Lehman, N., 2006. Self-Assembly of a Group I Intron from Inactive Oligonucleotide Fragments. *Chemistry and Biology*, 13(8), pp.909–918.

Hickman-Lewis, K. et al., 2018. Most Ancient Evidence for Life in the Barberton Greenstone Belt: Microbial Mats and Biofabrics of the ~3.47 Ga Middle Marker Horizon. *Precambrian Research*.

Higgs, P.G. & Lehman, N., 2015. The RNA World: Molecular cooperation at the origins of life. *Nature Reviews Genetics*, 16(1), pp.7–17.

Hinshelwood, C.N., 1952. On the chemical kinetics of autosynthetic systems. *Journal of the*

Chemical Society (Resumed), p.745.

Hordijk, W. et al., 2014. An investigation into irreducible autocatalytic sets and power law distributed catalysis. *Natural Computing*, 13(3), pp.287–296.

Hordijk, W., 2013. Autocatalytic Sets. *BioScience*, 63(11), pp.877–881.

Hordijk, W., Smith, J.I. & Steel, M., 2015. Algorithms for detecting and analysing autocatalytic sets. *Algorithms for Molecular Biology*, 10(1), pp.1–16.

Hordijk, W. & Steel, M., 2013. A formal model of autocatalytic sets emerging in an RNA replicator system. *Journal of Systems Chemistry*, 4(1), p.3.

Hordijk, W. & Steel, M., 2015. Autocatalytic sets and boundaries. *Journal of Systems Chemistry*, 6(1), p.1.

Hordijk, W. & Steel, M., 2004. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology*, 227(4), pp.451–461.

Hordijk, W., Steel, M. & Dittrich, P., 2018. Autocatalytic sets and chemical organizations: Modeling self-sustaining reaction networks at the origin of life. *New Journal of Physics*, 20(1).

Hordijk, W., Steel, M. & Kauffman, S., 2012. The Structure of Autocatalytic Sets: Evolvability, Enablement, and Emergence. *Acta Biotheoretica*, 60(4), pp.379–392.

Horning, D.P. & Joyce, G.F., 2016. Amplification of RNA by an RNA polymerase ribozyme. *Proceedings of the National Academy of Sciences*, 113(35), pp.9786–9791.

Huang, W. & Ferris, J.P., 2006. One-Step, Regioselective Synthesis of up to 50-mers of RNA Oligomers by Montmorillonite Catalysis. *Journal of the American Chemical Society*, 128(27), pp.8914–8919.

Huang, W. & Ferris, J.P., 2003. Synthesis of 35–40 mers of RNA oligomers from unblocked monomers. A simple approach to the RNA world. *Chem. Commun.*, 3(12), pp.1458–1459.

Issac, R. & Chmielewski, J., 2002. Approaching Exponential Growth with a Self-Replicating

Peptide. *Journal of the American Chemical Society*, 124(24), pp.6808–6809.

Jablonka, E. & Szathmáry, E., 1995. The evolution of information storage and heredity. *Trends in Ecology & Evolution*, 10(5), pp.206–211.

Jain, S. & Krishna, S., 2001. A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proceedings of the National Academy of Sciences*, 98(2), pp.543–547.

Jain, S. & Krishna, S., 2002. Large extinctions in an evolutionary model: The role of innovation and keystone species. *Proceedings of the National Academy of Sciences*, 99(4), pp.2055–2060.

Jayathilaka, T.S. & Lehman, N., 2018. Spontaneous Covalent Self-Assembly of the Azoarcus Ribozyme from Five Fragments. *ChemBioChem*, 19(3), pp.217–220.

Joyce, G.F., 2002. The antiquity of RNA-based evolution. *Nature*, 418(6894), pp.214–221.

Kan, C.W. et al., 2004. DNA sequencing genotyping in miniaturized electrophoresis systems. *Electrophoresis*, 25(21–22), pp.3564–3588.

Kauffman, S.A., 1986. Autocatalytic sets of proteins. *Journal of Theoretical Biology*, 119(1), pp.1–24.

Kauffman, S.A., 1971. Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. *Journal of Cybernetics*, 1(1), pp.71–96.

Kauffman, S.A., 1992. *The Origins of Order: Self-Organization and Selection in Evolution*,

Kazantsev, A. V et al., 2005. Crystal structure of a bacterial ribonuclease P RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38), pp.13392–7.

von Kiedrowski, G., 1986. A Self-Replicating Hexadeoxynucleotide. *Angewandte Chemie International Edition in English*, 25(10), pp.932–935.

von Kiedrowski, G., 1993. *Minimal Replicator Theory I: Parabolic Versus Exponential Growth*, Bioorganic Chemistry Frontiers.

von Kiedrowski, G. et al., 1991. Parabolic Growth of a Self-Replicating Hexadeoxynucleotide Bearing a 3'-5'-Phosphoamidate Linkage. *Angewandte Chemie International Edition in English*, 30(4), pp.423–426.

Kim, D.-E. & Joyce, G.F., 2004. Cross-Catalytic Replication of an RNA Ligase Ribozyme. *Chemistry & Biology*, 11(11), pp.1505–1512.

Kivioja, T. et al., 2012. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1), pp.72–74.

Klein, A.M. et al., 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), pp.1187–1201.

Klein, D.J. & Ferré-D'Amaré, A.R., 2006. Structural basis of \textit{glmS} ribozyme activation by glucosamine-6-phosphate. *Science (New York, N.Y.)*, 313(5794), pp.1752–1756.

Kosikova, T. & Philp, D., 2017. Exploring the emergence of complexity using synthetic replicators. *Chemical Society Reviews*, 46, pp.7274–7305.

Kozlov, I.A. & Orgel, L.E., 2000. Nonenzymatic Template-directed Synthesis of RNA from Monomers. *Molecular Biology*, 34(6), pp.921–930.

Kruger, K. et al., 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1), pp.147–157.

Kuhsel, M.G., Stkickland, R. & Palmer, J.D., 1990. An ancient group I intron shared by eubacteria and chloroplasts. *Science*, 250(4987), pp.1570–1573.

Kun, Á. et al., 2015. The dynamics of the RNA world: Insights and challenges. *Annals of the New York Academy of Sciences*, 1341(1), pp.75–95.

Kun, Á., Santos, M. & Szathmáry, E., 2005. Real ribozymes suggest a relaxed error threshold. *Nature Genetics*, 37(9), pp.1008–1011.

Lee, D.H. et al., 1996. A self-replicating peptide. *Nature*, 382(6591), pp.525–528.

Lee, D.H. et al., 1997. Emergence of symbiosis in peptide self-replication through a hypercyclic

network. *Nature*, 390(6660), pp.591–594.

Lehman, N., 2003. A case for the extreme antiquity of recombination. *Journal of Molecular Evolution*, 56(6), pp.770–777.

Lehman, N. et al., 2011. Complexity through recombination: From chemistry to biology. *Entropy*, 13(1), pp.17–37.

Leman, M. et al., 2015. Droplet-based microfluidics at the femtolitre scale. *Lab Chip*, 15(3), pp.753–765.

Levy, M. & Ellington, A.D., 2001. The descent of polymerization. *Nature Structural Biology*, 8(7), pp.580–582.

Li, Z. et al., 2015. Step-emulsification in a microfluidic device. *Lab Chip*, 15(4), pp.1023–1031.

Lilley, D.M.J. & Eckstein, F., 2007. *Ribozymes and RNA Catalysis*, The Royal Society of Chemistry.

Lincoln, T.A. & Joyce, G.F., 2009. Self-sustained replication of an RNA enzyme. *Science*, 323(5918), pp.1229–1232.

Link, D.R. et al., 2004. Geometrically Mediated Breakup of Drops in Microfluidic Devices. *Physical Review Letters*, 92(5), p.054503.

Liu, Y. et al., 2014. Crystal structure and mechanistic investigation of the twister ribozyme. *Nature Chemical Biology*, 10(9), pp.739–744.

Liu, Y., Wilson, T.J. & Lilley, D.M.J., 2017. The structure of a nucleolytic ribozyme that employs a catalytic metal ion. *Nature Chemical Biology*, 13(5), pp.508–513.

Luther, A., Brandsch, R. & Von Kiedrowski, G., 1998. Surface-promoted replication and exponential amplification of DNA analogues. *Nature*, 396(6708), pp.245–248.

Macosko, E.Z. et al., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), pp.1202–1214.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–380.

Mark, D. et al., 2010. Microfluidic lab-on-a-chip platforms: Requirements, characteristics and applications. *NATO Science for Peace and Security Series A: Chemistry and Biology*, (3), pp.305–376.

Markovitch, O. & Lancet, D., 2012. Excess Mutual Catalysis Is Required for Effective Evolvability. *Artificial Life*, 18(3), pp.243–266.

Matsumura, S. et al., 2016. Transient compartmentalization of RNA replicators prevents extinction due to parasites. *Science*, 354(6317), pp.1293–1296.

Mazutis, L., Araghi, A.F., et al., 2009. Droplet-Based Microfluidic Systems for High-Throughput Single DNA Molecule Isothermal Amplification and Analysis. *Analytical Chemistry*, 81(12), pp.4813–4821.

Mazutis, L. et al., 2013. Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols*, 8(5), pp.870–891.

Mazutis, L., Baret, J.-C. & Griffiths, A.D., 2009. A fast and efficient microfluidic system for highly selective one-to-one droplet fusion. *Lab on a Chip*, 9(18), p.2665.

McDonald, J.C. et al., 2000. Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis*, 21(1), pp.27–40.

Melin, J. & Quake, S.R., 2007. Microfluidic Large-Scale Integration: The Evolution of Design Rules for Biological Automation. *Annual Review of Biophysics and Biomolecular Structure*, 36(1), pp.213–231.

Miller, S.L., 1953. A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science*, 117(3046), pp.528–529.

Monnard, P.A., Kanavarioti, A. & Deamer, D.W., 2003. Eutectic Phase Polymerization of Activated Ribonucleotide Mixtures Yields Quasi-Equimolar Incorporation of Purine and Pyrimidine Nucleobases. *Journal of the American Chemical Society*, 125(45), pp.13734–13740.

Mustoe, A.M., Brooks, C.L. & Al-Hashimi, H.M., 2014. Hierarchy of RNA Functional Dynamics.

Annual Review of Biochemistry, 83(1), pp.441–466.

Nghe, P. et al., 2015. Prebiotic network evolution: six key parameters. *Mol. BioSyst.*, 11(12), pp.3206–3217.

Niu, X. et al., 2009. Electro-coalescence of digitally controlled droplets. *Analytical Chemistry*, 81(17), pp.7321–7325.

Nowak, E. et al., 2016. Effect of surfactant concentration and viscosity of outer phase during the coalescence of a surfactant-laden drop with a surfactant-free drop. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 505, pp.124–131.

Obexer, R. et al., 2017. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nature Chemistry*, 9(1), pp.50–56.

Orgel, L.E., 1968. Evolution of the genetic apparatus. *Journal of Molecular Biology*, 38(3), pp.381–393.

Orgel, L.E., 2004. Prebiotic chemistry and the origin of the RNA world. *Critical Reviews in Biochemistry and Molecular Biology*, 39(2), pp.99–123.

Oró, J., 1961. Mechanism of Synthesis of Adenine from Hydrogen Cyanide under Possible Primitive Earth Conditions. *Nature*, 191(4794), pp.1193–1194.

Patel, B.H. et al., 2015. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nature Chemistry*, 7(4), pp.301–307.

Paul, N. & Joyce, G.F., 2002. A self-replicating ligase ribozyme. *Proceedings of the National Academy of Sciences*, 99(20), pp.12733–12740.

Peselis, A. & Serganov, A., 2014. Themes and variations in riboswitch structure and function. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1839(10), pp.908–918.

Pfeiffer, F. & Mayer, G., 2016. Selection and Biosensor Application of Aptamers for Small Molecules. *Frontiers in Chemistry*, 4(June), pp.1–21.

Plaxco, K.W. & Gross, M., 2006. *Astrobiology: a brief introduction* J. H. U. Press, ed.,

Powner, M.W., Gerland, B. & Sutherland, J.D., 2009. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature*, 459(7244), pp.239–242.

Pressman, A., Blanco, C. & Chen, I.A., 2015. The RNA world as a model system to study the origin of life. *Current Biology*, 25(19), pp.R953–R963.

Rajamani, S. et al., 2008. Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Origins of Life and Evolution of Biospheres*, 38(1), pp.57–74.

Reinhold-Hurek, B. & Shub, D. a, 1992. Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature*, 357, pp.173–176.

Riley, C.A. & Lehman, N., 2003. Generalized RNA-Directed Recombination of RNA. *Chemistry and Biology*, 10(12), pp.1233–1243.

Rupert, P.B. & Ferré-D'Amaré, A.R., 2001. Crystal structure of a hairpin ribozyme|[ndash]|inhibitor complex with implications for catalysis. *Nature*, 410(6830), pp.780–786.

Ryckelynck, M. et al., 2015. Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions. *Rna*, 21(3), pp.458–469.

Saghatelian, A. et al., 2001. A chiroselective peptide replicator. *Nature*, 409(6822), pp.797–801.

Salehi-Ashtiani, K. & Szostak, J.W., 2001. In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, 414(6859), pp.82–84.

Satterwhite, L.E., Yeates, J.A.M. & Lehman, N., 2016. Group I intron internal guide sequence binding strength as a component of ribozyme network formation. *Molecules*, 21(10), pp.1–13.

Sciambi, A. & Abate, A.R., 2015. Accurate microfluidic sorting of droplets at 30 kHz. *Lab Chip*, 15(1), pp.47–51.

Scott, W.G., Finch, J.T. & Klug, A., 1995. The crystal structure of an AII-RNAhammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell*, 81(7), pp.991–1002.

Segré, D. et al., 1998. Graded Autocatalysis Replication Domain (GARD): kinetic analysis of self-replication in mutually catalytic sets. *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, 28(4), pp.501–514.

Segré, D., Ben-Eli, D. & Lancet, D., 2000. Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proceedings of the National Academy of Sciences of the United States of America*, 97(8), pp.4112–7.

Serganov, A. & Nudler, E., 2013. A Decade of Riboswitches. *Cell*, 152(1–2), pp.17–24.

Severin, K. et al., 1997. A synthetic peptide ligase. *Nature*, 389(6652), pp.706–709.

Severin, K. et al., 1998. Dynamic Error Correction in Autocatalytic Peptide Networks. *Angewandte Chemie International Edition*, 37(1–2), pp.126–128.

Sievers, D. & von Kiedrowski, G., 1994. Self-replication of complementary nucleotide-based oligomers. *Nature*, 369(6477), pp.221–224.

Sievers, D. & Von Kiedrowski, G., 1998. Self-replication of hexadeoxynucleotide analogues: Autocatalysis versus cross-catalysis. *Chemistry - A European Journal*, 4(4), pp.629–641.

Song, H., Chen, D.L. & Ismagilov, R.F., 2006. Reactions in droplets in microfluidic channels. *Angewandte Chemie - International Edition*, 45(44), pp.7336–7356.

Stan, C.A., Tang, S.K.Y. & Whitesides, G.M., 2009. Independent Control of Drop Size and Velocity in Microfluidic Flow-Focusing Generators Using Variable Temperature and Flow Rate. *Analytical Chemistry*, 81(6), pp.2399–2402.

Sullenger, B.A. & Nair, S., 2016. From the RNA world to the clinic. *Science*, 352(6292), pp.1417–1420.

Suslov, N.B. et al., 2015. Crystal structure of the Varkud satellite ribozyme. *Nature Chemical Biology*, 11(11), pp.840–846.

Szathmary, E., 2006. The origin of replicators and reproducers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1474), pp.1761–1776.

Tabeling, P., 2005. *Introduction to Microfluidics*, Oxford University Press.

Tang, S. & Whitesides, G., 2010. Basic microfluidic and soft lithographic techniques. *Optofluidics: Fundamentals, Devices, and Applications*, pp.7–32.

Theberge, A.B. et al., 2010. Microdroplets in microfluidics: An evolving platform for discoveries in chemistry and biology. *Angewandte Chemie - International Edition*, 49(34), pp.5846–5868.

Umbanhowar, P.B., Prasad, V. & Weitz, D.A., 2000. Monodisperse emulsion generation via drop break off in a coflowing stream. *Langmuir*, 16(2), pp.347–351.

Vaidya, N. et al., 2012. Spontaneous network formation among cooperative RNA replicators. *Nature*, 491(7422), pp.72–77.

Vaidya, N., Walker, S.I. & Lehman, N., 2013. Recycling of Informational Units Leads to Selection of Replicators in a Prebiotic Soup. *Chemistry & Biology*, 20(2), pp.241–252.

Vasas, V. et al., 2012. Evolution before genes. *Biology Direct*, 7(1), p.1.

Vasas, V., Szathmary, E. & Santos, M., 2010. Lack of evolvability in self-sustaining autocatalytic networks constraints metabolism-first scenarios for the origin of life. *Proceedings of the National Academy of Sciences*, 107(4), pp.1470–1475.

Wachowius, F., Attwater, J. & Holliger, P., 2017. Nucleic acids: Function and potential for abiogenesis. *Quarterly Reviews of Biophysics*, 50, pp.1–37.

Westall, F. et al., 2001. Early archean fossil bacteria and biofilms in hydrothermally-influenced sediments from the Barberton greenstone belt, South Africa. *Precambrian Research*, 106(1–2), pp.93–116.

Wochner, A. et al., 2011. Ribozyme-catalyzed transcription of an active ribozyme. *Science*, 332(6026), pp.209–212.

Wu, D., Qin, J. & Lin, B., 2008. Electrophoretic separations on microfluidic chips. *Journal of Chromatography A*, 1184(1–2), pp.542–559.

Yan, C. et al., 2015. Supplementary Materials for Structure of a yeast spliceosome at 3 . 6-angstrom resolution. *Science*, 349(6253), pp.1182–1191.

Yao, S. et al., 1997. A pH-modulated, self-replicating peptide. *Journal of the American Chemical Society*, 119(43), pp.10559–10560.

Yao, S. et al., 1998. Selective amplification by auto- and cross-catalysis in a replicating peptide system. *Nature*, 396(6710), pp.447–450.

Yeates, J.A.M. et al., 2016. Dynamics of prebiotic RNA reproduction illuminated by chemical game theory. *Proceedings of the National Academy of Sciences*, 113(18), pp.5030–5035.

Yeates, J.A.M., Nghe, P. & Lehman, N., 2017. Topological and thermodynamic factors that influence the evolution of small networks of catalytic RNA species. *RNA*, 23(7), pp.1088–1096.

Zagnoni, M., Baroud, C.N. & Cooper, J.M., 2009. Electrically initiated upstream coalescence cascade of droplets in a microfluidic flow. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(4), pp.1–9.

Zagnoni, M. & Cooper, J.M., 2009. On-chip electrocoalescence of microdroplets as a function of voltage, frequency and droplet size. *Lab on a Chip*, 9(18), p.2652.

Zaher, H.S. & Unrau, P.J., 2007. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *Rna*, 13(7), pp.1017–1026.

Zaug, A. & Cech, T., 1986. The intervening sequence RNA of Tetrahymena is an enzyme. *Science*, 231(4737), pp.470–475.

Zonta, E. et al., 2016. Multiplex Detection of Rare Mutations by Picoliter Droplet Based Digital PCR: Sensitivity and Specificity Considerations. *PLoS ONE*, 11(7), pp.1–20.