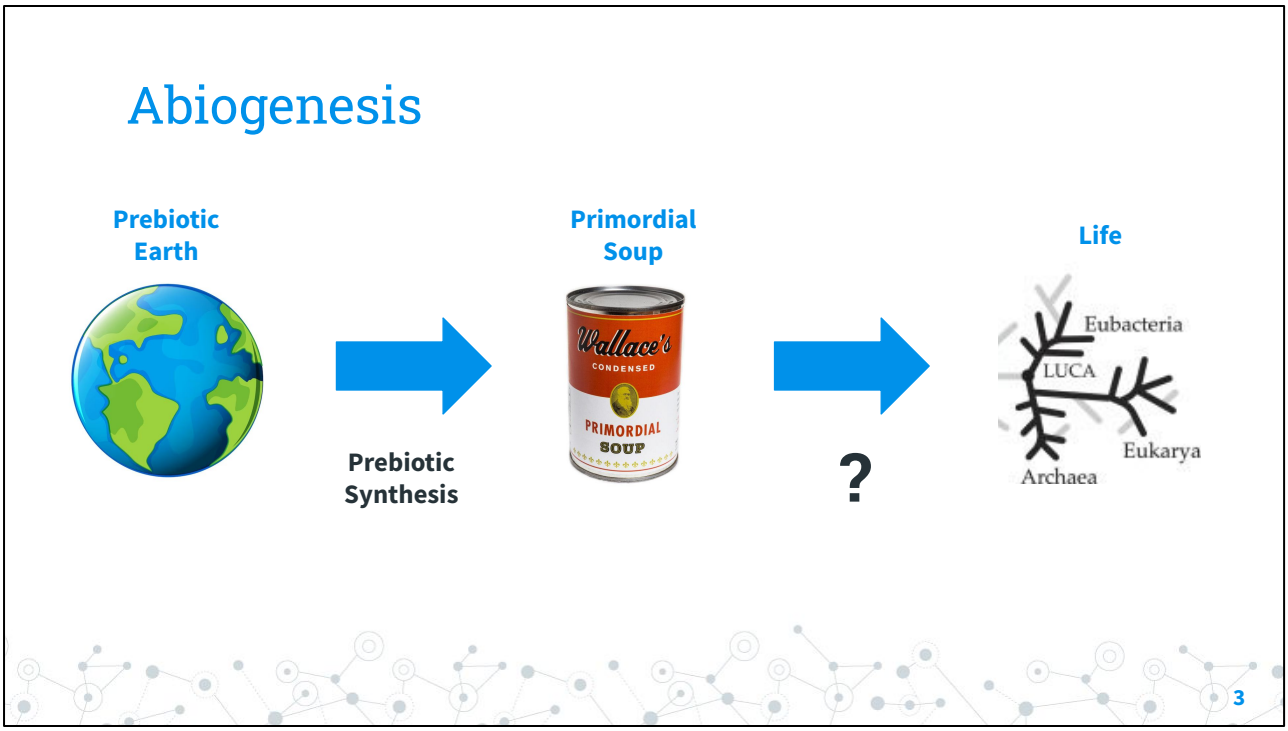
Emergence of Autocatalytic Reaction Networks

My name is Varun and today I'm here to talk about my work this semester with Prof. Jun Korenaga on the Emergence of Autocatalytic Reaction Networks

1.
Motivation

What is the origin of life?

Before we jump into the weeds of my project, I'd like to start by talking a little bit about why it matters and what we are trying to do. At a high level, this project is focused on answering the age old question: how did life originate?



At a high level, the origin of life or abiogenesis can be understood in 3 phases. I will briefly introduce them now, but we will go in more depth in follow slides. First, we have the prebiotic earth. As the name implies, this is the point in earth's history before life. Understanding the atmosphere remains an open research question but current theories characterize it as a "weakly reduced" atmosphere containing CO2, N2, and small amounts of O2. The sun was approximately 30% dimmer with higher amounts of ultraviolet and x-ray radiation than we see now.

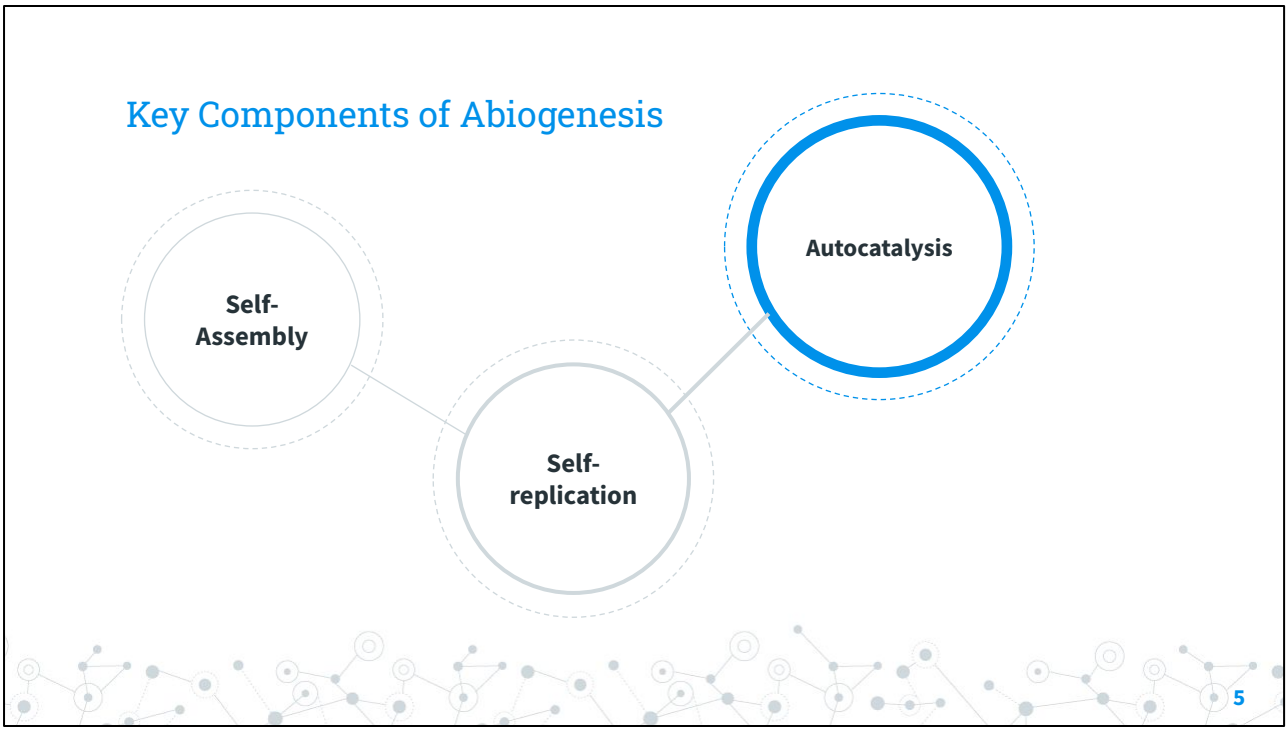
Through a process known as prebiotic synthesis, this phase slowly involved into something we call "primordial soup". Prebiotic synthesis is essential the production of biotic building blocks from traditional abiotic molecules. One of the famous experiments in this field was the Miller-Urey experiment in 1953. In this experiment Miller and Urey simulated the prebiotic atmosphere by applying sparks to a chamber of methane, oxygen, and nitrogen among other prebiotic gases. The experiment found that this simple model was able to produce well over 20 different amino acids, one of the key building blocks in proteins.

Over time as prebiotic synthesis accumulated biotic molecules, we enter a time period known as the primordial soup. This stage is characterized by a large "soup" of essential elements of life. Eventually, from this soup life emerged. In the field we refer to this original form of life as LUCA or the last

universal common ancestor. Essentially, the progenitor of life as we know it. Darwin's theory of evolution gives us a pretty solid understanding of how life as we know it today developed from LUCA, but the how life emerged from the primordial soup still remains an open question. The brunt of my research falls under the broad scope of answering this question.

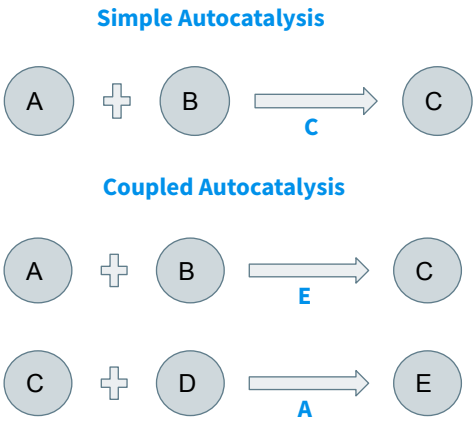
2.
Background

Before we jump into the weeds of my project, I'd like to start by talking a little bit about why it matters and what we are trying to do. At a high level, this project is focused on answering the age old question: how did life originate?



Currently, science's best understanding of this phenomena is through a combination of self-assembly, self-replication, and autocatalysis. My research in particular is focused on the idea of autocatalysis. So what is Autocatalysis?

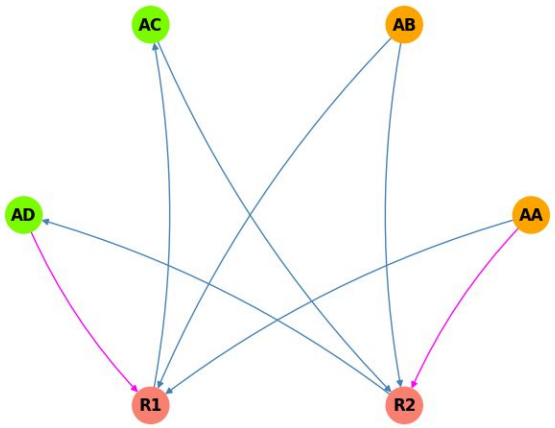
Autocatalysis



Autocatalysis refers to the phenomena where a product of a reaction catalyses itself. In the simplest case, the product itself catalyzes the reaction; however, in more complex examples, the reaction in question could be catalyzed by another coupled reaction. Either way, the main idea is that the reaction eventually produces a species that catalyzes the original reaction. Therefore, as time progresses, the reaction speeds up as more catalysts are produced.

We currently haven't discovered an abiotic reaction that is capable of simple autocatalysis, so research is focused on understanding coupled autocatalysis. In particular, we study these systems through reaction networks

Reaction Networks



We currently haven't discovered an abiotic reaction that is capable of simple autocatalysis, so research is focused on understanding coupled autocatalysis. In particular, we study these systems through reaction networks like the one here. For the time being we can ignore the he color difference between orange and green vertices, but we have two types of vertices. First, we have molecules or species which are represented by the letters AA through AD. Second, we have red vertices labeled R1 and R2 which represent reactions. Arrows going into the reaction vertices refer to reactants while arrows leaving the reaction represent products. So, in this example we have AA + AB produces AC through R1. The last thing to notice are the pink edges. These refer to catalyzing relations. So in this example reaction 1 is catalyzed by AD.

Kauffman Model

- X: All possible molecules
- F: Food Set
- R: All Possible Reactions
- C: Available Catalyzing Reactions

With this in mind, we can introduce the Kauffman Model. Developed by Stuart Kauffman in 1986 this model is a representation of prebiotic chemistry. A substantial portion of work done in this field is built upon or uses this model. I like to think of it as describing the chemistry in a small universe. So with that in mind let's quickly run through the main parameters and what this model is.

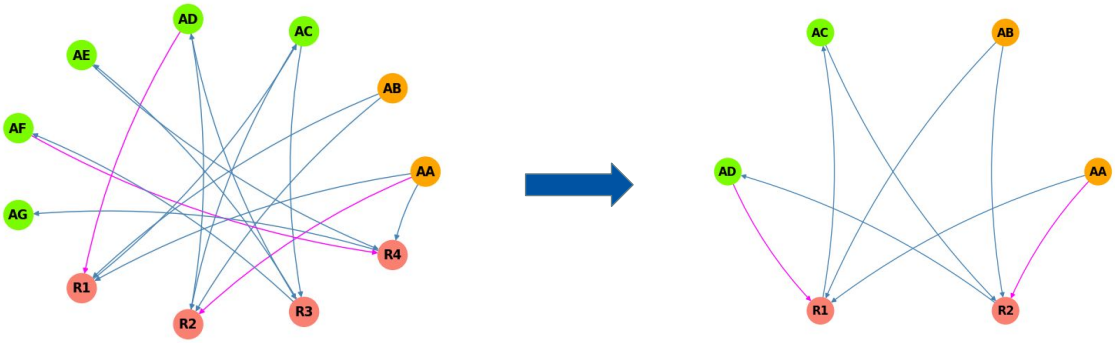
The model essentially defines a set X of all possible molecules. This is usually specified by a max sequence length n and k letter alphabet. For example, if n =2 and k =2, we have all possible combinations up to a length of 2 of A and B. This means A, B, AB, BA, AA, and BB. It's pretty easy to see that as we increase n and k our model size exponentially grows.

The second parameter specified is the food set. This refers to the set of molecules that are "ambient in the environment" We can think of these as existing in a sink. We don't need to produce them for them to be available to react.

The model also specifies a set of all possible reactions. Usually this is done as some kind of concatenation/splitting of molecules. In our earlier example, this refers to A + A producing AA or the reverse of AA splitting to form 2 As.

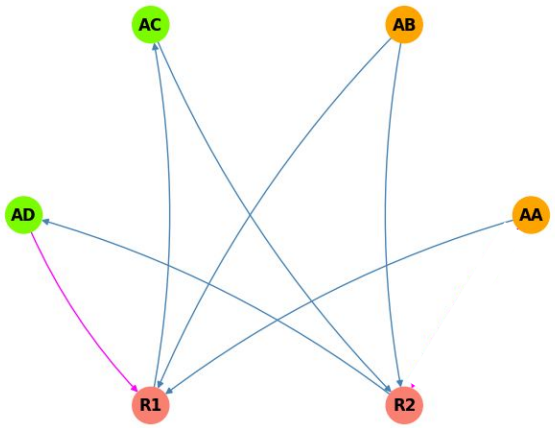
Finally, the model specifies catalyzing reactions. This is simply a list of reactions and which molecules catalyze them.

Autocatalytic Sets and RAF Sets



The kauffman model provides us a framework for chemical reactions, but doesn't tell us anything about our goal: autocatalytic sets. As we talked about earlier, autocatalytic sets are self-catalyzing reactions. Depending on how the kauffman model is set up it can potentially contain autocatalytic sets within it. For example, here we see a kauffman model with an autocatalytic set hidden within it. To add another layer of complexity, we can also require that this reaction set is self-sustaining. I.e. in a given environment the autocatalytic set would be able to survive. One way of ensuring this is to make sure every molecule in the set is either from the food set (the set of ambient molecules) or produced by the food set through a catalyzed reaction. We call such a set RAF which stands for reflexively autocatalytic and food-generated set.

RAF Example



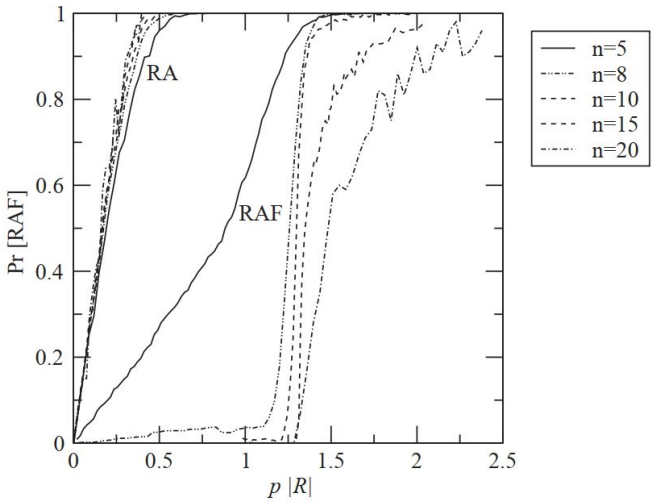
Let's take a quick look at an RAF set. Notice here we have two molecules in our food set: AA and and AB. They combine to form AC. AC then reacts with AB to form AD which catalyzes the original R1. Similarly, R2 is catalyzed by AA. Notice how every reaction is catalyzed by a molecule that is present in our set. Furthermore, every molecule present is either in the food set or produced by them. Therefore, in an environment where we have AA and AB present, we would be able to produce and sustain this reaction network.

Now, let's consider what happens if AA didn't catalyze R2. We see that this network quickly falls apart. Since R2 is no longer catalyzed, it can't produce AD. Since AD no longer exists, we can't catalyze R1 and AC can't be produced. Removing just one catalyzing relationship collapsed this entire network.

3.
Previous Work

Now that we understand the Kauffman model and RAF sets, we can ask the question, how prevalent are RAF sets? How can we study them? My work is aimed at answering these questions, but before we jump into my work, let's quickly talk about what is known about these sets

RAF Dependence on Probability of Catalyzation



From our previous example, it is pretty clear that the presence of an RAF set is dependent on catalyzing relations. With this in mind, we ask the question, what happens as we vary how many catalyzing relations there are in our reaction?

A paper published in 2004 tackled this question and found that if keep the network structure constant (set n) and add each possible catalyzing relation with probability p, we observe that as p increases the probability of RAF increases. More interestingly, they found that a phase transition occurs as we vary this probability parameter p. In fact, when the expected number of catalyzing relations per molecule is around 1.25 we see a jump from almost no RAF sets to a very high probability of finding one.

4.
Emergence of RAF Sets

Finally, with that background we can begin to discuss my work this semester. Funnily enough, a majority of my semester was dedicated to replicating the result on the previous page, but in the last few weeks I have been working on better characterizing the emergence of RAF sets.

So we understand that RAF sets are highly connected to probability of catalysis, but what else? In particular, is there anything about the network structure that gives us insight into the emergence of RAF sets? This is the question I have set out to answer.

Classification

This semester I attempted to answer that question in two ways. First, via classification. In this approach I tried to find other network characteristics that were indicative of the presence of an RAF set.

Classification Outline

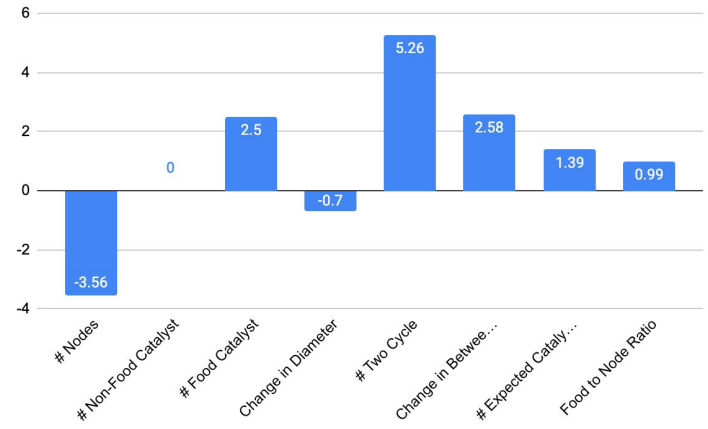
- Data Size: 11,100 Networks (7095 w/ RAF ~64%)
- Classification Parameters
  - # Nodes
  - # Non-food Catalyzation
  - # Food Catalyzation
  - Change in Network Diameter
  - # Two Cycles
  - Change in Betweenness Centrality
  - # Expected Catalyzation
  - Food to Node Ratio
- Classification Model: Logistic Regression

Since we didn't have data to classify, the first step of this process was creating a dataset of Kauffman models and determining whether they had RAF sets present. In total, we created a dataset of size 11,100 networks with a success of RAF presence rate of about 64%. The size of the networks varied between n =2 and n = 7. In addition to testing whether the network had an RAF set, we also checked for graph properties of interest. In particular, we looked at # nodes, …..

Finally, once we had our data created and processed, we ran a logistic regression classification algorithm to see which features were most predictive of an RAF set.

Classification Results

Score: 80%



16

Our logistic regression was decently accurate with a test score of 80%. More importantly, logistic regression yields easily interpretable coefficients and measures of influence for each factor. Since we normalized the data before regression, our interpretation of the the coefficients as log odds is slightly skewed, but we still get a reliable ranking of their relative importances.

As you can see in the chart our three most prominent features are #two cycles, change in betweenness centrality, and number of food catalysts. Additionally, we see that the number of nodes is inversely related to the presence of an RAF set. Furthermore, we see that #two cycles is about twice as influential as either food catalysts or change in betweenness centrality.

Classification Analysis

- Simple Autocatalytic reactions are highly influential in RAF sets
- Proximity to food set is important for RAF sets
- Probability of RAF decays with larger models
- Catalysts from the food set are relatively more important than other catalysts

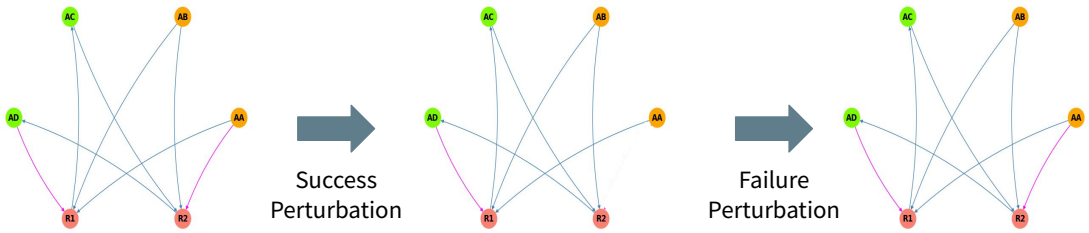
So what are our takeaways from this approach?

Stability Testing

From previous work in the field, we see that there is a probability at which RAF occurrence is about 50%. So what determines success or failure here? In this phase of the project we analyzed this regime to understand the underlying dynamics of RAF occurrence. In particular, we investigated how stable each of these failure/success configurations were.

RAF Stability

- Success (RAF)
  - Perturbation given by random removal of catalyst
- Failure (No RAF)
  - Perturbation given by random addition of a catalyst



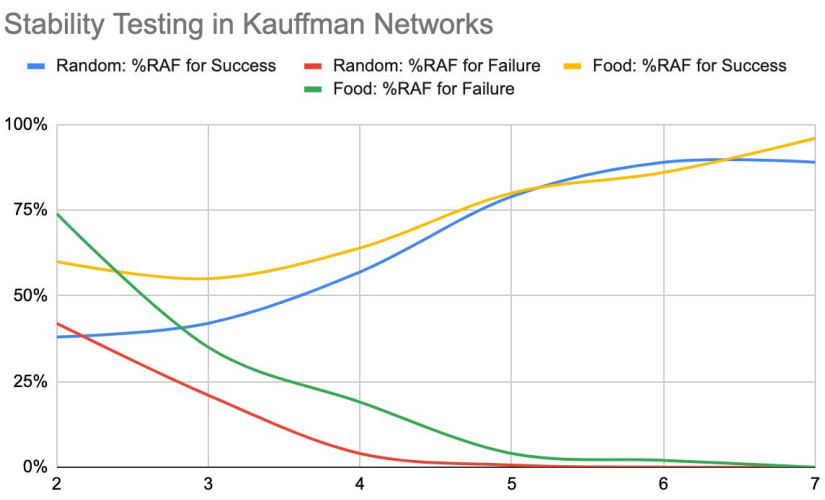
An important question here is what we mean by stability. In our work, we define stability of an RAF set as its ability to remain RAF after removing a catalyst. Similarly, in cases where an RAF is not present, we define the stability of the position as its ability to become RAF with the addition of a catalyst. Essentially, we define stability as the ability to maintain or become RAF given a perturbation of catalyst.

Stability Testing Simulation

Max Molecule Size (n)	# Trials
2	5000
3	2500
4	1000
5	1000
6	500
7	300

In testing stability, we found the probability value that yielded about a 50% chance of success and then ran varying numbers of trials to see how the networks responded to perturbations. Additionally, given our findings in the classification approach, we attempted a second perturbation regime where only food-based catalysts were added or removed.

Perturbation Results



As we can see here, as we increase n the influence of the perturbation decreases. We can also see that the food set perturbation is more influential than a random perturbation which corroborates our findings from the classification analysis. Furthermore, we can see that the influence of the food set perturbation is more pronounced at low n and both level off as n increases. It appears that the random perturbation for success cases approaches an asymptotic bound, but more testing is needed to confirm.

Stability Analysis

- Influence of food-set catalysts is pronounced in smaller networks
- As network size increases, the stability of the success/failure case improves
- Potentially an asymptotic bound on the probability of remaining RAF given a random perturbation

So what does this tell us about our networks?

Takeaways

- Autocatalytic/RAF sets are important in our understanding of abiogenesis
- The emergence of RAF sets follows a phase transition based on a probability parameter p
- The presence of an RAF sets is influenced by proximity to the food set
- Influence of the food set becomes less pronounced as the network size grows

Future Work

- Testing how many peturbations are required to change the RAF success/failure state of the network
- Further analysis/classification of graph properties
- Relaxing the reflexive catalyst assumption

Thank you!

Questions?