Genetics based AGI

Tom Rochette <tom.rochette@coreteks.org>

December 21, 2025 — 77e1b28a

0.1 Context

0.2 Learned in this study

0.3 Things to explore

- DNA is the software of life. If that is true, who wrote the code?
- How does the reproduction of cells (of an embryo) works in term of computation?
- If DNA is considered as the storage/tape of a Turing machine, can be it considered to be expandable? What are the other similar properties of DNA and Turing machines?
- Isomorphism between DNA and programs
- DNA is code, and it most likely didn't start the length it is now.
  - In order to lengthen, smaller parts could have merged together (working programs merged/concatenated to one another)
  - https://en.wikipedia.org/wiki/Origin_of_replication (What are the origin of DNA?)
- Can we build a source tree ala git that would explain our evolution?

1 Overview

One may extrapolate that the big bang is similar to the generation of random code. Everything that followed it is simply random permutation/mutation of the randomness that ended up into something that is coherent/structured. Like a well programmed neural network, with enough time, randomness will at some point have to generate patterns. However we know from the study of undirected/blind (as opposed to directed/guided) program generation, such as through a linear method, that the space of programs strings (chain of character that may or may not produce an executable program) is immense in comparison of the space of valid programs. It can easily be compared to the problem of finding a needle in a haystack.

In a similar way, given hundred of thousands of valid programs, if we want to find a particular one that does a number of specific things, then we're making it harder for a search/filtering algorithm to return us an appropriate one. If one can express what he wants the program to do in a boolean fashion, such as "the program does/does not do this", then for every expression, the number of potential program doubles. For instance, if you want 8 specific things, then there are 2^8 potential programs, but only 1 that does what you want. You want to add 1 more thing? Then you've effectively doubled the number of potential programs, while still looking for a single specific program.

We, as human beings, are a gigantic assembly of billions of cell-sized machines. Each and every cell contains its own copy of the program (DNA) executed by each and everyone of these machines and which is itself about 3 billion nucleotides. As there are 4 valid nucleotides/base, there are $(2^2)^{(3\times10^9)}$ possible combinations/programs. However, considering that we (humans) all share a ton of similar attributes (we all have two eyes, two ears, two arms, two legs and so on), it makes sense to assume that a lot of our DNA code is shared.

In the body, or more specifically in each cell, the DNA is used as the source from which ARNm is transcribed and then translated into a protein.

Our genetic software (DNA) itself only changes/evolve through the combination of two parent chromosomes.

Some properties:

- Code rarely changes (only when a new "program"/human is created)
- New code is the combination of two existing codes
- Two code bases are mostly similar (99.5% similar[1])
- A large amount of DNA is shared with other animals[2], which could imply that we either developed shared code base (evolved from similar ancestors) or that we ended up developing similar code bases independently
- A certain amount of DNA is considered noncoding, meaning that they do not "execute" into proteins, and can be considered to be passive data
  - A certain amount of this noncoding DNA is considered Junk DNA, the equivalent of dead code/data
- The 5' and 3' sections of the mRNA could be compared to the preambule and epilogue of functions in assembly, they serve to indicate the beginning and end of blocks of information/instructions
- If DNA is a string, then it most likely has a grammar (and its own language)
- Must follow some syntactic rules or else it is incorrect (see protein folding)
- Evolution is Nature's nondeterministic way to test out DNA machines, some survive (are born, live and die of old age), others don't (are not born)

2 Reasoning on AGI

Given that programs are part of the integer space (with their length in the same space), one can be certain that a program within this infinite space will be considered an AGI. It is highly likely that many such programs exists. In fact, if one such program exist, an infinity of its variants will exist as well (longer programs containing this "seed AI"). Assuming that programs containing this seed AI code can also exhibit the same functionalities, the assumption that an infinity of those program exists holds. Otherwise, it means that even though the seed AI code is present within the program's string (the integer represented as a sequence of integers in the 0-9 range), it cannot be activated/executed. For example, given a C program that is a seed AI, any junk at its beginning or end may render the program uncompilable (or unexecutable given the integer could be a executable binary). For the sake of analysis, we'll prefer to work within a language that considers this seed AI string as active wherever it is found.

If we accept that "a seed AI program exists" as a fact (human beings being an instance), then the obvious next question is "what is the length of this shortest seed AGI?" The answer is likely to be language specific. For instance, our DNA is believed to be the equivalent to Nature's AGI program. DNA is itself about 3 billion nucleotides. As there are 4 valid nucleotides/base, there are $(2^2)^{(3\times10^9)}$ possible combinations/programs, a single program being approximately 3 Gbases, 6 Gbits or 750 MBytes (approximately 3.75 MBytes being different between individuals). What those 750 MB of code and data allow us to do is to construct a huge variety of cells/proteins that end up having lives of their own.

If we take this amount of information as a basis to determine the size of a potential human-like AGI, we have to ask ourselves if what we "really" need is a subset of this information, or all of it is needed. In the former case, then we can hope to reduce our search space considerably, in the latter, it means that we at least have an upper bound for something that should produce human-like intelligence levels, given the appropriate environment simulation.

This "upper bound" or threshold has a couple of interesting properties. Let's consider the smallest AGI being a program of length l. This means that for all programs p smaller than l, in other words $|p| < l$ (where $|p|$ is the length of program p), the probability that we execute a program p_{AGI} that is AGI is $P(\text{p exhibits AGI} \mid |p| < l) = 0$, in other word we will at best observe sub-AGI intelligence but not AGI itself.

On the other hand, for any program larger or equal to l, we may assume that it is sufficient for a program p to contain the program p_{AGI} somewhere in its string definition. In other terms, if this program p contains the substring (from index a to b) $p_{a,b}$ that is the AGI program p_{AGI} ($p_{a,b} = p_{AGI}$), then $P(\text{p exhibits AGI} \mid |p| \geq l \wedge p_{a,b} = p_{AGI}) = 1$. Finally, we may ask ourselves what is the probability of finding an AGI program,

[1] https://en.wikipedia.org/wiki/Human_genetic_variation
[2] http://education.seattlepi.com/animals-share-human-dna-sequences-6693.html

given a program of length $|p|$ and a know seed AGI program p_{AGI} which is a subprogram/substring of p, $P(\text{p exhibits AGI} \mid |p| \geq l \wedge p_{AGI}) = ?$. More interestingly, we can ask what is the probability of finding an AGI program, given that we "know" the minimal program length of an existing AGI but do not have the code, $P(\text{p exhibits AGI} \mid |p| \geq l) = ?$.

Since we said that for a program to exhibit AGI it would have to contain a seed AGI as a substring of itself, we can simplify p exhibits AGI as $p_{a,b} = p_{AGI}$, in other words, let p be the shortest AGI program. $P(\text{p exhibits AGI} \mid |p| \geq l \wedge p_{a,b} = p_{AGI}) = P(\text{p}_{\text{a,b}} = \text{p}_{\text{AGI}} \mid |p| \geq l \wedge p_{a,b} = p_{AGI}) = 1$ is now obvious, since the evidence contains $p_{a,b} = p_{AGI}$. One of the questions we asked becomes $P(p_{a,b} = p_{AGI} \mid |p| \geq l) = ?$, which means "given that our program p is longer than an expected seed AI of length l, what is the probability that a part of its code (a substring) is p_{AGI}?" As the program length l increases, the probability decreases.

2.1 Observations

- If we assume there is only 1 program of length l that exhibits AGI, then all "variants" of the programs must be of length $l_{variant} > l$, in other words they must be at least one symbol longer and containing code that isn't part of the AGI program (dead code, similar to "junk DNA").

2.2 Applications

- Given that we established there are $(2^2)^{(3\times10^9)}$ potentials programs of the same length as the human genome that also appears to generate variants of similar programs, namely different individuals with varying capabilities, and that we estimate there have been about 108 billion individuals that lived so far (assuming they all had unique DNA and that their DNA was of this exact length, which isn't the case)[3]

$\frac{108\times10^9}{(2^2)^{(3\times10^9)}} = \frac{108\times10^9}{4^{3\times10^9}} = \frac{1.08\times10^{11}}{9.6357\times10^{1806179973}} \approx 10^{-1806179962}$

3 Questions

- Is the DNA/genome the same length for all individuals?
  - If it is the case
    - How is that possible?
    - Why does it have to be the exact same length?
  - If it is not the case
    - What is the impact of the missing/added parts?

Answer: DNA length varies amongst individuals. This is mostly due to the large amount of non-coding DNA.

> A major discrepancy in DNA length would cause infertility even if we assume it does not cause somatic defects.

Source:

- https://www.quora.com/Do-all-members-of-a-species-have-the-same-length-of-DNA-in-matching-chromosomes
- https://www.quora.com/Do-all-people-have-the-DNA-of-exactly-the-same-length
- https://en.wikipedia.org/wiki/Variable_number_tandem_repeat

4 See also

- Theory of self-reproducing automata

[3] http://www.prb.org/Publications/Articles/2002/HowManyPeopleHaveEverLivedonEarth.aspx

5 References

- https://en.wikipedia.org/wiki/Chromosomal_crossover
- https://en.wikipedia.org/wiki/Gene
- https://en.wikipedia.org/wiki/Genetic_recombination
- https://en.wikipedia.org/wiki/Human_genome#Information_content
- https://en.wikipedia.org/wiki/Messenger_RNA
- https://berthub.eu/amazing-dna/
- http://www.hxa.name/articles/content/genetics-basics_hxa7241_2003.html
- Ji, Sungchul. "The linguistics of DNA: words, sentences, grammar, phonetics, and semantics." Annals of the New York Academy of Sciences 870.1 (1999): 411-417.