On Information

Our goal is to invent a language that somehow "adequately" defines molecular information, and by extension can be used to define bio-information. So, we start with something seemingly simple: what is information? There are just too many different definitions, but "information entropy" is generally thought of as the uncertainty of knowing a random variable, where information entropy $H = \log_2(n) = 1$ bit for $n=2$ (a fair coin toss) or 3 bits for a fair roll of an 8-sided die, etcetera. Information in this sense is the relationship between a particular choice and all possible choices, or more precisely, it is the probability that a particular choice is made.

Thermodynamic entropy uses similar equations but involves slightly different concepts. Thermodynamic entropy is the tendency of a physical system to smooth out any differences between specific states and an average state of randomly behaving sub-units. The gist of thermodynamic entropy is that we cannot really ever know any particular choice of specific configurations, but all specific instances of this sort can be ignored, or considered informatively equivalent. Information entropy, on the other hand, gives us the precise values for each possible choice. These concepts are similar, but in some ways opposite.

If we imagine a Gaussian distribution or bell curve we might say that thermodynamic entropy is the tendency of a system to seek the middle of the curve. Information entropy gives us the probability value of having landed at any specific place on the curve. Note that any particular system will exist somewhere on the curve at any given instant, and outliers are possible, but they cannot be expected to persist for any significant length of time; otherwise, they can no-longer be considered random. Knowing the precise state of any system at an instant in time provides a measure of "information gain" that will be found in the difference between the middle of the curve and the state of the system at any particular instant. This still isn't very useful.

We might now generalize these two different concepts of entropy and say that they each involve a comparison or a relationship between two things, the first thing being a specified subset of the second thing. We might also define the word "meaning" as any relationship that persists in a system that differs from a totally random relationship. So a single fair coin toss has no meaning in and of itself. However, if we bet $1 on one toss, then heads means a $2 swing compared to tails for each of us, but now all of the "meaning" here lies in the bet and none in the coin itself. The coin is merely a meaningless cog in the larger "meaning machine" or the process that leads to some particular meaning in this particular simple system. We can expect, however, that if we continue tossing the coin long enough within this simple system that no money will ultimately change hands. It is not an inherently meaningful system in and of itself. However, if somebody were to win a bunch of money, we'd surely want to check the coin. Keep in mind, not all coins are equal.

So, now onto something hard: what is life?

Life is basically a rigged game. It is large and unimaginably complex, but at bottom it is nothing but a rigged game played over a long long long time. There currently is no accepted definition of life, but there is a vague human understanding of when

something is alive and when it is not. We might choose to see it as a binary state between Alive and Not Alive. So at the level of an organism there is now only one bit worth of information in being alive. (It hardly seems worth it.) However, if we took a single cell and all of the atoms in it, we could count the number of ways that we might rearrange all of the atoms in that cell and yet still consider it to be alive. There obviously are a huge number of ways that it could be arranged and still be considered alive. We might, for instance, rearrange the atoms in such a way that this one cell could change species, or phyla, or even change kingdoms – lots and lots of choices. We could also rearrange the atoms in many ways that would hardly even draw our attention – purely random, "meaningless" variations. However - and this is the important part - for all the ways that the atoms of this cell might be arranged into a living cell, there are virtually infinite numbers of ways to group each of them with arrangements not considered to be living. In other words, for each and every way there is to be Alive there are an infinite number of ways to be Not Alive. The unavoidable conclusion is that the meaningful information difference between Alive and Not Alive is considerable. It is perhaps just one bit – the Alive bit - but the meaning behind that one bit is somehow nearing infinity. Yet the meaning of that bit and of the cell itself is not contained in the atoms per se but somehow in the systems that arranged them in the first place. One cell, in isolation, is once again meaningless. The meaning of all life, so to speak, at the highest level can only be derived by comparisons of Alive bits for all cells through all of time. After all, as complex as the relationship between atoms of one cell may be, there is equal complexity in the relationship of all cells to each other. Life is layered meaning in layered complexity.

One might say in these particularly crude terms that the ultimate purpose of life is to generate more meaning and higher levels of meaning through time, or what I have been calling an accelerated process of information accumulation. Bio-information is increasing at accelerating rates on earth. Systems of meaning are built and leveraged to build higher-level systems of meaning. Negentropy!

So now we have sketched out the basic idea that there is information and there is meaningful information. Information is the probability of knowing something precisely and meaning is perhaps the value of knowing it. For instance, if you told me the particular arrangement of molecules in a gas at a particular time it would take a tremendous amount of bits to do so in any information system we might imagine; however, the value of knowing it would be nil - meaningless. After all, I could recreate the equivalent information with a few bits and a good random number generator because all of the gas molecule configurations are equivalently random, as far as the functional time course of any gas is concerned. This is true unless we were to repeat the descriptive process and find that the system persists in deviating from a random distribution in some particular way. The gas itself is still meaningless, but the consistent deviation will surely catch our attention. Perhaps there is "meaning" in the consistent deviation, or perhaps meaning can somehow be derived from it.

By way of example, if you told me that you found a coin that had been flipped 100 times and it had come up heads 96 of those times, the information entropy of the coin goes down considerably – from one bit to a fraction of a bit - but the meaning of this particular coin goes way above an average coin. It is potentially a highly meaningful coin. Who wouldn't want this coin? So, in some ways meaning and information entropy

can be inversely related. It depends on whether or not there is a system in place to cash in on the meaning of an unfair coin.

Hemoglobin should never happen in a universe full of nothing but fair coins. Hemoglobin never will happen in a universe devoid of any mechanism able to pick an unfair coin from a pile of fair ones.

Life looks for and cashes in on a rigged game whenever it can. Life is a collection of highly biased sets. It is the extreme biases of life that makes life alive. There are inherent differences on the most basic levels between sets of living molecules and sets of random molecules. A random gas and a random bio-molecule are hardly comparable even though they both obey thermodynamic laws. Now think of "all possible proteins." In this imaginary domain life obviously plays with a stacked deck. There is a huge conceptual difference between a random protein and any protein that life might ever come up with. The proteins in life are exquisitely biased in every way imaginable. Life uses random but life is never purely random. A salt crystal is the direct result of purely random atomic collisions, but how random is a salt crystal? As a computer programmer, we can call a function, call it "random()" but we know that there is no such thing as a random number in a computer. That's how life programs. Life calls random() all the time, but there is very little that is actually random about life in the same way that there is very little that is random about salt.

In this context we can merely think of the molecules of the genetic code as a breathtaking collection of unfair coins. Ask yourself this: in an environment of total uncertainty, what could have more meaning than certainty? How do molecules "know" what to do? How do molecules read a spreadsheet? Where did they get the spreadsheet in the first place? Life has been busily flipping and selecting a shiny pile of unfair coins on an unimaginable level of nested complexity. Fortunately, at the very core we can now see the game rather clearly. We can even paint a three-dimensional picture of it, but we still can't put it adequately into words. Sometimes words fail. It's a good thing we've got pictures. Think of it this way: the genetic code brings with it the certainty of making proteins and the certainty of finding new proteins. The genetic code therefore packs an extraordinary amount of inherent molecular meaning. What on earth is more efficient and meaningful in this context? How could the current model of the genetic code portrayed as inefficient and arbitrary be any less accurate? The genetic code is surely the most efficient and least arbitrary thing on the planet. Name a close second.

Hopefully, by now we can start to get a feel for the difficulties and complexities involved in defining any type of information, the most difficult type being bio-information. Yes, it is pure semantics, duh, but that is the point of any language. In this case there exists a hierarchical relationship of various complexities of information. There is physical, chemical, structural, functional, systemic and organism level or survival level information. At the very highest level of information complexity lies all of life. It represents a single level of information on earth: bio-information. All bio-information is related at some level, but how?

Temperature is a good example of information at a purely physical level. The kinetic states of a solid, liquid or gas contain an abundance of information but very little

meaning. A temperature is just one generic kinetic state relative to all other levels of generic kinetic states. Chemical information is well represented by pH, or equilibrium equations and binding constants. There is an infinitude of equivalent ways for water to have pH = 7, but the only "meaning" is in the ratio of H^+ and ^-OH. Structural information begins to enter the bimolecular domain of DNA and protein. Functional information is less a matter of precisely how something gets done than a matter of if it gets done, how fast, how much, and how reliably. Systemic information is the gray area between functional and cellular. Survival is the gray area between unicellular and multi-cellular. Life subsumes all else. What's the general equation?

Now note that at every level there are buffers. A buffer is a resistance to change from a set point. Life embraces change in general but greatly resists change toward becoming un-alive. We might then say that being Alive means that something contributes to the increase in bio-information. If we look again at the example of all the atoms in a cell, perhaps we could pick a single arrangement of all atoms and claim that it is the most representative arrangement of being Alive. All of the other arrangements would stray from this is some regard, but life is able to buffer bio-information in such a way to keep the cell close to the "ideal" arrangement of Alive, The cell generally does not change from being Alive to being Not Alive despite innumerable possible and unpredictable perturbations. Bio-information basically buffers itself from becoming non-bio-information. Further note that the general trend of life through time appears to be one of buffer enhancement. In other words, with time, life becomes more robust at every level. This could be said of each individual scale and of all scales taken together.

What then could be more informatively robust than the genetic code? Perhaps this is why superficial variations of it are ubiquitous and profound variations of it are rare. In other words, the concept of last common ancestor is meaningless, yet the concept of a first common ancestor is not. We all share ancestors at some level, but few of us share the exact same ancestors, and such cases are trivial anyway. What cell on earth today will claim all descendants on earth tomorrow? What cell on earth yesterday could do the same today? Why would we ever visualize sharing the exact same single-cell ancestor on any conceptual level? It is absurdly illogical. It is contrary to any reasonable foundation for properly understanding the basic processes of life. We can, however, share common components of systems of all types. It is components not ancestors that we all share. That will perhaps be the topic of my next rant.

Nonetheless, at some point we need to realize that the vast majority of the words we are using in biology are utterly backwards. At some point we need to find some new words. That's perhaps why we need a new language.