What is Life?

David Martin Degner
Degner Scientific and Engineering
Anchorage, Alaska
davidmartindegner@gmail.com

A static and dynamic physical model is presented for the Gram (+) prokaryotic cell, the hydrogen of biology. It is discovered that cells are chemical computers and the fundamental Turing machine architecture of biology is laid out.

1 The Nucleic Acid Mainframe 3
2 The Constant Code and the Thermal Code 7

1 The Nucleic Acid Mainframe

Abstract: The condensed DNA of prokaryotic cells forms a separate phase, the nucleoid phase, surrounded by the cytoplasm phase. NTP monomers flow into the nucleoid from the cytoplasm and single strand RNAs flow out of the nucleoid into the cytoplasm. I call this vectorial flow process the **nucleic acid mainframe**. While inside the nucleoid the newly synthesized RNAs cannot fold to form secondary and tertiary structure due to restriction by the DNA matrix. At the surface of the nucleoid single strand RNAs can fold to form secondary and tertiary structure and interact with proteins and ribosomes in the cytoplasm. I call this dynamic surface region where RNAs are folding and interacting with proteins and other RNA the **RNA processing zone**.

A precise physical description of a cell would be to specify the structure and position of every atom, ion, and molecule and their evolution over time. The position means position in space and changes over time. In biology classical space and classical time are fully able to define position and motion. The structure is the quantum mechanical structure. Atoms, ions, and molecules exhibit two phenomena of primary interest – diffusion and reaction. Atoms, ions, and molecules diffuse slowly to a specific configuration, such as a small molecule to the active site of a protein, and then a reaction occurs fast. Diffusion and reaction are the game of cells.

The most important level of organization of the cell is the phase organization. The definition of phase I use is a region of space that has constant composition net of thermal fluctuations. The information of a phase is the size and shape of the phase in space and the list of concentrations of the component atomic and molecular species. Transport in a phase, in the absence of a macroscopic electric or magnetic field defined on the phase, is strictly limited to random walk diffusion. Transport between phases can be vectorial, non-random, into one phase and out of the other phase. The transport between phases in cells is vectorial.

The gram (+) prokaryotic cell, the hydrogen of cells, is composed of 4 phases:

1. nucleoid
2. cytoplasm
3. membrane
4. cell wall

Cell phases have different inside and outside surfaces. The nucleoid has no inside surface, it is at the center, and has a nucleoid/cytoplasm interface. The cytoplasm has an inner interface with the nucleoid and an outer interface with the inner membrane surface that are different. The membrane has inner and outer surfaces that are different. The cell wall has inner and outer surfaces that are different. So there is organization over and above the strictly phase definition. This enables and reflects vectorial flow of ions and molecules between phases.

A fully quantitative description of the nucleoid cannot yet be made. The size and structure of the nucleoid, the DNA supercoil density and dynamics, the DNA helix/coil equilibrium, the number of RNA polymerases, the number of H, HU, and polyamines, the role of protein/DNA binding, protein/DNA bending, DNA/RNA hybrids, triplexes, etc., are not known with quantitative precision. It is a complex system. In this system resides the Turing machine of life. The DNA, the finite one-dimensional list of the Turing machine, is searched to determine a transcription profile. The transcription profile is simply the list of RNAs being synthesized at an instant in time. What is the spatial/temporal description for this system?

What is the composition of the nucleoid? Ions, small molecules, proteins, and RNAs partition between the cytoplasm and the nucleoid. The replication machinery and transcription regulatory proteins are in the nucleoid but here I am going to ignore them. Leaving out the replication machinery and transcription regulatory proteins the composition of the nucleoid is a relatively simple four-part system composed of DNA, RNA polymerase, NTP monomers, and freshly synthesized RNA single strand. The RNA polymerase holoenzyme, $\alpha_2\beta\beta'\sigma^{70}$, is big, $100\times100\times160$ Å, and massive, 450KDa, the same mass as ~ 680 basepairs of DNA. There are 10^{3-4} RNA polymerases per nucleoid. A significant fraction of the volume and mass of the nucleoid is RNA polymerase. In an E. coli with 3000 polymerases and 4.6×10^{6} bp of DNA the mass of the RNA polymerases is 44% of the mass of the DNA. Because the polymerase is compact and the DNA is extended the polymerase is a smaller volume fraction of the nucleoid than mass fraction. A typical bacterial nucleoid contains 10^{6-7} base pairs, and 10^{3-4} genes. In the nucleoid are several types of DNA-binding proteins and small molecules such as H, HU, and polyamines that stabilize supercoils. In the nucleoid are the topoisomerases that put in and take out supercoils. In the nucleoid are the dNTP and NTP monomer pools. Mg^{++} is in the nucleoid counterbalancing all the phosphate negative charge. Not in the nucleoid are all the translation and metabolism machinery found in the cytoplasm.

The DNA in the nucleoid is often, but perhaps not in all bacteria, negatively supercoiled. Whether negatively supercoiled or not the DNA in the nucleoid of bacterial cells is highly condensed. The forces that give rise to the nucleoid phase are not known. I think a significant force for condensation is that between the charged phosphates of the backbone and Mg^{++}. NTP and dNTP monomers carry significant charge and are closely associated with Mg^{++}. Proteins with positive charge that can interact with the negatively charged phosphate backbone partition into the nucleoid. Typical negative supercoil densities predict supercoils every $100-300$ base pairs. The supercoils are no smaller than 110 Å in diameter. There are on the order of 10^{4-5} supercoils in a nucleoid with 10^{6-7} basepairs DNA.

Where does transcription take place? The RNA polymerases are distributed randomly throughout the nucleoid. Transcription takes place inside the nucleoid at random sites throughout the nucleoid. The idea that transcription takes place on the surface of the nucleoid at the interface of the nucleoid and cytoplasm is not physically plausible. The RNA polymerases open up a hole in the nucleoid DNA matrix and are confined to a fixed position in the nucleoid by contacts on all sides with the DNA matrix. The DNA contacts exert a pressure on the surface of the RNA polymerases. During transcription elongation the polymerase must remain fixed in place and the DNA translates through the polymerase/DNA/RNA ternary complex. The RNA polymerase is very massive so it does not translate, rather the DNA that has low linear mass density must account for the relative motion between RNA polymerase and the DNA. Since the DNA is a helix it must also twist though the fixed polymerase/DNA/RNA ternary complex. Collapse of a supercoil in front of the polymerase/DNA/RNA ternary complex coupled to formation of a supercoil behind the complex allow the DNA to twist through the fixed complex when transcribing in one direction and a vice versa migration for transcription in the opposite direction. The supercoil density remains constant net of thermal fluctuations but the supercoils are moving around on the DNA.

The RNA single strand leaving the polymerase/DNA/RNA ternary complex does a random walk to the surface of the nucleoid. While on this random walk it is prevented from folding by the nucleoid DNA matrix. When the single strand RNA gets to the surface of the nucleoid it can fold into secondary and higher structures. NTP monomers do a random walk in the nucleoid to the polymerase/DNA/RNA ternary complexes where synthesis is taking place. Overall two vectorial flows occur between nucleoid and cytoplasm associated with transcription: 1. NTP monomers diffuse into the nucleoid from the cytoplasm. 2. Single strand RNA diffuses out of the nucleoid into the cytoplasm. I call this flow system the **nucleic acid mainframe**. It is what is called in non-equilibrium thermodynamics a dissipative structure. The high energy of the NTP monomer pools is converted to lower energy RNA polymers and phosphates. At the interface between nucleoid and cytoplasm is where the RNA single strands fold to form secondary and higher structures and can begin to interact with proteins and ribosomes in the cytoplasm. I call this dynamic surface region that arises when transcription is taking place the **RNA processing zone**.

There are important molecular evolution and origin of life implications of the nucleic acid mainframe model. Starting with DNA and counterions the nucleoid phase can self-assemble and NTP monomers spontaneously partition into the nucleoid. Polymerization of RNA using non-coded for protein catalysts as crude RNAPs can lead to primitive transcription. The associated transcription initiation models, The Constant Code and the Thermal Code, have implications to a primitive transcription system. Of course, given the complexity of a cell, it is always a bit difficult to understand how such a system could arise.

The nucleic acid mainframe is not available in current in vitro systems. The nucleoid is a highly concentrated and exquisitely balanced system only found in vivo. The concentration of biomolecules is much higher in vivo than in vitro. The forces giving rise to the nucleoid are unknown. It may be difficult or impossible to assemble and run a nucleoid in vitro, then again, it may not be that difficult starting out with a qualitative model.

2 The Constant Code and the Thermal Code

Abstract: The **constant code** model postulates the existence of a transcription initiation/rejection of initiation code based on DNA sequence alone at constant supercoil density. The **thermal code** model postulates that the state of the cell is encoded in the list of NTP monomer and modified monomer concentrations. The thermal code model further postulates that in the nucleoid of prokaryotic cells interaction of NTP monomers with denaturation bubbles in the DNA lead to dynamic, hybrid helices of monomers/DNA single strand. These hybrid helices are thought to be of length $3-12$, on one or possibly both strands of the denaturation bubble, and located between the -35 upstream homology and the -10 Pribnow box. In the thermal code scenario these structures are thought to be the critical intermediate for transcription initiation in the absence of regulatory proteins. The thermal code is coupled to the binary fission growth cycle.

The general transcription regulation problem in bacteria is how to turn on and off 10^{3-4} genes in a precise quantitative way. There is both the problem of setting the global transcription rate by a nucleoid and the differential transcription pattern. The time period of regulation spans at least $3-4$ orders of magnitude from initiation every ~ 1 second to initiation every 10^{3-4} seconds.

How the DNA is searched to arrive at the transcription pattern determines how powerful of a Turing machine the cell is. The key reaction is transcription initiation. The transcription pattern is determined by differential transcription initiation for different genes. Therefore transcription initiation determines how powerful a computer the cell is. How is DNA searched to arrive at a transcription pattern?

Replication, transcription, and recombination require unwinding of the DNA double helix. In all three processes one or both strands must have the H-bonding faces of the bases rotated out from the helix configuration to facing into the surrounding solution. This is required for the DNA single strand to serve as a template where it can H-bond with incoming bases either as monomers in transcription and replication or with another single strand DNA in recombination. Where on the genome is the DNA helix unwound and why? Is it only unwound in combination with proteins?

Consider collisions between RNA polymerase holoenzyme and DNA. Of course, collisions between polymerase and DNA are complex and not simple like a collision of billiard balls. There are two possible outcomes to a collision – initiation or rejection of initiation. At constant supercoil density there are five possibilities that can determine collision outcome between RNA polymerase and DNA:

1. Sequence determines the collision outcome. This would involve specific sequence recognition in the major and/or minor groove.
2. A structural feature of DNA such as a disrupted helix determines the collision outcome and this structural feature is determined by sequence alone.
3. A structural feature that is a product of the interaction of DNA with a regulatory protein determines the collision outcome – this is obviously very important.
4. A structural feature that is the product of interaction of DNA with NTP monomers and modified monomers determines the collision outcome – what I am proposing in this paper.
5. Some combination of the above.

For scenarios one and two there is no information processing. The DNA sequence alone determines the collision outcome. The RNA coding regions of genes give a rejection of initiation. The promoters have a probability of initiation associated with the DNA holoenzyme collision. This requires a code. I will call this code, based on sequence only at constant supercoil density, the **constant code** because there is no way to turn genes on and off – there is no way to adjust the global transcription rate or to change the differential pattern of expression. The constant code is guaranteed to exist since it determines the in vitro strength of promoters, the open complex formation pattern, in the absence of regulatory proteins.

The most important aspect of the genetic code is degeneracy – multiple codons per amino acid. Codon usage statistics reveal that most codons are used at significant levels. Wobble allows a smaller number of classes of tRNAs than codons to be used for translation. The reason degeneracy is the most important feature of the genetic code is that it allows a transcription initiation/rejection of initiation code to co-exist with the amino acid specifying function of DNA. Protein coding regions must be coded differently than promotor sequences and that is precisely what degeneracy allows. In addition to protein coding sequences the leader sequences, tRNA sequences, and rRNA sequences must be coded not to be promoters – to give rejection of initiation on collision with polymerase. The words in the code are sequences, N-mers, of unknown length N. N is probably more than 5 bases and less than 20 bases long and maybe is variable in length along the DNA sequence. The number of words grows fast as 4^N. RNA coding sequences of the DNA are the subset of the 4^N large set of words that give rejection of initiation on collision with polymerase. Promoters are the subset of the 4^N large set of words that give initiation on collision with polymerase. Degeneracy and codon usage are strong evidence for a transcription initiation/rejection of initiation code in DNA primary sequences.

Consider the relative motion of polymerase and DNA in the nucleoid. The problem of searching the DNA to determine global and differential expression is either done by the RNA polymerase diffusing along the DNA or the DNA diffusing to the polymerase. Because of the large mass, 450KDa, of the polymerase and the low mass of DNA, $660\text{Da} / \text{bp}$, the polymerase does not do a topologically complex one-dimensional diffusion along the DNA helix. In the nucleoid the polymerase is relatively fixed in space by numerous contacts with supercoiled condensed DNA. In the nucleoid the DNA helix diffuses to the largely stationary polymerase. The DNA is twisting and writhing, transmitting energy and forces along the backbone, and supercoils are migrating around. The DNA is in motion – the polymerase is stationary. As in elongation where the ternary complex is fixed in the space of the nucleoid and the DNA translates through the ternary complex for initiation the polymerase is fixed in the space of the nucleoid and the supercoiled condensed DNA does the diffusion to the polymerase. In the constant DNA code scenario and for the possibility that the sequence only and not a structural feature of DNA determines the collision outcome the collisions between polymerase and DNA would be random – all sequences would have the same average number of collisions over time with polymerase. For the possibility that a structural feature of DNA determines the collision outcome it is possible that different structures have different diffusion rates to the polymerase so the collisions between a given sequence and polymerase would not be random over time.

The model for initiation I am proposing postulates the existence of a DNA structural feature, namely, hybrid helices of NTP monomers and single strand DNA. I believe that in the nucleoid there are denaturation bubbles migrating around as the DNA twists and writhes and supercoils migrate. NTP monomers can interact with these denaturation bubbles forming dynamic hybrid helices of length $3-12$ bases on one or both strands of the denaturation bubble. The NTP monomers can make both stacking interactions and conventional H-bonding in these hybrid helices leading to favorable energetics. I think these structure are located between the upstream -35 homology and the -10 Pribnow box but they might include the Pribnow region. These structures are dynamic in that they are bendy regions of DNA and the monomers are making and breaking H-bonds with the single strands of DNA. At any instant in time the number of pairings is stochastic—they have an average structure of so many base pairings. It is this average structure that I call a hybrid helice. To form hybrid helices requires twisting of the single stranded regions and this can be accomplished topologically by rotation of the sugar phosphate backbone at the edge of the hybrid helices. **I believe the rate limiting step for initiation in the nucleoid is the diffusion of these structures to the polymerase**. I believe the time period for diffusion to the polymerase is tunable from ~ 1 to 10^{3-4} seconds depending on the length of the hybrid helice. The longer the length of the hybrid helices the faster the diffusion to the polymerase. A hybrid helice is a bendy stretch of DNA and the rest of the DNA helix is rod-like. Bendy regions are segregated to the polymerases in the nucleoid. Rod-like regions segregate to between polymerases in the nucleoid. The walk of a hybrid helice to the polymerase is a random walk as diffusion must be in a phase. Different lengths of hybrid helice have different average time periods to do this random walk. When one of these structures gets to the polymerase initiation occurs fast—all that is required is rotation of the polymerase. The formation of these hybrid helices is a function of the DNA sequence and the NTP concentrations. At lower NTP concentrations fewer and shorter hybrid helices occur – turning down the transcription rate. At higher NTP concentrations more and longer hybrid helices occur – turning up the transcription rate. The equilibrium between DNA helix and DNA helix plus hybrid helices is adjusted by the NTP monomer pool levels establishing the global rate and differential pattern:

$$\text{DNA helix} \xleftrightarrow{\text{monomers}} \text{DNA helix} + \text{hybrid helices}$$

Hybrid helices are a small fraction of the DNA sequence – probably no more than 1% of the DNA sequence. The global transcription rate is determined by the absolute level of NTP monomer pools. The differential transcription pattern is determined by the relative NTP monomer concentrations. I call this model where NTP monomers interact with DNA sequence to determine global and differential transcription the **thermal code model** because it runs on thermal energy.

I believe the critical initiation step once one of these structures is impinging on a polymerase is getting into the DNA helix – forming the open complex. In vivo without monomers to form hybrid helices the DNA helix cannot be opened up in a collision with polymerase – the energy required to open up the helix is too high. Monomers lower this energy barrier by forming hybrid helices.

The information channel width can be defined as the number of words in the code. This depends in a simple way on the number of monomers in a hybrid helice. If the longest hybrid helice is 12 bases there are 4^{12} words in the code. Because there are probably $3-12$ monomers in a hybrid helice this mechanism provides the coding possibility for exquisite quantitative control of global and differential transcription.

I believe the upstream homology is a DNA helix making an edge to a hybrid helice and is presumably recognized through sequence specific interaction in the major groove. So the requirement for initiation is a sequence specific recognition of the upstream homology adjacent to a hybrid helice. This two part mechanism gives rise to correct strand selection for $5' \to 3'$ synthesis.

I have different roles for the Pribnow homology between in vitro and in vivo. In vitro, I believe, the Pribnow region is the region of destabilized helix, low T_M because all A and T, where the strands separate and polymerase gets into the helix, between the strands, forming the open complex. In vivo, I believe, the polymerase gets into the helix between the upstream and Pribnow homology and that the Pribnow homology defines a start site by sequence recognition of the Pribnow region once inside the helix. In vivo the Pribnow region gives rise to a well defined start base.

If diffusion to the polymerase is not the rate limiting step in transcription initiation then a collision outcome based theory can be formulated. In this scenario the outcome of a collision depends on the length of hybrid helice – the longer the hybrid helice the higher the probability of formation of the open complex. The only requirement to have the thermal code model is sequence specific interaction of monomers with DNA.

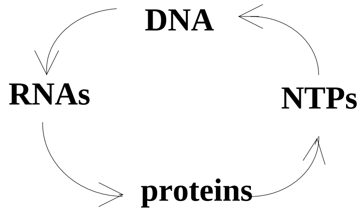
The constant code and thermal code are co-linear – they exist on top of each other in roughly the same places on the DNA sequence. RNA coding regions not only reject initiation in the absence of monomers but also in the presence of monomers. Promoters function in vitro forming the open complex in the absence of monomers. In vitro promoters have strengths based on the constant code alone. The quantitative relation for the relative roles played by the constant and thermal codes will be difficult to unravel.

I believe hybrid helices also occur in replication. Ahead of the DNA polymerase are hybrid helices of dNTP monomers, single strand binding proteins, and DNA single strands. This explains the high synthesis rate of $800-1000$ bases/second.

For the thermal code to work the monomer pool levels must encode the state of the cell. The division of labor among monomers in the cell accomplishes this – UTP used for cell wall, CTP used for cell membrane, GTP used for protein synthesis, and ATP used for general synthesis. A critical role for modified monomers such as cAMP and ppGpp in the thermal code model can easily be imagined. Modification can affect the partitioning between cytoplasm and nucleoid of monomers or affect the formation and diffusion of hybrid helices.

There is a pervasive role of nucleic acid monomers in recognition and energetics in cells. NTP monomers are high energy and can be recognized and dock in a precise way with proteins and do work either phosphorylating other biomolecules or transferring energy to other biomolecules, or forming polymers and phosphate. Monomers regulate the activity of many proteins in addition to being a substrate.

With the nucleic acid mainframe and the thermal code the central dogma becomes circular rather then linear – NTP monomers feeding back on DNA to determine the RNA synthesis profile:



The Turing machine model level is important to understanding the cell as an information processing machine. Turing machines in biology may seem a little esoteric but are very simple. Answering the questions how is DNA searched and how is the transcription pattern arrived at are the Turing machine description. Only small molecules like monomers can make the state of the cell available throughout the nucleoid. The entire DNA is continuously searched through interaction with monomers. Continuous search of the entire DNA is the most powerful Turing machine model of DNA possible. A large channel width – number of distinct words – also is essential to be a powerful Turing machine model. If transcription initiation, in the absence of regulatory proteins, is only regulated by the constant code the cell is a dumb Turing machine. An example of the type of calculations the cell Turing machine must do is to calculate the surface area to volume ratio given the shape of the cell and to calculate linear combinations of this ratio. In molecular terms an example of such a calculation is of how many lipids are needed for the membrane or how many of a membrane protein are needed for a given size cell. The nucleic acid mainframe and thermal code provide for information definition, transport, and processing in the cell – the basic smart Turing machine architecture.

The thermal code drives the binary fission growth cycle of prokaryotes. For eukaryotes where cells are in steady states for much of the life cycle the thermal code would not seem to work. But we should look in eukaryotes for a monomer-based cell cycle regulation.

The thermal code is not available in current in vitro systems. The nucleoid is a highly concentrated and exquisitely balanced system only found in vivo. The concentration of biomolecules is much higher in vivo than in vitro. It may be difficult or impossible to get the thermal code to work in vitro, then again, starting with a qualitative model, it may be possible.

There are important molecular evolution and origin of life implications of the constant code and thermal code model. The constant code and, if it exists, the thermal code must have preceded the genetic code. Before proteins are coded for in DNA transcription must have been occurring and a transcription code must have existed. The thermal code model allows for transcription to occur with non-coded for protein catalysts as crude RNAPs where self-assembly of the hybrid helices is the site specific precursor step to polymer bond formation. Of course, given the complexity of a cell, it is always a bit difficult to understand how such a system could arise.

To elucidate the thermal code at a fully quantitative level is a difficult task. If the thermal code does not exist and only the constant code exists even that code may not be easy to elucidate in a quantitatively precise way. Getting good quantitative data on monomer pool levels in the nucleoid will be very difficult because it is not known how they partition between nucleoid and cytoplasm phases. Modeling the interaction of monomers with DNA will be difficult. A molecular simulation and kinetics analysis required to get the thermal code out will be difficult. Then again, it may not be too hard to do starting out with a qualitative model.

There are important implications of the thermal code model to eukaryotic biology. At the root of cancer and developmental biology is the cell cycle. Possibly at the root of the cell cycle in eukaryotes is some kind of monomer-based code. The nucleic acid mainframe and the RNA processing zone would seem to be important in all cells.