Connecting Transistors and Proteins

Claudio Mattiussi[1], Dario Floreano[1]

[1] Autonomous Systems Laboratory, ASL-I2S-STI-EPFL, 1015 Lausanne, Switzerland
claudio.mattiussi@epfl.ch

Abstract

We connect transistors and proteins in two ways. The first is by showing that they have much in common as fundamental devices of electronics and life. The second is by describing how an evolvable wiring of electronic devices can parallel the wiring of proteins into genetic regulatory networks. We then transform this connection into a methodology for the study of the evolutionary properties of circuits. The approach is based on the use of analog electronic circuit simulators. We present an example of implementation with the first results obtained.

Introduction

Many functions within living cells are performed by proteins in their role as catalysts (Alberts et al. 2002, Creighton 1993). In the simplest scenario, a chemical substance generically called the substrate must be converted into another substance called the product. The free energy of the substrate is higher that that of the product. Hence, the former would convert spontaneously in the latter. However, the conversion requires the passage through a less favorable transition state. In the absence of the catalyst, the barrier constituted by the transition state keeps the reaction rate low. The effect of the catalyst is to lower the barrier and thus accelerate the reaction rate.

The operation of active semiconductor devices such as transistors is conceptually similar. For example, a bipolar junction transistor (BJT) is composed by three adjacent regions of semiconductor having different physical characteristics (Cooke 1990). These regions are called the emitter, the base, and the collector. In the typical circuit configuration, the voltages applied to emitter and collector, make energetically favorable the flowing of current carriers from emitter to collector. This current, however, must pass through the base. When the base is left unconnected, it acts as a barrier to the current flow, which is therefore small. A suitable voltage applied to the base lowers this barrier with the effect of increasing the current flowing from emitter to collector.

These descriptions reveal a striking analogy in the operation of proteins and transistors (Figure 1). Evolution designed the basic devices of life just as engineers designed the basic devices of electronics. Both kinds of devices permit the variation of the rate of some physical process. In other words, they are the key to the implementation of constraints to the spontaneous dynamics of those physical processes. As argued by Pattee (Pattee 1995), natural selection leads indeed to the formation of structures whose presence influences the dynamics of the surrounding space-time in ways that favor the persistence and, eventually, the self-reproduction of these structures. If Pattee's intuition is correct, we should therefore expect to observe the emergence of devices performing this kind of function within our synthetic experiments on the evolution of life. Thus, we can take a major evolutionary shortcut if we adopt directly these structures as our basic building blocks. Note that this does not spoil our inquiry since, as we will argument below, there remains to study the crucial aspect of the establishment of the connectivity between the structures.

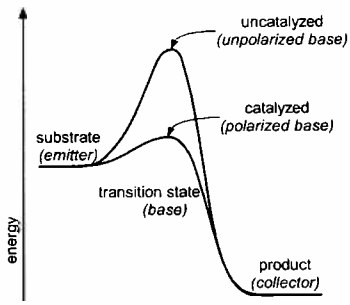

Figure 1: The analogy between the stages of a chemical reaction, and the regions of a bipolar transistor. The vertical axis represents the free energy of the substances (substrate, transition compound, product) in the course of the chemical reaction or the energy of the current carriers in different regions (emitter, base, collector) of the transistors body. The presence of the catalyst decreases the height of the barrier that hinders the transformation of the substrate into the products, just as a suitable polarization of the transistors base decreases the barrier that hinders the flow of current carriers from the emitter to the collector.

If we follow this approach, we can take advantage of the existence of analog electronic circuits simulators (Vladimirescu 1994). In these simulators, the physics of the devices and of their interaction is modeled at high level,

through a set of algebraic and differential equations, which embed the relevant conservation laws of physics. The resulting implementation is efficient and physically sound. Besides, by using the models of energy storing components such as capacitors, this approach allows the modeling of delays to the propagation of signals, a phenomenon that affects also chemical signals that must diffuse across spatially extended structures in cells.

Signals and connections

A collection of unconnected devices performs no function. The task of the engineer and of evolution consists in finding how to connect the available components to obtain the desired behavior.

In the biological case the connectivity corresponds to the network of interactions between the elements. For genetic regulatory networks and omitting many details (for the full picture see for example Alberts et al. 2002) the interactions can be schematized as follows (Figure 2). A protein interacts with a stretch of a DNA, and activates a transcription machine called RNA polymerase. Each cell contains many copies of a few types of these transcription machines. The output of the machine is a molecule of RNA, which, after a number of further steps, leads to the synthesis of a protein. This protein can in turn interact with the DNA, to activate or repress the transcription of another sequence of DNA, and so on. The identity of the connected elements and the strength of the interaction depend on the chemical nature of the participants in a way that we will describe below.

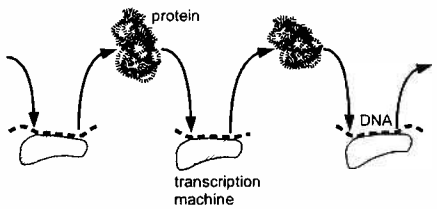

Figure 2: A very schematic representation of the interactions that compose a genetic regulatory network. The mediation is due to proteins that can activate or repress the functioning of specialized transcription machines.

In the case of analog electronic circuits, the connection between devices is determined by conducting wires that guide the signals. The strength of the connection can be varied by changing the value of resistance of the connection, from a minimal value of zero, which corresponds to the maximum strength of the interaction, to the absence of direct interaction, which corresponds to a virtual infinite-valued resistance (Figure 3).

Inspired by this similarity, we could thus imagine to evolve an electronic circuit by determining the connections between the electronic devices in a way that reminds that used in biological systems. Since in the biological case it is

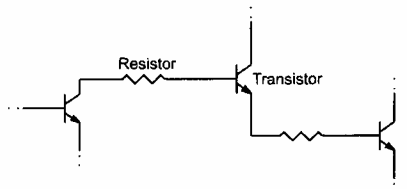

Figure 3: The interactions between electronic devices are mediated by conducting wires connecting the terminals. The value of resistance of the connection determines the strength of the interaction.

the characteristics of two DNA sequences that determine the existence and the strength of the interaction (Figure 2), we could imagine to associate a sequence of characters to each terminal of the circuit components. Then we could define a mapping of pairs of sequences in order to determine the existence of a connection and its strength. By coding those sequences in an artificial genome, we could then parallel the process of evolution of biological circuits (Figure 4).

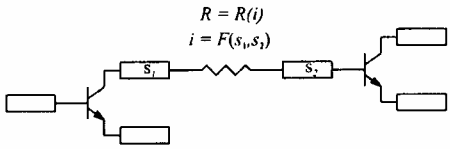

Figure 4: The strength of the connection between devices can be assigned by associating sequences of characters to the terminals, defining a mapping $i = F(s_1, s_2)$ from pairs of sequences to values of connection strength, and a further mapping $R(i)$ that gives the values of resistance.

The protein folding objection

Before we proceed to detail the nature of the mapping outlined above, there is a major objection to the whole program that must be addressed. After the transcription of the DNA sequence into an RNA molecule, the latter undergoes a series of transformations that convert it into a chain of amino acids. To become a functional protein, this chain must fold into a precise three-dimensional shape (Alberts et al. 2002). In living cells the folding usually proceeds effortlessly. However, the simulation of this process appears computationally daunting. Thus, if the decoding of sequences into interaction strengths requires the computation of an equivalent of the folding process, our suggested approach becomes computationally impractical.

Some authors (for example, Conrad 1999) have argued that the characteristics of the folding process are unique in determining the evolvability of living systems. We can think of this processing as a mapping from the space of sequences to the much higher dimensional space of protein shapes. This mapping provides redundancy to the evolutionary process, thanks to its being potentially many-to-one; it brings a degree of smoothness to the discrete universe of DNA sequences, but still allows abrupt discontinuities in shape with only a few nucleotides substitutions. Moreover,

it determines the shaping of proteins that gives them their specificity, and leads to the phenomenon of allostery (a phenomenon that consists in the change of shape of the protein in presence of physical or chemical signals).

Fortunately, molecular biologists have discovered that at the level of genetic regulatory network, things seem to be simpler than was previously imagined (Ptashne and Gann 2002). In the case that interests us, the process of transcription proceeds as illustrated schematically in Figure 5. Here a single protein (called activator) is assumed to be in charge of the activation of the transcription of a certain DNA sequence. This protein recognizes a sequence of nucleotides along the DNA and binds to it in a well-defined position. Then it recruits the transcription machine to the DNA through another binding interaction, which is sufficient for the transcription to start and proceed autonomously. It turns out that in most cases, the activator does not need to alter – as was instead previously imagined – the transcription machinery, for example with some complicated allosteric interaction. The only specific interaction is the readout of the sequence of DNA that binds the activator. The surprise of molecular biologist at this finding is witnessed by the following extract (Ptashne and Gann 2002, p. 176)

> That so much of the specificity of regulation and hence so much of development and evolutionary change depends on simple binding interaction is (or we think should be) hard to swallow. It certainly is for us. We, and we suspect many others, had expected that the meanings of biological signals would have been, somehow, more solidly based.

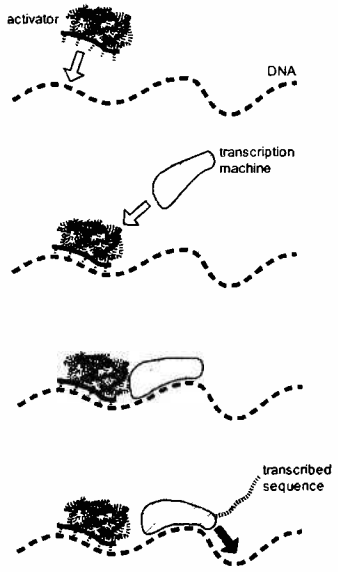

Figure 5: In many cases, the regulation of the transcription of DNA sequences depends on simple binding interactions which correspond to the recognition of a sequence of nucleotides, and to a generic adhesive recruitment of the transcription machinery.

The consequence of this finding is that we can hope to obtain an equivalent of the protein-mediated interaction in terms of a mapping from pairs of sequences which does not imply the complexity of protein folding. Note that we are not saying that protein folding, three-dimensional shape and allostery play a minor role in the existence of living beings. These phenomena are obviously essential in a world where significant physical and chemical signals, the energy sources, the strength of materials, the dynamics of motion, and many other essential aspects, are imposed from the outside and must be complied with in order to survive. What we observe is merely the contingent fact that where living systems "talk to themselves" and are free to define their own language for example in exchanging internal signals across genetic regulatory networks they appear to employ forms of interaction where allostery and three-dimensional shape play at most a generic role, and where the communication can be interpreted as a sequence-to-sequence correspondence (Figure 6).

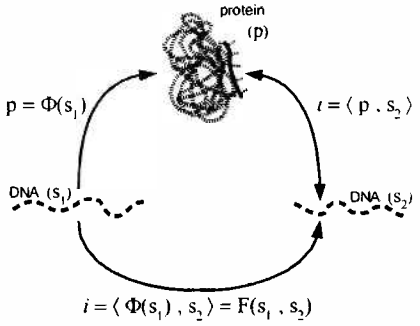

Figure 6: The interaction between two DNA sequences within genetic regulatory networks can be schematized by composing the mapping $F(s_1)$ that converts a sequence s_1 into a folded protein p, and the mapping $\langle p,s \rangle$ representing the interaction of the protein p with another DNA sequence s_2. Since the resulting process is based on simple binding interactions, we can hope to model it with a computationally tractable mapping $F(s_1,s_2)$ that gives the strength of interaction i.

Defining and decoding the genome

We can now proceed to the definition of an evolutionary system based on the ideas presented in the previous section. The first thing that we need to specify is the structure of the genome. The genome must contain at least the description of the devices and the sequences of characters associated with the terminals of the devices, which determine the strength of the connections (Figure 4). It is useful to have also the possibility to evolve the value of some parameters associated with the devices, for example the capacitance value of a capacitor, or some parameter of a transistor.

To fulfill these requirements, we use a genome constituted by a sequence of characters. Each kind of device is identified by a token of a few characters, for example "NBJT"

for an NPN BJT, and "CAPA" for a capacitor. We define two other tokens: one relative to the terminals, for example "TERM", and one to the parameters of the device, for example "PARM". Note that we chose human-readable tokens just to facilitate the visual inspection of the genomes.

The decoding proceeds as follows. We search in the genome the first token identifying a device, which signals the start of the fragment of genome coding for that device. Each kind of device has a characteristic number of terminals and evolvable parameters, and we search that number of terminal and parameter tokens in that fragment. If all the required tokens are found before the end of the chromosome (or before the next device token, if no overlap of device descriptors is allowed), a device – for the moment, unconnected – is created in the circuit. The sequences of characters delimited by the tokens are associated with the terminals and parameters of the device (Figure 7). When the terminals of the device are not interchangeable, an order is specified for them and the association of extracted sequences follows that order. Once a device has been decoded we proceed to search the next device token in the genome, and so on until all the genome has been examined.

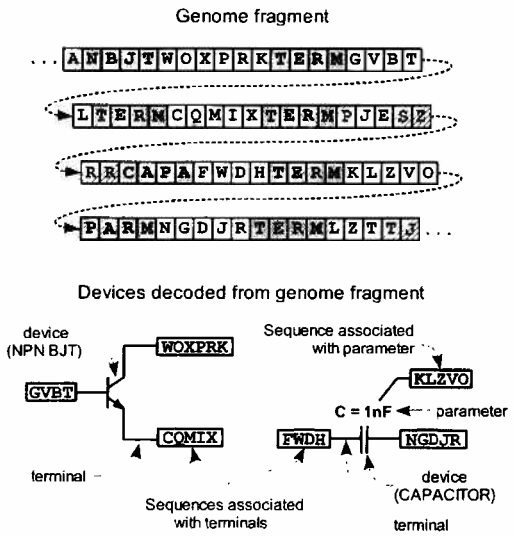

Figure 7: A fragment of genome (top) and the corresponding devices decoded from it (bottom). A series of tokens (shaded) identify the start of coding regions and delimit the sequences of characters associated with the terminals and evolvable parameters of the devices. The hatched characters correspond to noncoding genome.

Connecting the evolved components

The result of the process just described is a collection of devices with sequences of characters associated with their terminals and parameters. To connect the devices we proceed as follows. We define a mapping $F(s_1, s_2)$ that transforms pairs of sequences in a scalar value i that represents an abstract interaction strength. For each pair of terminals of the collection of devices extracted from the genome, we calculate the interaction strength determined by their associated sequences. Then, we transform i into a resistance value with a predefined mapping $R(i)$, and we insert in the circuit a resistor connecting the two terminals and having $R(i)$ as resistance value (Figure 4). We will give below an actual example of the mappings $F(s_1, s_2)$ and $R(i)$.

An analogous process assigns the values to the parameters of the devices. With each evolvable parameter is associated a sequence s of characters extracted from the genome (Figure 7). We define a fixed sequence v that will be used for all the devices parameters, we evaluate $i = F(s, v)$ and transform the result into the parameter value with another mapping, for example $C(i)$ for the capacitance of a capacitor.

External connections

A living system is connected to the external world, to absorb energy and matter, expel waste, exchange signals. The evolution of these interactions is actually a topic of major interest for ALife (Bedau et al. 2001). In the case of an electronic circuit, this corresponds to the presence of external devices or circuits, such as power supplies, signal generators and output loads. We must thus specify how our decoded circuit connects to these external parts and how these connections can evolve. The simplest solution consists in associating predefined fixed sequences to the terminals of the external devices that must connect to the evolved circuit (Figure 8), so that the connection strategy described for the devices decoded from the genome can be extended to the external devices, and evolution of these connections is possible. For more complex approaches to the establishment of external connections, see (Mattiussi and Floreano 2004).

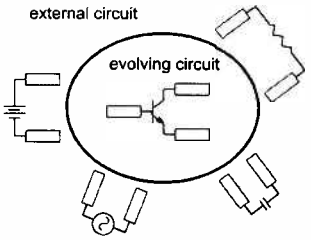

Figure 8: By associating sequences of characters to the terminals of the external devices, connections from the circuit specified by the genome (shaded region) to the external devices can be subjected to evolution.

Compartments, modules and hierarchies

With the decoding strategy described above, the strings associated with the terminals implicitly determine the connections between all the terminals in the external and decoded circuit. This frees the genome from the necessity of specifying explicitly all those connections. However, this comes at the cost of calculating the value of the mapping $F(s_1, s_2)$ for all pairs of terminals. The number of evaluations grows

quadratically with the number of devices in the circuit. At the same time, the function $F(s_1,s_2)$ cannot be too simple without compromising the evolvability of the system. Therefore, the computational cost of the decoding could become intolerable as the complexity of circuits grows.

A solution to this problem is the inclusion in the tokens for terminals, of an evolvable marker for the compartment to which the terminal belongs. In this way, only the connections for pairs of terminals belonging to the same compartment would have to be considered. At the same time, the system would have the possibility of evolving a compartmentalized or modular architecture, with all the advantages that this entails (Kazadi et al. 2000). Minor elaborations on this strategy, may allow the evolution of hierarchical and multicellular structures.

Genetic operators

The genome as defined above can be composed of several distinct sequences of characters that we can call chromosomes. The structure of the genome and of the decoding process permits the execution of many genetic reorganization operations that are known to apply to biological genomes (Graur and Li 2000) but are seldom used in artificial evolution experiments because they usually make the genome undecodable. In our case, besides the usual substitution of single characters, we can perform operations such as insertion and deletion of them; operations on chromosome fragments, such as duplication, deletion, transposition, recombination of pairs of chromosomes, and insertion of component descriptors; operations on whole chromosomes, such as duplication and deletion; and the duplication of the whole genome. The possibility of performing such operations is important, since they are assumed to play a crucial role in the evolution and complexification of living systems (Graur and Li 2000). Note that from the point of view of the genetic operators each chromosome is just a sequence of characters where the tokens for devices, terminals and parameters (Figure 7) have no special meaning. Therefore, the tokens are *not* protected from the action of the genetic operators, whose action can invalidate any device descriptor present in the genome, making that particular descriptor undecodable.

An example of implementation of the mapping

So far, we have described only in abstract terms the mapping $F(s_1,s_2)$ that transforms pairs of sequences into interaction strengths i, and the function $R(i)$ that gives the value of connecting resistors. We will describe now briefly the characteristics of that mapping for an actual implementation of the system, along with some results obtained performing evolutionary runs with the implementation.

The genome is composed by sequences of uppercase alphabetic ASCII chars. To derive a connection strength from pairs of sequences extracted from the genome, we use the

sequence alignment (Sankoff and Kruskal 1983). The basic idea is that subsequences of one sequence can be put in correspondence with subsequences of the other through operations of insertion, deletion and substitution of characters (Figure 9). To each operation is assigned a score that rewards close matches and the absence of insertions and deletions. The value of the local alignment score $i = F(s_1,s_2)$ of two sequences s_1 and s_2 is defined as the maximum value of the sum of the scores that can be obtained putting in correspondence a subsequence of s_1 with one of s_2. Some favorable properties of this mapping are its high redundancy, the possibility to operate with sequences of variable length, and the possibility due to the locality of the alignment of matching several distinct sequences with a single one.

The values of i obtained are non-negative integers. These are transformed in resistance values through a table of correspondences. This means that there is a finite set of possible values, but this is not a serious limitation; for example, the number of commercial values available to engineers for their designs is also finite. A whole range of values of i corresponds to the zero-valued resistor (direct connection), and another one to the infinite-valued resistor (no direct connection).

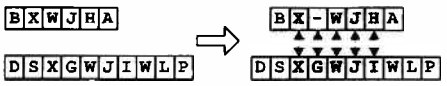

Figure 9: The local alignment of sequences is based on the establishment of correspondences between the subsequences of two sequences of characters, using operations of insertion, deletion, and substitution of characters.

Experiments

We ran a first series of experiments of circuit evolution using SPICE as simulator (Vladimirescu 1994). The experiments were targeted at the synthesis of a circuit giving a constant voltage as output, in presence of a variable input voltage and environment temperature. This problem is interesting in an ALife perspective, since the solution implies the evolution of capabilities of measurement and control (Pattee 1995). We obtained good results, while observing biologically evoking phenomena such as phenomena of gene overlapping (Graur and Li 2000)) and the appearance of vast zones of noncoding genome.

A logarithmically distributed set of resistance values was used to connect the components using the string alignment technique. This resistance set covers 6 decades with 8 values per decade, from 1Ω to 1MΩ. The 1Ω resistance value corresponds to an alignment scores of 20, whereas the 1MΩ value is associate with a score of 68. The whole range of scores below 20 is associated with an infinite-valued resistor (no connection) and that above 68 is associated with a zero-valued resistor (direct connection).

The external circuit is the one represented in Figure 10, where the voltage of the power supply (left) can vary from 4V to 6V, the source resistance is 1kΩ, and the load resistance (right) is 1kΩ. The goal is the generation of a constant 2V voltage across the load when the temperature varies from 0°C to 100°C. To this end, the decoded circuits where simulated with a power supply voltage varying from 4V to 6V in steps of 0.1V, with simulation temperatures of 0°C, 25°C, 50°C, 75°C, and 100°C. For each power supply voltage and circuit temperature, the square of the difference between the actual voltage on the load and the required output voltage was computed. The fitness was defined as the opposite of the sum of all these squares, so that the goal was the maximization of the fitness, with optimal value zero.

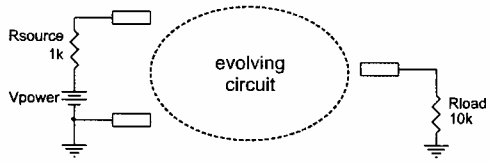

Figure 10: The components of the external circuit for the voltage reference evolutionary experiment.

The genome of all the individuals of the initial population was constituted by one chromosome containing as devices 10 NPN BJT descriptors. The terminal sequences for all the devices had an initial length of twenty characters, randomly filled with elements of the genetic alphabet. We used a genetic algorithm with tournament selection, tournament size of 5, and elitism. The size of the population was 100. The probabilities of nucleotide insertion, deletion, and substitution, those of chromosome duplication and deletion and the probability of genome duplication were set to 0.001. The probabilities of chromosome fragment duplication, deletion, transposition, and that of chromosome single point crossover were set to 0.01. Chromosome fragment reorganization was performed by selecting two random points in the source chromosome to define the fragment, and one random point in the target chromosome, when required. Figure 11 illustrates the course of the best of four runs of 10000 generations evolution. The evolved circuit gives an output voltage that stays within ±1.5% of the prescribed value in the whole temperature and input voltage range. For further details see (Mattiussi and Floreano 2004).

Conclusions

We have presented a methodology to genetically represent and evolve collections of interconnected elements. The technique allows the variation in the course of evolution of the number of elements and of the number and strength of the connections between them. The approach is biologically motivated by the interaction of genes and proteins within genetic networks but does not imply the implementation or the mimicking of the details of protein folding and binding

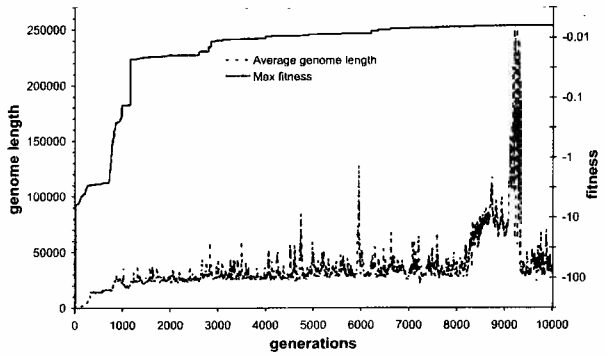

Figure 11: The progress of an evolutionary run aimed at the synthesis of a voltage reference circuit.

reactions. The resulting genome tolerates drastic reorganizations such as duplications and transpositions, which appear instrumental to the open-endedness of the evolutionary process. The first results obtained with this representation witness the evolutionary potential of the proposed approach.

Acknowledgments

This work was supported by the Swiss National Science Foundation, grant no. 620-58049.

References

Alberts, B, et al. 2002. Molecular Biology of the Cell (4th edition). New York : Garland Science.

Bedau, M.A. et al. 2001. Open Problems in Artificial Life. Artificial Life 6:363-376.

Conrad, M. 1999. Molecular and evolutionary computation: the tug of war between context freedom and context sensitivity. Biosystems 52: 99-110.

Cooke, M.J. 1990. Semiconductor Devices. New York: Prentice Hall.

Creighton, T.E. 1993. Proteins: Structures and Molecular Properties (2nd edition). New York: W.H. Freeman.

Graur, D., and Li, W.-H. 2000. Fundamentals of Molecular Evolution. Sunderland, Mass: Sinauer.

Kazadi, S. et al. 2000. Levels of Compartmentalization in Alife. In Proceedings of the 7th International Conference on Artificial Life, 81-89. Cambridge, Mass: MIT Press.

Mattiussi, C., and Floreano, D. 2004. Evolution of Analog Networks using Local String Alignment on Highly Reorganizable Genomes. in Proceedings of the 2004 NASA/DoD Conference on Evolvable Hardware. Forthcoming.

Pattee, H.H. 1995. Artificial Life Needs a Real Epistemology. In F. Moran et al. eds. Advances in Artificial Life, 23-38. Berlin: Springer.

Ptashne, M. and Gann, A. 2002. Genes & Signals. Cold Spring Harbor, New York: Cold Spring Harbor Lab. Press.

Sankoff D., and Kruskal J.B. eds. 1983. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Reading, Mass: Addison-Wesley.

Vladimirescu, A. 1994. The SPICE Book, New York: Wiley.