Some Differences between Genetic Algorithms in Nature and in Computer Science

Wolfgang Maass

Institut für Grundlagen der Informationsverarbeitung
Technische Universität Graz, Austria
Motivation for this talk

• There is some interest in Genetic Algorithms (GAs) and Artificial Life at our institute. We also teach that.

• We are about to experiment with GAs and other optimization methods to optimize circuits consisting of biologically realistic models for neurons and synapses.

• The analysis of computations that emerge in biological organisms is a research area of our institute (and neural computation is just a small part of biological computation).

Ref. Hartwell et al, Genetics: From Genes to Genomes; 2nd ed., 2004
Bower, Bolouri: Computational Modeling of Genetic and Biochemical Networks, 2001
What is DNA?

(a) In a leap of imagination, Watson and Crick took the known facts about DNA’s chemical composition and physical arrangement in space and constructed a wire-frame model that not only united the evidence but also served as a basis for explaining the molecule’s function.
1. Original double helix.

2. Strands separate.

3. Complementary bases align opposite templates.

4. Enzymes link sugar-phosphate elements of aligned nucleotides into a continuous new strand.

Figure 6.14 The model of DNA replication postulated by Watson and Crick. Unwinding of the double helix allows each of the two strands to serve as a template for the synthesis of two new strands by complementary base pairing. The end result: A single double helix becomes two identical daughter double helices.
What is the information content of DNA?

DNA is a long sequence of letters ("base pairs", abbreviated bp) A, C, G, T. Each letter contains 2 bits of information. Each triple ("codon") of letters encodes one of 20 amino acids, or encodes a stop signal.

DNA is copied into messenger-RNA (mRNA) (with the letter T replaced by the letter U).
mRNA leaves the nucleus, and attaches to ribosomes, where the encoded sequence of amino acids is assembled. The resulting sequences of amino acids are called proteins (= polypeptides). These proteins are the building blocks of all biological organisms.
Thus each segment of DNA, from some start so some stop-codon, encodes a particular protein (to be more precise, it can also encode rRNA or tRNA). Each such coding segment of DNA is called a gene.

A single string of DNA (i.e., a chromosome) is usually on the order of a meter long, and consists of a few 100 or a few 1000 genes.

The DNA strings of the genome are intricately folded so that they have just 1/10,000th of their natural length.

The genome does not only exist of genes. For example the human genome consists of $3 \cdot 10^9$ nucleotide pairs, but just 3% of them are known to be part of genes.

Figure 12.4 Models of higher-level packaging. (a) Electron micrographs contrasting the 100 Å fiber (left) with the 300 Å fiber (right). The line drawings show the probable arrangement of nucleosomes (with green cores) in these structures. (b) The radial loop-scaffold model for yet higher levels of compaction. According to this model, the 300 Å fiber is first drawn into loops, each including 60–100 kb of DNA, that are tethered at their bases by nonhistone scaffold proteins (brown and orange) including topoisomerase II (brown). (c) Additional nonhistone proteins might gather several loops together into daisy-like rosettes and then compress the rosette centers into a compact bundle.
The non-coding regions of DNA are called **introns**, the coding regions **exons** (there exist additional non-coding regions at the end of a chromosome and at the centromer).

**Figure 8.15  The human dystrophin gene: An extreme example of RNA splicing.** Though the dystrophin gene is 2500 kb (or 2.5 Mb) long, the dystrophin mRNA is only 14 kb long. More than 80 introns are removed from the 2500 kb primary transcript to produce the mature mRNA (which is not drawn to scale).
The complete genetic information (genome) of each organism is encoded by a certain number of chromosomes, for example 46 for humans (containing together about 40,000 genes). If one stains them, each chromosome exhibits dark and light bands. Ordered according to their size, they provide the karyotype (= visual appearance of a complete genome).

**Figure 4.6  Karyotype of a human male.** Computer-enhanced photos of metaphase human chromosomes are paired and arranged in order of decreasing size. In a normal human male karyotype, there are 22 pairs of autosomes, as well as an X and a Y (2n = 46).
Mutations occur at a rate of $\approx 10^{-6}$ per gene generation (p. 193).
Recombination ("Crossover")

Before egg- or sperm-cells are created, homologous chromosomes from both parents are brought closely together in the cell. Then they are attached to each other at "recombination modules", and corresponding sections of chromosomes from both parents are interchanged.

**Figure 4.14** Prophase I of meiosis at very high magnification.
So far this process looks exactly like GAs in computer science, but ...

1. We must have missed something important: The genome is obviously not just the code for some 10,000 proteins. It also encodes how to build a biological organism out of these proteins.

But where and how?
2. Genes are sometimes changed in the DNA through mutations and crossover, so that they encode a different protein. On the other hand such changes in the code for a protein are usually fatal, or they have very little impact for the organism since such local "errors" are often automatically corrected, or somewhere else in the genome there exists another copy of the same gene that still encodes the same old protein.

In fact, proteins have not changed too much during evolution. 46% of the proteins encoded in the human genome are closely related to proteins in yeast (a single cell eukaryotic organism) and only 7% of human proteins are specific for vertebrates (these 7% are especially used by the immune system, and by the nervous system).

Hence the generation of new proteins does not appear to be the main driving force of evolution.
How were the genomes for new species really created during evolution?

To understand that, we need to look at gene regulation,

and at the actual differences in the genomes for two different species, e.g. the genomes of mice and humans.
Gene Regulation

In eukaryotic cells (which are the cells of all plants and animals, except for bacteria) the DNA remains in the nucleus of the cell. Hence in order to produce a protein according to the code contained in a gene, the information contained in this gene has to be transported from the cell nucleus to the ribosomes, which produce the proteins. The transportation of information from the cell nucleus to the ribosomes is carried out by mRNA ("messenger-RNA").

Thus the transcription of information from a gene in the DNA to a mRNA is a necessary first step in the production of the associated protein.
For each of the \( \approx 40,000 \) proteins encoded in the human DNA an intricate mechanism decides when and in which cell such mRNA for that protein is transcribed from the DNA, or more precisely with what rate the corresponding mRNA is produced in that cell at each particular time. This mechanism is referred to as gene regulation.

These gene regulation mechanisms are also encoded by the genome (where else?), but in a much more subtle way. Furthermore most biologists believe that the most important changes of the DNA during evolution have affected these gene regulation mechanisms, not the proteins (or rRNA, tRNA) themselves.
Figure 17.2  The three RNA polymerases of eukaryotic cells have different functions and recognize different promoters.

(c) RNA polymerase II recognizes promoters associated with all of the diverse protein-encoding genes in the genome.
Figure 17.4 Basal factors bind to promoters of all protein-encoding genes. (a) Schematic representation of the binding of the TATA box–binding protein (TBP) to the promoter DNA, the binding of two TBP-associated factors (TAFs) to TBP, and the binding of RNA polymerase (pol II) to these basal factors. (b) Computer-generated image of the actual molecular structure of basal factors complexed to the promoter.
There exist thousands of proteins called transcription factors (TFs) whose presence at or near the promoter region immediately before a gene increases or decreases the rate of transcription for that gene.

Figure 3.4
Summary of Endo16 cis-regulatory organization. The 2300-base pair cis-domain of Endo16 comprises 55 protein binding sites (red boxes on horizontal black line denoting DNA) organized into six modular subregions labeled A – G at the top of the figure. The boundaries of the modules are delineated by the thin vertical lines incident on the DNA. Bp represents the basal promoter and the region to its right, marked by the bold arrow, marks the coding region of the gene. The large colored symbols above the DNA line indicate the unique binding sites within each module. The smaller colored symbols below the DNA line mark nonunique binding sites as labeled. (See plate 4.)
Figure 3.10
Computational model for module A regulatory functions. Schematic diagram of interrelations and functions. Interrelations between upstream modules (G to B; figure 3.4) and specific module A target, and among the module A target sites, are indicated beneath the double line at top, representing the DNA. The region from module G to module B is not to scale. Each circle or node represents the locus in the system of a specific quantitative operation, conditional on the state of the system. Operations at each node are carried out on inputs designated by the arrows incident on each circle, and produce outputs designated by arrows emergent from each circle. In the logic statements, “inactive” indicates a site or module site that has been mutationally destroyed or is inactive because its factor (or factors) is missing or inactive; “active” indicates that the site or module is present and productively occupied by its cognate transcription factor. For the case of modules F, E, and DC, a Boolean representation is chosen because ectopic expression is essentially zero (beyond technical background) in endoderm and mesenchyme when these modules (together with module A) are present in the construct (Yuh and Davidson 1996); otherwise, ectopic expression occurs. Similarly, when they are present, LiCl response occurs; otherwise, it does not (Yuh and Davidson 1996) (figure 3.13). The logic sequence specifies the values attained at each operation locus, either as constants determined experimentally and conditional on the state of the relevant portions of the system, or in terms of time-varying, continuous inputs designated by the symbol (t). The values γ=2 or γ=1 derive from the experiments of figure 3.14 (see text). The kinetic output of modules B and G and of the Otx site are represented as $B(t)$, $G(t)$, and $Otx(t)$, respectively. The input $B(t)$ can be observed as the CAT activity profile generated by module B over time in figure 3.11A (see also figure 2 in Yuh et al. (1998) and figure 3 in Yuh and Davidson (1996)). $G(t)$ is shown in the same figures in Yuh et al. (1996) and Yuh and Davidson (1996). $Otx(t)$ is the time course generated by the construct OtxZ in figure 12B (unpublished observations). The final output, $\Theta(t)$, can be thought of as the factor by which, at any point in time, the endogenous transcriptional activity of the BTA is multiplied as a result of the interactions mediated by the cis-regulatory control system.
\[ \alpha: \text{if (F or E or DC are active) and (Z is active) then} \]
\[ \alpha = \text{active} \]
\[ \text{else} \quad \alpha = \text{inactive} \]

\[ \beta: \text{if (P is active and CG1 is active) then} \]
\[ \beta = 2 \]
\[ \text{else} \quad \beta = 0 \]

\[ \gamma: \text{if (CG2 is active and CG3 is active and CG4 is active) then} \]
\[ \gamma = 2 \]
\[ \text{else} \quad \gamma = 1 \]

\[ \delta: \quad \delta(t) = B(t) + G(t) \]

\[ \varepsilon: \quad \varepsilon(t) = \beta \times \delta(t) \]

\[ \zeta: \text{if (\varepsilon(t) is inactive (i.e. =0)) then} \]
\[ \zeta(t) = \text{Otx}(t) \]
\[ \text{else} \quad \zeta(t) = \varepsilon(t) \]

\[ \eta: \text{if (\alpha is active) then} \]
\[ \eta(t) = 0 \text{ (repressed)} \]
\[ \text{else} \quad \eta(t) = \zeta(t) \]

\[ \Theta: \quad \Theta = \gamma \times \eta(t) \]

\[ \alpha: \text{the repressive action of modules F, E, and DC mediated by site Z.} \]

\[ \beta: \text{both P and CG1 are needed for amplification of module B influence.} \]

\[ \gamma: \text{final, non-specific boosting of transcription level.} \]

\[ \delta: \text{positive input from modules B and G.} \]

\[ \varepsilon: \text{amplification of module B effect by CG1-P subsystem.} \]

\[ \zeta: \text{switch determining whether (a) Otx in module A or (b) modules B & G control} \]
\[ \text{level of activity of Endo16 (triggered from (a) to (b) at gastrulation).} \]

\[ \eta: \text{complete inhibition of Endo16 transcriptional activity due to repressive effects of} \]
\[ \text{modules DC (in the primary/skeletogenic mesenchyme cells) and F and E (in the} \]
\[ \text{vegetal ectoderm) before gastrulation (mediated by site Z as in a above).} \]

\[ \Theta: \text{total regulatory effect communicated to the basal transcription apparatus} \]
\( Y_{AB} = \frac{A.K_{R_A} B.K_{R_B} K_q}{D_N + A.D_N + B.D_N + A.K_{R_A} D_N + B.K_{R_B} D_N + A.K_{R_A} B.K_{R_B} K_q} \)

\( I = M.(1 - e^{-k_2 X_{AB}/M}) \)

\( \frac{d}{dt}(mRNA) = I - k_{de} \cdot mRNA \)

\( \frac{d}{dt}(mRNA) = I_{t1+\Delta t} - k_{de} \cdot mRNA \)

\( \frac{d}{dt}(P) = k_t \cdot mRNA_{t1+\Delta t} - k_{de} \cdot P \)

- \( K_R \) = relative equilibrium constant
- \( K_q \) = coefficient of cooperativity
- \( K_D \) = CRM activating strength
- \( Y_{AB} \) = DNA occupancy by A & B
- \( I \) = no. of transcription initiations per minute
- \( M \) = max. no. of initiations per minute
- \( I_{t1+\Delta t} \) = initiations per minute, \( \Delta t \) mins delayed
- \( mRNA_{t1+\Delta t} \) = mRNA molecules, \( \Delta t \) mins after transcription
- \( k_t \) = rate of mRNA translation
- \( k_d \)'s = rates of degradation
Plate 3 (opposite, top) The pattern of Endo16 expression during S. purpuratus embryonic growth. Numbers of hours post fertilization are shown at the top left of each figure. Side views are shown. Cells labeled brown indicate the daughters of a single macromere; cells stained blue are those expressing Endo16. Note Endo16 expression is always confined to a fraction of the macromere descendants. Adapted from Ransick and Davidson (1995). (See chapter 3.)
Fig. 1. (A) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein $X$ binds regulatory DNA regions of a gene to regulate the production rate of protein $Y$), through cells (neuron $X$ is synaptically connected to neuron $Y$), to organisms ($X$ feeds on $Y$). (B) All 13 types of three-node connected subgraphs.

Network motifs:
Simple building blocks of complex networks.
Other mechanisms of gene regulation: Many proteins (and still not understood processes) are involved in unfolding just at the right time certain segments of the DNA, so that genes on that segment can be transcribed (provided the right TFs etc. are present).

![Diagram](image)

**Figure 12.12 DNA compaction and transcription.** (a) The sensitivity of DNA to DNase I measures levels of chromosome compaction.
Another example for the way in which gene regulation networks cause the emergence of spatial structure during development

Carroll, Grenier, and Weatherbee: From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design

Figure 3.4 The Drosophila life cycle

In *Drosophila*, the development of the segmented, motile larva from a fertilized egg takes about one day. After two larval instars, the imaginal discs of the late third larval instar that will give rise to adult tissues are well developed. Morphogenesis and differentiation of adult tissues take place during the pupal stage, before the adult emerges (eclosion).
Carroll, Grenier, and Weatherbee: From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design

Figure 3.5
The segmentation genetic regulatory hierarchy

(left) The expression patterns of five classes of anteroposterior axis patterning genes are depicted in embryos at different stages.
(right) Selected members of these classes are shown and the regulatory interactions between these genes are indicated. An arrow indicates a positive regulatory interaction; a line crossed at its end indicates a negative, repressive regulatory relationship.
Figure 3.6
Maternal gradients and gene activation

(a) The bicoid and nanos mRNAs are localized to the anterior and posterior poles, respectively, whereas the hunchback and caudal mRNAs are found throughout the syncytial embryo. (b) Diffusion of the Bicoid and Nanos proteins leads to the formation of concentration gradients. Bicoid activates hunchback transcription and represses Caudal translation; Nanos represses hunchback translation, leading to graded distributions of all four proteins. (c) The hunchback gene contains several binding sites for the Bicoid protein upstream of the promoter. Occupancy of these sites leads to activation of the hunchback gene in the anterior half of the embryo.


Carroll, Grenier, and Weatherbee:
From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design
Regulation of a pair-rule stripe: combinatorial control of an independent cis-regulatory element

The regulation of the *eve-skipped* cis-regulatory element controls the formation of the second stripe in the early embryo. (a) The stripe 2 element controls just one of seven stripes of *eve* expression. (b) The stripe forms within the domain of the Bicoid and Hunchback proteins and at the edge of the Giant and Krüppel gap protein domains. The former are activators, and the latter are repressors, of *eve stripe 2* expression. (c) The *eve stripe 2* element spans from about 1.7 to 1.0 kilobases upstream of the *eve* transcription unit. (d) Within this element, several binding sites for each regulator exist. The net output of the combination of activators and repressors is expression of the narrow *eve stripe 2*.

Another important role of gene regulation: Creation of temporal patterns

Example: Circadian clock

Forger et al. 2003, Development and Validation of Computational Models for Mammalian Circadian Oscillators
Clock Protein Families

<table>
<thead>
<tr>
<th>Protein</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLOCK</td>
<td>[MLL] [PAS] Srs2</td>
</tr>
<tr>
<td>MOP4</td>
<td>[MLL] [PAS]</td>
</tr>
<tr>
<td>BMAL1</td>
<td>[MLL] [PAS]</td>
</tr>
<tr>
<td>MOP9</td>
<td>[PAS]</td>
</tr>
<tr>
<td>PER1</td>
<td>[PAS]</td>
</tr>
<tr>
<td>PER2</td>
<td>[PAS]</td>
</tr>
<tr>
<td>PER3</td>
<td>[PAS]</td>
</tr>
<tr>
<td>CRY1</td>
<td>[FLAV] [PAS]</td>
</tr>
<tr>
<td>CRY2</td>
<td>[PAS]</td>
</tr>
<tr>
<td>CKI8</td>
<td>[ATP] [INACT]</td>
</tr>
<tr>
<td>CKI8</td>
<td>[ATP] [INACT]</td>
</tr>
</tbody>
</table>

Clock Gene Mutations

<table>
<thead>
<tr>
<th>Gene</th>
<th>Type</th>
<th>Behavioral Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock</td>
<td>Deletion</td>
<td>Long period to arrhythmic</td>
</tr>
<tr>
<td>BMAL1</td>
<td>Null</td>
<td>Arrhythmic</td>
</tr>
<tr>
<td>Per1</td>
<td>Null</td>
<td>Var. period to arrhythmic</td>
</tr>
<tr>
<td>Per2</td>
<td>Null</td>
<td>Var. period to arrhythmic</td>
</tr>
<tr>
<td>Per3</td>
<td>Null</td>
<td>Short period</td>
</tr>
<tr>
<td>Per1/2</td>
<td>Null/Null</td>
<td>Arrhythmic</td>
</tr>
<tr>
<td>Cry1</td>
<td>Null</td>
<td>Short period</td>
</tr>
<tr>
<td>Cry2</td>
<td>Null</td>
<td>Long period</td>
</tr>
<tr>
<td>Cry1/2</td>
<td>Null/Null</td>
<td>Arrhythmic</td>
</tr>
<tr>
<td>CKI8</td>
<td>Missense</td>
<td>Short period (tau hamster)</td>
</tr>
<tr>
<td>CKI8</td>
<td>Not Done</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

FIG. 3. Molecular components of the mammalian circadian clockwork. (Left) Four families of proteins are known to be important for circadian rhythms in mammals. Within the basic helix-loop-helix-PAS family, CLOCK and BMAL1 are most important for behavioral rhythms, while their close homologs, MOP4 (also called NPAS2) and MOP9 (also called BMAL2) appear less important within the central clock. The Period gene family has three members, mPer1, mPer2, and mPer3. The Cryptochrome family has two members, mCry1 and mCry2. Two casein kinases, CK1 epsilon and CK1 delta, are capable of phosphorylating PER proteins as well as other circadian proteins. Signature structural domains for each gene family are indicated. (Right) Mutations alter circadian behavioral rhythms in mice. For the Period and Cryptochrome gene families, partial redundancy of function is apparent. No mutations of CK1 delta have been reported. (Adapted from Reppert and Weaver, 2002.)

Forger et al. 2003, Development and Validation of Computational Models for Mammalian Circadian Oscillators
FIG. 5.  Comparison of model predictions with experimental data. Experimental data (points) for the time profile of total mPER levels were generated from data on the mPER1 and mPER2 rhythms in whole cell extracts from mouse liver, determined by quantitative western blotting (Lee et al., 2001). The model predictions are from the mammalian model developed by D.B. Forger and C. Peskin (2003). Shown is model prediction (line) and experimental data (points) in darkness after entrainment to a lighting cycle consisting of 12-h light and 12-h darkness per 24 h. The amplitude and shape of the profile are correctly predicted, and the phase of entrainment of the model accurately reflects the experimental data.
Fig. 2. Circadian oscillations in DD (A and B) and entrainment by LD cycles (C and D). (A) The mRNA of Bmal1 oscillates in antiphase with respect to the mRNAs of Per and Cry. (B) Corresponding oscillations of the PER, CRY, and BMAL1 proteins. (C) Oscillations of the mRNAs after entrainment by 12:12 LD cycles. The peak in Per mRNA occurs in the middle of the light phase. (D) Oscillations are delayed by 9 h and the peak in Per mRNA occurs in the dark phase when the value of parameter $K_{2c}$ is decreased from 0.6 to 0.4 nM. Other parameter values correspond to the basal set of values listed in Table 1. In C and D, the maximum value of the rate of Per expression, $v_{2c}$, varies in a square-wave manner such that it remains at a constant low value of 1.5 nM/h during the 12-h-long dark phase (black rectangle) and is raised up to the high value of 18 nM/h during the 12-h-long light phase (white rectangle). The curves have been obtained by numerical integration of Eqs. 1-16 (see Supporting Text) of the model without REV-ER$eta$.

Figure 6.3. The dynamics within the DR induced by teacher-forcing the sinewave $d(n)$ in the output unit. 50-step traces of the 20 internal DR units and of the teacher signal (last plot) are shown.

The difference between the genome of mice and humans

One can break the human genome into 170 fragments (of average length 17.6 Mb) and paste them together in a different order to arrive at a "reconstruction" of the mouse genome. In addition, many genes are moved to different positions within a chromosome, and they are duplicated (often thousands of times). It is commonly believed that these structural changes primarily change the regulation of genes, not so much genes (i.e., the encoded proteins) themselves.
Some operations different from mutation and crossover are known to occur on DNA-strings:

Unequal crossing over and movement of shorter or longer strings:

Figure 7.10  How unequal crossing-over and the movement of transposable elements (TEs) change DNA’s information content. 
(a) If two nearby regions contain a similar DNA sequence, the two homologous chromosomes may pair out of register during meiosis and produce gametes with either a deletion or a reciprocal duplication of the intervening region. Colorblindness in humans can result from unequal crossing-over between the nearby and highly similar genes for red and green photoreceptors. (b) TEs move around the genome. Some TEs copy themselves before moving, while others are excised from their original positions during the transposition process. Insertion of a TE into a gene often has phenotypic consequences.
**Duplications:**

<table>
<thead>
<tr>
<th>Types of duplications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tandem duplications</strong></td>
</tr>
<tr>
<td>Normal chromosome</td>
</tr>
<tr>
<td>Same order</td>
</tr>
<tr>
<td>Reverse order</td>
</tr>
<tr>
<td><strong>Nontandem (dispersed) duplications</strong></td>
</tr>
<tr>
<td>Same order</td>
</tr>
<tr>
<td>Reverse order</td>
</tr>
</tbody>
</table>

*Figure 13.8 Duplications: Structure, origin, and detection.*

(a) In tandem duplications, the repeated regions lie adjacent to each other in the same or in reverse order. In nontandem duplications, the two copies of the same region are separated.

---

**Inversions:**

Inverted segments are protected from changes by recombinations, since most crossing-overs with chromosomes that do not have the same inversion do not produce organisms that can survive.
Creation of super families through repeated composition of duplications with crossing-over, mutations etc.:
Review of some obvious structural differences between GAs and the creation of new species through the evolution of genomes:

• The coding strings that are considered in the context of GAs in Computer Science usually encode objects, whereas a genome encodes a process.

• It makes a big difference whether one changes stochastically the code for an object, or the code for a process.
• Duplications of genes are meaningful for the genome, since they change the transcription rate for those genes.

• **Movements** of genes are meaningful for the genome, since they place the same gene into a different regulation context and thereby change the transcription-rate for that gene in dependence of numerous external and internal signals.

• Large non-coding segments of the genome provide an interesting playground for mutations etc., where no current function is required (which is good since a mutation of a coding segment usually does no longer encode a useful protein), but where some segment of it may later become relevant when it encodes – after longer series of mutations etc., – again a protein, provided that this new protein (and its accidental regulation context) supports the survival of an organism.