Programming of Life: Bioinformatics Basics

Don Johnson
Ph.D. Chemistry: Michigan State Univ.
Ph.D. Computer & Information Sciences: U of Minn

Topics for This Presentation

What is a computer?
The nature of data versus 3 types of bio-information
The roles of chance and probability
Information and its processing systems in every cell
Prescriptive and information theory ramifications
Information and evolution
Unanswered problems

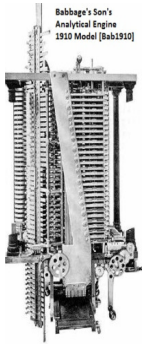
Life as Computer System?

Mechanical computer designed 1837

"The machine code of the genes is uncannily computer-like. Apart from differences in jargon, the pages of a molecular biology journal might be interchanged with those of a computer engineering journal." Dawkins River Out of Eden, p17

"Human DNA is like a computer program but far, far more advanced than any software we've ever created." *Bill Gates, The Road Ahead, p. 228.*

"Life is basically the result of an information process, a software process. Our genetic code is our software." Craig Venter, 2010 Guardian interview.



What is A Computer?

Necessary and sufficient requirements for a functional computer (mechanical, electronic, or biological) are:

- **Input (or embedded data)**
- **Memory and internal data transfer**
- **An instantiated algorithm (program)**
- **Processing capability**
- **Capability to produce meaningful output**

The Atanasoff-Berry (first electronic) Computer Couldn't be reprogrammed and had no branching instructions

Electronic and biological computers have multiple components

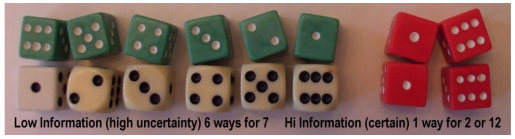
- **DNA/RNA can store program instructions to be executed**
- **Proteins can be processing and communication components**
- **Proteins and cellular controls are examples of output**

Data vs 3 Kinds of Information

Data may or may not have meaning
Binary is the smallest base to hold data in a bit
A binary digit (bit) can represent any 2 possibilities
married/single resident/nonresident male/female
If 110 is married nonresident female,
001 is single resident male (arbitrary)
Information: contingency ruling out other possibilities

- **Functional: useful/purposeful/meaningful**
- **Prescriptive: instructional/algorithmic choices**
- **Shannon: reduction of possibilities or uncertainty (no functionality required)**

Shannon Information (information theory)



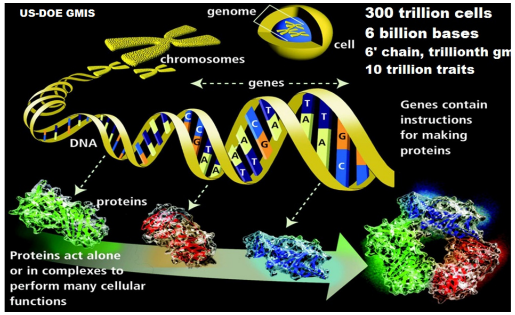
Purely probability-based – functionality not required
Redundant patterns provide no additional information
"junkjunkjunkjunk": only1st is fully informational
Shannon info defines limits on info storage or transmission
e.g. – Zip compresses file retaining Shannon info
Random data (0 functional info) has maximum Shannon info
Cannot be compressed using a more concise alphabet

Examples of Coded Functional Information & Data

- **Random coin tosses/1010100 = ASCII 'T' (head=1)**
- **Random die throws/Minneapolis area code**

Chance can't produce functional coded information
Pattern match probability: 1/128 (coins), 1/216 (dice)
Biosemiotics: cybernetic sign-systems (>20) in life
Arbitrary conversion by mutually agreed-to protocol rules (Rules not determined by law or physicality)

Simplified Genetic Code for Protein Construction



Simplified View of Life's Incredible Complexity

25,000 genes (many overlapping to produce >100,000 proteins)

"A single gene can potentially code for tens of thousands of different proteins... It's the way in which genes are switched on and off, though, that has turned out to be really mind-boggling, with layer after layer of complexity emerging" Le Page, "Genome at 10," New Scientist, 6/16/10.

Genome: Digital (base 4) self-correcting encoded information

- **Group of 3 1-of-4 bases (ACGT) : 4^3 (= 64) possible codons**
- **20 amino acids for proteins redundantly codon-specified**
- **Information in 1 teaspoon of DNA: all people + all books**
- **Information density is 1.88 X 10^{21} bits/cm^3**

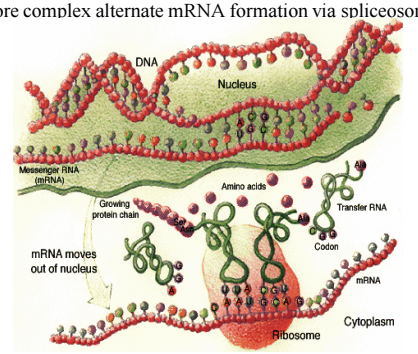
Even "simplest" organism's DNA has >150,000 nucleotides

DNA, proteins, etc. must be fully-formed/functional

- **>2000 enzyme proteins enable reactions**
- **Slowest non-enzymatic reaction would take a trillion yrs**

Simplified DNA Transcription/Translation Process

(more complex alternate mRNA formation via spliceosomes)



Information Systems in Life

- **Genetic system is a preexisting operating system**
- **Specific genetic program set (genome) is application**
- **Native language has codon-based encryption system**
- **Enzyme-based computers (with own OS) read codes**
- **Enzyme's output is to another OS in a ribosome**
- **Codons-specified amino acid is transported to protein construction site**
- **In each cell, there are multiple OSs, multiple programming languages, encoding/decoding hardware and software, specialized communications systems, error detection/correction mechanisms, specialized input/output channels for cell component control and feedback, and variety of specialized "devices" to accomplish the tasks of life.**

A comparative approach for the investigation of biological information processing, D'Onofrio & An, Theoretical Biology and Medical Modeling, 1/21/10

Disk/DNA properties & functional equivalences are compared

Chromosome/partition, file/gene, fragmentation/epigenome

"The cell is viewed as a complete computational machine in terms that are akin to a multi-core computer cluster… as a system with centralized memory with multi-access capability leading to distinct computing units."

Biosemiotics: Arbitrary cybernetic sign-system

Information transfer from protein to RNA is impossible (20 to 64 symbols exceeds Shannon channel capacity)

Life's initial alphabet was at least that of codon alphabet

(See Appendix D in PoL or PDF's extra slide for technical details)

Algorithmic Prescriptive Information (PI) in Life

DNA gene sequences are real computer programs
Chance & law can't explain decision nodes (choice)

PI is intrinsically formal, but implemented physically
Abel, "The Biosemiosis of Prescriptive Information," Semiotica:174-1, 2009, p1-19

A nucleotide can be in multiple prescriptions

"No rational scientific basis exists for blindly believing in a relentless uphill push by mere physicality toward formal algorithmic optimization" Abel & Trevors, "Self-Organization vs Self-Ordering events in Life-Origin Models," Physics of Life Rev:3, 2006, p211-228.

"The Origin-of-Life Prize® ... will be awarded for proposing a highly plausible natural-process mechanism for the spontaneous rise of genetic instructions in nature sufficient to give rise to life."

Neo-Darwinian Biology: Random mutation/Selection

Richard Dawkins: "Each nucleus ... contains a digitally coded database larger, in information content, than all thirty volumes of the Encyclopedia Britannica." "Each successive change in the gradual evolutionary process was simple enough, relative to its predecessor, to have arisen by chance… Even if the evidence did not favour it [evolution by natural selection], it would still be the best theory available!" "Mutation is not an increase in true information content, rather the reverse." ***Climbing Mount Improbable, Blind Watchmaker, Information Challenge***

"The failure to observe even one mutation that adds information is more than just a failure to support the theory. It is evidence against the ... neo-Darwinian theory." ***Spector, Not By Chance, p160***

Information Increase Moving up Tree

- **The simplest life has only 267,000 information bits**
- **Human DNA has over 6 billion information bits**
- **Based on functional information, simplest life is $10^{300,000,000}$ more probable than man**
- **No mechanism to produce ANY net info increase**

  New functionality offset by functionality loss, e.g.--

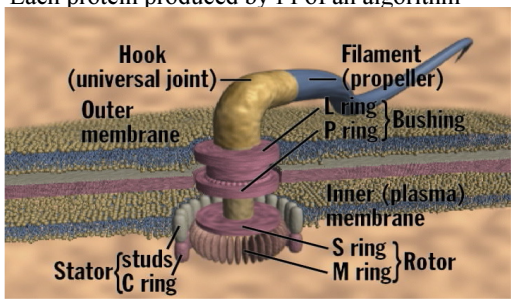
  Single mutation causes sickle cell anemia

  Nylon-eating bacteria: frame-shift/plasmid transpose
- **"We must concede there are presently no detailed Darwinian accounts of the evolution of any biochemical or cellular system, only a variety of wishful speculations"** Harold, The Way of the Cell, 2001, p205.

Bacterial Flagellum: Irreducibly Complex

48+ proteins (>30 unique): <1 in $10^{5,250}$ probability

Each protein produced by PI of an algorithm



Computer Simulations & Artificial Life

Dawkins *(Scientific American, 6/88)* randomly changed:

"WDLTMNLT DTJBKWIRZREZLMQCO P"

to produce on the 43rd try:

"METHINKS IT IS LIKE A WEASEL"

He knew the goal in advance and stopped mutation if correct proving that programmers can solve problems using computers.

"Everywhere on the apparatus and in the 'genetic algorithms' appear the scientist's fingerprints: the 'fitness functions' and 'target sequences.' These algorithms prove what they aim to refute: the need for intelligence and teleology [targets] in any creative process." George Gilder, "Evolution and Me," National Review, 7/17/06

"Neglect of key factors or unrealistic parameter settings permit conclusions to be claimed which merely reflect what the decision maker intended a priori." Royal Truman, "Evaluation of Neo-Darwinian Theory Using the Avida Platform," PCID 3.1.1, 11/04.

Darwinism Doubted by Thousands of Scientists

"**The complexity of biology has seemed to grow by orders of magnitude… Biology's new glimpse at a universe of non-coding DNA — what used to be called 'junk' DNA — has been fascinating and befuddling… the signaling information in cells is organized through networks of information rather than simple discrete pathways. It's infinitely more complex**." Erika Hayden, "Life is Complicated," Nature, 4/10, p664-667

"**Much of the vast neo-Darwinian literature is distressingly uncritical... Natural selection has shown insidious imperialistic tendencies."** Fodor & Piattelli-Palmarini, "Survival of the fittest theory: Darwinism's limits," New Scientist, 2/3/10

"Natural selection is not a mechanism, it's the process by which the results of evolution are sorted." Bruce Runnegar, p188 of The Altenberg 16: An Exposé of the Evolution Industry, 2010 (Mazur)

Evolution via Natural Genetic Engineering

"Molecular cell biology has revealed a dense structure of information-processing networks …The natural genetic engineering functions that mediate genome restructuring are activated by multiple stimuli…One of the traditional objections to Darwinian gradualism has been that it is too slow and indeterminate a process to account for natural adaptations, even allowing for long periods of random mutation and selection … natural genetic engineering … employs a combinatorial search process based upon DNA modules that already possess functionality … Such a cognitive component is absent from conventional evolutionary theory because 19th and 20th century evolutionists were not sufficiently knowledgeable about cellular response and control networks." James A Shapiro, "Mobile DNA and evolution in the 21st century," Mobile DNA 1/25/10

Science Needs to Provide Plausible Mechanisms to Explain **How did nature:**

write the prescriptive programs needed to organize life's metabolism?

formally solve life's other complex problems and write the programs?

develop the operating systems and programming languages?

develop the arbitrary protocols for communication and coordination among the thousands (or millions) of computers in each cell?

develop alternative generation of prescriptive messages using techniques such as overlapping genes, messages within messages, multi-level encryption, and consolidation of dispersed messages?

defy computer science principles by avoiding software engineering's top-down approach required for complex programming systems?

produce complex functional programs without planning by randomly modifying existing algorithms?

simultaneously modify multiple such programs to result in the production of irreducibly complex structures?

(from "Programming of Life" -- www.djpol.info)

Defeating Creationism in the Courtroom, But Not in the Classroom

Berkman & Plutzer, 1/28/11 Science. P404-405

Only "28% of all biology teachers consistently implement the major recommendations and conclusions of the National Research Council"

Recommended fix is for those "who cannot accept evolution as a matter of faith to pursue other careers."

Their plan "would reduce the supply of teachers who are especially attractive to the most conservative school districts."

Such statements destroy the credibility of science.

Science evaluates evidence (not dogmatic doctrine)

Genetic Problems not Usually Considered

Harmful mutations limit life's existence

current 60 per newborn human -- extinction in <10ky

Fitness declines by 1-2% per generation (1995 J. Theo. Biol. Paper title: "Why have we not died 100 times over?")

>300 generations would cause certain extinction!

Each mutation causes a guaranteed net information loss in the genome (DNA), changing the prescriptive program

"Stunningly, information has been shown not to increase in the coding regions of DNA with evolution. Mutations do not produce increased information... the amount of coding in DNA actually decreases with evolution" [David Abel, "The GS (genetic selection) Principle," Frontiers in Bioscience (14), 1/1/09, p2959-2969]

Summary

Life has the necessary & sufficient computer requirements

Data is functional information only if it can be used

Life is incredibly complex and information rich

Is information science incorrect?

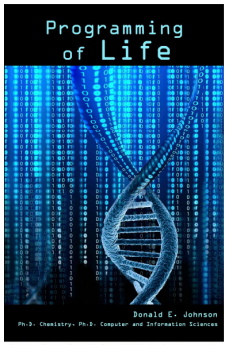
- **Can chance produce complex functional information?**
- **Can multiple mutational information losses cause gain?**
- **Can chance produce codes or formal protocols?**
- **Was life's first code simpler than the current codon code?**
- **Can chance write prescriptive algorithms (programs/OSs)?**
- **Can chance create genetic engineering capability?**

Scenarios proposed inadequately address information

- **Assertions for origins of life & species need verification**
- **Other avenues may provide more fruitful paths**

Science speculation is inappropriate for non-scientists

POL highlights the informational aspects of life that are usually overlooked or ignored in chemical and biological evolutionary. Each cell of an organism has thousands (or millions) of interacting computers reading and processing digital information using algorithmic digital programs and digital codes to communicate information. Most scientists have been attempting to use physical science to explain life's information domain, a practice which has no scientific justification. For more info see www.djpol.info and video at www.programmingoflife.com



Amazon Bestsellers Rank: #2,685 in Books (after 4 weeks)
#1 in Books > Science > Evolution > Molecular Biology
#14 in Books > Science > Biological Sciences > Biology

Shannon Channel Capacity (maximum mutual entropy)

Mutual entropy between input (x) and output (y) channels

$\mathbf{I(B;A) = I(A;B) = H(x) - H(x|y)}$ (for alphabets A & B)

has conditional (x_i given y_j received) entropy

$\mathbf{H(x|y) = -\sum_{ij} p_j\, p(i|j)\, \log_2 p(i|j)}$ and information entropy

$\mathbf{H(x)_n = -\sum_{i=1}^{n} p_i \log_2(p_i)}$ with probability vector elements

$\mathbf{p_j = \sum_{i}^{} p_i p(j|i)}$ (relates to conditional probability matrix)

The DNA-mRNA-protein system is:

discrete because all symbols in the alphabet are defined,

memoryless with no dependence on previous symbols, and

unconstrained since any symbol may follow any symbol.

Therefore, a particular DNA message can be treated as one member of a stochastic ensemble generated by a stationary Markov process, completely characterized by probability space coding $[\Omega, A, \mathbf{p_A}]$, $[\Omega, B, \mathbf{p_B}]$ with A(& any predecessors) $\geq$ B (or channel capacity would be exceeded).