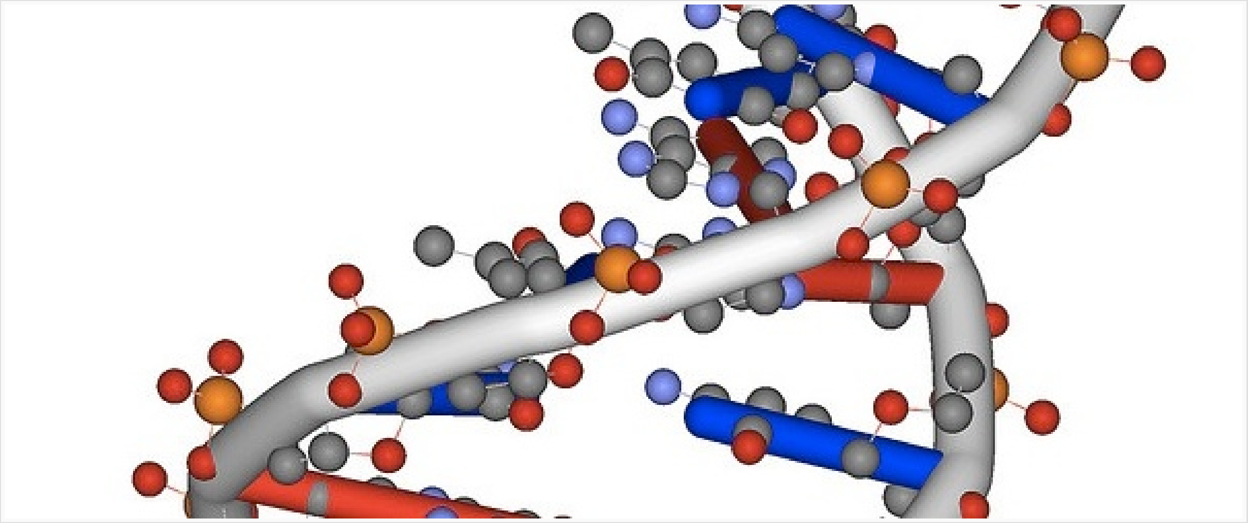
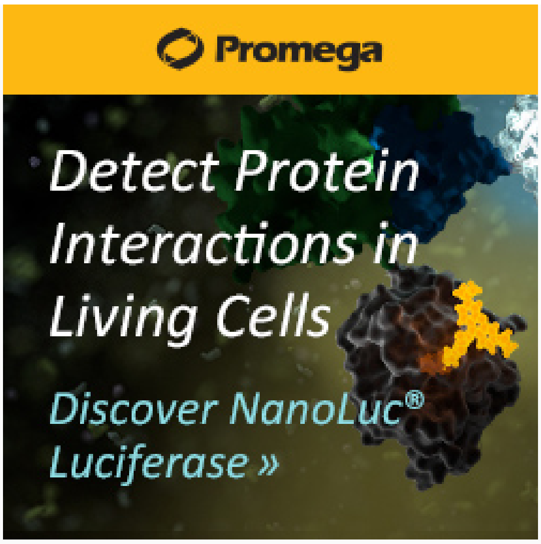
How Much Information is Stored in the Human Genome?

by Yevgeniy Grigoryev on 16th of March, 2012 in Inspiring & Thought Provoking



The other day I was having a conversation with a friend of mine who had some background in computer science. The conversation shifted towards my research and the following question came up: What is the amount of digital information stored in a human genome? I started searching in the deep dark corners of my brain, but I realized that I simply did not know the answer. So I decided to do the math to estimate how much information is stored in our genome.

Laying out the information storage capacity of the


Promega
Detect Protein Interactions in Living Cells
Discover NanoLuc® Luciferase »

genome

The human genome contains the complete genetic information of the organism as DNA sequences stored in 23 chromosomes (22 autosomal chromosomes and one X or Y sex chromosome), structures that are organized from DNA and protein. A DNA molecule consists of two strands that form the iconic double-helix "twisted ladder", whose backbone, which made of sugar and phosphate molecules, is connected by rungs of nitrogen-containing bases. DNA is composed of 4 different bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). These bases are always paired in such a way that Adenine connects to Thymine, and Cytosine connects to Guanine. These pairings produce 4 different base pair possibilities: A-T, T-A, G-C, and C-G. The haploid human genome (containing only 1 copy of each chromosome) consists of roughly 3 billion of these base pairs grouped into 23 chromosomes. A human being inherits two sets of genomes (one from each parent), and thus two sets of chromosomes, for a total of 46 chromosomes, representing the diploid genome, which contains about 6×10^9 base pairs.

Comparing the genome to computer data storage

In order to represent a DNA sequence on a computer, we need to be able to represent all 4 base pair possibilities in a binary format (0 and 1). These 0 and 1 bits are usually grouped together to form a larger unit, with the smallest being a "byte" that represents 8 bits. We can denote each base pair using a minimum of 2 bits, which yields 4 different bit combinations (00, 01, 10, and 11). Each 2-bit combination would represent one DNA base pair. A single byte (or 8 bits) can represent 4 DNA base pairs. In order to represent the entire diploid human genome in terms of bytes, we can perform the following calculations:

6×10^9 base pairs/diploid genome x 1 byte/4 base pairs = 1.5×10^9 bytes or 1.5 Gigabytes, about 2 CDs worth of space! Or small enough to fit 3 separate genomes on a standard DVD!

Data storage across the whole organism

Some interesting question could follow. For example, how many megabytes of genetic data are stored in the human body? For simplicity's sake, let's ignore the microbiome (all non-human cells that live in our body), and focus only on the cells that make up our body. Estimates for the number of cells in the human body range between 10 trillion and 100 trillion. Let us take 100 trillion cells as the generally accepted estimate. So, given that each diploid cell contains 1.5 GB of data (this is very approximate, as I am only accounting for the diploid cells and ignoring the haploid sperm and egg cells in our body), the approximate amount of data stored in the human body is:

1.5 Gbytes x 100 trillion cells = 150 trillion Gbytes or 150×10^12 x 10^9 bytes = 150 Zettabytes (10^21)!!!

Sexual information exchange

Along the same lines, how much genetic data is exchanged during human reproduction?Each sperm cell in a human male is heterogametic and haploid, meaning that it contains only one of two sex chromosomes (X or Y) and only one set of the 22 autosomal chromosomes. Thus, each sperm contains about 3 billion bases of genetic information, representing 750 Mbytes of digital information. The average human ejaculate contains around 180 million sperm cells. So, that's 180 x 10^6 haploid cells x 750 Mbytes/haploid cell = 135 x10^9 Mbytes=135000 Terabytes!!!! Following this idea even further, while 13500 Tbytes are transferred, only one sperm cell will fuse with an egg, using only 750 Mbytes of data, combining it with another 750 Mbytes of data from the egg. Thus, essentially 99.9999…% of the data transferred during sexual reproduction is lost in the pipeline … Whether the remaining fraction of information will result in anything constructive is up to good parenting.

Having worked out the above numbers, a whole bunch of other curious questions can be asked. Have you ever wondered about the data capacity of our biological organism? What is the rate of data transmission during cell division? The rate of data transmission during gamete fusion? The rate of data transmission when human lymphocytes circulate through the bloodstream? What amount of data is destroyed daily by apoptosis? What amount of data is created daily? How does this compare to the rate of data transfer via an optical fiber?

Please feel free to contribute your own dubious calculations and questions below!

Photo Credits

- ghutchis

Related Content

Love Is In The Air

How to Format Your Manuscript

Keeping Your Science Out of the (Junk) Headlines

9 thoughts on "How Much Information is Stored in the Human Genome?"

Balaclava says: March 20, 2012 at 4:28 pm

You are completely ignoring the fact that information in DNA is not only present in the base sequences but also (and maybe more so!) in its methylation profile. So when you are trying to represent a single specific genome in the form of bytes you have to take that into consideration.

Furthermore, when you are calculating the amount of data stored in the entire human body you have to take into account that not all promoters are in the same state in every cell (that's why every cell is different..!)

So how would you represent acetylations and methylations?

Log in to Reply

Yevgeniy Grigoryev says: March 20, 2012 at 8:16 pm

Dear Balaclava,

Your observation about the methylation and acetylation profile are very valid, such epigenetic factors make the coding capacity of our genome almost infinite! However, what I tried to calculate was mere data "stored", not expressed by the human genome. My calculations are oversimplified, of course. I doubt that currently there is a way to calculate the amount of data "expressed" in the genome, that also factors in all the genome-wide epigenetic modification events such as methylation and acetylation.

You are also absolutely right about the varying promoter states across different cells. However, the promoter expression does not affect the data stored. I think it would be virtually impossible to calculate the data actually expressed at any moment, taking into account any all the possible promoter states and epigenetic events. Of course, all attempts at such daring tasks are more than welcomed here, this is the point of this post, after all.

Log in to Reply

Emily Crow says: March 20, 2012 at 8:44 pm

What do you think epigenetics would equate to in computer terms – processing power? Could different promoters be thought of as the amount of operations that can be carried out simultaneously??

Log in to Reply

Yevgeniy Grigoryev says: March 21, 2012 at 4:45 pm

So my knowledge of how computer works is rather limited but I think processing power is a good comparison. While the genetic data stored can be compared to RAM (Random access memory, a form of computer data storage). The amount of data expressed can be compared to processing power or computing power. The epigenetics machinery and promoters can be compared to computer processors that handle massive calculations. Alternatively, they can probably also be compared to the CPU's electronic clock, that creates a series of electrical pulses at regular intervals. This allows the computer to synchronize all its components and determine the speed at which the computer can pull data from its memory and perform calculations. In a cell that would equate to all the active transcriptional states at any given time.

Log in to Reply

1857 says: March 23, 2012 at 8:34 am

hi, this is leili, i would tahnk you if you add more last seminars and congress in biology, please,,,,,,,,,,,

Log in to Reply

DougB says: April 11, 2012 at 5:48 pm

I have say much credit is due for even attempting the calculation. How about the data storage capacity of cellular mitochondria? It seems this also would add to the total. I am curious to know how your friend responded to the storage capacity of

DNA.

Log in to Reply

chrismckay says: September 27, 2012 at 2:00 am

Your calculations are based on the chromosome, I pose the question would 150 Zettabytes be enough to store the human body as a binary form and then be able to reintegrate it?

Log in to Reply

smartmoves says: July 5, 2013 at 2:52 am

Give or take an order of magnitude, a 1Ghz async network would take 1,750 years to transfer that much data. I guess "beam me up Scotty" is out to lunch for a while…

Log in to Reply

Mike Levens says: February 8, 2014 at 5:15 pm

All of this is a good discussion of the amount of data in the human genome, but it's not really the same as discussing the amount of *information.* Storing an arbitrary 150-zettabyte value as a human would require the possibility of having human beings who have a completely different genome in each and every cell, which is obviously crazy. So how much redundancy is there, and what is the Shannon entropy of the data being stored?

Suppose my sci-fi future self wanted to write a compression algorithm for storing and transmitting a particular human genotype, say for molecular reassembly and cloning at a remote location or something. What data compression ratio could I achieve in theory? How big would this 150 zettabytes look like if LZW'd?

Log in to Reply

Leave a Reply

You must be logged in to post a comment.
Click below to connect to Bitesize Bio using your existing social network to start a new subscription or update your subscription preferences. If you already have a Bitesize Bio login please enter your details below to add new subscriptions.

SEARCH

Share your tips, advice & wisdom on Bitesize Bio

We gather the best tips, advice and wisdom from you guys at the bench and publish them to help each other improve in the lab.

What could you add to this collection? Click here to explore how you could contribute to Bitesize Bio.

Recent Posts

A Primer on Designing Site-Directed Mutagenesis Primers

Gel Electro-For-Whatsit? Breaking Down How Gel Electrophoresis Works

Analysing Microscopy Images? What You Should Know About Dynamic Range: Part 2

Get Your Proteins! Hot Proteins Here! Radioactively Labeled Proteins!

Chasing the Pot of Gold at the End of the Rainbow: Choosing the Right Fluorochromes for Your Flow

Cytometry

Categories

Analytical Chemistry

Basic Lab Skills & Know-how

Bioinformatics

Career Development & Networking

Cell / Tissue Culture

Chemistry for Biologists

Cloning & Expression

Dealing with Fellow Scientists

Epigenetics

Equipment Mastery & Hacks

Flow Cytometry

Fun Stuff

Genomics

Getting Funded

History of Biology

Inspiring & Thought Provoking

Lab Safety

Lab Statistics & Math

Microbiology

Microscopy & Imaging

Next Generation Sequencing

Nucleic Acid Purification and Analysis

Of Interest

Organization & Productivity

PCR & Real-time PCR

Personal Development

PhD Survival

Protein Analysis, Detection & Assay

RNAi

Science Communication & Ethics

Software & Online Tools

Survive & Thrive

Taming the Literature

Writing, Publishing & Presenting

Navigation

Channels

Articles

Webinars

Questions

Products

Contact

Contribute

Latest Posts

A Primer on Designing Site-Directed Mutagenesis Primers

Gel Electro-For-Whatsit? Breaking Down How Gel Electrophoresis Works

Analysing Microscopy Images? What You Should Know About Dynamic Range: Part 2

Get Your Proteins! Hot Proteins Here! Radioactively Labeled Proteins!

Chasing the Pot of Gold at the End of the Rainbow: Choosing the Right Fluorochromes for Your Flow Cytometry

Recent Questions

Volunteering in a lab

Can a protein precipitate from a RNA islation assay be used for protein analysis

Calculation of number copies/ soil from cDNA

O.D at 600 nm

Standard curve in qPCR

Tags

Alkaline lysis Antibiotics Bio-audio Bitesize Bio Books Cloning Co-IP cryo sectioining Disasters DNA extraction DNA ligation Electroporation ELISA Gene Synthesis GMO Good practice GottaDance H&E Harlem Shake histology History How stuff works how to immunoprecipitation IP Lab Hacks Look after yourself Mac microbiology microscopy paraffin embedding PC PCR & Real-time PCR Protein Biochemistry protein detection protein quantification Pubmed qPCR resin embedding special stains Tech Guides Tech Tips Tech Tips and Guides Western blot Western Blotting

CHANNELS ARTICLES WEBINARS QUESTIONS PRODUCTS CONTACT CONTRIBUTE