

Computer Science and Biology

Jun 8, 2007
KangKang Yin

before we talk about science

GATTACA
(1997 Sony Pictures)

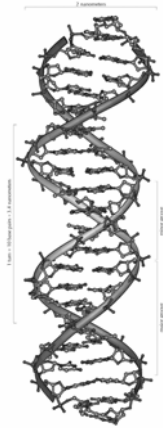
The Human Genome Project

- ▶ to decode (i.e. sequence) more than three billion nucleotides
- ▶ Started in 1990
 - by the U.S. Department of Energy and National Institutes of Health
 - international and academic
 - was expected to take 15 years
- ▶ Celera changed the game
 - private firm
 - started in 1998 and claimed completeness in 2 years

DNA: The Famous Double Helix

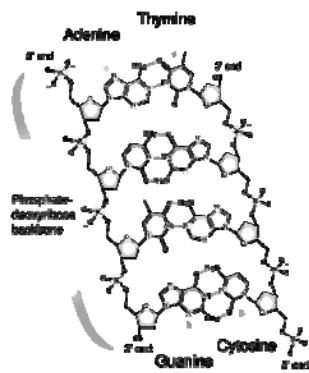
3 billion base pairs
<http://en.wikipedia.org/wiki/DNA>

Genes - segments of our DNA



Bases

four bases: A,C,G,T
two pairs: A-T, G-C
<http://en.wikipedia.org/wiki/DNA>



Proteins

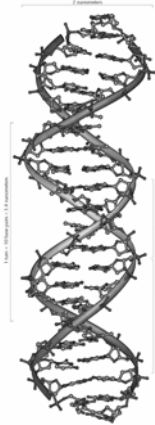


body's activists: carry blood, digest food, form hair...
3D complex shapes that can fold
shape is key, and determined by genes
<http://en.wikipedia.org/wiki/Proteins>

Genes

- ▶ segments of our DNA
- ▶ we share
 - 99.9% of our genome with each other
 - 98% of our genome with chimpanzees
 - 50% of our genome with bananas
- ▶ *mutations* – changes in the bases of a gene – can cause genetic diseases

<http://en.wikipedia.org/wiki/DNA>



decode the Genome letter sequence

- ▶ the foundation to understand and cure genetic diseases
- ▶ Cracking the code of life
(<http://www.pbs.org/wgbh/nova/genome/NOVA>)
 - How it's done in the old days (2. Getting the Letters Out)
 - How it's done nowadays (4. The Sequencing Race Begins)

giant biological jigsaw puzzle

- grab a piece of the genome



- make copies



- chop into fragments



- sequence fragments

CAAGACAA CAACAATA TTACGGGCC

- record, assemble fragments on a computer, to produce sequence:

CAAGACAA TTACGGGCC
CAACAATA
CAAGACCAACAATAACGGGCC

Sequence for Yourself

- ▶ <http://www.pbs.org/wgbh/nova/genome/sequencer.html>

To be continued