Information Storage in DNA

Nick Goldman
European Molecular Biology Laboratory
European Bioinformatics Institute
Hinxton, Cambridge, UK

@EBIgoldman
goldman@ebi.ac.uk

IC Research Day – 2016, EPFL, Lausanne, 30.vi.16
Claude Shannon
PhD thesis, 1940,
MIT and
Cold Spring Harbor
Laboratory

“An algebra for
theoretical genetics”
Populations: African Ad Mixed East American European Asian South Asian


Victoria Tower, London: British law recorded on vellum
Qianlong Stone Steles, Beijing Guozijian (Imperial Academy), Temple of Confucius
Damascus steel: 3rd – 18th Centuries, Damascus, Syria
European Molecular Biology Laboratory
Member States

The five branches of EMBL

- Heidelberg, DE
  - Basic research in molecular biology + Administration
- Hamburg, DE
  - Structural biology
- Hinxton, UK
  - Bioinformatics
- Grenoble, FR
  - Structural biology
- Monterotondo, IT
  - Mouse biology

- 1500 staff
- >60 nationalities
What is EMBL-European Bioinformatics Institute?

- International, non-profit research institute
- Part of the European Molecular Biology Laboratory
- Europe’s hub for biological data services and research
- 580 members of staff from >50 nations.

OUR MISSION

To contribute to the advancement of biology through investigator-driven research in bioinformatics
Research themes

Genes
- Nick Goldman
- Ewan Birney
- Paul Flicek

Expression
- Anton Enright
- John Marioni
- Oliver Stegle
- Alvis Brazma

Chemical biology
- Christoph Steinbeck

Proteins & structures
- Janet Thornton
- Pedro Beltrao
- Alex Bateman

Systems biology
- Moritz Gerstung

Tree of Life

You are here
Making sense of incongruent trees from different loci


OUR MISSION

To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress
Infrastructures are critical…

Types of data

- Genomes
- Nucleotide sequence
- Gene expression
- Protein families, domains and motifs
- Protein-protein interactions
- Pathways
- Literature
- Protein sequence
- Proteomes
- Protein structure
- Chemical entities
- Systems
- Literature
You are here

Tree of Life

DNA

4 Å = 0.4 nm = 0.4 x 10^-9 m

Nucleobases
Base pair
Helix of sugar-phosphates

Cytosine
Guanine
Adenine
Thymine

Nucleobases of DNA
High-throughput DNA sequence readers (DNA sequencing)

High-throughput DNA sequence writers (oligo synthesis)
High-throughput DNA sequence writers (oligo synthesis)
- EBI.jp2 — 184264 bytes — JPEG 2000 format image file
  colour photograph (16.7M colours, 640 x 480 pixel resolution) of the
  European Bioinformatics Institute (authors’ own picture)

- MLK_excerpt_VBR_45-85.mp3 — 168539 bytes — MP3 format sound file
  26 second-long extract from
  Martin Luther King’s
  “I Have A Dream” speech
  (from http://www.americanrhetoric.com/
  speeches/mlkihaveadream.htm,
  modified to achieve higher
  compression: variable bit rate,
  typically 48–56kbps; sampling
  frequency 44.1kHz)
Watson and Crick’s (1953) publication describing the structure of DNA (from the Nature website, http://www.nature.com/nature/dna50/archive.html, modified to achieve higher compression and thus smaller file size)

all 154 Shakespeare sonnets (from Project Gutenberg, http://www.gutenberg.org/ebooks/1041)
1 MB of computer files = almost invisible dust of DNA
Jarvis, Goldman et al., Music of the Spheres, www.artforeating.co.uk
1 test-tube full of DNA = 1 000 000 CD ROMs

Encoding technique scales up to and beyond global data volumes
With DNA it is easy to make many copies of your digital data.

People will always be interested in DNA.
What are the properties of our communication channel? or

What happens to our fragments of DNA?

**Synthesis:**

- deletion
- insertion
- substitution

Example DNA sequence: `CGATCGTAGACGTATACGTA`
What are the properties of our communication channel?
or
What happens to our fragments of DNA?

Synthesis:

Storage:

Sequencing:
deletion
insertion
substitution
deletion
(another kind)
and additional thanks to:

**Cambridge:**
- David MacKay
- Graeme Mitchison

**EMBL-EBI:**
- Kevin Gori, Daniel Henk*, Remco Loos*, Ari Löytynoja*, Hazel Marsden*, Tim Massingham*, Sarah Parks*

**EMBL-Heidelberg:**
- Vladimir Benes, Đinko Pavličić, Jonathon Blake

**EMBL-EM:**
- Birgit Kerber, Boris Bryk

**Art for Eating:**
- Charlotte Jarvis

---

* = now moved to new positions
Your DNA will be your data.