

3-3-2021

Exploring Biological Information: Document and discover Descriptions of Life

Chao Cai
Purdue University, caic@purdue.edu

Follow this and additional works at: https://docs.lib.purdue.edu/lib_fspres



Part of the [Library and Information Science Commons](#), and the [Life Sciences Commons](#)

Recommended Citation

Cai, Chao, "Exploring Biological Information: Document and discover Descriptions of Life" (2021).
Libraries Faculty and Staff Presentations. Paper 169.
https://docs.lib.purdue.edu/lib_fspres/169

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries.
Please contact epubs@purdue.edu for additional information.



Exploring Biological Information:

Document and Discover Descriptions of Life

Chao Cai, Ph.D.

Assistant Professor

Plant Sciences Information Specialist

Purdue Libraries and School of Information Studies

March 3rd, 2021 PULSIS Brown Bag

Outline

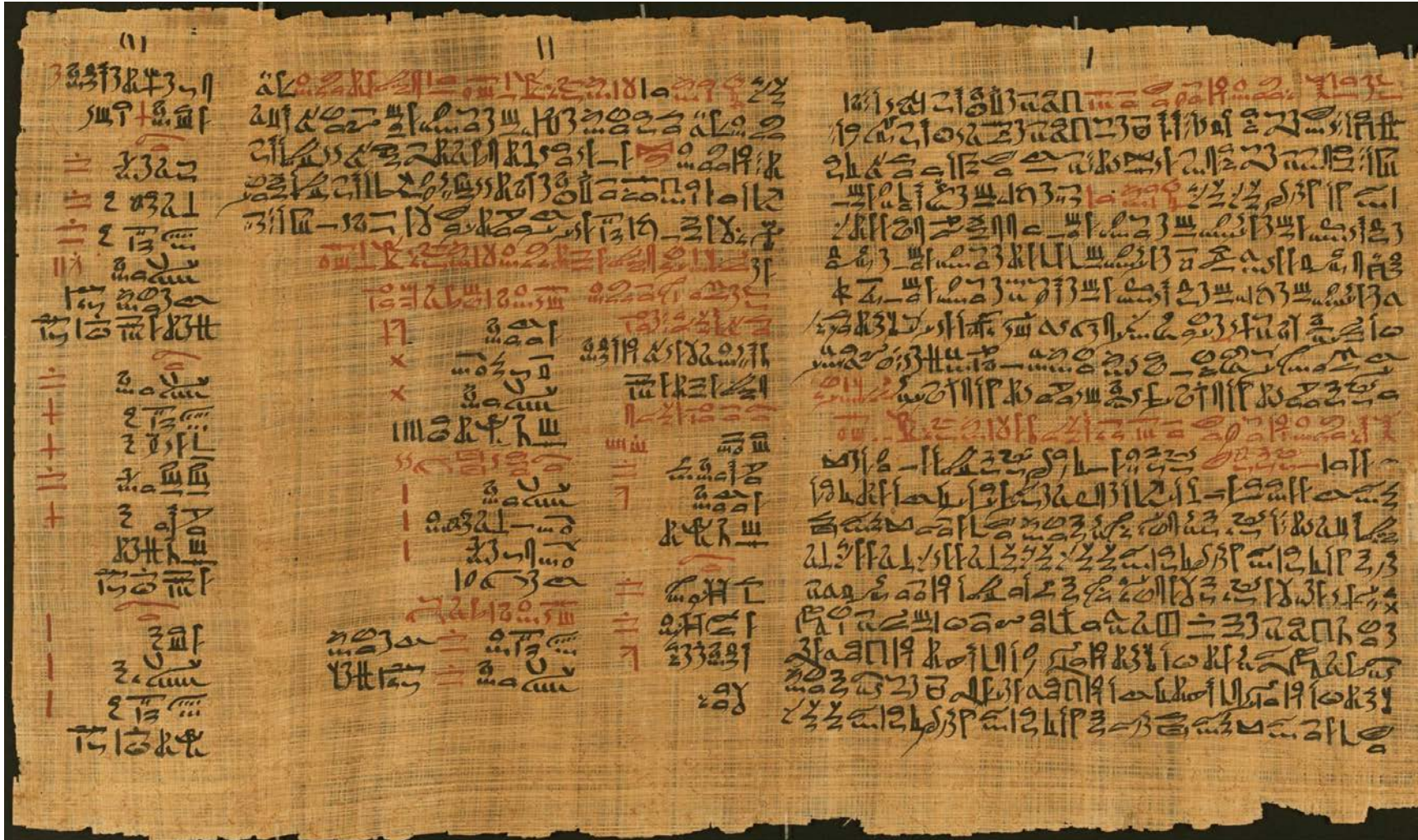
- **Brief history of human documentation of living organisms**
- **Biological information in the genomic era**
- **New tools for biological information search/discovery**

Humans have been documenting living organisms since paleolithic



Cave paintings of animals from 40,000 years ago, found in Borneo.

Invention of writing systems made documenting life possible



Written in hieratic around 3500 years ago, **Papyrus Ebers** is one of the earliest medical documents, which described extensive anatomical understanding and surgical practices by ancient Egyptians.

Invention of writing systems made documenting life possible



The brutal penalty for medical malpractice boosted the careful documentation of medical practice.

If the doctor shall treat a gentleman and shall open an abscess with the knife and shall preserve the eye of the patient, he shall receive ten shekels of silver. If the doctor shall open an abscess with a blunt knife and shall kill the patient or shall destroy the sight of the eye, his hands shall be cut off or his eye shall be put out.²

(Anthony Serafini, 1993, [The Epic History of Biology](#))

Preserved specimens are critical tools for documenting biological information



(Mummified Ibis of ancient Egypt, [Brooklyn Museum](#))



(Specimen FMNH PR 2081, [Field Museum of Natural History](#))



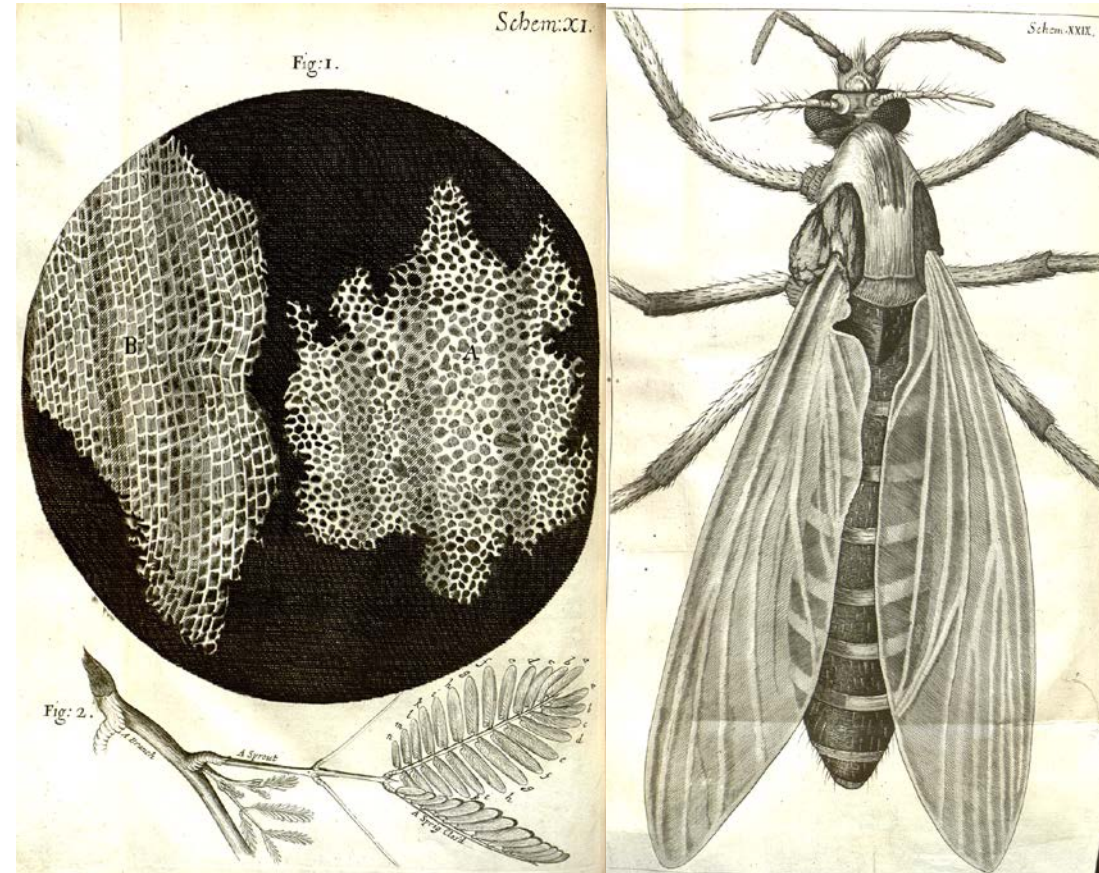
(Specimen of lily, from [Purdue Herbaria](#))

Text, illustration and preserved specimen remain to be major methods for documenting biological information

- Biology became a science in ancient Greece (βίος - λογία)

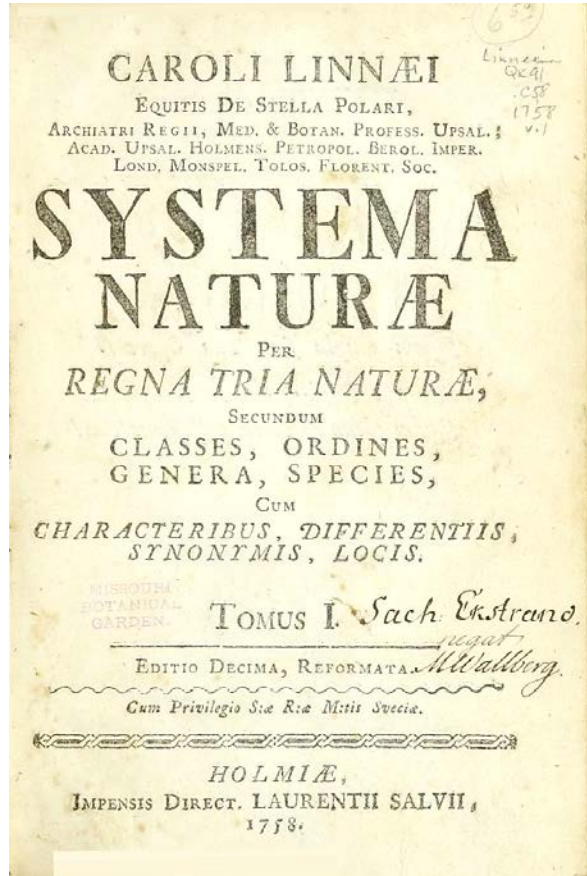


(William Harvey, 1628, De Motu Cordis)



(Robert Hooke, 1665, Micrographia)

Text, illustration and preserved specimen remain to be major methods for documenting biological information



(Carl Linnaeus, 1758, Systema Naturae)

IMPERIUM N
HOMO sapiens, creatorum opti & summum, in Telluris cortice, iudicis obsequio, constitutus, fecundum, admirans pulchritudinem, progrediendo per generationem rarum; ad utrumque invitans Nexus, Fines, Commoda. Si plenus est (m), dum omnia creat hominem (n), qui ex inertis humo, platur auctoris sui Majestatem ei ipse constitutus Summi Entis præevadit præsentia voluptatis caelestis luce obambulat & tanquam in te Nec pietas adversus Deum, nisi sine explicatione Naturæ in
SAPIENTIA, divina particula au Hominis Sapiens. Primus Sapiens Notitia consistit in vera idea objectibus distinguuntur notis propriis, hanc notitiam ut cum aliis comm confundenda singulis diversis inveniunt, perit & rerum cognitio. Et quibus destitutus nemo naturam iure proprio, nulla descriptio quam demonstrat, sed plerumque fallit
METHODUS, anima Scientiæ, inditas systematice digestas; S; subdividitur: sic
Classis, Ordo, Genus
Genus sum. G. intermedium, G. pro Provincia, Territoria, Pars
Legiones, Cohortes, Mani
Nisi enim in ordine redigantur distribuuntur tumultu & pueri cessat esse (r).
NOMINA respondeant Methodo
Nomina Classium, Ordinum, Gen
Character. Classium, Ordinum, Gen
Differentis definita, nam nomina
A 4
(m) Eft. V: 4. (n) David. CXVIII.
(o) Casalpini. (r) Casalpini.

period, the homologous chromosomes unite in pairs. There has been much controversy as to how this union takes place, but in some cases at least, the uniting chromosomes twist around each other as they come together. This is illustrated to the left in Fig. 24. As a consequence, parts of one chromo-

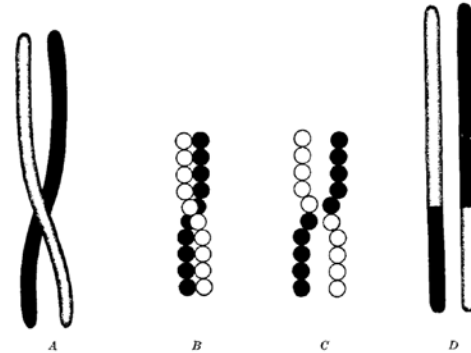
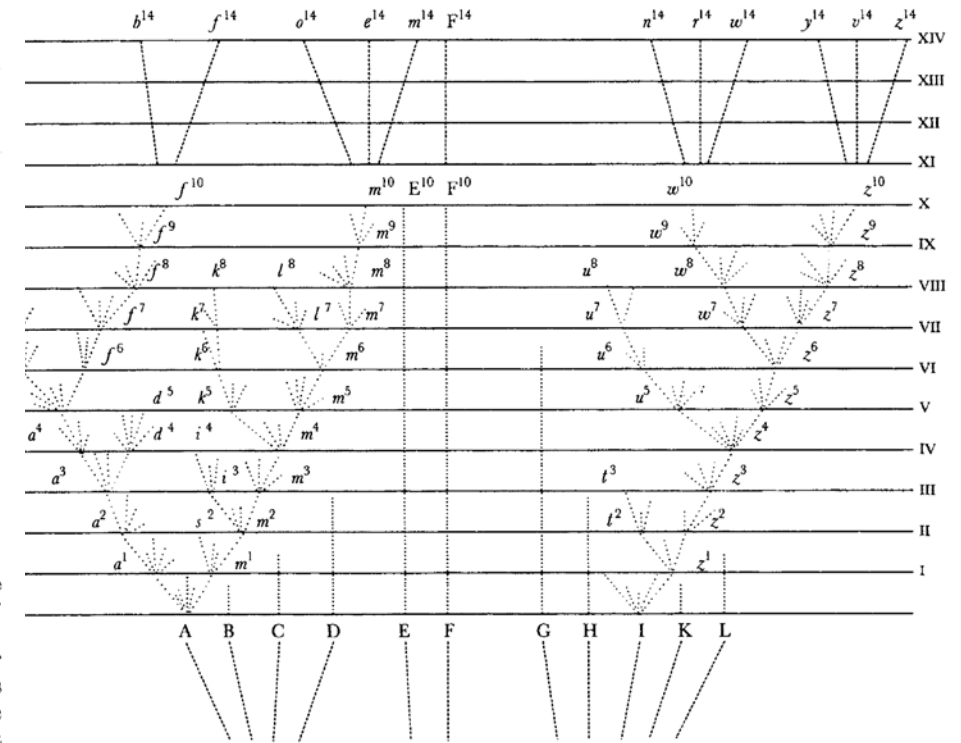


FIG. 24.—Diagram to represent crossing over. At the level where the black and the white rod cross in A, they fuse and unite as shown in D. The details of the crossing over are shown in B and C.

some will come to lie now on one, now on the other side of the mate. If when the twisted chromosomes separate, the parts on the same side go to the same pole the end result will be that shown to the right in Fig. 24. Each chromosome has interchanged a part with its mate. This process has been called crossing over. It is, of course, also possible that the twisted chromosomes do not break and reunite where

(Morgan et al., 1915, The Mechanism of Mendelian Heredity)



(Charles Darwin, 1859, The Origin of Species)

Molecular biology opens an era for biological information

No. 4356 April 25, 1953

NATURE

737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

¹Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
²Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
³Von Arx, W. S., *Woods Hole Papers in Phys. Oceanog. Meteor.*, **11** (3) (1950).
⁴Franklin, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey¹. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β -D-deoxy-ribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's² model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

is a residue on each chain every 3.4 Å. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 Å. The distance of a phosphorus atom from the fibre axis is 10 Å. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally^{3,4} that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

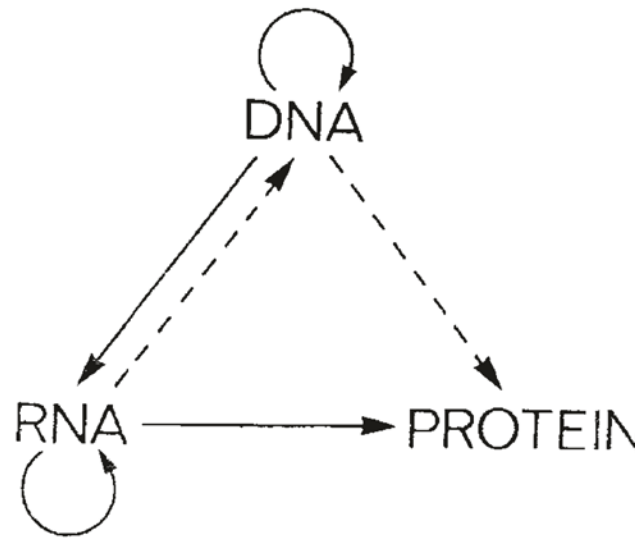
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data^{3,4} on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at



"Modern biology is a science of information."

-- David Baltimore

Central Dogma of Molecular Biology

by
 FRANCIS CRICK
 MRC Laboratory of Molecular Biology,
 Hills Road,
 Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

"The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable over-simplification."

This quotation is taken from the beginning of an unsigned article¹ headed "Central dogma reversed", recounting the very important work of Dr Howard Temin² and others³ showing that an RNA tumour virus can use viral RNA as a template for DNA synthesis. This is not the first time that the idea of the central dogma has been misunderstood, in one way or another. In this article I explain why the term was originally introduced, its true meaning, and state why I think that, properly understood, it is still an idea of fundamental importance.

The central dogma was put forward⁴ at a period when much of what we now know in molecular genetics was not established. All we had to work on were certain fragmentary experimental results, themselves often rather uncertain and confused, and a boundless optimism that the basic concepts involved were rather simple and probably much the same in all living things. In such a situation well constructed theories can play a really useful part in stating problems clearly and thus guiding experiments.

The two central concepts which had been produced, originally without any explicit statement of the simplification being introduced, were those of sequential information and of defined alphabets. Neither of these steps was trivial. Because it was abundantly clear by that time that a protein had a well defined three dimensional structure, and that its activity depended crucially on this structure, it was necessary to put the folding-up process on one side, and postulate that, by and large, the polypeptide chain folded itself up. This temporarily reduced the central problem from a three dimensional one to a one dimensional one. It was also necessary to argue that in spite of the miscellaneous list of amino-acids found in proteins (as then given in all biochemical textbooks) some of them, such as phosphoserine, were secondary modifications; and that there was probably a universal set of twenty used throughout nature. In the same way minor modifications to the nucleic acid bases were ignored; uracil in RNA was considered to be informationally

analogous to thymine in DNA, thus giving four standard symbols for the components of nucleic acid.

The principal problem could then be stated as the formulation of the general rules for information transfer from one polymer with a defined alphabet to another. This could be compactly represented by the diagram of Fig. 1 (which was actually drawn at that time, though I am not sure that it was ever published) in which all possible simple transfers were represented by arrows. The arrows do not, of course, represent the flow of matter but the directional flow of detailed, residue-by-residue, sequence information from one polymer molecule to another.

Now if all possible transfers commonly occurred it would have been almost impossible to construct useful theories. Nevertheless, such theories were part of our everyday discussions. This was because it was being tacitly assumed that certain transfers could not occur. It occurred to me that it would be wise to state these preconceptions explicitly.

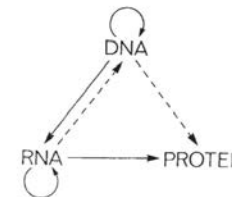


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

A little analysis showed that the transfer could be divided roughly into three groups. The first group was those for which some evidence, direct or indirect, seemed to exist. These are shown by the solid arrows in Fig. 2. They were:

- I (a) DNA → DNA
- I (b) DNA → RNA
- I (c) RNA → Protein
- I (d) RNA → RNA

The last of these transfers was presumed to occur because of the existence of RNA viruses.

Next there were two transfers (shown in Fig. 2 as dotted arrows) for which there was neither any experimental evidence nor any strong theoretical requirement. They were

- II (a) RNA → DNA (see the reference to Temin's work²)
- II (b) DNA → Protein

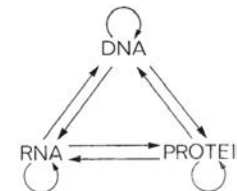
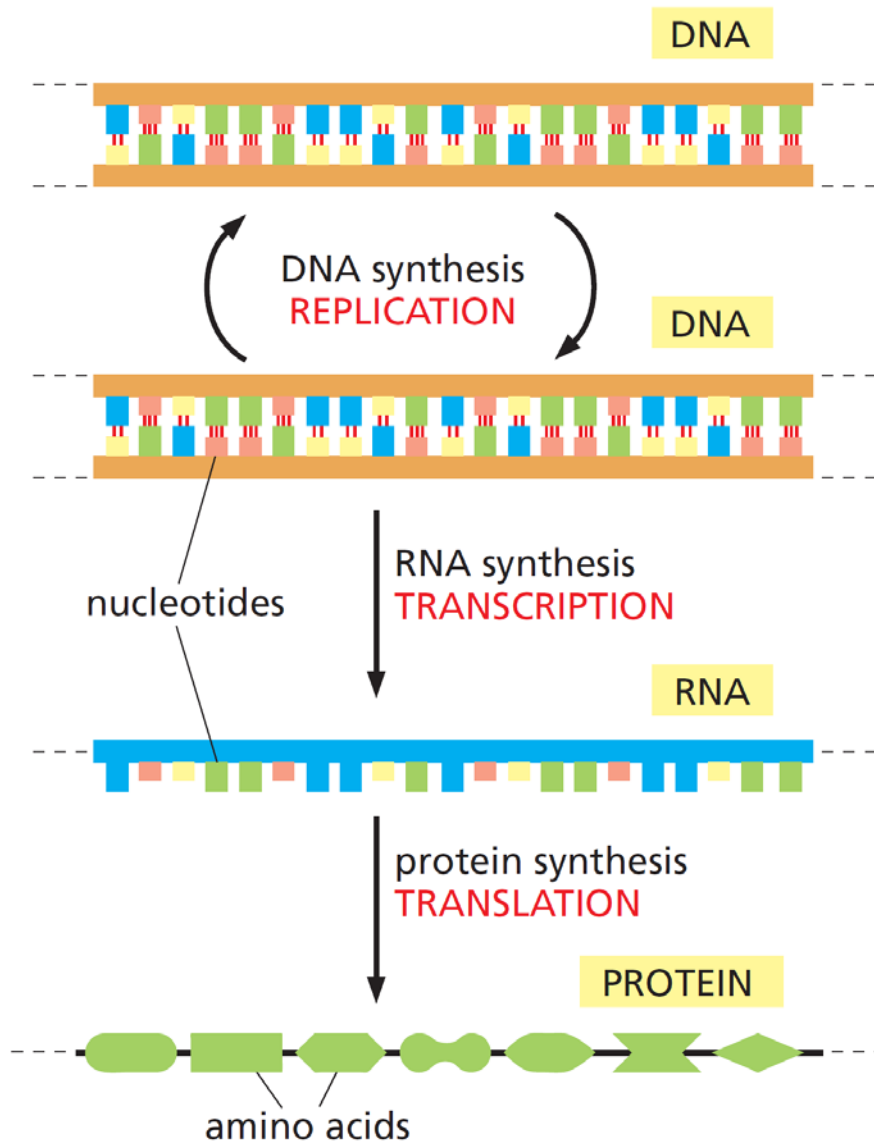


Fig. 1. The arrows show all the possible simple transfers between the three families of polymers. They represent the directional flow of detailed sequence information.

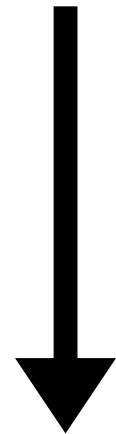
NATURE VOL. 227 AUGUST 8 1970

561

The flow of genetic information



```
>NM_000518.5 Homo sapiens hemoglobin subunit beta (HBB), mRNA
ACATTTGCTTCTGACACAACCTGTGTTCAC TAGCAACCTCAAACAGACACCATGGTGCACTGTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAAAGTGGATGAAAGTTGGTGGTGGAGCCCTGGGC
AGGCTGCTGGTGGTCTACCCCTGGACCCAGAGGTTCTTTGAGTCTTTGGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGC
TCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACCTGTGACAAGCTGCACGTGGAT
CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGGCTAATGCCCTGGCCACAAGTATCA
CTAAGCTCGCTTCTTGCTGTCCAATTTCTATTAAGGTTCTTTGTTCCCTAAGTCCAACACTAAACT
GGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAA
```



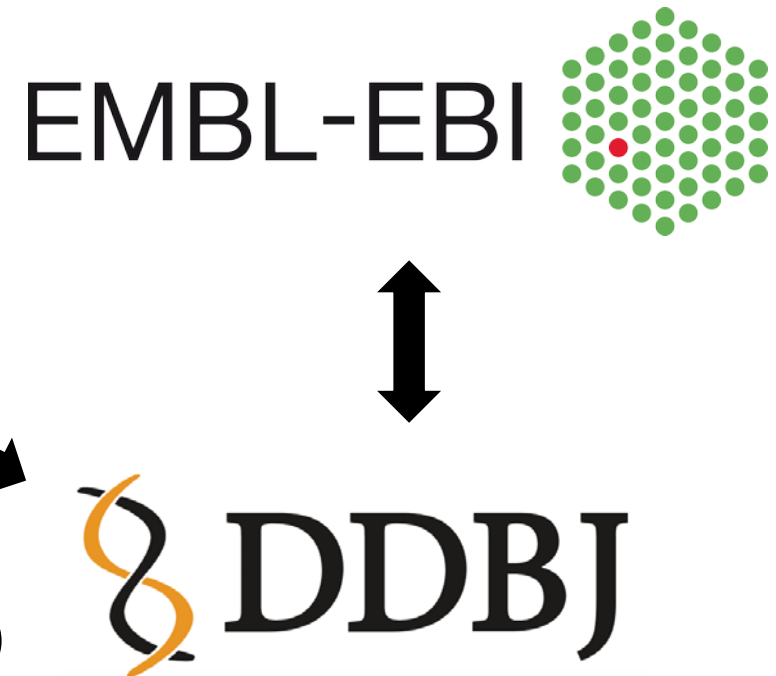
```
>NP_000509.1 hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDVEVGGELGRLLVVYPWTQRFEFESFGDLSTPDAVMGNPKVKAHGKKV L G
AFSDGLAHLNDLNKGTFFATLSSE LHC DKLHVDPENFRLLGNLVLCVLAHHPGKEFTPPVQAA YKVVAGVAN
ALAHKYH
```



Databases for molecular biology information



“GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.” ([link](#))



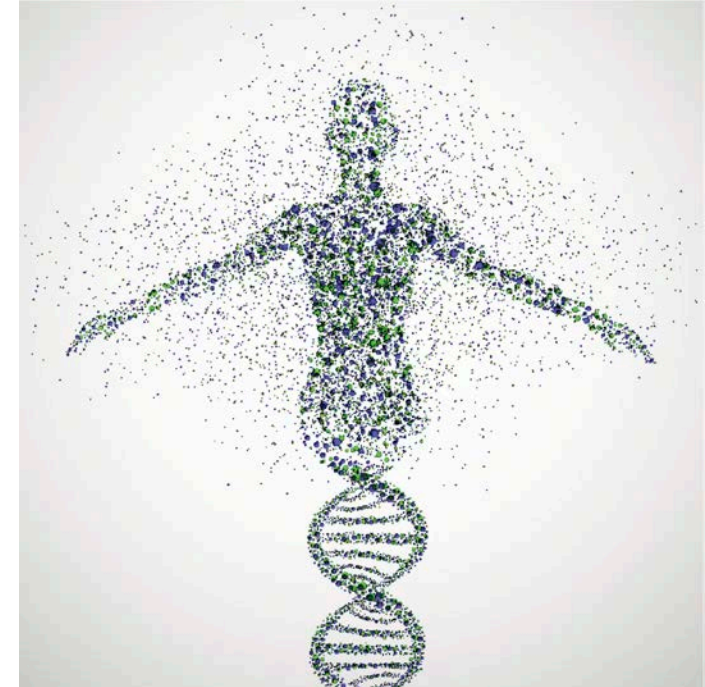
Genomic information took off with the initiation of the Human Genome Project



Manhattan Project (1940s)
Atomic bomb



Apollo Program (1960s)
Landing on the moon



Human Genome Project (1990s)
Sequenced >3 billion base pairs of human genome DNA

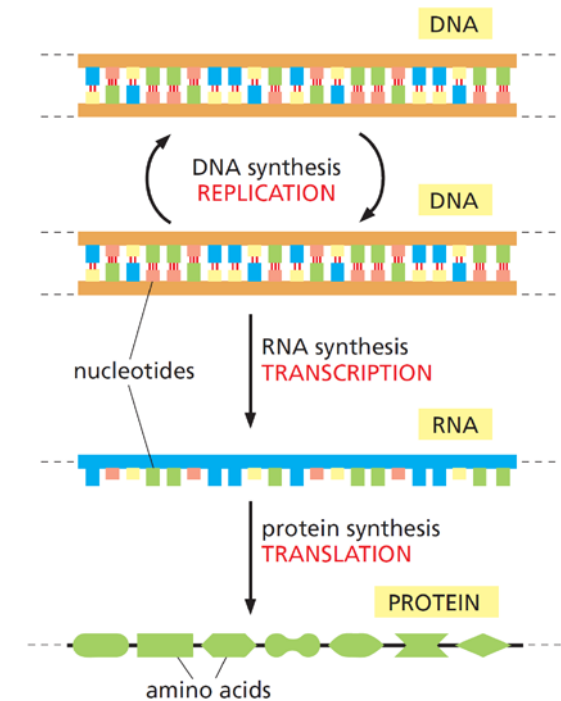
- Cost ~3 billion dollars of US taxpayers' money
- Took 13 years to finish (1990~2003)

BLAST: the “Google” of Biological Research

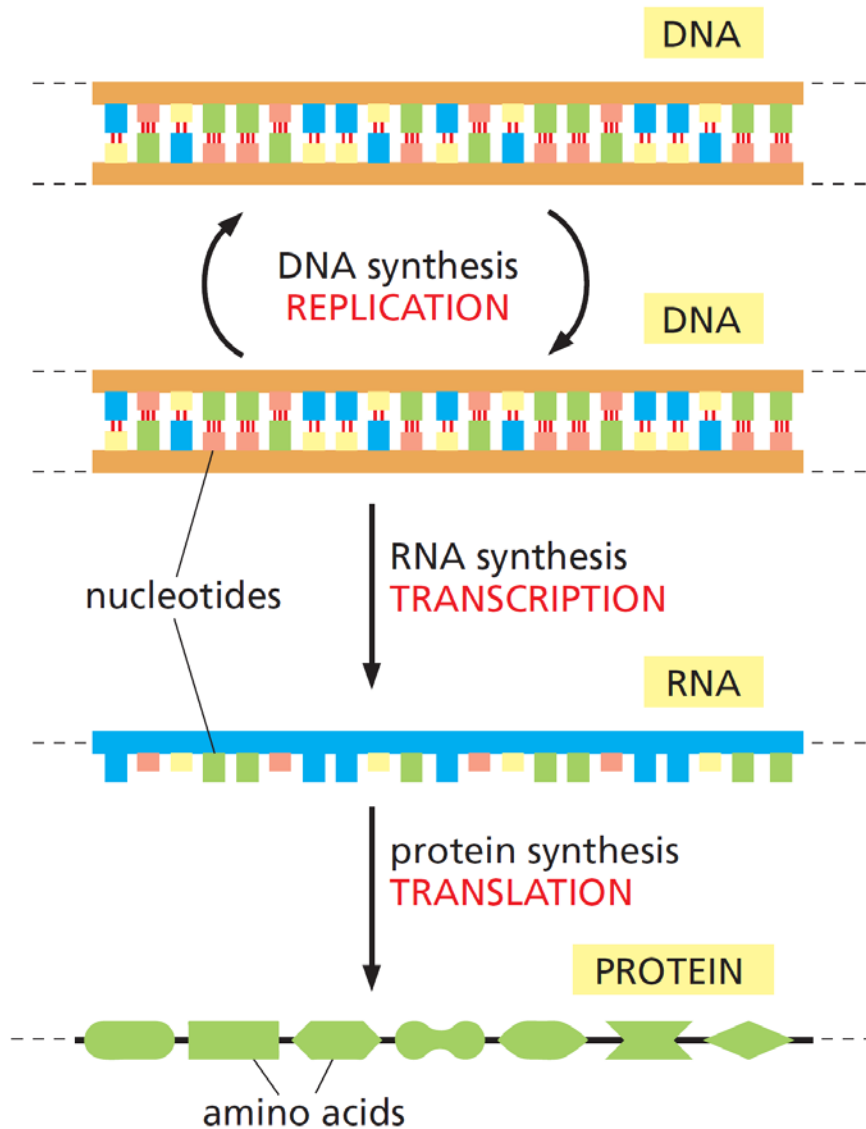
- [Basic Local Alignment Search Tool](#): BLAST is a tool for searching similar DNA or protein sequences given a query sequence.

The screenshot shows the BLAST web interface with the 'blastn' tab selected. The interface is divided into several sections:

- Enter Query Sequence:** Includes a text input for 'Enter accession number(s), gi(s), or FASTA sequence(s)', a 'Query subrange' section with 'From' and 'To' fields, and an 'Or, upload file' section with a 'Browse...' button and 'No file selected.' status.
- Choose Search Set:** Includes a 'Database' section with radio buttons for 'Standard databases (nr etc.)', 'rRNA/ITS databases', 'Genomic + transcript databases', and 'Betacoronavirus'. A dropdown menu shows 'Nucleotide collection (nr/nt)'. There is also an 'Organism' section with a text input and an 'Add organism' button, and an 'Exclude' section with checkboxes for 'Models (XM/XP)' and 'Uncultured/environmental sample sequences'.
- Program Selection:** Includes an 'Optimize for' section with radio buttons for 'Highly similar sequences (megablast)', 'More dissimilar sequences (discontiguous megablast)', and 'Somewhat similar sequences (blastn)'. There is also a 'Choose a BLAST algorithm' link.




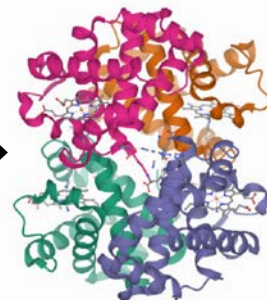
Genomic data are mostly relational



GenBank

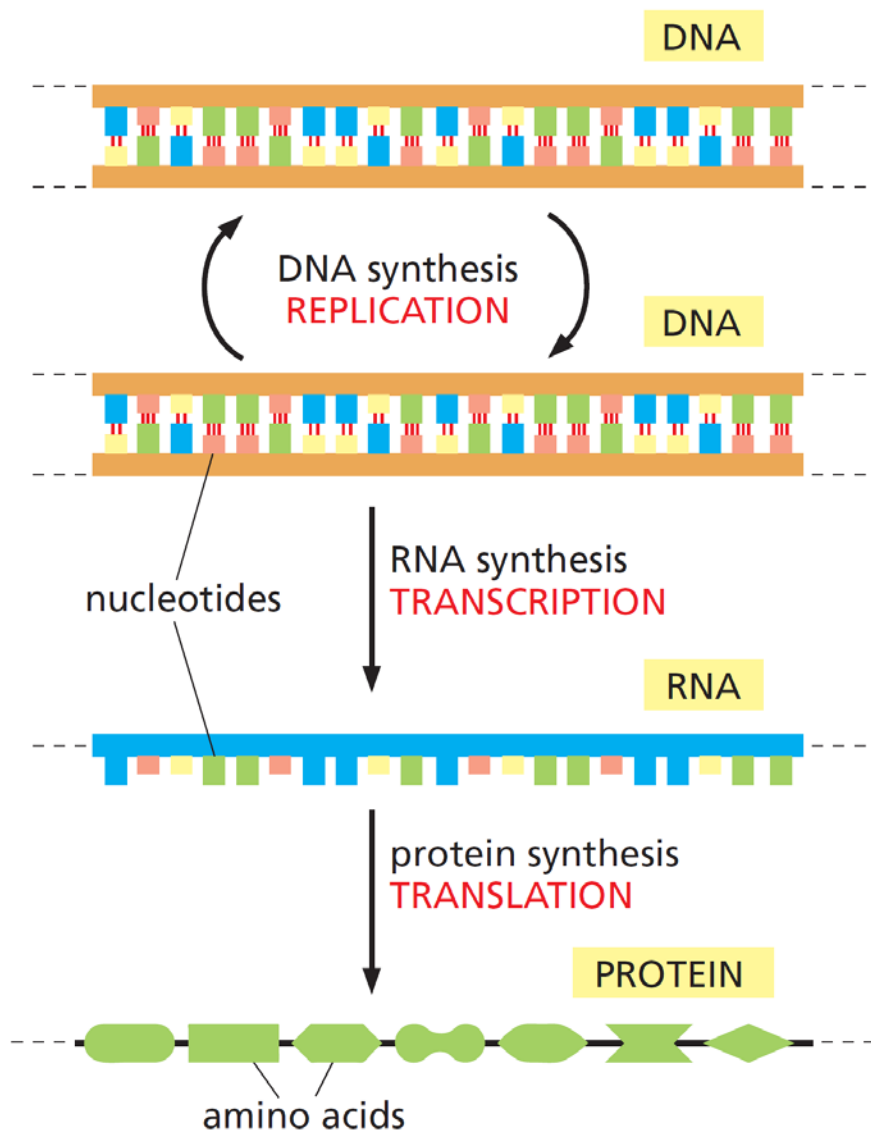
RCSB **PDB**
PROTEIN DATA BANK

 **GENEONTOLOGY**
Unifying Biology
Controlled vocabulary for gene functions

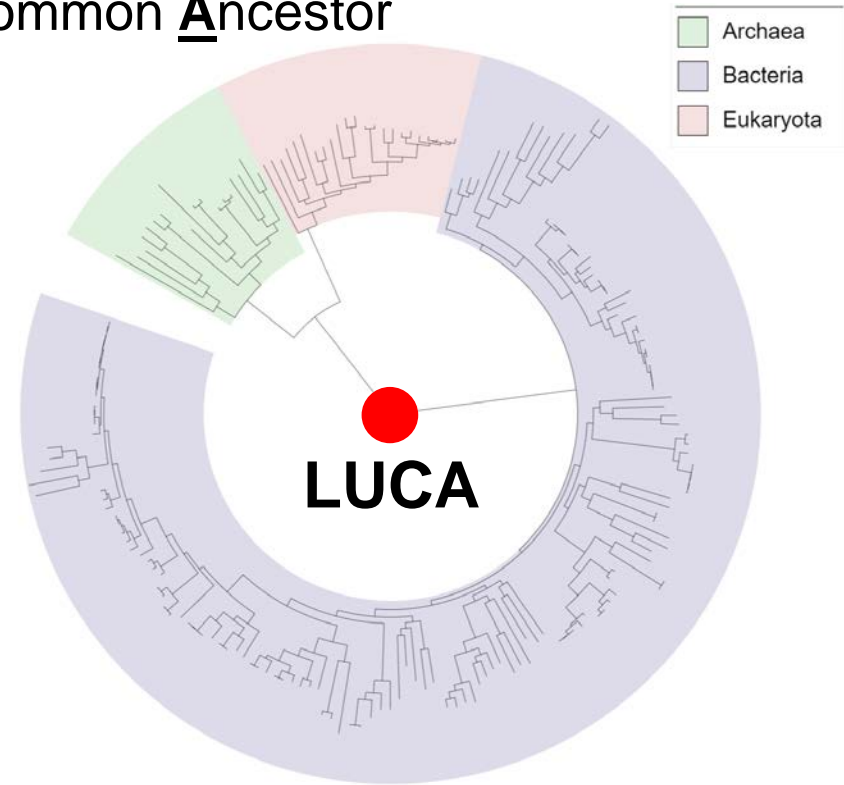


Function of genes

Genomic data are mostly relational

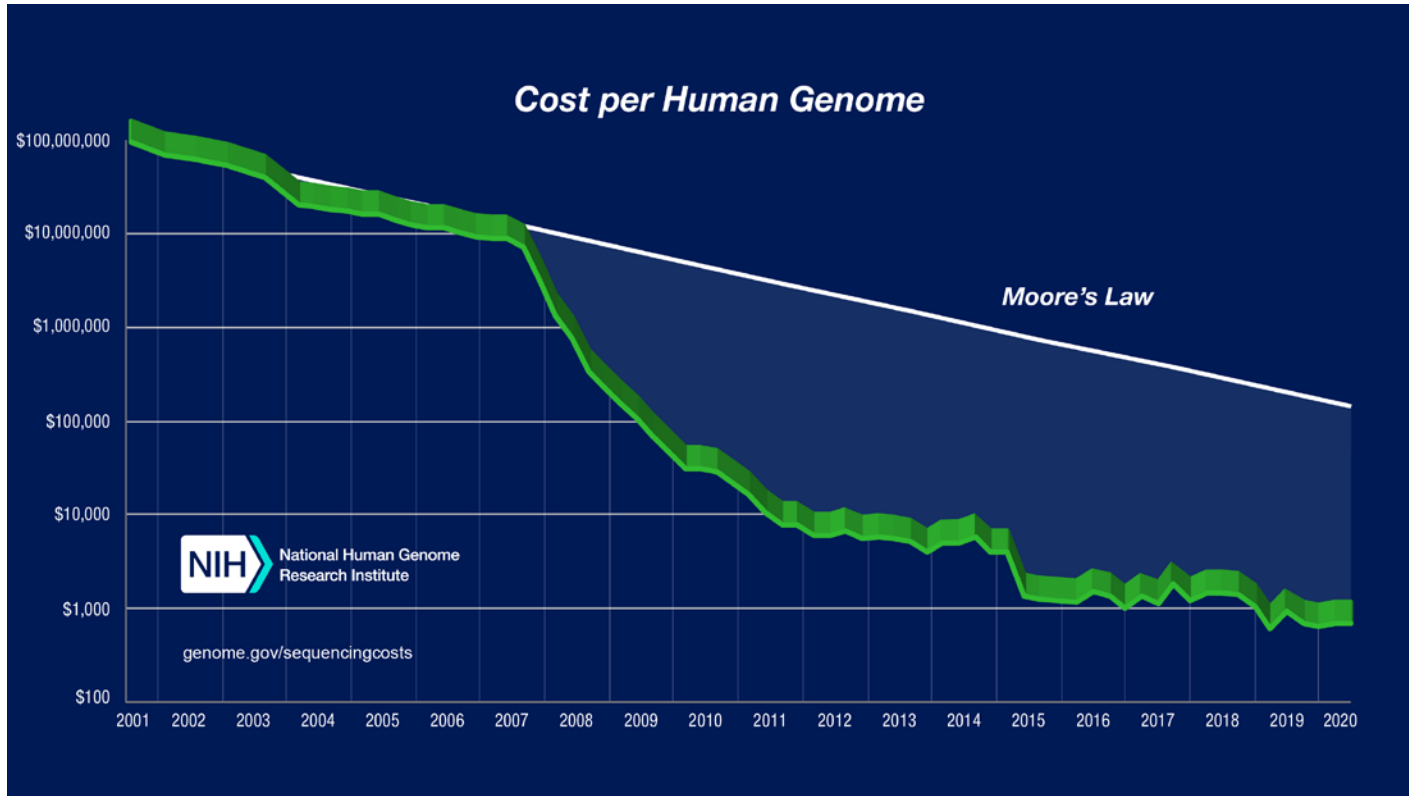


Last Universal Common Ancessor



Genetic information (genes) could be related to each other as they share a common ancestor.

Next generation and third generation sequencing technology introduce new challenges



See also [\[edit \]](#)

- [1000 Genomes Project](#) – International research effort on genetic variation
- [100,000 Genomes Project](#) – A UK Government project that is sequencing whole genomes from National Health Service patients
- [Chimpanzee genome project](#) – effort to determine the DNA sequence of the chimpanzee genome
- [ENCODE](#) – Research consortium investigating functional elements in human and model organism DNA
- [Physiome](#)
- [HUGO Gene Nomenclature Committee](#)
- [Human Brain Project](#)
- [Human Connectome Project](#)
- [Human Cytome Project](#)
- [Human Epigenome Project](#)
- [Human Microbiome Project](#)
- [Human proteome project](#)
- [Human Variome Project](#)
- [List of biological databases](#)
- [Neanderthal genome project](#) – effort to sequence the Neanderthal genome
- [Wellcome Sanger Institute](#) – British genomics research institute
- [Genographic Project](#)

- The development of the next generation and third generation sequencing technology significantly reduce the cost for collecting biological information.

Next generation and third generation sequencing technology introduce new challenges



Opinion

Anticipating the \$1,000 genome

Elaine R Mardis

Address: Genome Sequencing Center, Washington University School of Medicine, 4444 Forest Park Boulevard, St. Louis, MO 63108, USA.
Email: emardis@wustl.edu

Published: 27 July 2006

Genome Biology 2006, 7:112 (doi:10.1186/gb-2006-7-7-112)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2005/7/7/112>

© 2006 BioMed Central Ltd



- [Whole Genome And Exome Sequencing Markets - By Research, Clinical, Direct to Consumer, AgriBio & Tumor with Executive and Consultant Guides 2020 to 2024](#)

Next generation and third generation sequencing technology introduce new challenges

- Computation:
- Storage:
- Archive:
- Maintenance:
- Secondary usage:
- Regulations/sensitive data

Thank you!