Special Topics on Genetics

Section 5: Comparative Genomics

Triantafyllidis A.

School of Biology

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

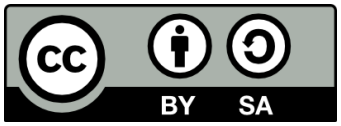
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

License

- The offered educational material is subject to Creative Commons licensing.
- For educational material, such as images, that is subject to other form of licensing, the license is explicitly referred to within the presentation.



Funding

- The offered educational material has been developed as part of the educational work of the Instructor.
- The project "Open Academic Courses at Aristotle University of Thessaloniki" has financially supported only the reorganization of the educational material.
- The project is implemented under the Operational Program "Education and Lifelong Learning" and is co-funded by the European Union (European Social Fund) and national resources.

Section Contents

- Comparative Genomics – An introduction
- Identification of gene function
- Study of synteny
- Genome duplications
- The minimal genome
- The origin of the eukaryotic genome
- Repetitive sequences
- Transposable elements
- Horizontal gene transfer
- Increasing the complexity of the proteome

Comparative Genomics – An introduction to life (1/2)

- Bacteria-like cells have existed for 3.5 billion years
- Eukaryotic cells have existed for 1.4 billion years

CONTINUOUS CHANGE is apparent in life

Comparative Genomics – An introduction to answers given (2/2)

- The discovery of new genetic markers (SSRs and SNPs), as well as the realization of the large degree of polymorphism present in organisms. In some cases, polymorphisms are associated with predisposition to diseases.
- Which are the genes that are common between different races / species? Which are the unique genes? Does the structural order of these genes vary?
- At protein level, which proteins are common and which are unique? What are the differences and similarities of protein interactions and regulation of their expression between races / species?
- Predictions can be made about the function of a protein in an organism according to the rules of homology, based on its sequence and its structure in another organism.

Identification of gene function (1/3)

- Discovery of genes and their function. By finding the homologies between genomes, the identification / documentation of gene function in other organisms is facilitated (annotation).
- Studying if two genes in two organisms are **orthologs** (have appeared after duplication in a common ancestor, and probably present the same function) or **paralogs** (there is no immediate common ancestor in a recent species, but they have arisen by gene duplication in an older species and perform a similar but not identical function). Mistakes are avoided by comparing the wrong genes when trying to understand their function.

Identification of gene function (2/3)

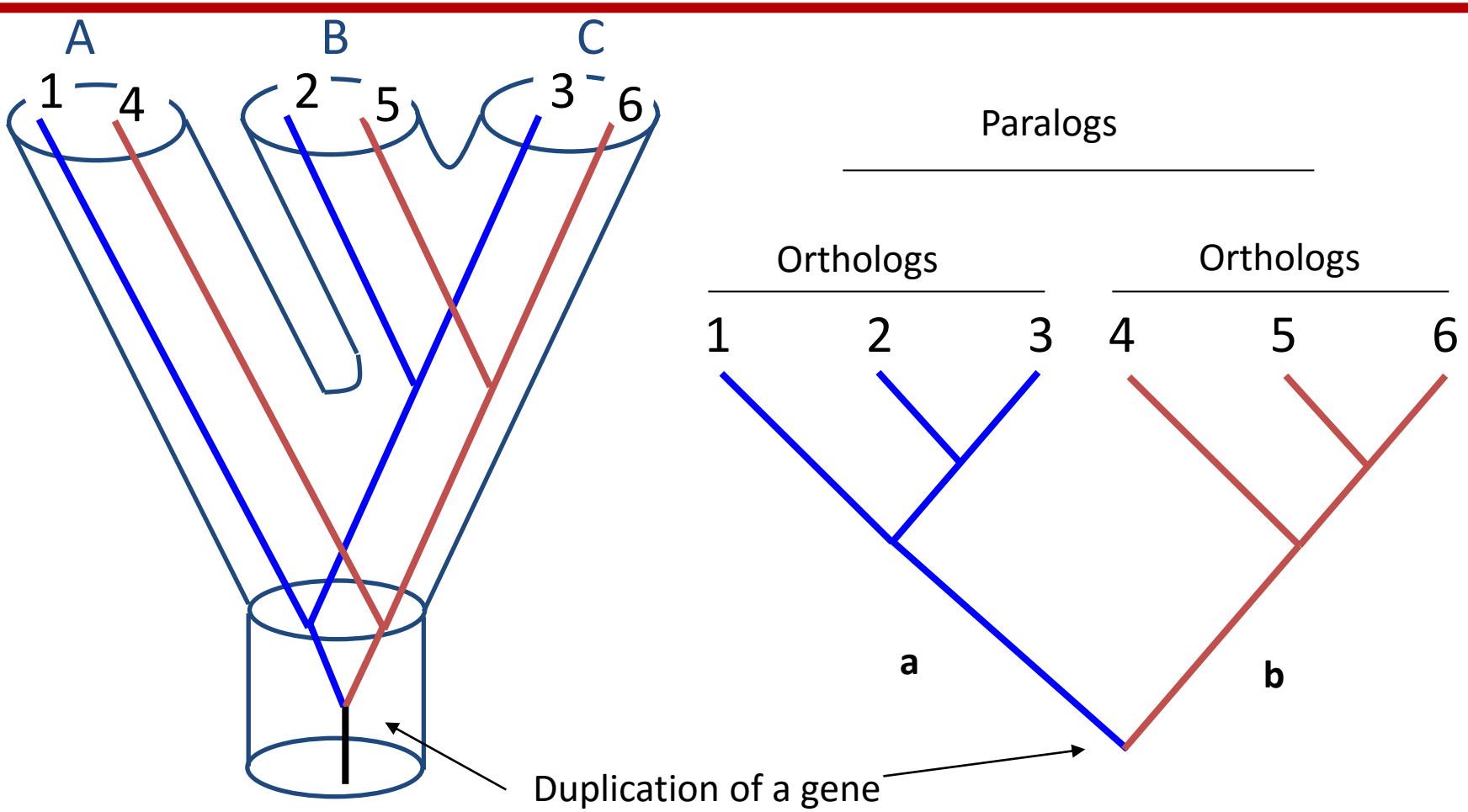


Figure 1: Ortholog genes are also homologs, but homolog genes are not necessarily orthologs.

Identification of gene function (3/3)

The link below presents the relation (orthologs/paralogs) of Hox genes in 5 different species (*Crassostrea gigas*, *Capitella teleta*, *Drosophila melanogaster*, *Branchiostoma floridae*, *Homo sapiens*):
http://www.nature.com/nature/journal/v490/n7418/fig_tab/nature11413_F2.html

Onolog genes! Paralog genes that have arisen after duplication of the whole genome.

Onolog genes in the zebrafish genome:
http://www.nature.com/nature/journal/v496/n7446/fig_tab/nature12111_F3.html

The domains of life

There are multiple similarities at DNA level between organisms (e.g. regions in ribosomal DNA genes that have remained conserved from the simplest single-cell organisms to more complex ones).

Life appeared once and evolved into different organisms.

The eukaryotic genome is a mosaic of DNA from Archaea and Eubacteria.

E.g. The eukaryotic genome rarely has operons but introns are present. These characteristics are also found in Archaea.

The origin of the eukaryotic genome

The basic assumption (often called the global tree or the tree of life) indicates, that archaea and eukaryotes have evolved independently of each other and that the base of this tree is the bacteria, ie each group is monophyletic.
The last few years more sophisticated methods of building phylogenetic trees have been developed. It is now probable that the tree of life as we know it, is not valid.
Based on new data, it seems that the group of archaea is paraphyletic, ie eukaryotes arose through a particular group of archaea, the group TACK, as it is called. Therefore only two major groups of organisms exist, from which life evolved, the bacteria and the archaea.

An archaeal origin of eukaryotes supports only two primary domains of life
Nature 2013, 504, 231-236
http://www.nature.com/nature/journal/v504/n7479/full/nature12779.html

The origin of the genome of organelles

MITOCHONDRIA
Symbiosis of an anaerobic archaebacterial host and an alpha-proteobacteria (such as *Rickettsia prowazekii*).

CHLOROPLASTS
Derived from cyanobacteria

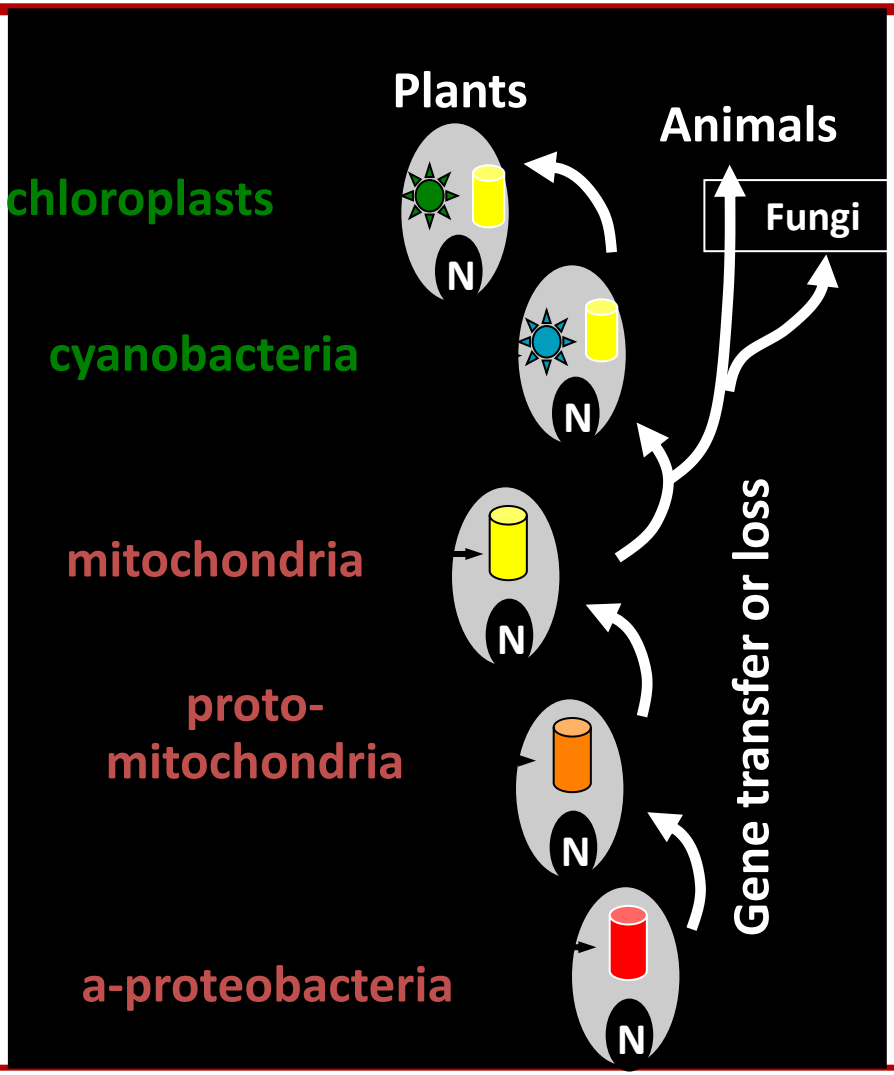


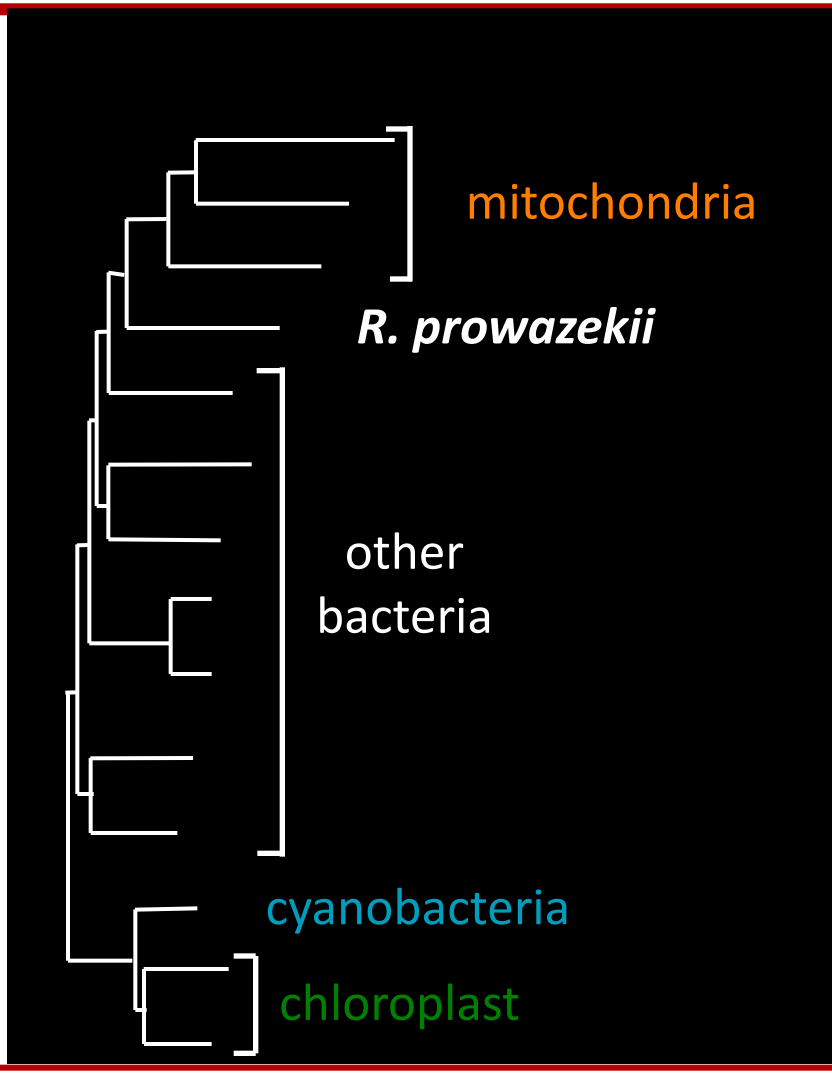
Figure 2: Origin of mitochondria and chloroplasts.

The origin of mitochondrial genome

Figure 3: Sequence analysis of small ribosomal RNA of *Rickettsia prowazekii*, showed that they have the highest similarity to the mitochondrial ribosomal RNA.

Same results obtained based on mitochondrial proteins

This event happened only once



The domains of life – 4th domain?

Metagenomic studies have been carried out in order to study the large viruses and to clarify whether they belong to a fourth Domain

Wu, D. *et al. PLoS ONE 6, e18011 (2011).*
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0018011

- **Figure 4** shows the phylogenetic tree constructed based on the sequence of the beta subunit of RNA polymerase II
- The nucleocytoplasmic large DNA viruses, NCLDVs are placed together in a possible fourth Domain

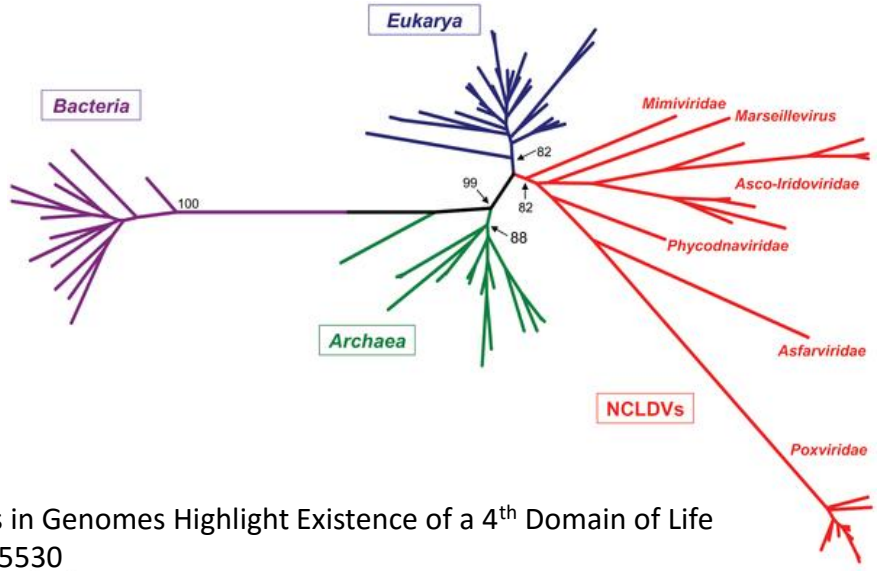


Figure 5: Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. 2010. Boyer M., *et al*. PLoS ONE, 5(12): e15530
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0015530

The minimal genome (1/6)

Which is the minimum number of genes for an organism to survive (in a conducive growth stress-free environment)?

M. genitalium has the smallest genome size: 0.58 Mb.
M. pneumoniae with a size 0.82 Mb has197 additional genes.

Some of the simplest self-replicating prokaryotic organisms. They have no cell wall and can cause diseases in various species.

The minimal genome (2/6)

Bioinformatics analysis indicates number of essential genes for various functions:

Function	*M. genitalium*	*H. influenzae*	*E. coli*
Protein genes	470	1727	4288
Replication, DNA Repair	32	87	115
Transcription	12	27	55
Translation	101	141	182
Regulatory proteins	7	64	178
Biosynthesis of amino acids	1	68	131
Biosynthesis of nucleic acids	19	53	58
Lipid metabolism	6	25	48
Energy metabolism	31	112	243
Transport proteins	34	123	427

The minimal genome (3/6)

Based on experimental approach

- Fraser and Hutchison caused knock-out mutations in almost all genes of the two *Mycoplasma* species and tested the survival of the strains
- Autonomous organisms need about 250-300 genes to survive
- These genes are involved in the following functions:

Translation, DNA replication, recombination, transcription machines, protective proteins, proteins involved in the glycolytic pathway, transport of proteins and a limited number of metabolites.

The minimal genome (4/6)

Synthetic Biology

In late 2002 C. Venter announced that he plans to build such an organism in the laboratory. In summer 2007 he submitted a patent application for *Mycoplasma laboratorium!*

In January 2008 they successfully synthetically assembled the genome (with some modifications to be able to identify it).

In May 2010 Gibson *et al*. were able to import a synthetic genome in a bacterial cell and reproduce it. It took 15 years and $ 40 million!

http://www.robaid.com/bionics/scientists-made-a-living-cell-entirely-controlled-by-synthetic-dna.htm

The minimal genome (5/6)

Synthetic Biology

09/2011 – Parts of two artificially eukaryotic chromosomes were constructed and successfully placed in yeast.

05/2014 – Construction of a complete yeast artificial chromosome of 272,181 bases which replaced its chromosome III (316,617 bases) – the synthesis procedure was different from the one used for Mycoplasma.

http://syntheticyeast.org/

http://syntheticyeast.org/build-a-genome/

The minimal genome (6/6)

Synthetic Biology

10 years ago synthesis service cost US $ 25 per base and sequencing cost $ 0.25. In 2010, prices are $ 0.35 per base synthesis and $ 0.00000317 for sequencing. Nevertheless new technologies give hope as regards drop of prices for synthesis services.

http://www.nature.com/nature/journal/v473/n7347/full/473403a.html

The field of Synthetic Biology is rapidly developing.
E.g. the Biofab company promises to construct regulatory elements that will be able to express proteins in cells with success > 90% !

http://www.biofab.org/

See a related biological comic
http://www.nature.com/nature/comics/syntheticbiologycomic/

Synthetic Biology

Synthetic-biology firms shift focus

Switch to food and fragrances risks consumer rejection.

Nature 29 January 2014

- Target: chemical additives (sweeteners), flavorings for cosmetics
- Development: faster, less costly, less risky
- "Natural" products, since they do not include the original synthetic organism
- Much better efficiency than biofuels, which have not yielded expected results

http://www.sciencemag.org/content/346/6211/1256272.abstract

Genomically encoded analog memory with precise in vivo DNA writing in living cell populations

Genomic duplications (1/5)

How can the complexity and large size of the eukaryotic genome be explained?

The differences in the number of genes between prokaryotes and eukaryotes are due to the **increase** in genome size.
The duplication of entire regions played an important role.
The appearance of the vertebrates was accompanied by a large increase in their genome size.

Genomic duplications (2/5)

At least 55 duplicated regions with a total of 376 genes are found in yeast. 50 of them have exactly the same relative position in different chromosomes.

Its genome arose through the duplication of 8 ancestral chromosomes and the subsequent loss of 90% of duplicated genes.

Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*
Kellis et al. 2004. Nature 428, 617-624

Display of duplicated regions in yeast chromosomes
http://www.nature.com/nature/journal/v428/n6983/fig_tab/nature02424_F3.html

Genomic duplications (3/5)

- Extended duplication phenomena (with rearrangements and loss of genes) occur also in the human genome.
- The duplications can be located in small regions for a small number of genes or may include also regions that are extended along the whole chromosome.
- The duplications include ~10,000 genes.
- The larger duplications appear to correlate with the appearance of vertebrates 500 million years ago.

Genomic duplications (4/5)

Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates

Yoichiro Nakatani, *et al. Genome Res. 2007. 17: 1254-1265*

- Scenario of chromosome evolution of vertebrates
- 2 rounds of whole genome duplication

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950894/figure/F1/

Three periods of regulatory innovation during vertebrate evolution

August 19, 2011 ,Science

The evolution of vertebrates occurred in 3 genetically distinct epochs.

The first, 500 million years ago, included an increase of regulatory regions that particularly affected embryonic development, affecting body shape.

The second, 300 million years ago, affected cell communication.

The third, 100 million years ago, affected intracellular communication.

http://www.sciencemag.org/content/333/6045/1019.full

Genomic duplications (5/5)

Ancestral polyploidy in seed plants and angiosperms

WGD phenomena occurred 319 and 192 million years ago in plants as well, creating suitable material for the development and deployment of seedlings and angiosperms, because of the increase of the regulatory elements associated with the creation of seeds and flowers.

Nature 473, 97–100 (05 May 2011) doi:10.1038/nature09916
http://www.nature.com/nature/journal/v473/n7345/full/nature09916.html

Study of synteny (1/4)

- Discovery of **syntenic regions** (large chromosomal regions that are conserved between organisms).
- There are 342 homologous regions between the mouse genome and the human genome of 300 Kb – 66 Mb, with a mean size of 16 Mb (corresponding to approximately 90% of the genome).
- As if the genome was broken into 342 pieces and reassembled randomly in the two species.
- Most rearrangements (~ 250) happened after rodents separated from other mammals.
- By comparing the syntenic areas, useful information is collected about the genes contained within them. The nature and the extent of synteny varies from chromosome to chromosome.

http://www.mun.ca/biology/scarr/MGA2-11-33smc.html

Study of synteny (2/4)

The existence of conserved synteny can be illustrated using the **chromosome painting** technique.

It is based on the **FISH technique**. However, sequential hybridizations occurs with labeled probes carrying fluorescent dyes that emit light at different wavelengths.

Multiplex FISH-CHROMOSOME PAINTING:

the 23 pairs of human chromosomes are depicted with different colors

http://www.cyto.purdue.edu/cdroms/micro1/7_spon/chroma/image4.htm

Study of synteny (3/4)

CHROMOSOME PAINTING

- DNA sequences from one species are used as probes to color the chromosomes of other species.
- Hybridizations occur in non - strict conditions. The detection of inter-specific hybridization between partially complementary homologous genes is allowed.
- Human sequences have been used and labeled with different colors in order to "paint" the chromosomes of other primate or mammalian species.
- This technique was used for example, to construct the gibbon karyotype using FISH hybridization with human chromosome fragments.

http://www.chrombios.com/cms/website.php?id=/en/index/anicyto/experiments/exp07.htm&sid=t33sf6oe5h8gqsfch6a70vt3o2

Study of synteny (4/4)

During the evolution of mammals, three different categories of conserved synteny can be observed:

1) conservation of whole chromosomes
2) conservation of large chromosome fragments and
3) Combination of parts of different chromosomes to create new syntenies.

Example of synteny betweem human, mouse and rat
http://www.nature.com/nature/journal/v428/n6982/fig_tab/nature02426_F4.html

Comparison of gibbon and human
http://www.nature.com/nature/journal/v513/n7517/fig_tab/nature13679_F2.html

Repetitive DNA

- **Transposable elements:** They constitute more than 45% of the human genome. This percentage is lower in mouse.
- **Pseudogenes:** They result from gene duplication or from replication of RNA into DNA and its integration into the genome (more than 10,000 such genes).
- **Simple Sequence Repeats** (SSRs) of size 2-5 bp (~ 3% of the genome).
- **Duplications of large areas** 10-300 Kb in size (~ 5% of the genome).
- **Structural repetitive regions:** centromeres, telomeres and other chromosomal regions. E.g. 250-1000 TTAGGG repeats are found at human telomeres.

Transposable elements (1/3)

Independent

- LINES, long interspersed elements: 6 – 8 Kb, ~850,000 copies in the genome. L1 elements are typical representatives. 20-50 adenines (A) exist at their 3‘ end
- SINES, short interspersed elements: 100-300 bp, ~1,500,000 copies in the genome. The *Alu* family is typical in mammals.
- Long terminal repeat retroposons. Repetitive regions of 340 bp are found at their ends, enabling their movement.
- DNA transposable elements.

The distribution of these different groups in mammalian genomes is 21%, 13%, 8% and 3% respectively

Nature 409, 860-921(15 February 2001)

http://www.nature.com/nature/journal/v409/n6822/fig_tab/409860a0_F17.html

Transposable elements (2/3)

- Based on available data, most transposable elements in humans have now lost their ability to be transcribed. They are "fossils" of an era when they were active.
- Only a small number of L1 and Alu elements seem to be still capable of translocation (about 100 per person).
- The presence of L1 has been associated with 14 human diseases (Hemophilia, Muscular Dystrophy).
- In mouse and rat there are 10,000 active transposable elements.

Transposable elements (3/3)

Significance

- At least 47 genes were formed via displacement of elements in the genome.
- They provide information on evolutionary mechanisms (mutations, selection). In *Drosophila* 50% of mutations are caused by translocations of transposable elements (Morgan was the first to discover the white eye mutation).
- They facilitate rearrangements of chromosomes by recombination between two identical transposable elements.
- Distribution of transposable elements is not random. Other regions possess large numbers, while others (such as in Hox genes) there is almost none.

PARASITES or BENEFICIAL?

Horizontal gene transfer (1/3)

Horizontal gene transfer is called the direct physical transfer of genes from one species in the reproductive cells of another species.
Horizontal gene transfer occurs in nature, possibly with the help of transposable elements or directly.

The developmental potential of organisms is unlimited.
Concerns are created about the safety of genetically modified organisms and genes that can be passed in this way from one species to another through the food chain.

Horizontal gene transfer (2/3)

✓ It has been well known (since the '70s) that the trypsin enzyme of *S. griseus* bacterium is more similar to cattle trypsin than to other bacterial ones.
✓ Evolution does not occur just through vertical transfer of genes from ancestors to offspring.
✓ The analysis of the human genome revealed that at least 100 genes appeared in humans through horizontal transfer from bacteria.
✓ Correspondingly eight genes of *M. tuberculosis* have originated from human.
✓ Such phenomena are more frequent in 'vulnerable' cancer cells.

Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples," *PLOS Computational Biology*, 2013.

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003107

Horizontal gene transfer (3/3)

Horizontal genome transfer as an asexual path to the formation of new species
Fuentes et al. 2014, NATURE, 511, 232-235

Through allopolyploidization mechanisms, which did not occur due to hybridization phenomena but due to asexual mechanisms, completely new species can be created in nature that can displace their ancestors.

The tree of Life…postgenome

From the Tree of Life to the Web of Life

• Horizontal gene transfer complicates relationships between species, as well as the understanding of the phylogenetic tree of life, where all species were simply believed to be derived from a common ancestor

• The genes in a genome can represent different evolutionary histories, since even a rare gene transfer can cause different molecular genealogical relations

• The realistic scenario does not actually correspond to the Tree of life, but to the much more complex Web of life.

http://www.texscience.org/reports/sboe-tree-life-2009feb7.htm

Increasing complexity of the proteome (1/2)

By studying the proteome it became obvious that the vertebrate proteome is more complex than the proteome of yeast, worm, *Drosophila* and other lower organisms.

How is this possible?

- *More genes.* The genome of vertebrates contain more genes than other organisms.
- *More 'paralog' genes*. Within gene families numerous duplications have occurred. E.g. ~ 1000 olfactory genes are found in humans whereas only 60 in fish.

http://www.pnas.org/content/101/8/2584/F2.large.jpg

Reorganization, increase and decrease of functional regions of proteins. Throughout evolution, new combinations have appeared that have significantly increased the total possible architectural structures of proteins.

Increasing complexity of the proteome (2/2)

Alternative splicing. E.g. 3 genes are responsible for the production of neurexin molecules that connect neurons and neural cells. Through differential splicing > 2,000 forms are produced (Genomics. 2002 Apr;79(4):587-97) . How many of these forms actually perform different functions?… still difficult to answer

- *Post-translational modifications of proteins*. There are more than 400 biochemical modifications for human proteins. 20,000 different mRNAs, but more than 200,000 different proteins.

Reference note

Licensing note

End of Section

Processing: Minoudi Styliani
Thessaloniki, Winter Semester 2014-2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Notes Preservation

Any reproduction or adaptation of the material should include:

- the Reference Note
- the Licence Note
- the Notes Preservation

accompanied with their hyperlinks.