Molecular Population Genetics

The 10th CJK Bioinformatics Training Course

in

Jeju, Korea

May, 2011

Yoshio Tateno
National Institute of Genetics/POSTECH

DDBJ
DNA Data Bank of Japan

Accession DNA Protein AllDBs Taxonomy Site Search

Accession numbers Go

DDBJ DRA UniProt PDB DAD PRF Patent >>more

HOME | Submission | How to Use | Search/Analysis | FTP/WebAPI | Report/Statistics | Contact Us | RSS | Japanese

- About DDBJ
- How to Use
- Q and A

Sequence Submission

- SAKURA
- Mass Submission
- Data Updates
- DDBJ Sequence Read Archive
- DDBJ Trace Archive

Search

- getentry
- ARSA
- TXSearch
- BLAST

Phylogenetics

- ClustalW

DDBJ: DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of the three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, which was established through cooperative work with EBI in Europe and NCBI in USA.

Hot Topics

More

- Apr. 14, 2011 Temporary delay of a part of the DDBJ services for consecutive holidays
- Apr. 12, 2011 Resumed BLAST and ClustalW
- Apr. 5, 2011 Release of new WGS of domesticated barley 8,583 entries
- Feb. 22, 2011 DDBJ will continue Sequence Raw Data Archiving

Maintenance

More

- Mar. 14, 2011 Suspension of a part of the DDBJ services due to the effects of recent disaster

Sequence Data Submission

Submit my sequences

Orientation for the data submission

FTP/Web API

FTP (ftp.ddbj.nig.ac.jp)

Download data files

DDBJ
DNA Data Bank of Japan

DDBJ Sequence Read Archive

» Login D-way

DDBJ Sequence Read Archive | DDBJ Trace Archive

Home | Documentation | Submission | Search | Download | Pipeline | About | ▶ Japanese

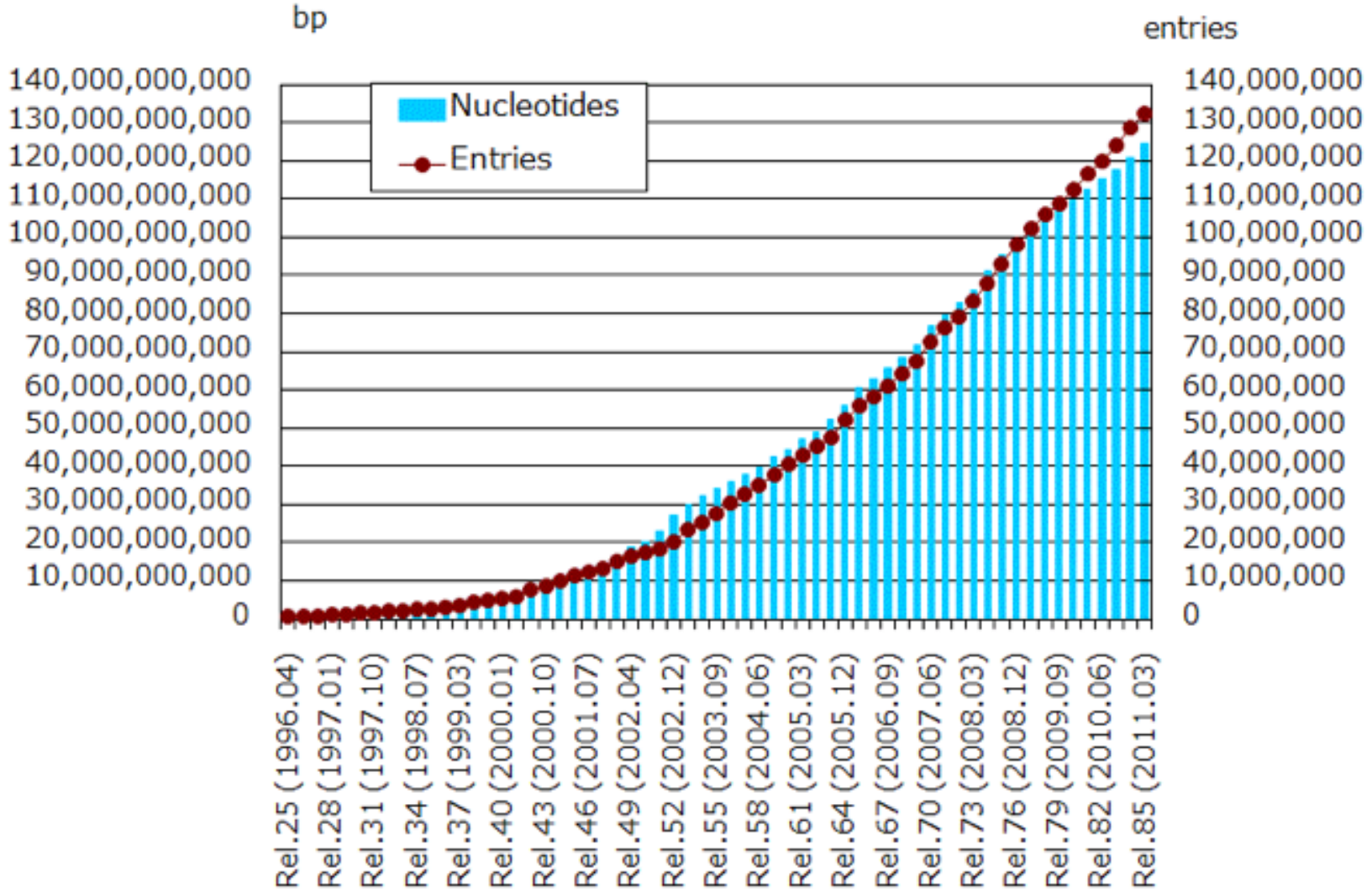
DDBJ Sequence Read Archive (DRA) is an archive database for output data generated by next-generation sequencing machines including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, and others. DRA is a member of the International Nucleotide Sequence Database Collaboration (INSDC) and archiving the data in a close collaboration with NCBI Sequence Read Archive (SRA) and EBI Sequence Read Archive (ERA). Please submit the trace data from conventional capillary sequencers to DDBJ Trace Archive.

- Data necessary for submission
- How to submit your data
- Search and downlooad data
- Analyze your data in the DDBJ Read Annotation Pipeline

DRA is part of the National project of integrating life science databases, and is supported by the Japan Science and Technology Agency Institute for Bioinformatics Research and Development.

Last modified: Feb. 16, 2011

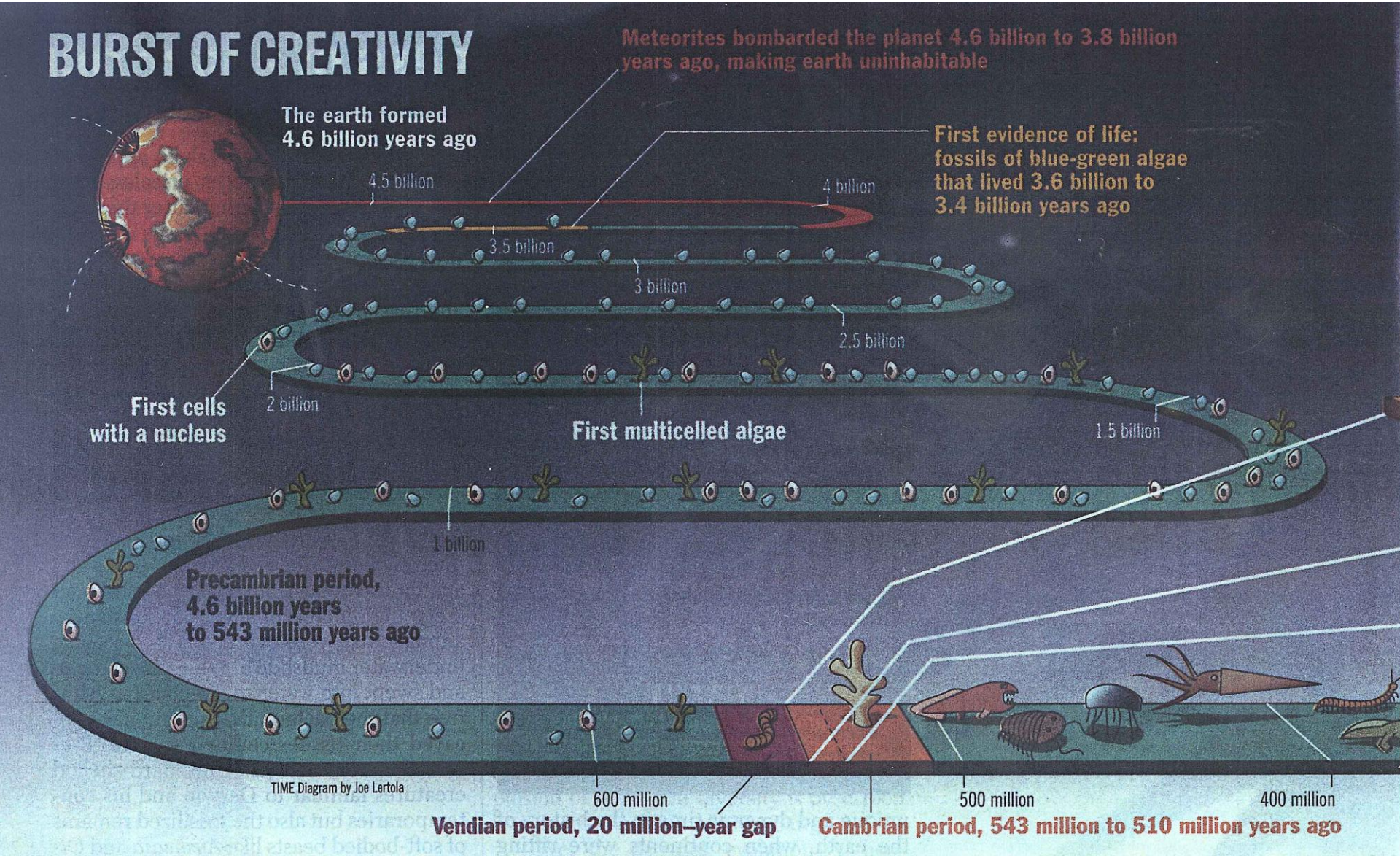
DDBJ/EMBL/GenBank database growth



Top 10 species in INSDC
(as of April, 2011)

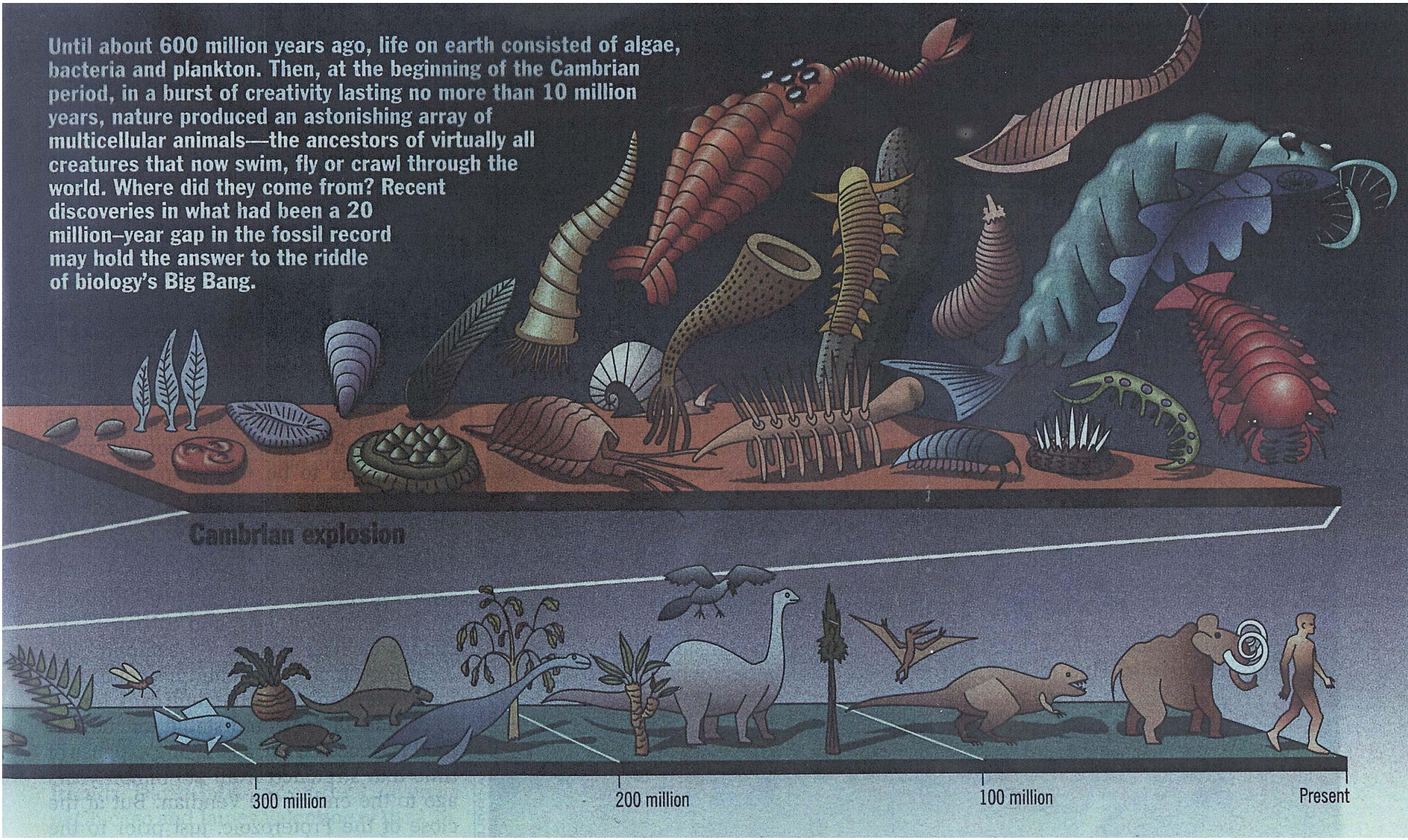
No.	Organisms	Nucleotides	Entries
001	Homo sapiens	14813854723 bp	15655926 entry
002	Mus musculus	8859499642 bp	7875214 entry
003	Rattus norvegicus	6444234541 bp	2184105 entry
004	Bos taurus	5361703017 bp	2190542 entry
005	Zea mays	5037654694 bp	3892585 entry
006	Sus scrofa	4784533986 bp	3218932 entry
007	Danio rerio	3136145051 bp	1697980 entry
008	Unknown.	2646099850 bp	5051961 entry
009	marine metagenome	2149495444 bp	2643001 entry
010	Strongylocentrotus purpuratus	1352920226 bp	228238 entry

CONTENTS


BURST OF CREATIVITY
Meteorites bombarded the planet 4.6 billion to 3.8 billion years ago, making earth uninhabitable
The earth formed 4.6 billion years ago
First evidence of life: fossils of blue-green algae that lived 3.6 billion to 3.4 billion years ago
4.5 billion
4 billion
3.5 billion
3 billion
2.5 billion
2 billion
First cells with a nucleus
First multicelled algae
1.5 billion
1 billion
Precambrian period, 4.6 billion years to 543 million years ago
TIME Diagram by Joe Lertola
600 million
500 million
400 million
Vendian period, 20 million–year gap
Cambrian period, 543 million to 510 million years ago

“We conclude that photosynthetic organisms had evolved
and were living in a stratified ocean supersaturated in dissolved silica 3,416 million (3.4 billion) years ago.”

M. M. Tice and D. R. Lowe (Stanford University, USA)
Nature 431: 549 - 552, 2004


Until about 600 million years ago, life on earth consisted of algae, bacteria and plankton. Then, at the beginning of the Cambrian period, in a burst of creativity lasting no more than 10 million years, nature produced an astonishing array of multicellular animals—the ancestors of virtually all creatures that now swim, fly or crawl through the world. Where did they come from? Recent discoveries in what had been a 20 million–year gap in the fossil record may hold the answer to the riddle of biology's Big Bang.
Cambrian explosion
300 million
200 million
100 million
Present

14–17 Nucleic Acids Research, 1997, Vol. 25, No. 1

DNA Data Bank of Japan in the age of information biology

Yoshio Tateno* and Takashi Gojobori

Center for Information Biology, National Institute of Genetics, Yata, Mishima 411, Japan

Received September 16, 1996; Revised and Accepted October 8, 1996

We believe that information biology will be one of the most important areas in biology, medicine and agriculture in the next century, and that molecular evolution will form a core in information biology. As mentioned earlier, a great number of genes and other DNA regions have been sequenced and have accumulated in the international DNA sequence databases. The biological functions of many sequences have, however, been unelucidated. The origins and functions of those genes and regions will be sought and solved by way of information biology in particular large scale data analysis using high performance computers. We are now in the position to analyze DNA and protein not only *in vivo* and *in vitro* but also *in silico*.

In bioinformatics, we deal with large quantities of DNA, RNA and protein sequences, gene expression, proteomics and pathways data.

The scientific background of dealing with such biological data is evolution or molecular evolution. This is because all the biological data are the products of evolution.

For example, BLAST is useless without this conception.

Genome

1. **is to design and control life activity in individuals,**

2. **is to be inherited to offspring - evolution.**

Therefore, it is important to recognize that genes, proteins and organisms are products of evolution, and pay attention to the evolutionary view when studying them.

Usage of Genome Data

1. To deduce the function of a DNA sequence by comparing to others whose functions are known,
2. To know the evolutionary origin and process of a gene or a species by constructing a phylogenetic t related (orthologous) genes,
3. To examine if a particular gene exists in a particular species by constructing a phylogenetic tree of the species and related species,
4. To find a route of virus or bacterial infections,
5. To understand regulation of gene expression (CAGE microRNA), and more…..

Evolution is quantitative and qualitative changes in genes over time.

Evolution is primarily driven by mutation.

No mutation, no evolution.

Mutation

1. Point Mutation

One nucleotide (base) is replaced by another.

2. Insertion

One or more nucleotides are inserted into the extant sequene.

3. Deletion

One or more nucleotides are deleted from the extan t sequence.

4. Inversion

A sequence of two or more nucleotides is replaced in the opposite direction to the extant sequence.

5. Recombination

A sequence in a chromosome is exchanged with another in the other homologous chromosome.

6. Translocation

A sequence in a chromosome is moved and located in another chromosome.

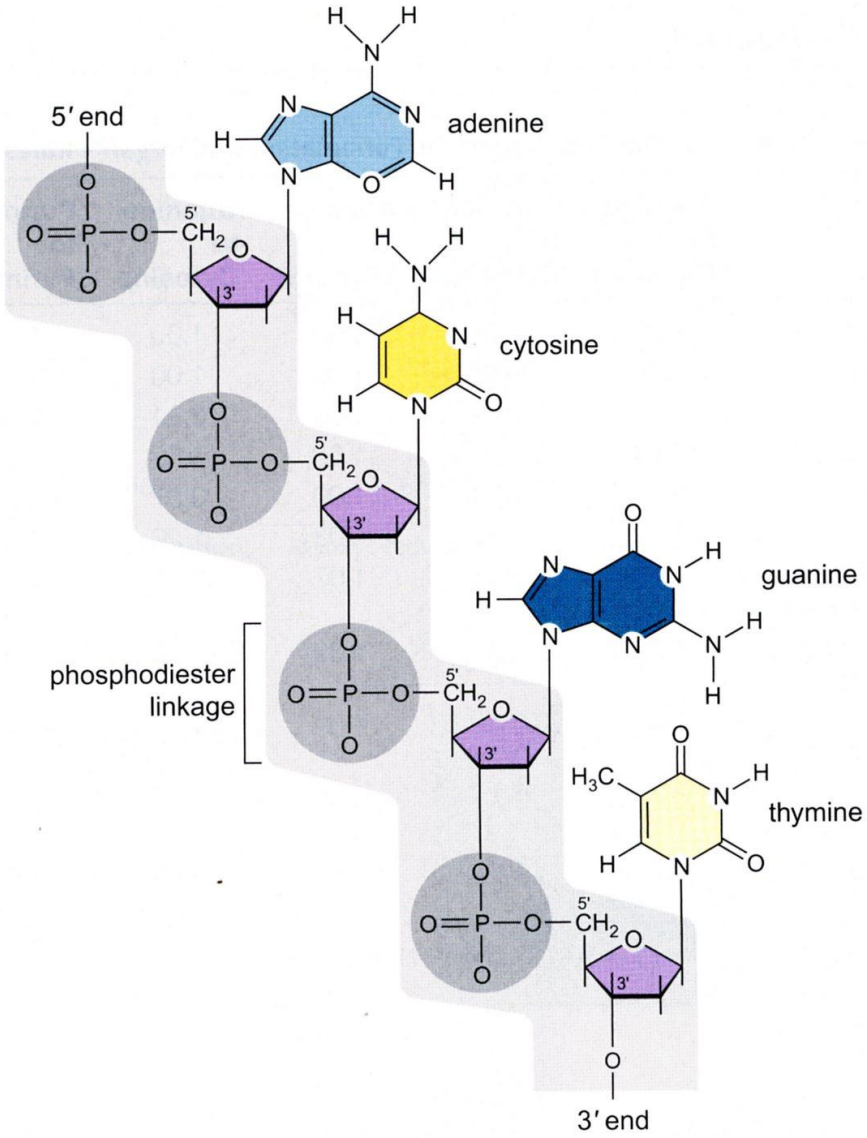


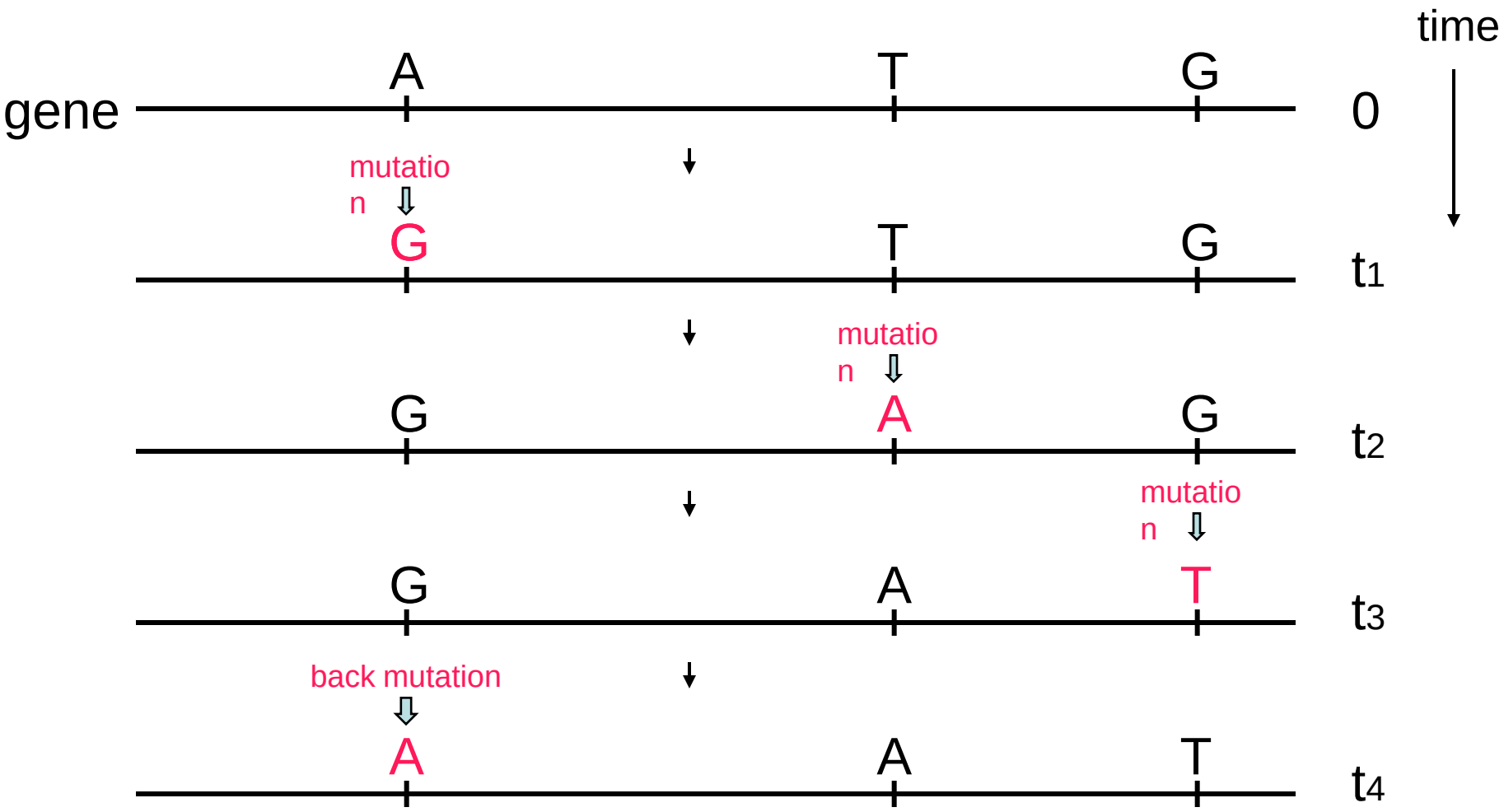
FIGURE 2-5 A portion of a DNA polynucleotide chain, showing the 3′ → 5′ phosphodiester linkages that connect the nucleotides. Phosphate groups connect the 3′ carbon of one nucleotide with the 5′ carbon of the next.

Spontaneous mutations

Natural miss-pairings between the two nucleotides

FIGURE 2-6 The replication of DNA.
The newly synthesized strands are shown in orange.

Gene Evolution by Mutation



Poisson Process

This process is characterized by that the chance of a new event in any short interval is independent, not only of the previous states of the system in question, but also of the present state. The process is called a random process.

We can thus assume that the chance of a new addition to the total account during a very short interval (Δt) is written as $\lambda \Delta t$ where λ is a rate of the addition ignoring the chances of two or more new additions.

If we define Pn ($t + \Delta t$) as the probability that the exactly n events have occurred by time $t + \Delta t$, we can derive the following equation.

$$P_n(t+\Delta t) = P_{n-1}(t)\cdot \lambda \Delta t + P_n(t)(1-\lambda \Delta t) \quad \cdots\cdots\cdots\cdots\cdots (1)$$

The first term on the right of the equation represents the probability that one new event added during Δt to the state that n - 1 events had occurred by time t. The second term means that no event was added during Δt to the probability that n events had occurred by time t.

Formula (1) is rewritten as

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \lambda P_{n-1}(t) - \lambda P_n(t)$$

This difference equation can be approximated as the following differential equation.

$$\frac{dP_n(t)}{dt} = \lambda \left\{ P_{n-1}(t) - P_n(t) \right\} \quad \cdots\cdots\cdots\cdots (2)$$

For $n = 0$, we have,

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \quad \cdots\cdots\cdots\cdots (3)$$

The solution of (3) is given by,

$\ln P_0(t) = -\lambda t + c$, or

$P_0(t) = ce^{-\lambda t}$

Since $P_0(0) = 1$, $P_0(t) = e^{-\lambda t}$ $\cdots\cdots\cdots\cdots (4)$

For $n = 1$, (2) is expressed as

$$\frac{dP_1(t)}{dt} = \lambda \left\{ P_0(t) - P_1(t) \right\} = \lambda \left\{ e^{-\lambda t} - P_1(t) \right\} \quad \cdots\cdots\cdots\cdots (5)$$

If we remember

$\frac{t}{dt}\left[f(t)g(t) \right] = f'(t)g(t) + f(t)g'(t)$, and

$\frac{d}{dt}\left[e^{at} \right] = ae^{at}$,

we can rewrite the right of (5) as

$t'\left(\lambda e^{-\lambda t}\right) + t\left(e^{-\lambda t}\right)'$.

Therefore,

$P_1(t) = \lambda t e^{-\lambda t}$. $\cdots\cdots\cdots\cdots$ (6)

Repeating this procedure, we can get

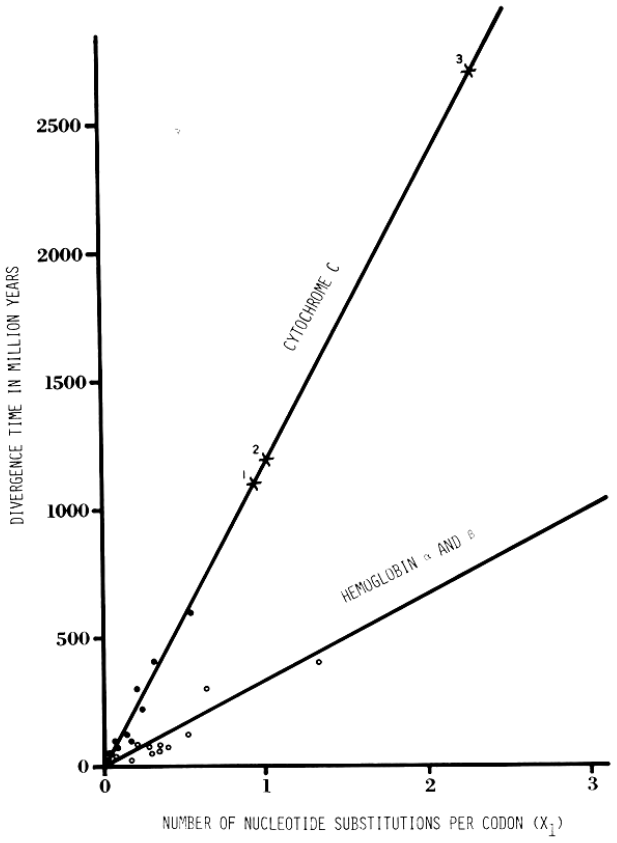
$$P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \quad \cdots\cdots\cdots\cdots (7)$$

which is the general expression of Poisson probability.

Random events in the time axis means that they occur constantly over time, or proportionally to time.

Fig. 2-2. Estimated number of nucleotide substitutions (X_1) and divergence time.

The open circles refer to hemoglobin α and β chains, and the closed circles to cytochrome c. Cross mark 1 represents the estimated divergence time between plants and animals, cross mark 2 that between yeasts and other eukaryotes, and cross mark 3 that between prokaryotes and eukaryotes.



From my thesis, 1978

Population Genetics and Molecular Evolution

- Study of gene frequency change over time in a population.

- Evolution is defined as gene frequency change over time.

W. Bateson (1902)

Alleles

A/A: homozygote
A/a: heterozygote

A pair of genes that are located in the two homologous loci

ÒHbmologousÓhere means that they are derived from the common ancestor.

In 1955 Tjio and Levan first reported that humans had 23 pairs of chromosomes.

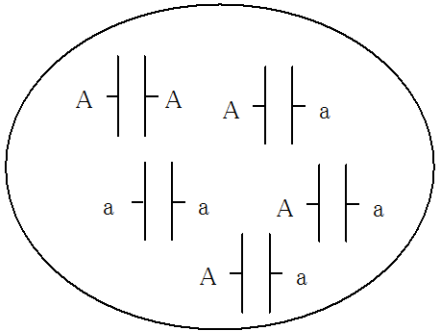
The chromosomes of humans, 2n = 46,

duploid < monoploid.

体外で培養したヒト胎児の肺線維芽細胞の中期. ヒト染色体数が 46 であるとした最初の報告による.

1955年、TjioとLevanはヒトの染色体数が46本であることを発見した。

Gene Frequency



Mendelian population
N=5, 2N=10

$$\text{Frequency of A gene} = \frac{\text{Number of A genes}}{2N} = \frac{5}{10} = 50\%$$

$$\text{Frequency of a gene} = \frac{\text{Number of a genes}}{2N} = \frac{5}{10} = 50\%$$

Or, you can also compute them using the frequencies of the individu als.

A A の頻度 $= \frac{1}{5}$

A a の頻度 $= \frac{3}{5}$

a a の頻度 $= \frac{1}{5}$

$$\text{Frequency of A gene} = \frac{1}{5} + \frac{1}{2} \cdot \frac{3}{5} = \frac{1}{5} + \frac{3}{10} = \frac{5}{10} = 50\%$$

$$\text{Frequency of a gene} = \frac{1}{5} + \frac{1}{2} \cdot \frac{3}{5} = 50\%$$

Quiz 1

Two equal-sized populations, 1 and 2, have frequencies p_1 and p_2 of the allele A. The populations are fused into a single randomly mating unit.

(1) What is the frequency of AA homozygotes in the mixed populations?

(2) What is the answer to (1), if population 1 is 4 times as large as population 2?

Four Major Evolutionary Factors Causing Gene Frequency Change

１．Natural Selection

２．Mutation

３．Random Genetic Drift

４．Migration

自然選択と適応度 (Natural Selection and Fitness)

Frequency of A gene : p

Frequency of a gene : q

(p + q = 1)

Genotype	AA	Aa	aa
Frequency	p^2	2pq	q^2
Fitness	$w_{11}=1$	$w_{12}=1-s_1$	$w_{22}=1-s_2$

Genotype
遺伝子型

s_1, s_2 are selection coefficients and range in the following regions.

$$0<|s_1|<1,\ \ 0<|s_2|<1$$

AA can leave ralatively $w_{11} p^2$ offspring in the next generation.

Aa can leave ralatively $2w_{12} pq$ offspring in the next generation.

aa can leave ralatively $w_{22} q^2$ offspring in the next generation.

When $s_1 = s_2 = 0$, the three genotypes take the same fitness,
and no selection operates; A and a are neutral genes .

Dominance and Recessiveness of Alleles

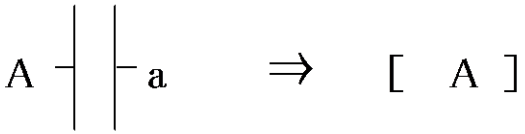
1. Genotype : Gamate types

2. Phenotype : Expressed genotypes in individuals

1) Dominant allele : Expressed allele type

2) Recessive allele : Allele type not expressed

Example,



Genotype Phenotype

A is dominant over a,
a is recessive to A.

3) Codominant alleles : Both A and a alleles are equally expressed.

Gene frequency at the next generation: p'

$$p' = \frac{w_{11}p^2 + w_{12}pq}{w_{11}p^2 + 2w_{12}pq + w_{22}q^2} \quad \cdots\cdots (1)$$

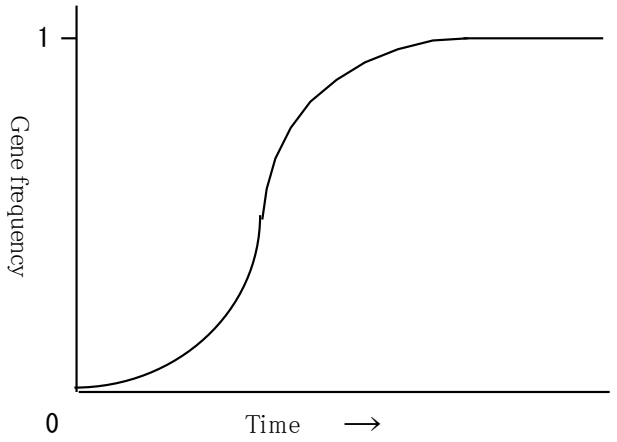
Difference in gene frequency change between the two generations : $\triangle p$

$$\triangle p = p' - p = \frac{pq\{p(w_{11} - w_{12}) + q(w_{12} - w_{22})\}}{w_{11}p^2 + 2w_{12}pq + w_{22}q^2} \quad \cdots\cdots (2)$$

1） Codominance Selection or Genic Selection

A A	A a	a a	
1	1 - s	1 - 2s	($0 < s < 1$)

$$\triangle p = \frac{spq}{1 - 2sq} > 0 \quad \cdots\cdots (3)$$



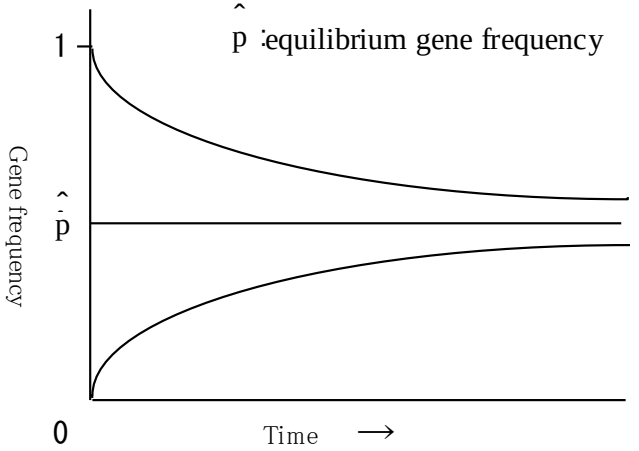
2） Overdominance Selection

A A	A a	a a	
1 - s	1	1 - t	($0 < s,\ t < 1$)

$$\triangle p = \frac{t - (s + t)p}{1 - sp^2 - tq^2} \quad \cdots\cdots (4)$$

$$\triangle p = 0 \Rightarrow \hat{p} = \frac{t}{s + t} \quad \cdots\cdots (5)$$

$\hat{p}$:equilibrium gene frequency



Quiz 2

Confirm formulas 3, 4 and 5 in the previous slide.

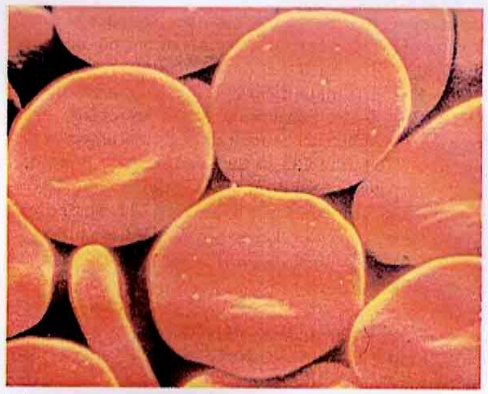
An Example of Over-dominance Selection

Sickle-cell Anemia

This is caused by a mutation at the 6th residue of the beta chain

GAG (Glu) —> GTG (Val)

Upper: Normal
Lower: Sickle-cell



(*a*)

(*b*)

FIGURE 6-11. Scanning electron micrographs of: (*a*) Normal human erythrocytes revealing their biconcave disklike shape. [David M. Phillips/Visuals Unlimited.] (*b*) Sickled erythrocytes from a patient with sickle-cell anemia. [Bill Longcore/Photo Researchers, Inc.]

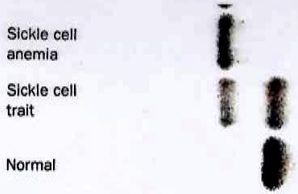


FIGURE 6-12. The electrophoretic pattern of hemoglobins from normal individuals and those with the sickle-cell trait and sickle-cell anemia. [From Montgomery, R., Dryer, R.L., Conway, T.W., and Spector, A.A., *Biochemistry, A Case Oriented Approach* (4th ed.), *p.* 87. Copyright © 1983 C.V. Mosby Company, Inc.]

Sickle cell and malaria

Erythrocytes of the heterozygote for the sickle cell mutation is a less favorable environment for malaria parasites than those of the normal homozygote.

Therefore,

1. **The normal homozygote is advantageous over the heterozygotes in no malarial regions.**
2. **The heterozygote is advantageous over the normal homozygote in malarial regions - over-dominance.**

This indicates that the advantage or disadvantage of a genotype depends on the environments.

Gene frequency change over time by mutation

$$A \underset{u}{\overset{v}{\rightleftharpoons}} a$$

u, v : Mutation rate per genearation

Frequency of A at present : p

Frequency of a at present : q

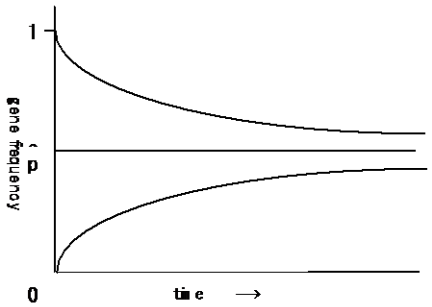
$$p + q = 1$$

Frequency of A at the next generation : p'

$$p' = (1 - v)p + uq$$

$$\triangle p = p' - p = (1 - v)p + uq - p$$
$$= uq - vp = u(1 - p) - vp$$
$$= u - (u + v)p$$

$$\triangle p = 0 \Rightarrow \hat{p} = \frac{u}{u + v}$$



Four Major Evolutionary Factors Causing Gene Frequency Change

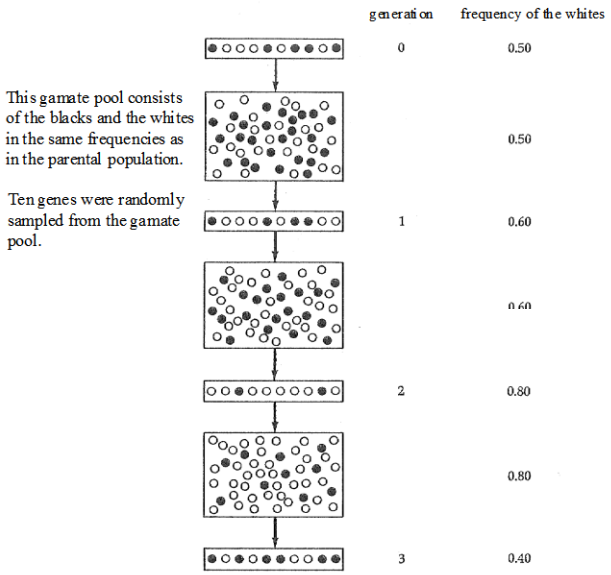
１．Natural Selection

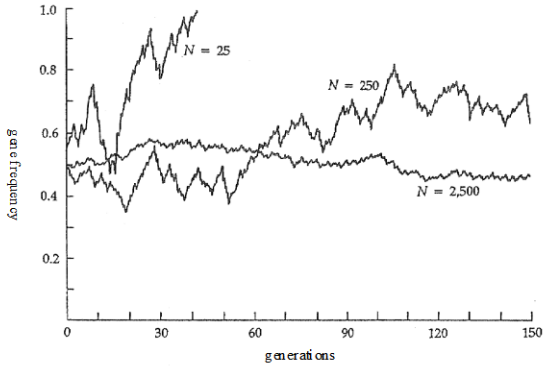
２．Mutation

３．Random Genetic Drift

４．Migration

Gene Frequency Change by Random Genetic Drift

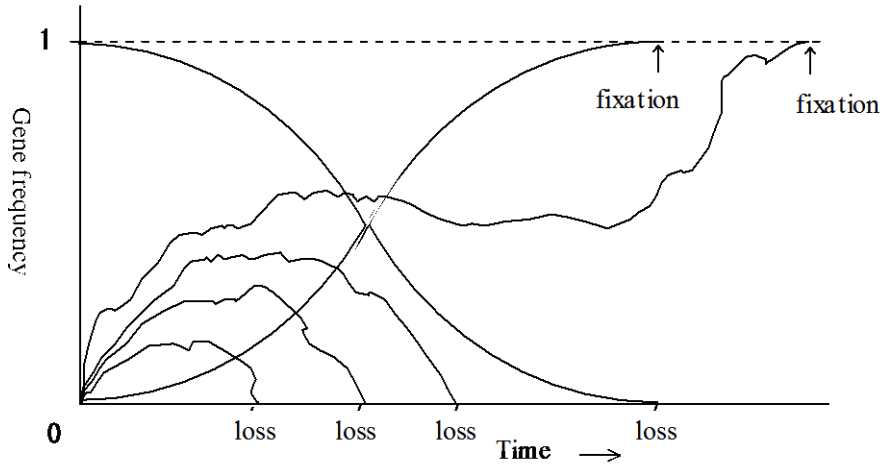




Gene frequency change by random genetic drift in populations of different sizes. Fixation occurs at the 42nd generation in the smallest population, whereas the two other populations are polymorphic even after 150 generations have passed.

Fixation

The situation in which only one type of alleles occupies the population as time elapses.

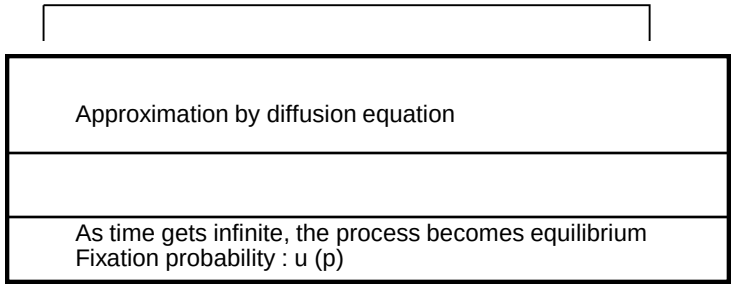


There is no genetic variability in the fixed population, monomorphism; while there is in a population where two or more allele types exist, polymorphism.

Fixation Probability

1) Gene frequency changes continuously over time

2) Time changes continuously



$$u(p) = p,$$

where p is the initial gene frequency, which is $\frac{1}{2N}$

evolutionary rate : k

mutation rate : v

k = the number of mutants × fixation probability

$$= 2Nv \times \frac{1}{2N} = v \qquad \text{(Kimura 1968)}$$

2. In case of co-dominant genes

Genotype	A A	A a	a a
Frequency	p^2	2pq	q^2
Fitness	1	1-s	1-2s

$$u(p) \fallingdotseq \underline{2s}$$

$$k = 2Nv \times 2s = \underline{\underline{4Nsv}}$$

Polymorphism

1. 1) Degree of polymorphism per locus (Heterozygosity per locus)

$$h = 1 - \sum_{i=1}^{n} p_i^2$$

p_i : gene frequency of the i-th allele

n : number of the alleles in the locus

2) Heterozygosity for multiple loci

$$H = \frac{1}{N} \sum_{i=1}^{N} h_i$$

h_i : heterozygosity of the i-th locus

N : number of the loci

2. For DNA sequences

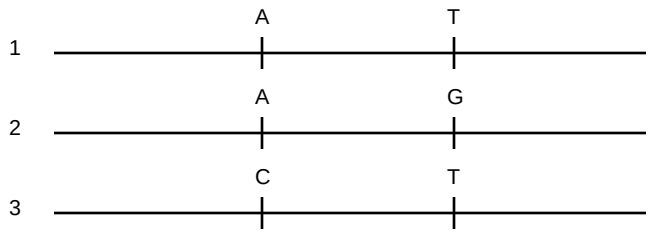
Nucleotide diversity (Nei & Li 1979)

$$\pi = \sum_{i \neq j} p_i p_j s_{ij}$$

p_i : frequency of the i-th sequence

s_{ij} : difference between the i-th and j-th sequences per nucleotide

Quiz 3: What is the nucleotide diversity in the following case where the total number of the nucleotides is 100 for each sequence, and there is no other defference than given here.



The rate of evolution (k) in case of neutral genes or mutations

$k = 2Nv \times (1/2N) = v$

2Nv: number of mutant genes

1/2N: fixation probability

k is constant over time if v is.

Motoo Kimura

木村資生　1924 年愛知県岡崎市生まれ。1947 年京都大学理学部植物学科卒業。1949 年国立遺伝学研究所へ。1968 年分子進化中立説を提唱。集団遺伝学の世界的権威。国立遺伝学研究所名誉教授。

Neutral mutations vs Selective mutations

Neutral mutations occur so that the protein product is not affected by them; ex. the 3rd position of a codon that will not change the corresponding amino acid.

Synonymous mutations

Selective mutations occur so that the protein product is affected by them; ex. the 1st and 2nd positions of a codon that will change the corresponding amino acid.

Non-synonymous mutations

Synonymous mutation

GTT(Val) -> GTC(Val)

Non-synonymous mutation

GTT (Val)-> GCT(Ala)

Natural Selection vs. Neutral Theory at Molecular Level

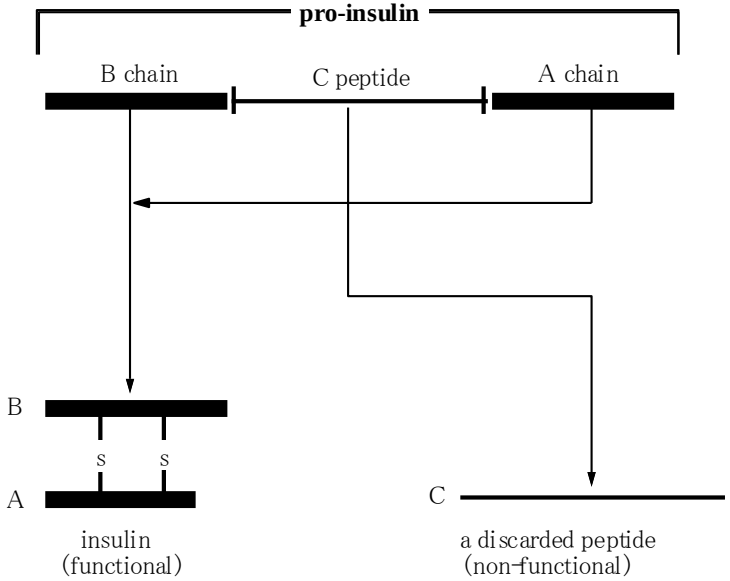
Natural selection

Advantageous genes evolve faster than the other ones.
Namely, functionally important regions in a gene evolve faster than the other regions.

Neutral theory at molecular level

Funcitonally important regions in a gene evolve slower than the other regions
due to functional restrictions on the former.

In case of pro-insulin



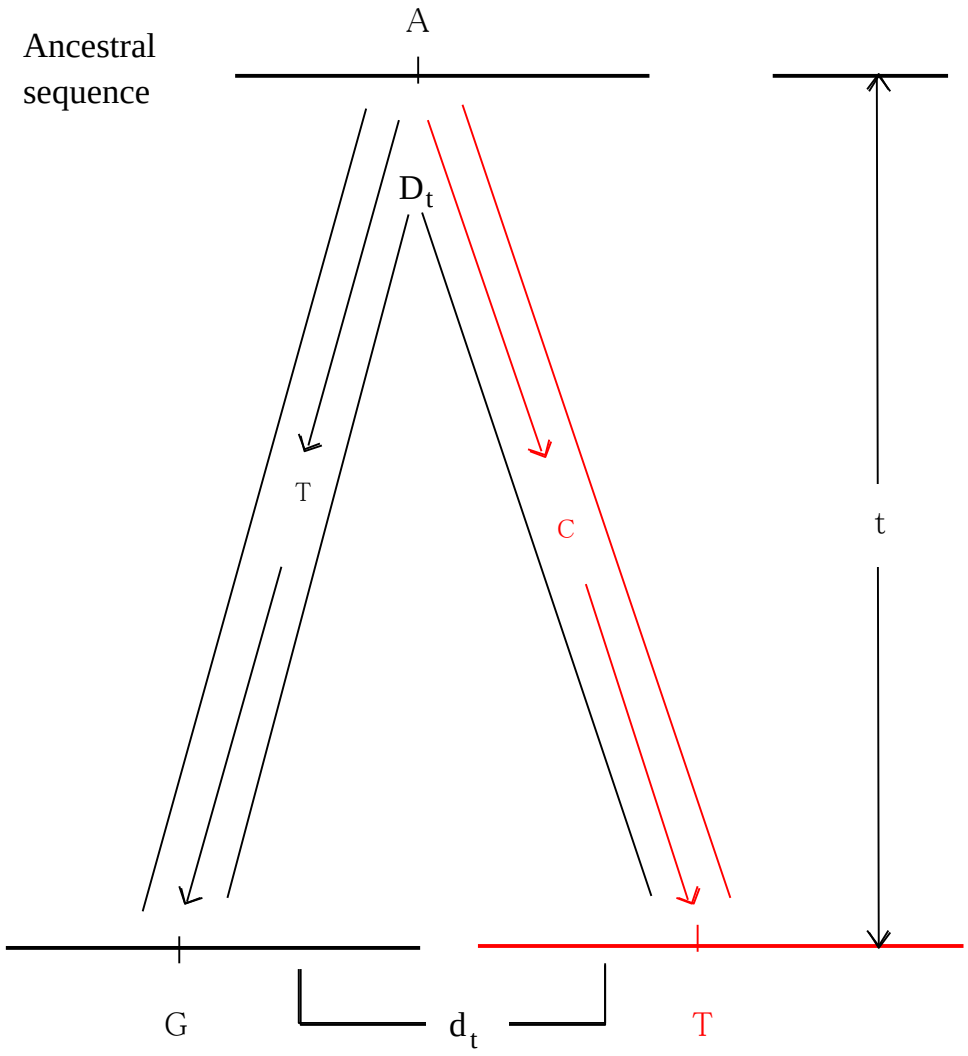
C peptide prevents diabetic vascular and neural dysfunction. (Ido, Y.*et al.*, Science 277: 563-566, 1997)

Table 1. Rates of synonymous and nonsynonymous substitutions in various mammalian protein-coding genes.[a]

Gene	L^b	Nonsynonymous rate ($\times 10^9$)	Synonymous rate ($\times 10^9$)
HISTONES			
Histone 3	135	0.00 ± 0.00	6.38 ± 1.19
Histone 4	101	0.00 ± 0.00	6.12 ± 1.32
CONTRACTILE SYSTEM PROTEINS			
Actin α	376	0.01 ± 0.01	3.68 ± 0.43
Actin β	349	0.03 ± 0.02	3.13 ± 0.39
HORMONES, NEUROPEPTIDES, AND OTHER ACTIVE PEPTIDES			
Somatostatin-28	28	0.00 ± 0.00	3.97 ± 2.66
Insulin	51	0.13 ± 0.13	4.02 ± 2.29
Thyrotropin	118	0.33 ± 0.08	4.66 ± 1.12
Insulin-like growth factor II	179	0.52 ± 0.09	2.32 ± 0.40
Erythropoietin	191	0.72 ± 0.11	4.34 ± 0.65
Insulin C-peptide	35	0.91 ± 0.30	6.77 ± 3.49
Parathyroid hormone	90	0.94 ± 0.18	4.18 ± 0.98
Luteinizing hormone	141	1.02 ± 0.16	3.29 ± 0.60
Growth hormone	189	1.23 ± 0.15	4.95 ± 0.77
Urokinase-plasminogen activator	435	1.28 ± 0.10	3.92 ± 0.44
Interleukin I	265	1.42 ± 0.14	4.60 ± 0.65
Relaxin	54	2.51 ± 0.37	7.49 ± 6.10
HEMOGLOBINS AND MYOGLOBIN			
α-globin	141	0.55 ± 0.11	5.14 ± 0.90
Myoglobin	153	0.56 ± 0.10	4.44 ± 0.82
β-globin	144	0.80 ± 0.13	3.05 ± 0.56

(Continued on next page)

Evolution of DNA sequences



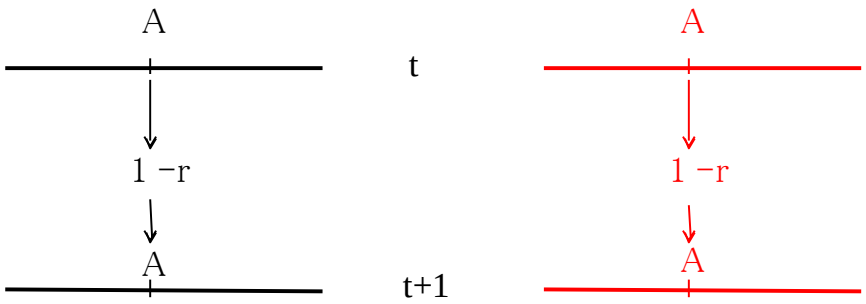
Jukes-Cantor Method (1969)

i_t : probability that a nucleotide at a site remains unchanged after t years

r : rate of nucleotide subsitution per year

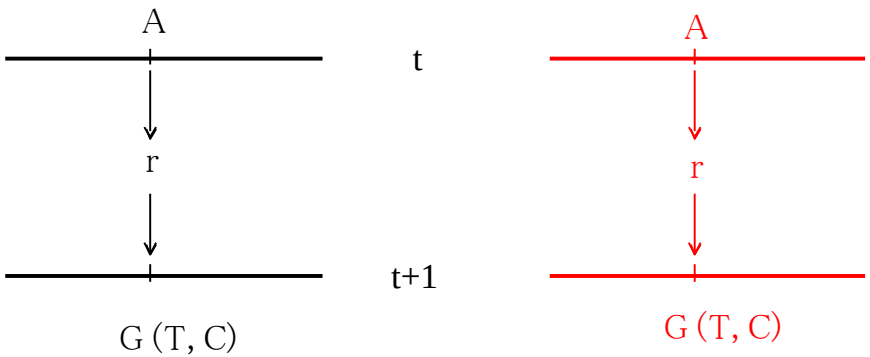
Same ⟶ Same

1.



$$(1-r)^2$$

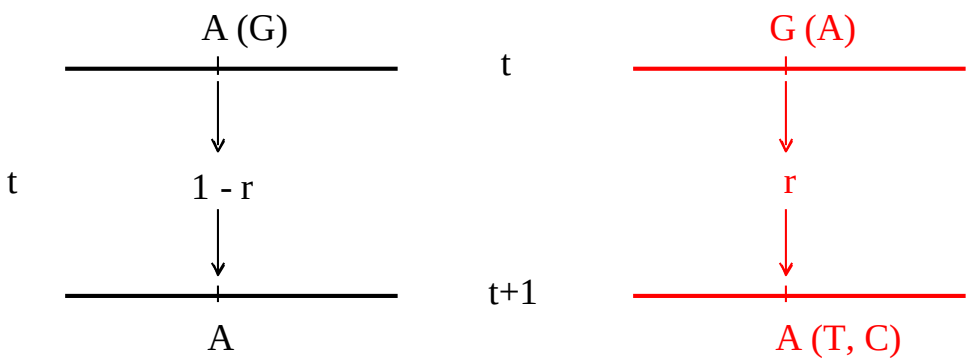
2.



$$3 \times \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) r^2 = \frac{1}{3} r^2$$

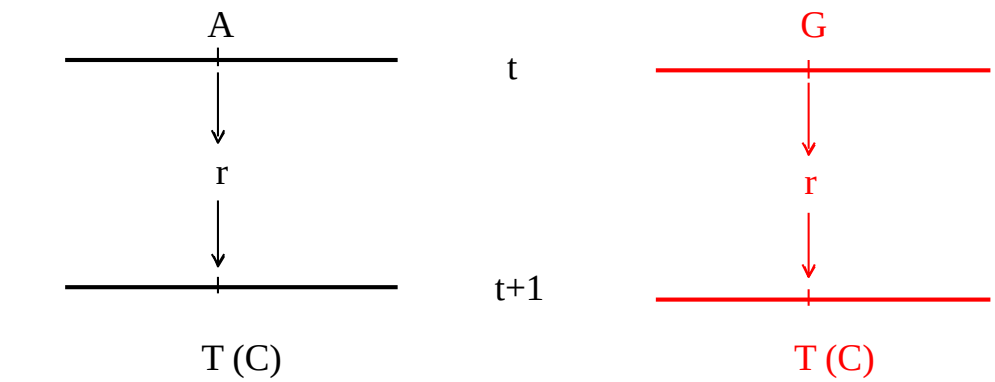
Different ⟶ Same

1.



$$2 \times r\ (1 - r) \times \frac{1}{3} = \frac{2}{3}\ r\ (1 - r)$$

2.



$$2 \times \frac{1}{3}\ r \times \frac{1}{3}\ r = \frac{2}{9}\ r^{2}$$

From the figures we can derive the following difference equation,

$$i_{t+1} = [(1-r)^2 + \frac{1}{3}r^2]i_t + [\frac{2}{3}r(1-r) + \frac{2}{9}r^2](1-i_t)$$

$$\fallingdotseq (1-2r)i_t + \frac{2}{3}r(1-i_t) \qquad (1)$$

Thus,

$$i_{t+1} - i_t = \frac{8}{3}r(\frac{1}{4} - i_t) \qquad (2)$$

The difference equation can be approximated as the differential equation such that,

$$\frac{di_t}{dt} = \frac{8}{3}r(\frac{1}{4} - i_t) \qquad (3)$$

The solution of (3) is given as,

$$\frac{1}{4} - i_t = ce^{-\frac{8}{3}rt} \quad \text{(c is a constant.)} \qquad (4)$$

Since $t = 0 \Rightarrow i_t = 1$

$c = -\frac{3}{4}$, and (4) is rewritten as,

$$\frac{1}{4} - i_t = -\frac{3}{4}e^{-\frac{8}{3}rt} \qquad (5)$$

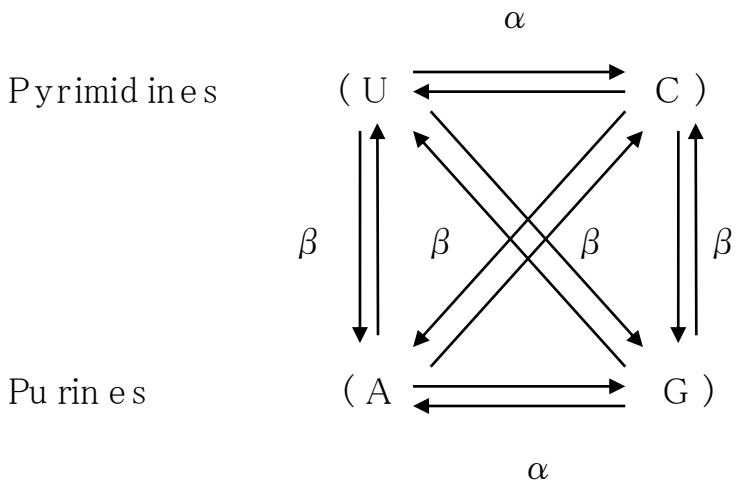
Put $d_t = 1 - i_t$, then (5) is expressed as

$$-\frac{8}{3}rt = \ln(1 - \frac{4}{3}d_t) \qquad (6)$$

Since $2rt = D_t$,

$$D_t = -\frac{3}{4}\ln(1 - \frac{4}{3}d_t) \qquad (7)$$

Estimation of Evolutionary Rates of Nucleotide Substitutions
M. Kimura (1980)



Same	UU	CC	AA	GG	Total
(Frequency)	(R_1)	(R_2)	(R_3)	(R_4)	(R)
Differenet,					
TypeI	UC	CU	AG	GA	Total
(Frequency)	(P_1)	(P_1)	(P_2)	(P_2)	(P)
Different,	UA	AU	UG	GU	
TypeII	(Q_1)	(Q_1)	(Q_2)	(Q_2)	Total
	CA	AC	CG	GC	(Q)
(Frequency)	(Q_3)	(Q_3)	(Q_4)	(Q_4)	

- Types of nucleotide base pairs occupied at homologous sites in two species. Type I difference includes four cases in which both are purines or both are pyrimidines (line 2). Type II difference consists of eight cases in which one of the bases is a purine and the other is a pyrimidine (lines 3 and 4).

$P_1(T+\Delta T) = [1-(2\alpha+4\beta)\Delta T]P_1(T) + \alpha\Delta T[R_1(T)+R_2(T)] + \beta\Delta T \cdot Q(T)/2$

$P_2(T+\Delta T) = [1-(2\alpha+4\beta)\Delta T]P_2(T) + \alpha\Delta T[R_3(T)+R_4(T)] + \beta\Delta T \cdot Q(T)/2$

Summing these two equations, and noting $P(T) = 2P_1(T) + 2P_2(T)$ and $R(T) = R_1(T) + R_2(T) + R_3(T) + R_4(T) = 1 - P(T) - Q(T)$, and writing $\Delta P(T) = P(T+\Delta T) - P(T)$, we get

$\Delta P(T)/\Delta T = 2\alpha - 4(\alpha+\beta)P(T) - 2(\alpha+\beta)Q(T).$ ----------(1)

Carrying out a similar series of calculations for base pairs of type II, we obtain

$\Delta Q(T)/\Delta T = 4\beta - 8\beta Q(T).$ ----------(2)

From these two finite difference equations (Eqs.1 and 2), we obtain the following set of differential equations

$$\frac{dP(T)}{dT} = 2\alpha - 4(\alpha+\beta)P(T) - 2(\alpha-\beta)Q(T)$$

$$\frac{dQ(T)}{dT} = 4\beta - 8\beta Q(T)$$ ----------(3)

The solution of this set of equations which satisfies the condition

$P(0) = Q(0) = 0,$ ----------(4)

i.e., no base differences exist at T = 0, is as follows.

$$P(T) = \frac{1}{4} - \frac{1}{2}e^{-4(\alpha+\beta)T} + \frac{1}{4}e^{-8\beta T}$$ ----------(5)

$$Q(T) = \frac{1}{2} - \frac{1}{2}e^{-8\beta T}.$$ ----------(6)

Writing P_T and Q_T for P(T) and Q(T), we get, from these two equations,

$4(\alpha+\beta)T = -\log_e(1-2P_T-Q_T)$ ----------(7)

and

$8\beta T = -\log_e(1-2Q_T),$ ----------(8)

so that

$4\alpha T = -\log_e(1-2P_T-Q_T) + (1/2)\log_e(1-2Q_T).$ ----------(9)

Since the rate of evolutionary base substitutions per unit time is

$k = \alpha + 2\beta,$

the total number of substitutions (including revertant and superimposed changes) per site which separate the two species (and therefore involve two branches each with length T) is

$K = 2Tk = 2\alpha T + 4\beta T,$

where αT and βT are given by Eqs.(8) and (9). Then, omitting the subscript T from P_T *and* Q_T, we obtain

$$K = -\frac{1}{2}\log_e\{(1-2P-Q)\sqrt{1-2Q}\}.$$ ----------(10)

PNAS
Proceedings of the National Academy of Sciences of the United States of America

Genomic evolution of MHC class I region in primates

Kaoru Fukami-Kobayashi *, Takashi Shiina †, Tatsuya Anzai †, Kazumi Sano †, Masaaki Yamazaki ‡, Hidetoshi Inoko †, and Yoshio Tateno §, ¶

+ Author Affiliations

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved May 17, 2005 (received for review February 9, 2005)

Abstract

To elucidate the origins of the MHC-B–MHC-C pair and the MHC class I chain-related molecule (MIC)A–MICB pair, we sequenced an MHC class I genomic region of humans, chimpanzees, and rhesus monkeys and analyzed the regions from an evolutionary stand-point, focusing first on LINE sequences that are paralogous within each of the first two species and orthologous between them. Because all the long interspersed nuclear element (LINE) sequences were fragmented and nonfunctional, they were suitable for conducting phylogenetic study and, in particular, for estimating evolutionary time. Our study has revealed that MHC-B and MHC-C duplicated 22.3 million years (Myr) ago, and the ape MICA and MICB duplicated 14.1 Myr ago. We then estimated the divergence time of the rhesus monkey by using other orthologous LINE sequences in the class I regions of the three primate species. The result indicates that rhesus monkeys, and possibly the Old World monkeys in general, diverged from humans 27–30 Myr ago. Interestingly, rhesus monkeys were found to have not the pair of MHC-B and MHC-C but many repeated genes similar to MHC-B. These results support our inference that MHC-B and MHC-C duplicated after the divergence between apes and Old World monkeys.

« Previous | Next Article »
Table of Contents

This Article

Published online before print June 20, 2005, doi: 10.1073/pnas.0500770102
PNAS June 28, 2005 vol. 102 no. 26 9230–9234

Abstract *Free*
Figures Only
» Full Text
Full Text (PDF)
A correction has been published

– Classifications

Biological Sciences
Evolution

– Services

Email this article to a colleague
Alert me when this article is cited
Alert me if a correction is posted
Similar articles in this journal
Similar articles in ISI
Similar articles in PubMed
Add to My File Cabinet
Download to citation manager

Search PNAS GO
advanced search »

This Week's Issue
April 26, 2011, 108 (17)

From the Cover

- The art of organic synthesis
- Arresting metastasis with contact inhibition
- Schistosome cell death pathway
- Fitness via genome duplication

Alert me to new issues of PNAS

Early Edition
Archives
Online Submission
Feature Articles
PNAS Plus
Commentaries
Letters

There are three classes of MHC genes

Class I: A, B, C, E, F……

Class II: DP, DQ, DR,…

Class III: C2, C4, TNF,…

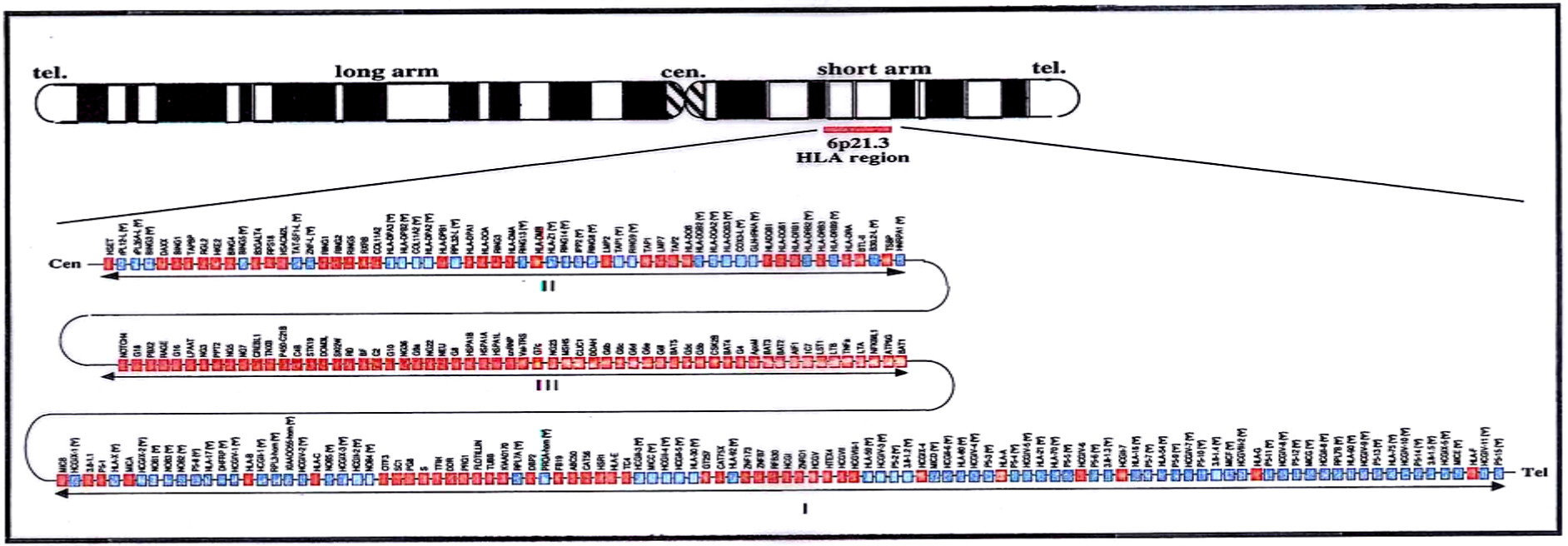
and

Class I related genes: MICA, MICB,……

Classes I and II were coined by Jan Klein in 1977.

(Klein, J. *In* The Major Histocompatibility System in Man and Animals, D. Gotze ed, Springer - Verlag, Berlin, 1977)

Whole Genome Structure of HLA on Chromosome 6



High Density of Genes: 235 Genes per 3.6 Mbp

Human LINE1

5' Region	875 bp
ORF1	1017 bp
IS	36 bp
ORF2	3852 bp
3' Region	192 bp

ORF 1: Gag-like protein?,
ORF2: EN (integration) and RT (reverse transcriptase)

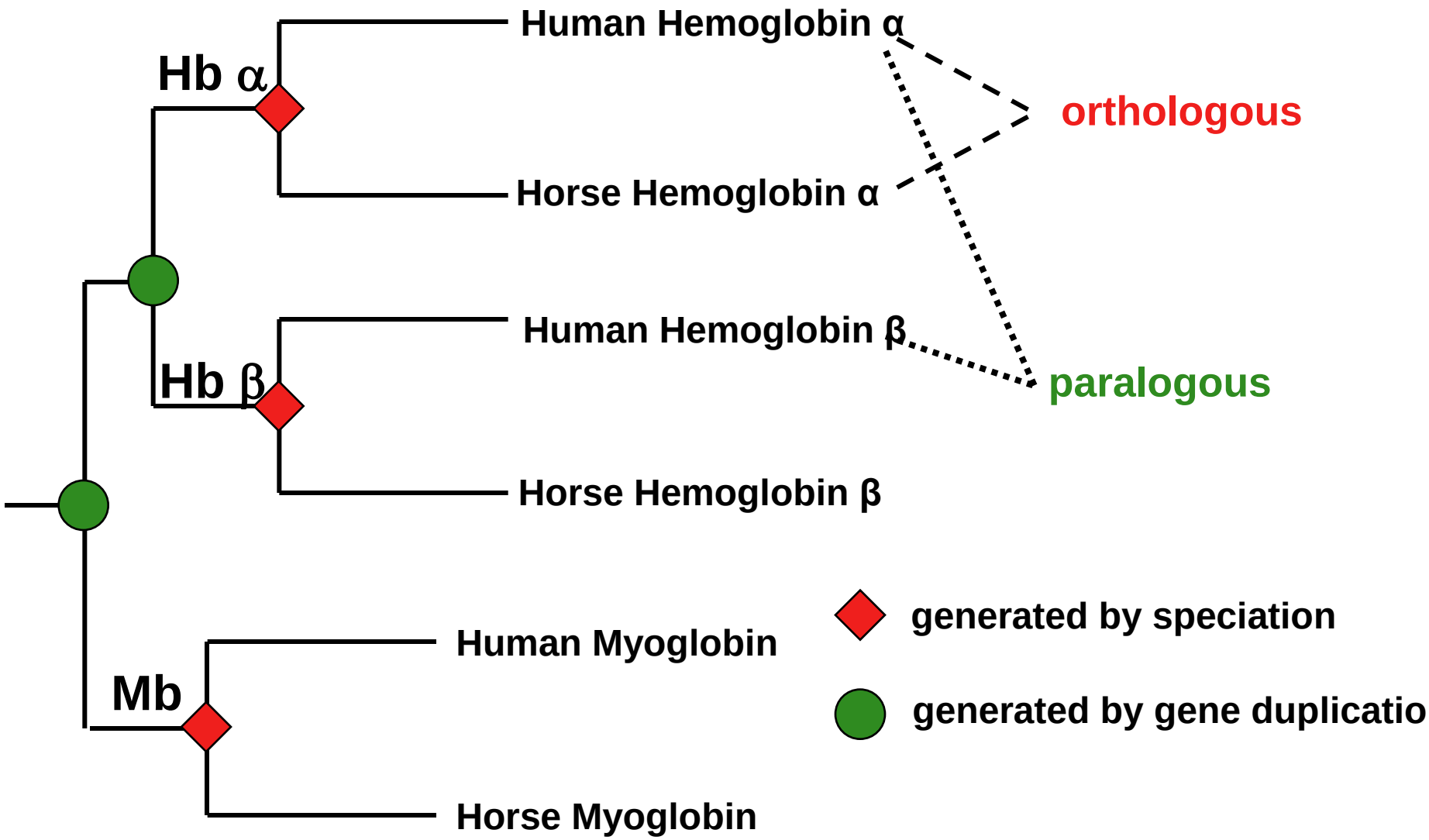
Our Approach

1. **Because MHC genes themselves are known to have been subject to strong positive natural selection, we instead focused on neutral or nearly neutral segments in the MHC class I genome regions. Neutral genes and genome segments are known to evolve constantly over time (Kimura 1968).**

2. **We also try to carry out genome-oriented analysis.**

3. **We thus selected incomplete (and thus neutral) LINE sequences that were orthologous among the species studied.**

Orthologous vs. Paralogous



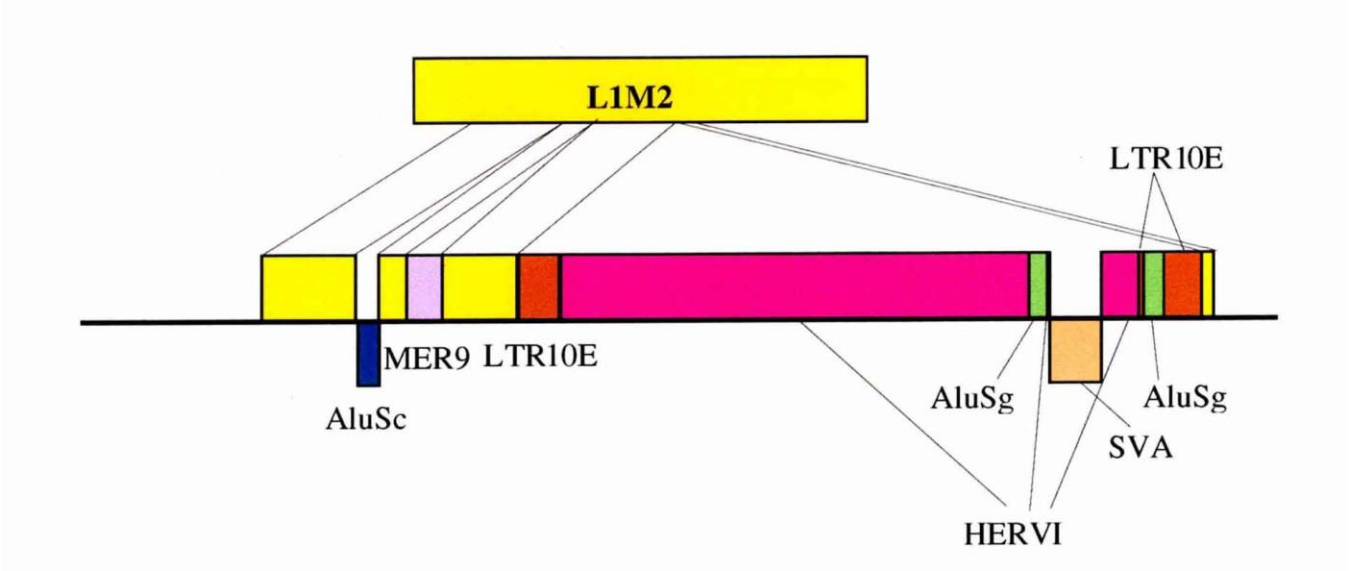
Orthologous Repeated Sequences among Spec

Generally, it is not easy to select repeated sequences that are orthologous to each other among the species studied.

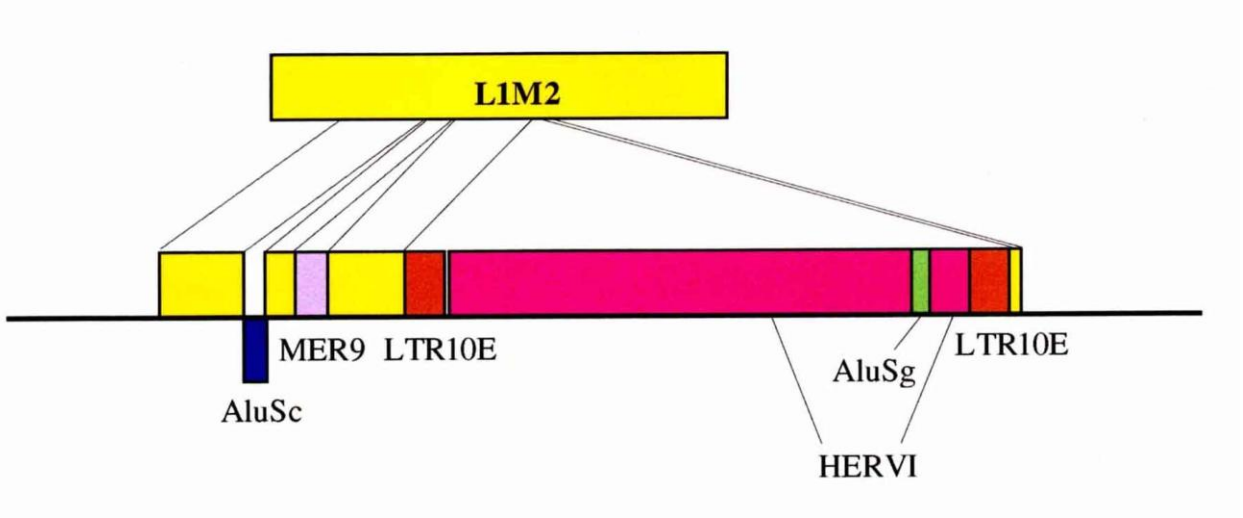
We confirmed they were orthologous among the species by examining the following four aspects.

1) **their relative locations to other genes and fragments,**
2) **5'-3' direction,**
3) **structures and**
4) **homology.**

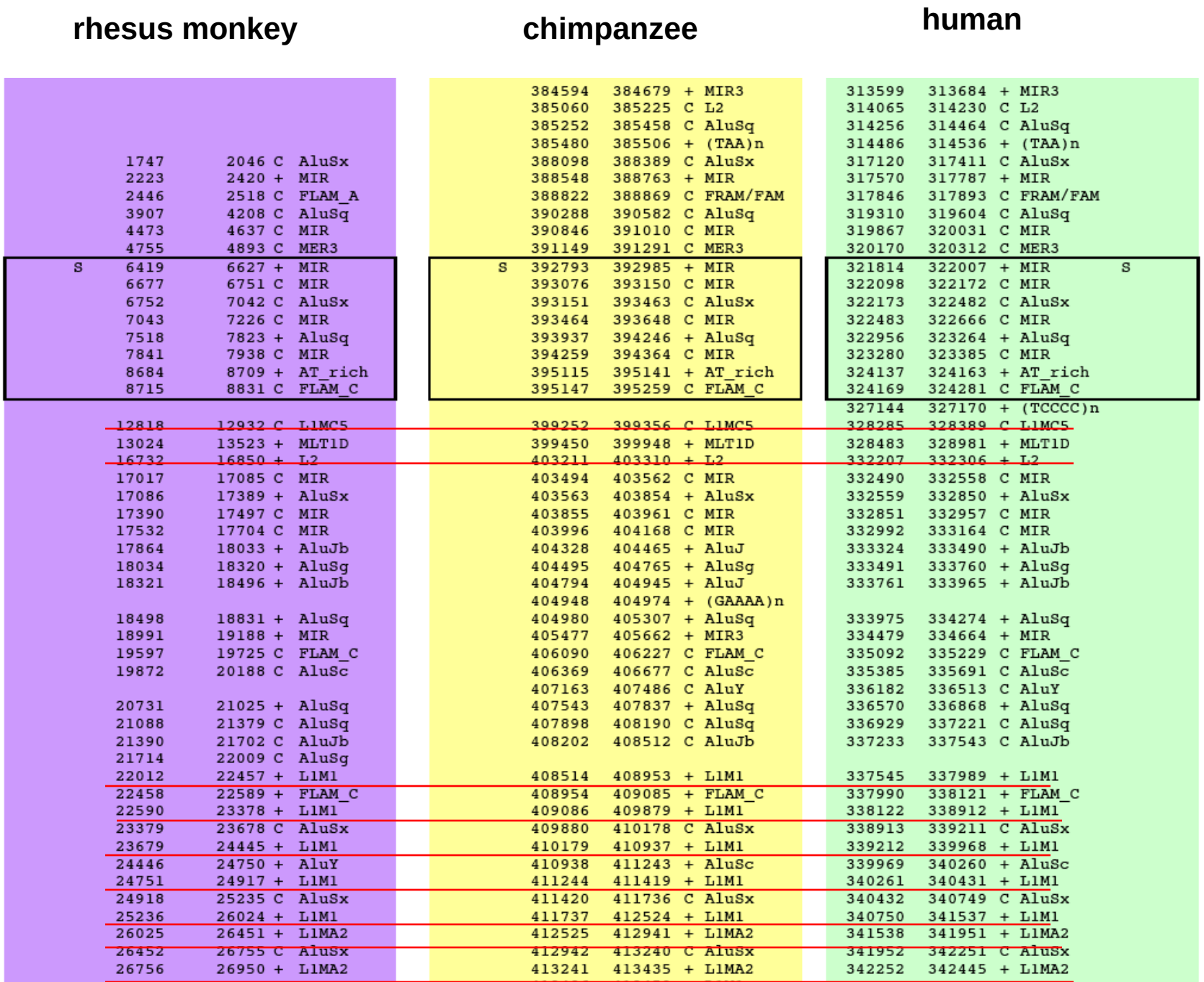
Man



Chimpanzee



A genomic region near *CDSN*



Number of LINE sites and %identity among human, chimpanzee and rhesus monkey

	Rhesus	Chimp	Human
Rhesus	—	29,344 (95)	29,602 (98)
Chimp	94.0	—	31,313 (98)
Human	94.0	98.8	—

Identity (%) is located below and to the left of the diagonal; number of sites (LINEs) is located above and to the right of the diagonal.

Evolutionary distance and divergence time among human, chimpanzee and rhesus monkey

	Rhesus	Chimp
Rhesus	-	
Chimp	0.0629 ± 0.0015	-
Human	0.0624 ± 0.0015	0.0125 ± 0.0006

Evolutionary distances (substitutions per site) were computed by Kimura's two parameter method.

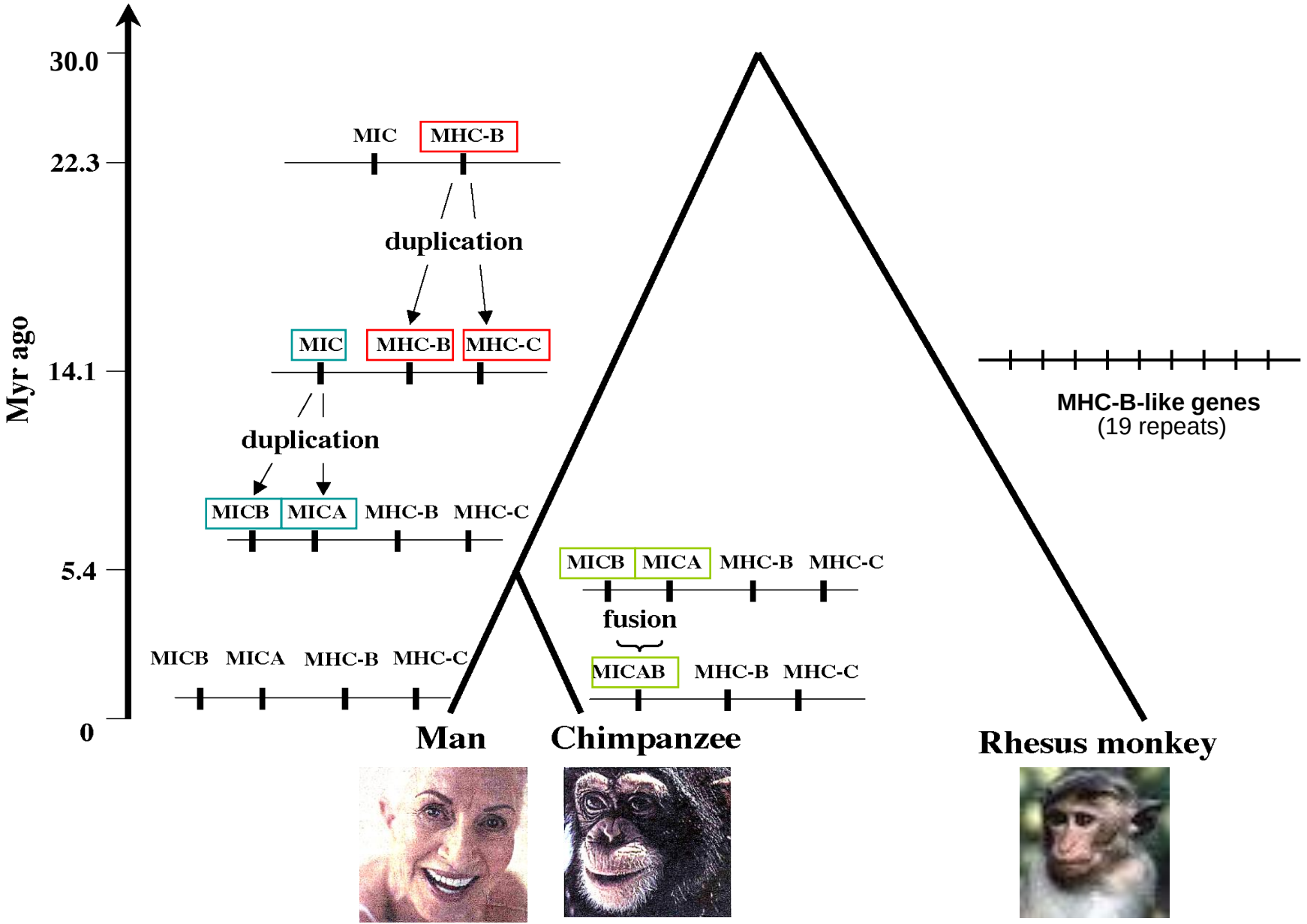
Rate of LINEs in this region

$$r = \frac{0.0125}{2 \times 6 \times 10^{6}} = 1.04 \times 10^{-9} \quad site / year$$

Divergence time between rhesus and (man, chim)

$$T = \frac{0,062 / 2}{1.04 \times 10^{-9}} = 30.0 \quad million \quad years$$

Duplications of MHC-B, C and MICA, B in reference to phylogeny of man, chimpanzee and Rhesus monkey




Thank you very much