생명정보학에서 쓰이는 컴퓨터 PerI 언어의 기초 교육

2017/9/9

Jongsun Park, Ph. D.

성신여자대학교 SUNGSHIN UNIVERSITY

서울시 강서구 화곡동 359-63 영주빌딩 201호
Tel 02 2698 1188 / Fax 02 6280 8821

www.infoboss.co.kr

InfoBoss
Infomations & Systems

Bioinformatics

From Wikipedia, the free encyclopedia

For the journal, see Bioinformatics (journal).

Bioinformatics /ˌbaɪ.oʊˌɪnfərˈmætɪks/ (listen) is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. Bioinformatics has been used for *in silico* analyses of biological queries using mathematical and statistical techniques.

Bioinformatics is both an umbrella term for the body of biological studies that use computer programming as part of their methodology, as well as a reference to specific analysis "pipelines" that are repeatedly used, particularly in the field of genomics. Common uses of bioinformatics include the identification of candidate genes and single nucleotide polymorphisms (SNPs). Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organisational principles within nucleic acid and protein sequences, called proteomics.[1]

From Wikipedia

- Bioinformatics is a field to **develop methods** and **software tools** for analyzing all kinds of biological data.
- Main way to analyze biological data (efferently or systematically) is writing a program.

- Bioinformatics started with needs to deal with huge amount of data (sequences) which human cannot handle.

- Human Genome Project (HGP), of which aim is unrevealing human genome sequences fully, generated a lot of data using Sanger Sequencing Method.

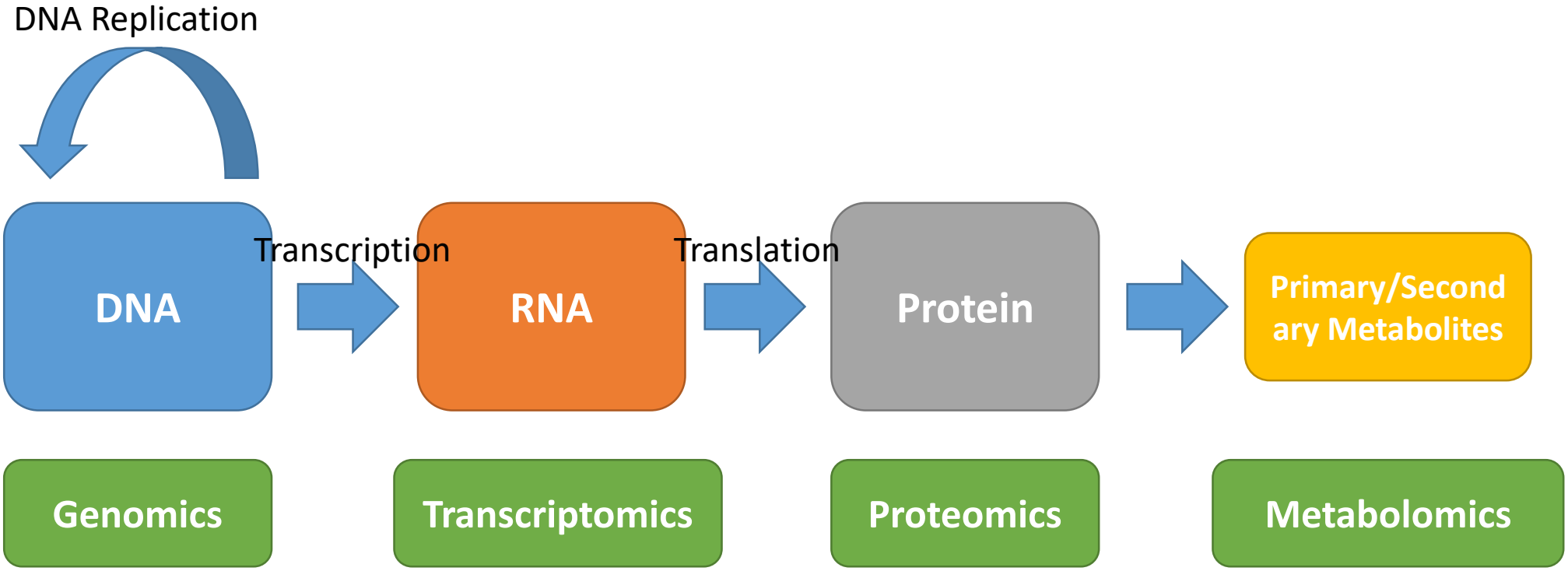
- Several programs, such as phred, phrap, and consed, had been developed for automation of sanger sequencing data.

- With the aid of program, in the late phase of human genome project, new sequencing method to uncover human genome sequences was addressed by Craig Venter.
- International group for human genome sequences keep going with traditional method and published paper in 2001.
- Craig Venter made his company (Celera Genomics) in 1999 and finished his genome sequences at the same time to international group.
- He published his genome sequence in nature and international group published in Science.

CELERA

Venter, J.C. et al. The sequence of the human genome. Science 291, 1304-1351 (February 16, 2001).
Lander, E.S. et al. The Genome International Sequencing Consortium. Intial sequencing and analysis of the human genome. Nature 409, 860-921 (February 15, 2001).

- Genomics is a starting point of bioinformatics in early phase.
- More technologies which can see global pictures at the diverse levels, DNA(=genome), RNA(=transcriptome), proteins (=proteomics), and primary and secondary metabolites (=metabolomics), have been developed with forming diverse bioinformatics area.

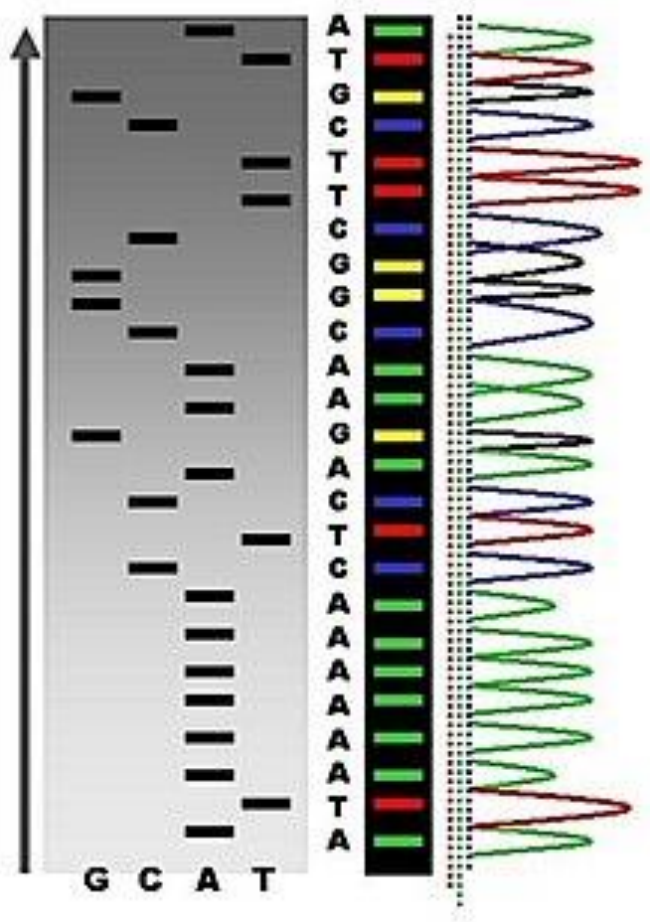


DNA Sequencing Technology (1)

Sequencing technology is for reading nucleotides one by one

InfoBoss
Infomations & Systems

- Sequencing technology is for reading nucleic acid one by one.
- Sanger developed Sanger's Method and published in Nature, 1977, with uncovering sequence of bacteriophage.
- This technique has been automated with florescence dyes (ABI3730).



Journal home > Archive > Article > Full Text

Article

Nature **265**, 687-695 (24 February 1977) | doi:10.1038/265687a0; Accepted 24 December 1976

Nucleotide sequence of bacteriophage ΦX174 DNA

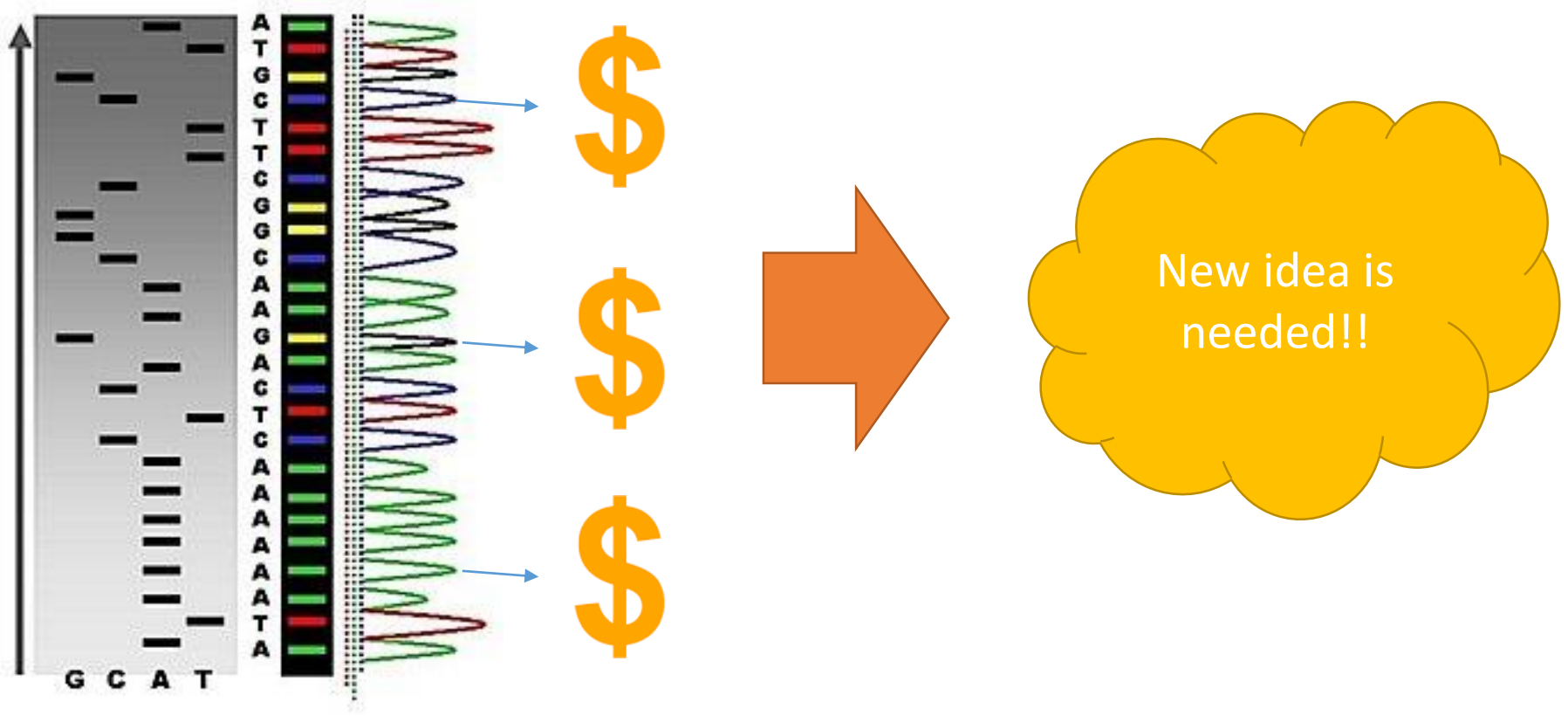
F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, III‡, P. M. Slocombe§ & M. Smith¶

1. MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK
2. Present addresses : *, John Curtin School of Medical Research, Microbiology Department, Canberra City ACT 2601, Australia, †, Department of Biochemistry, University of Bristol, Bristol BS8 1TD, UK, ‡, Department of Bacteriology and Immunology, University of North Carolina, Chapel Hill, North Carolina 27514, §, Max-Planck-Institut für Molekulare Genetik, 1 Berlin 33, FRG, ¶Department of Biochemistry, University of British Columbia, Vancouver BC, Canada V6T 1W5,

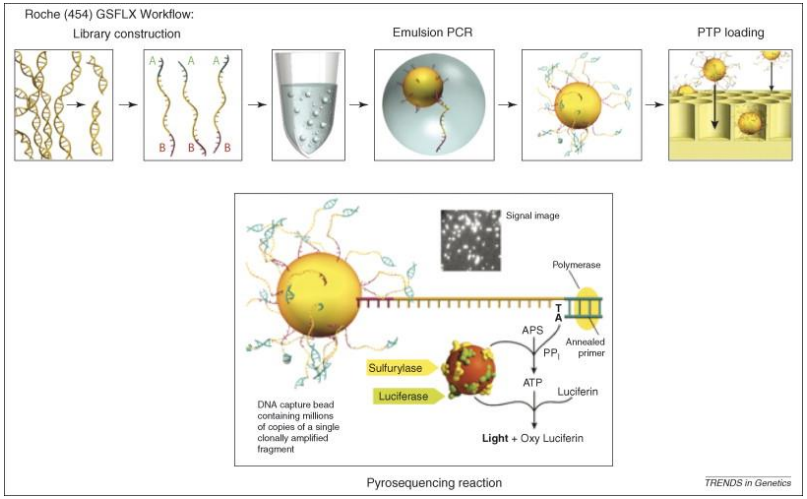
A DNA sequence for the genome of bacteriophage ϕX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

Top

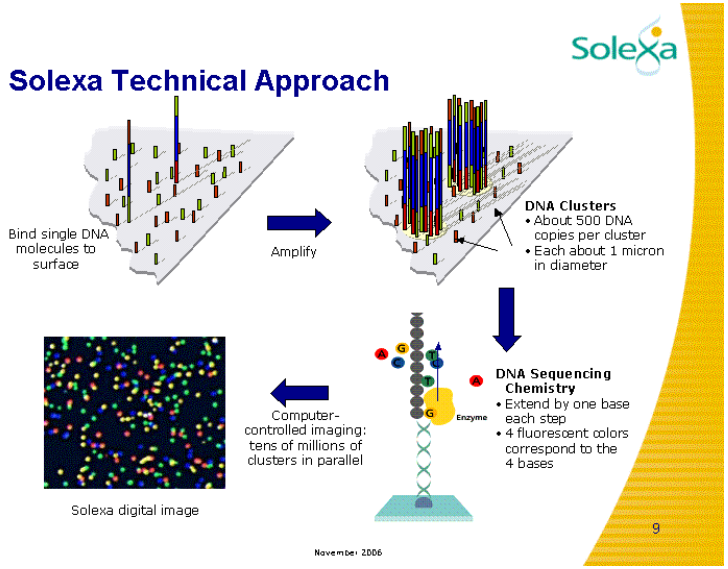
- Sanger Sequencing technology allows us to read DNA sequences up to 1,000 bp per one reaction.
- However, chromosome or most of our targets are more than 1 kb, so that we need to make idea to get the data.
- In addition, through the automation, the cost of sequencing is not cheap due to the limitation of technology.

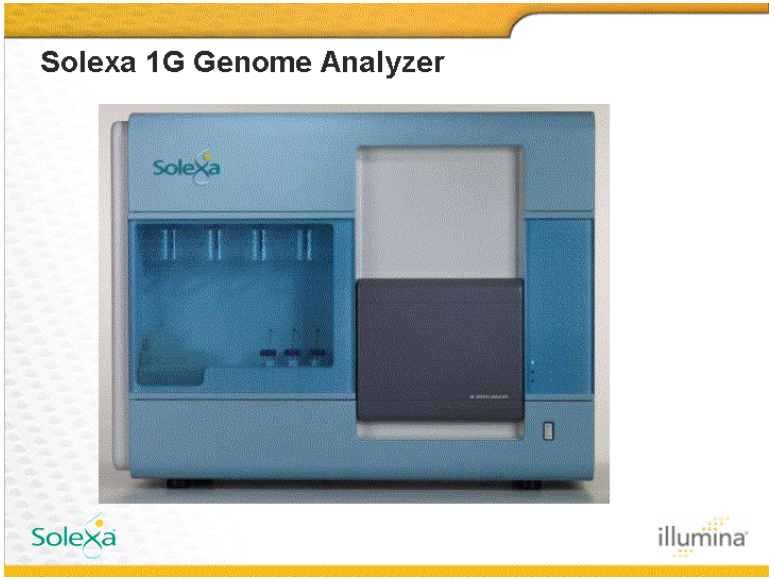


- Next Generation Sequencing Technology was commercially launched in 2005: 454 Life Science.



- Nextly, Solexa sequencer was launched in 2007.

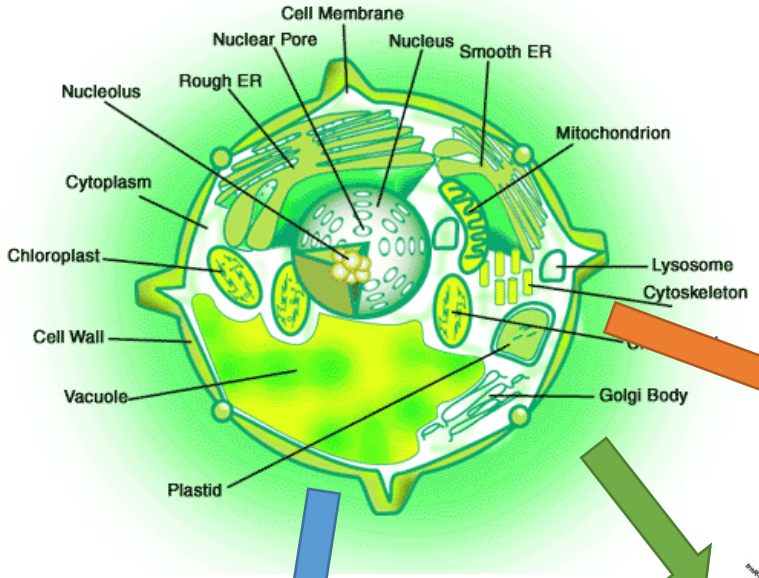




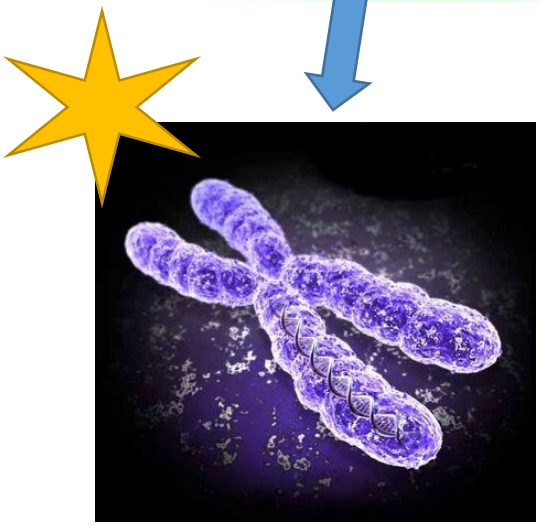
- Data format of NGS sequencer results is simple text-based one (fastq).

```
[starflr@Caturra 1093]$ head 2044_1.fq
@ERR024191.34 EAS131_89:5:1:1048:7930 length=100
AGTAACCACTATTCTTCTGTATTTTCAGGCAATCNACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+ERR024191.34 EAS131_89:5:1:1048:7930 length=100
BBCD5A-5AADDD?DA=D?BC:DDDDDABA####!##!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@ERR024191.36 EAS131_89:5:1:1049:12752 length=100
CCTTCTTGAATTTAATTTTTTTTTCAACTGCCCCAGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+ERR024191.36 EAS131_89:5:1:1049:12752 length=100
FFF?EFDE=DEA:EBDEDE=DCCCED:D:?=CA@CA#!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@ERR024191.37 EAS131_89:5:1:1049:2843 length=100
TGTATCTCTTGGGTTTATGCAGGATAAATTTCTAGATGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

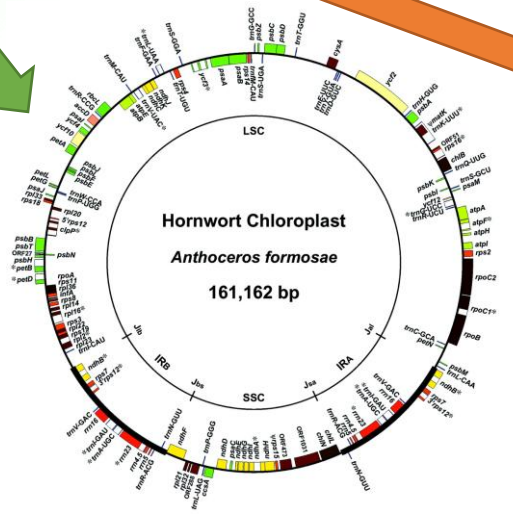
- It contains read name (=number), reads (nucleotide sequence), quality information (Q-33).
  - Based on ASCII code, quality data can be calculated.



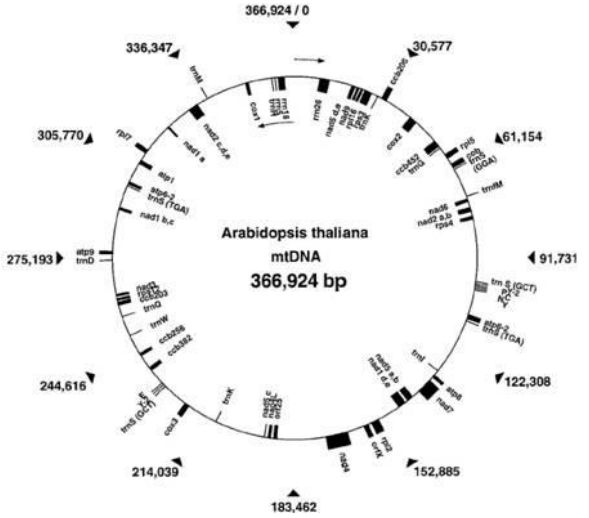
- A **genome** is an organism's **complete set of DNA**, including all of its genes.
- Each genome contains all **information** needed to build and maintain that organisms.



Chromosomes



Chloroplast



Mitochondrion

- Usually, plant cell has three types of chromosomes including two organelle genomes: chloroplast and mitochondria.

Data Structure of Genome Sequence

Informatic aspect of genome sequences

InfoBoss
Infomations & Systems

Genome sequence (Nucleotide sequence)

Gene

3’ UTR

ATGAGCAACAACAGCTCCATCGTGGGTGCTCCAAAGCTTGTAGCACGCGCGCAGCTGCGGAAGCTCC
TGAAGCGGTGGAGCGGACACATCCCGGTATCGATGAAGTACGAGCGGGGGAAGAAGGAAGCAGTGC
CTATCCCTGTCCCTGATGCCCAGCCTAAGCCTCTTAGTGCGGCAGAACGCGAGACGCTCAGCACTCT
CGCCCCGTGCGTCTACAAACTCGTCAACTACAGCGACCCCAACCACGCAAGCCTGTGCCCCAACCCT
GCGTACCGGAAGCTGTTCTACCACCTGGGCTTCATGCACAGCGACGCCGAGCTCCTCCGGCACGGG
GCTCAGGACACGGTCGTGCA

5’ UTR

Gene

Transcript

Alternative transcripts

Open reading frame

MSNNSSIVGAPKLVARAQLRKLLKRWSGHIPVSMKYERGKKEAVPIPVPDAQPKPLSAAERETLSTLAPCVY
KLVNYSDPNHASLCPNPAYRKLFYHLGFMHSDAELLRHGAQDTVVQILTTLKLSAGAPHIEIECEWDQALHTY
APCLQAVLLSGIDRDQTRKGFRKSGSTWKSREKRQGTGYKVDANLLYALAPLLQHLLRLSFWETDYIANHQA
SALLCEALGVERGEICWGCLPGDRGSQRVGKDVYFDRVPELVDEIRELMSGPCRGTSDGILAGEAYVANQH
RLSSSNVYHRVFRDVPQDAGKDKPNGKCIGGERDVKCNEETDDGCPELVSDEVAKKGAKKGGKKVRCQKA
ERSDTDGVMIVTCKHLFIYYYHVLLRHESLHDVFTFLVTRFEHRPPKFIIYDNACELARYCRARSPEFFARTMF
LTDVFHGVTHKCSDEFKYHAHLLKEGREFQGPRVTGPQEVINGVMKRFDSTQAASIX

Protein

Current Status of Plant Genomes

How many plant genomes are available at Augustus, 2017.

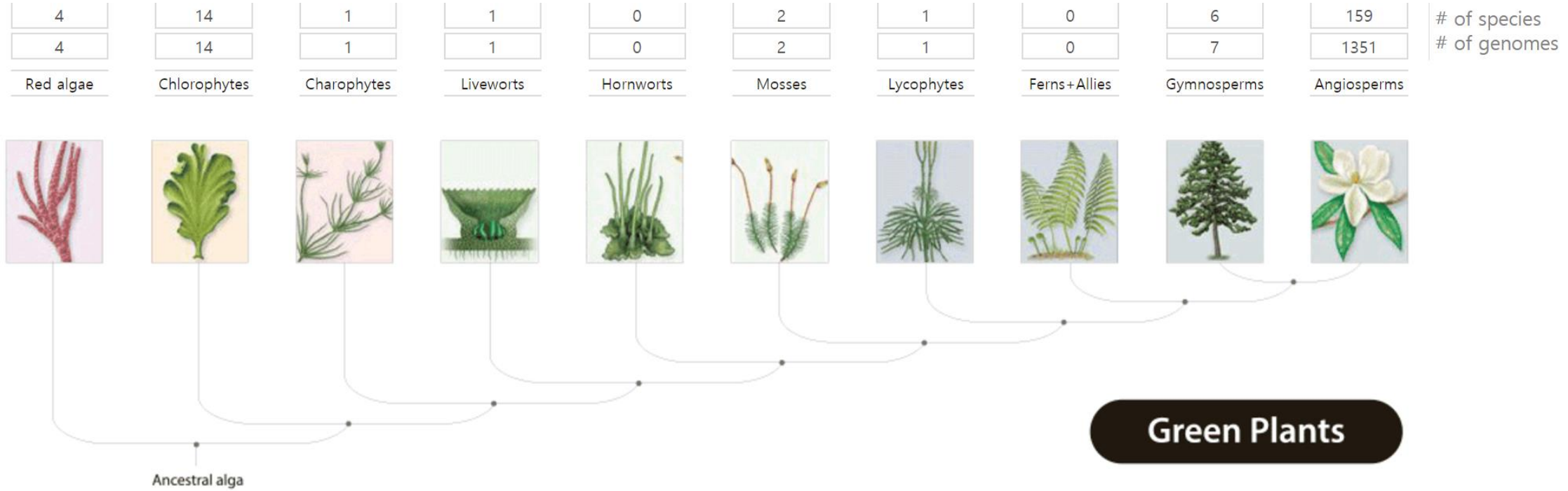
InfoBoss
Infomations & Systems

Genome Statistics

Plant Genome Release 2.3

Plant Genome Database

# of plant genomes / with gene model
1,381 / 165
Total length of plant genomes (bp)
418,219,776,616
# of plant genes / ORFs
5,529,476 / 7,223,612



- **1,381** plant genomes including red algae originated from **188** species are available.
- **418 Gbp genome sequences** and **5.5 million plant genes (7.2 million transcripts)** are ready to be analyzed.

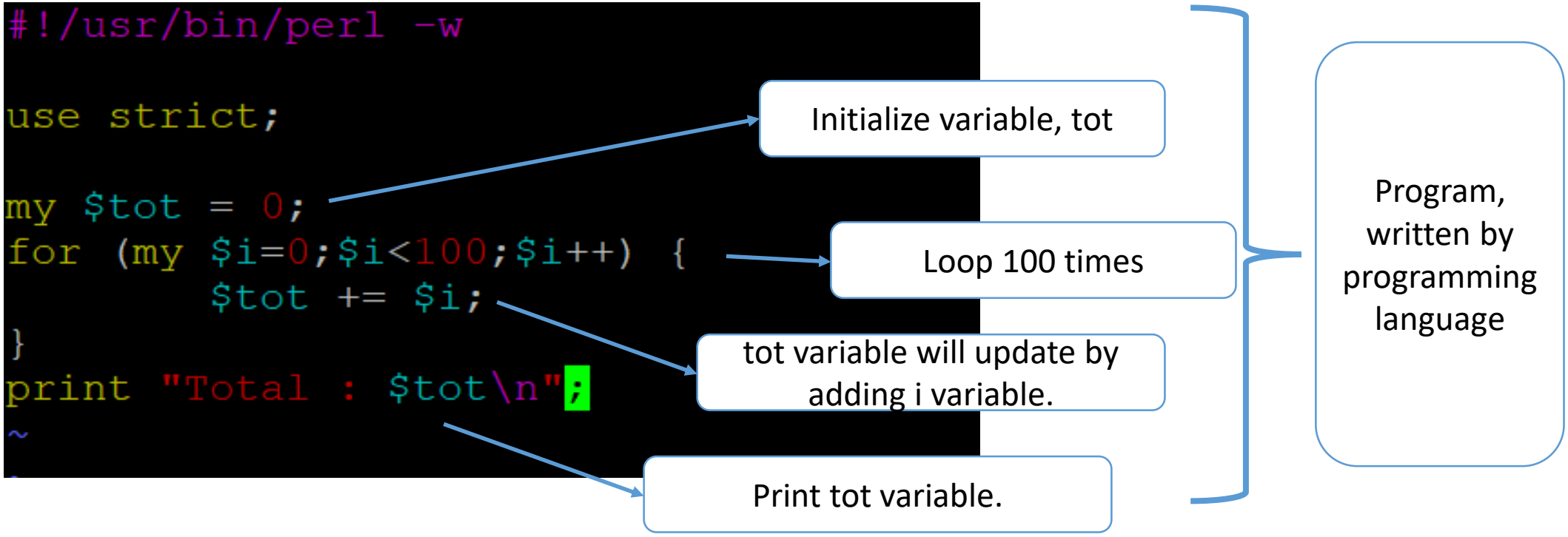
Computer program

From Wikipedia, the free encyclopedia

For the TV programme, see The Computer Programme.

A **computer program** is a collection of instructions[1] that performs a specific task when executed by a computer. A computer requires programs to function and typically executes the program's instructions in a central processing unit.[2]

- Program is a collection of instructions to make computer works.



A **programming language** is a formal language that specifies a set of instructions that can be used to produce various kinds of output. Programming languages generally consist of instructions for a computer. Programming languages can be used to create programs that implement specific algorithms.

```c
/* This line basically imports the "stdio" header
 * the standard library. It provides input and outp
 * to the program.
 */
#include <stdio.h>

/*
 * Function (method) declaration. This outputs "Hello, world" to
 * standard output when invoked.
 */
void sayHello() {
    // printf() in C outputs the specified tex
    // formatting options) when invoked.
    printf("Hello, world!");
}

/*
 * This is a "main function". The compiled pro
 * defined here.
 */
void main() {
    // Invoke the sayHello() function.
    sayHello();
}
```

Programming Language, C

```php
<?
#!--- Apply4 header
#!System name: INFOBOSS
#!File type: PHP web page
#!File name: index.php
#!File contents: This page displays the main page of INFOBOSS system
#!Current version: 1.2.2
#!Comment: Add the code for showing temperature of office in main page
#!--- Apply4 header

include "/home/infoboss/lib/php/eqinit.ph";

if (empty($a)) { $a = "view"; }

switch ($a) {
        case "view":
                checkSimPagePermission($a);

                $data = new stdclass();

                $connect = connectDB("INFOBOSS");

                $cnt = 0;
                $sql = "SELECT * from Project
```

Programming Language, PHP

- Perl is a family of high-level, general-purpose, interpreted, dynamic programming languages.
- Perl is OS-independent: it can be used under Windows and Linux(UNIX).
- Perl was originally developed by Larry Wall in 1987 as a general-purpose Unix scripting language to make report processing easier. (Regular expression is the best component in Perl.
- Perl syntax is similar to that of C, AWK, and sed:

Perl

```
#!/usr/bin/perl -w

use strict;

my $tot = 0;
for (my $i=0;$i<100;$i++) {
        $tot += $i;
}
print "Total : $tot\n";
~
```

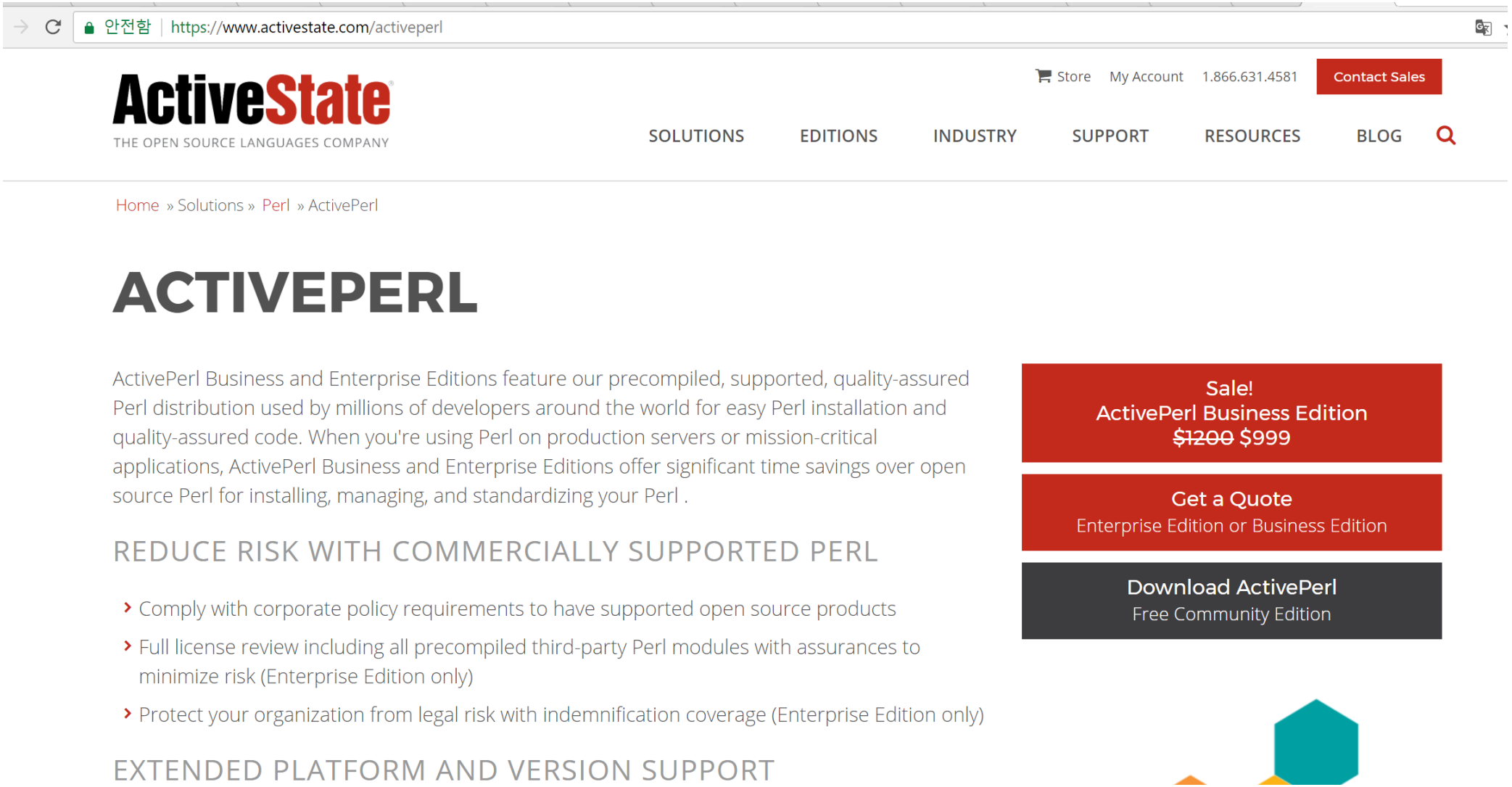
C

```
#include <stdio.h>

int main() {
        int tot = 0;
        for (int i=0;i<100;i++) {
                tot += i;
                }
        printf("Total : %d\n", tot);
}
~
```

- In Windows, you have to install Perl Interpreter.

- ActiveState Perl (http://www.activestate.com/activeperl/) can be downloaded freely.

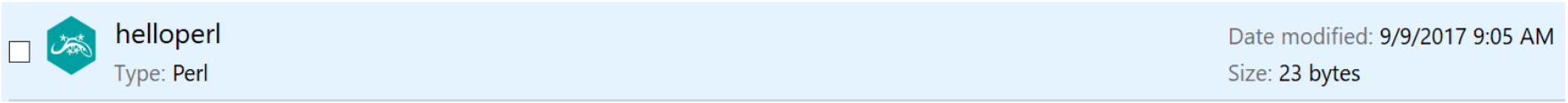


- Here is simple Perl code.

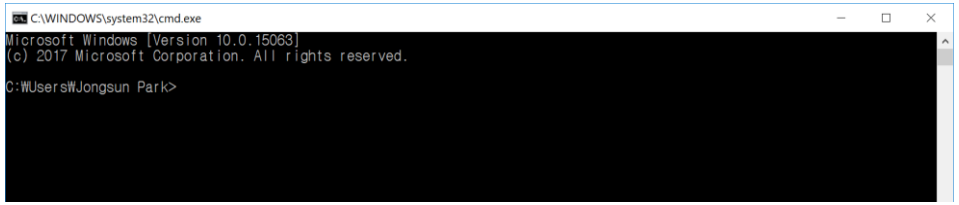
```
print “Hello, perl.\n”;
```

- Running this Perl script like below:

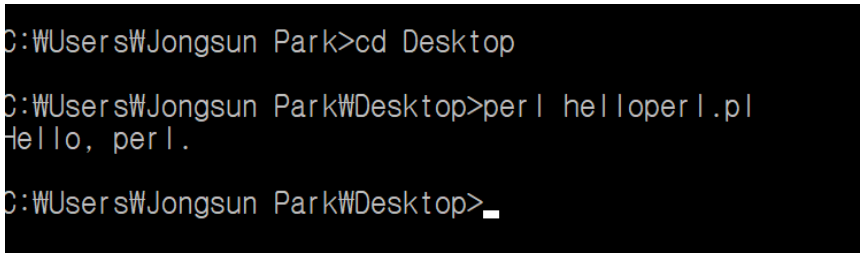
  - Make helloperl.pl on the Desktop.



  - Open ‘cmd’ via Windows+R key.



  - Run the program



- Variable is a space for storing value in memory.

- In Perl, variable can be defined using ‘my’ command:

```
my $a;
```

- ‘$’ means scalar variable in Perl, which is normal variable.

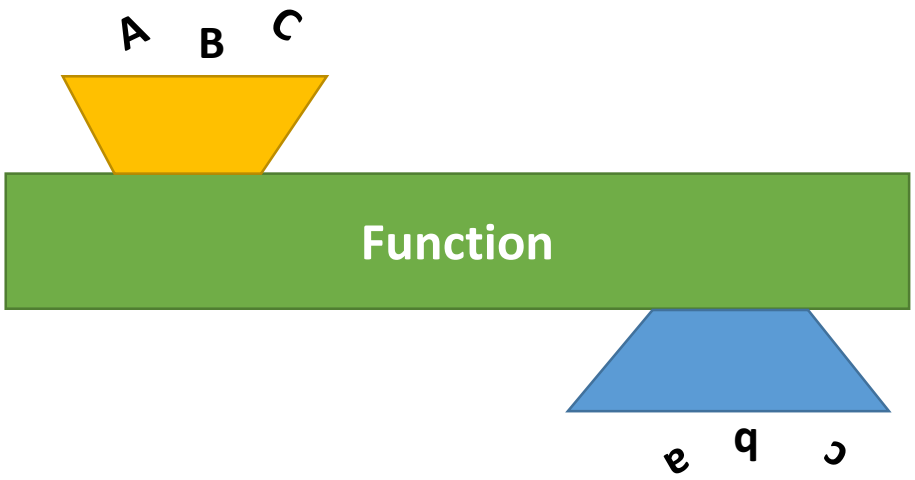
- After defining variable, you can ‘assign’ value on it because it is a space for storing value.

```
$a = 1;
```

```
$a = “ABC”;
```

- Value can be numeric as well as string because Perl allow any types of value be stored in variable.

- Function is a set of commands with arguments and return values.



- Perl provide built-in functions in interpreter: for example, ‘print’.

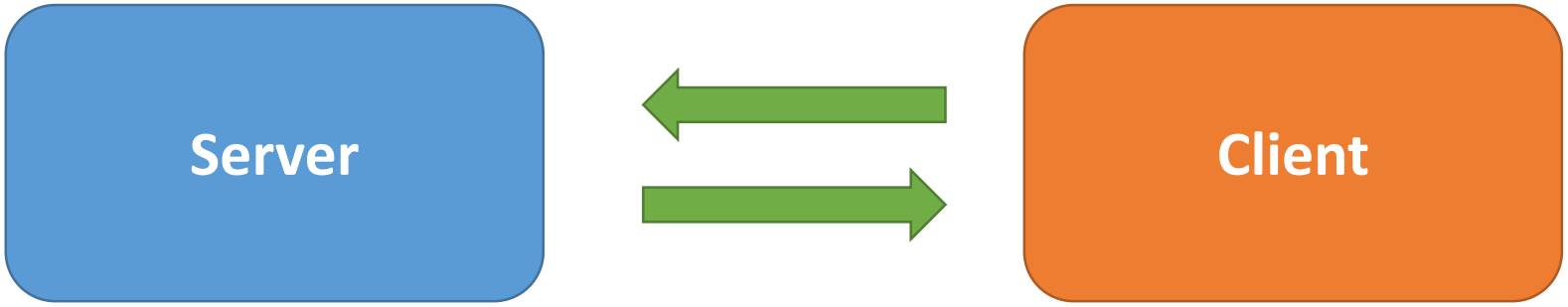
```
print “Hello perl!\n”;
print $a;
```

- print function display string or variables on the screen.

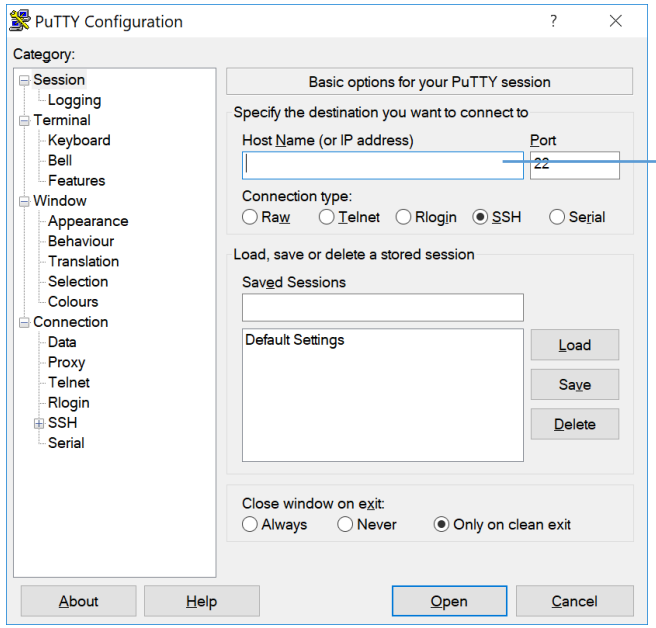
```
print “a valie is “.$a.”\n”;
```

- Server-client is important concept for bioinformatic analysis.

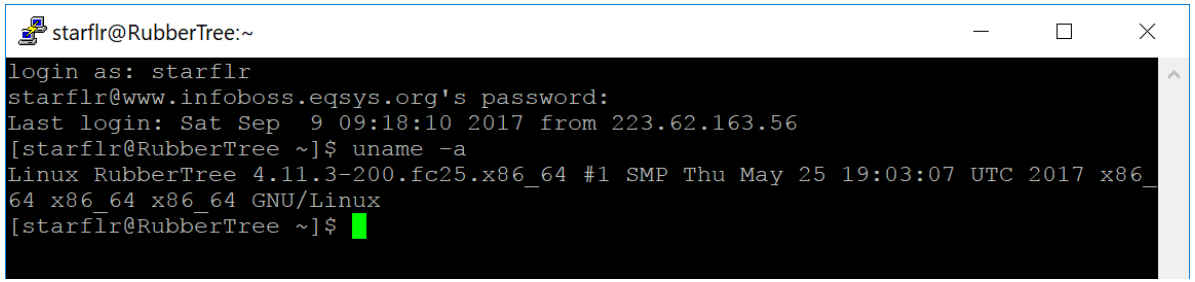
- Operating system is Linux (=UNIX) not Windows.



- 'Putty' is program to access server.



IP Address / Port
Account (ID / Pasword)



Thank you for your attention!

If you have question,
please ask! =)

starflr@infoboss.co.kr

InfoBoss
Infomations & Systems