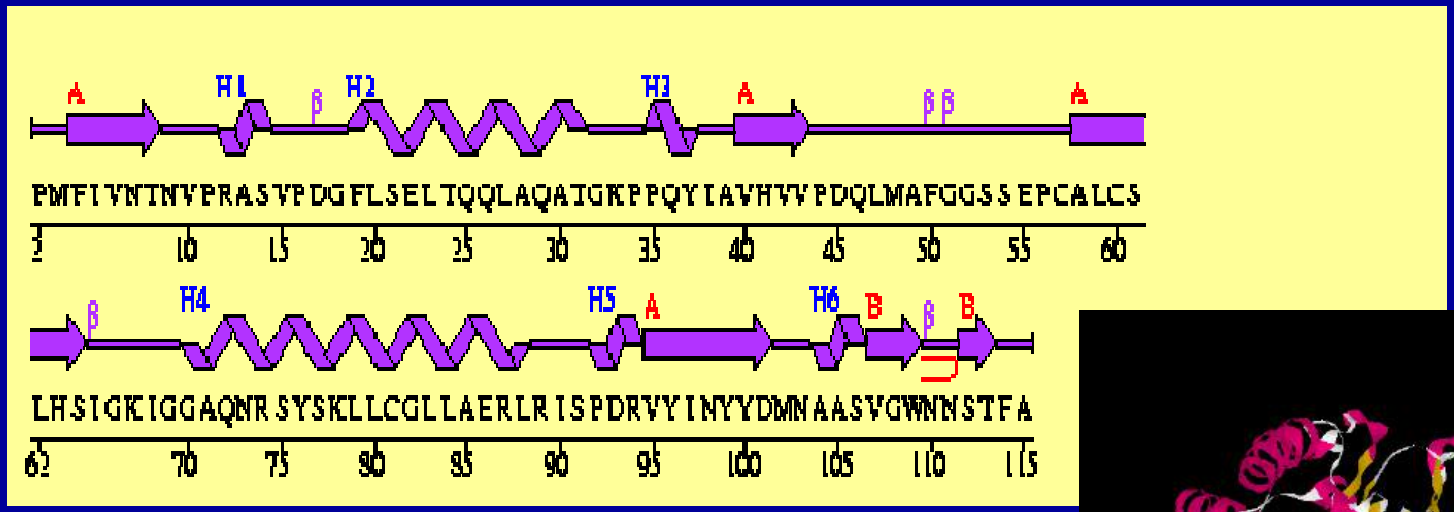
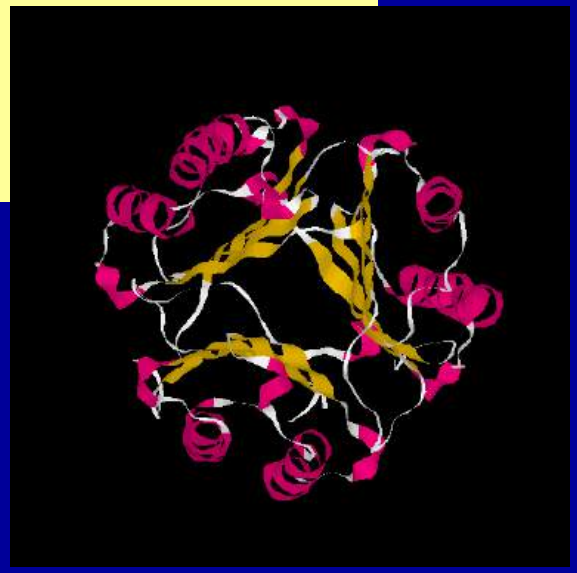
M.Sc. Bioinformatics

UNIT - I

Introduction to bioinformatics





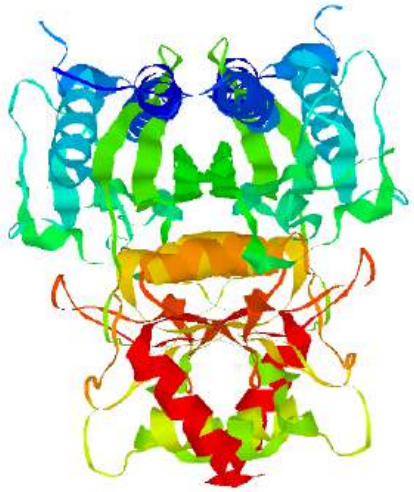
What is bioinformatics?

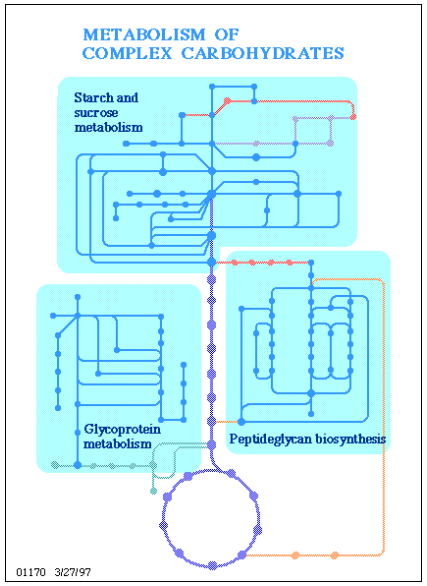
- an emerging interdisciplinary research area
- deals with the computational management and analysis of biological information: genes, genomes, proteins, cells, ecological systems, medical information, robots, artificial intelligence...

The Core of Bioinformatics to date

- **Relationships between**

TDQAAFDTNIVTLTRFVMEQ
GRKARGTGEMTQLLNSLCTA
VKAISTAVRKAGIAHLYGIA
GSTNVTGDQVKKLDVLSNDL
VINVLKSSFATCVLVTEEDK
NAIIVEPEKRGKYVVCFDPL
DGSSNIDCLVSIGTIFGIYR
KNSTDEPSEKDALQPGRNLV
AAGYALYGSATMLV





sequence **3D structure** **protein functions**

- **Properties and evolution of genes, genomes, proteins, metabolic pathways in cells**

- **Use of this knowledge for prediction, modelling, and design**

“The holy grail of bioinformatics”

GCTCCTCACTGTCTGTGTTTATTC
TTTTAGCTTCTTCAGATCTTTTAG
TCTGAGGAAGCCTGGCATGTGCA
AATGAAGTTAACCTAA...

> 500, 000 genes sequenced to date

Expected number of unique protein structures:

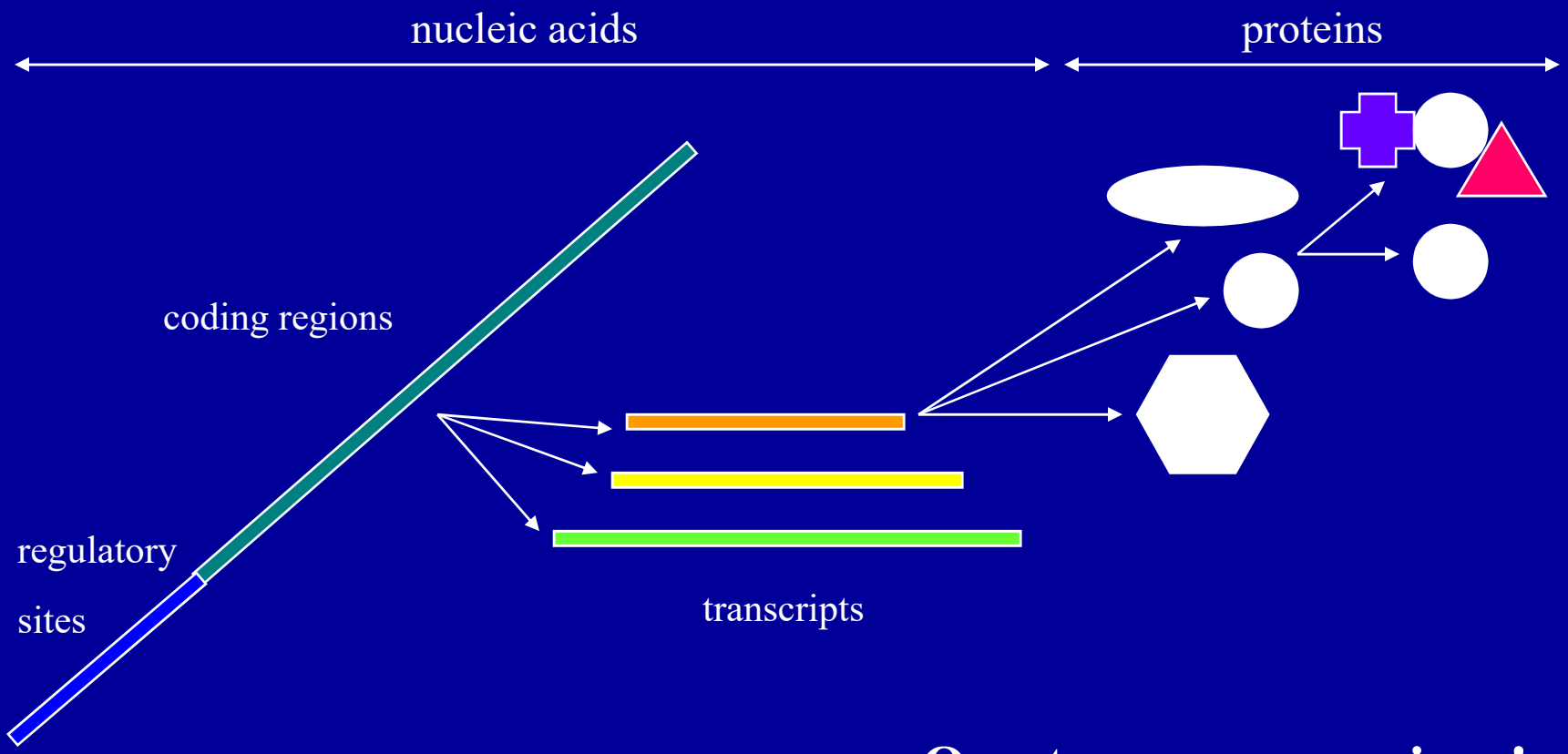
~ 700-1, 000

Basic concepts

- conceptual foundations of bioinformatics:
  evolution
  protein folding
  protein function

- bioinformatics builds mathematical models of these processes -
  to infer relationships between components of complex biological systems

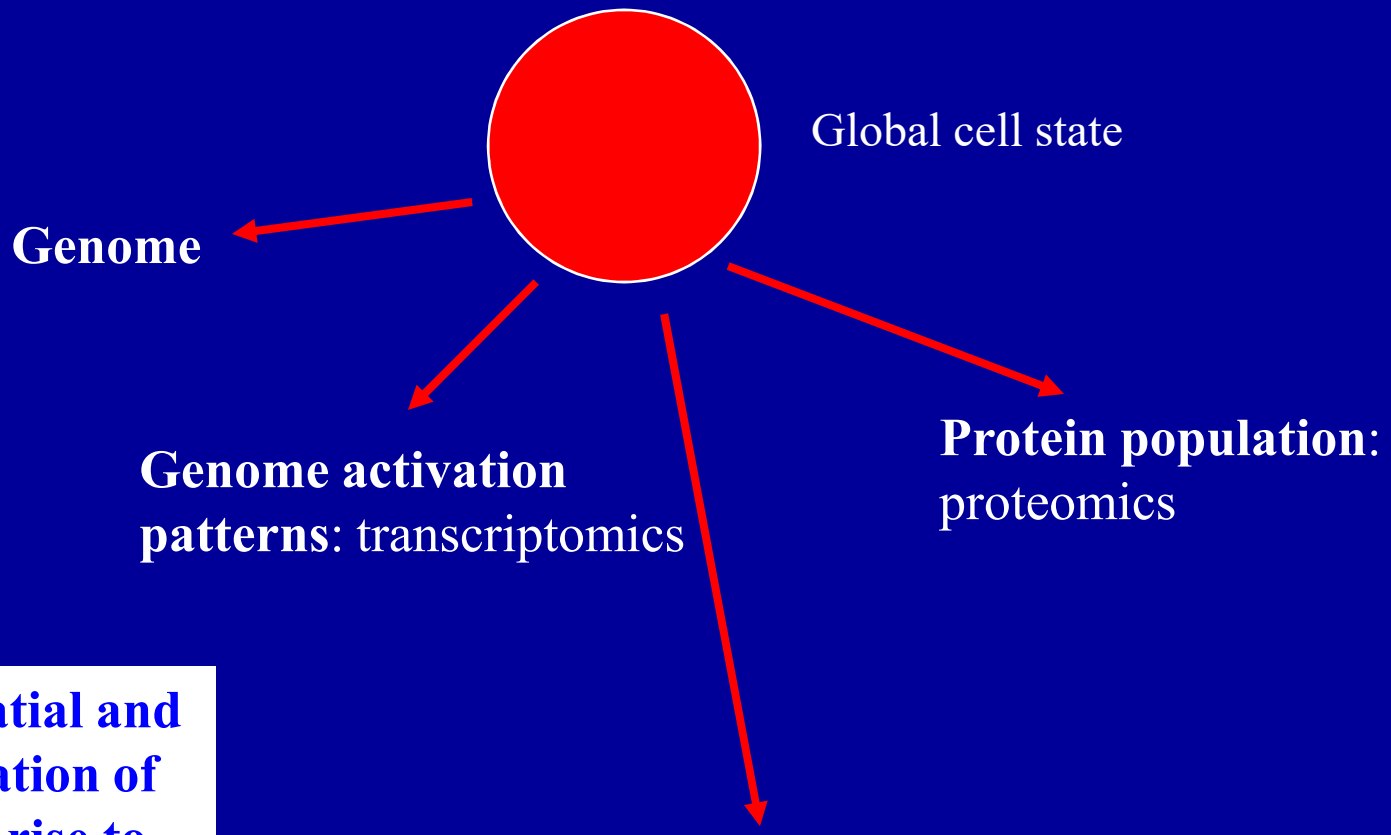
Information processing in cells



One-to-many mappings!

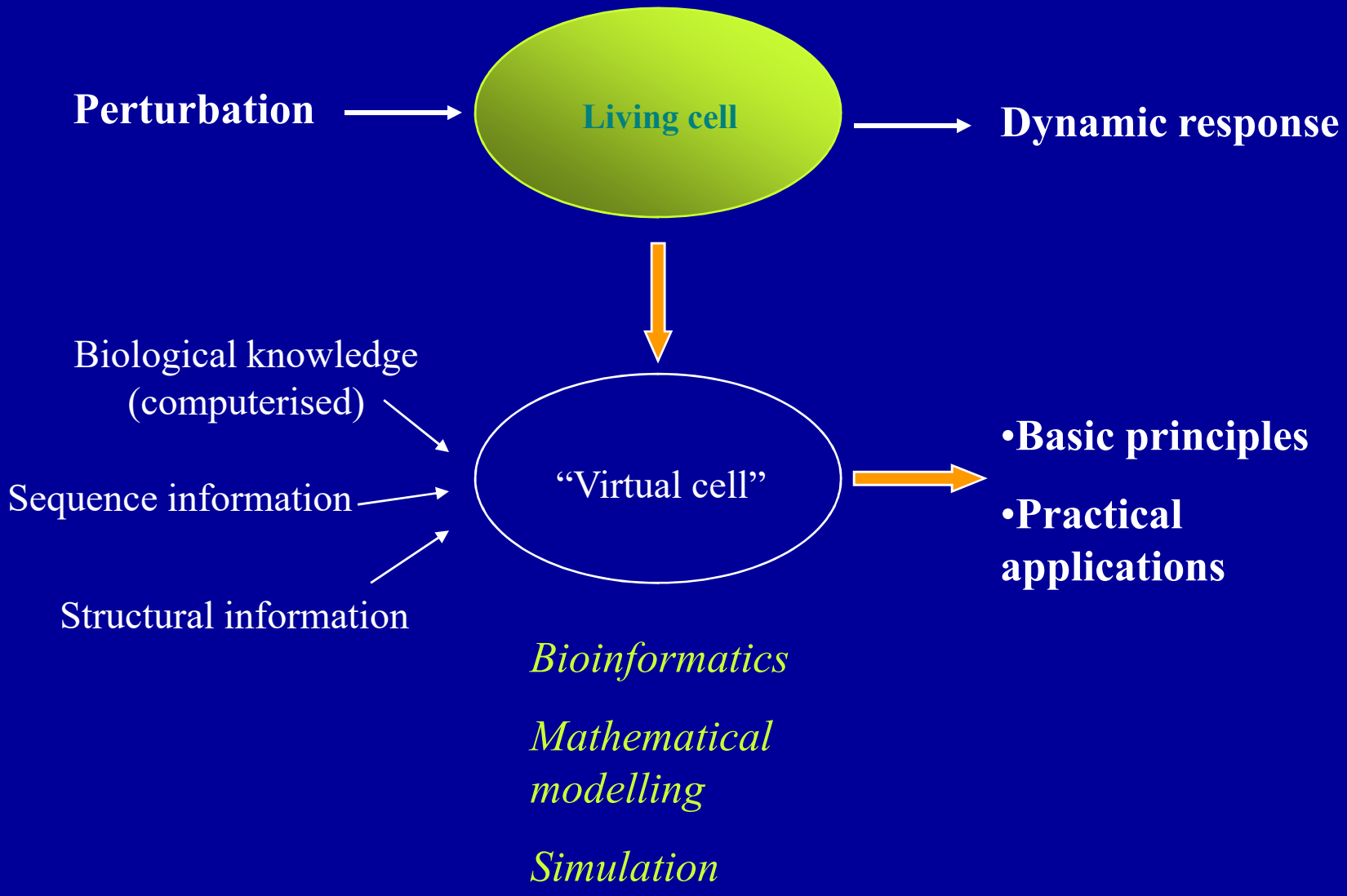
Context-dependence!

Global approaches: Toward a new Systems Biology

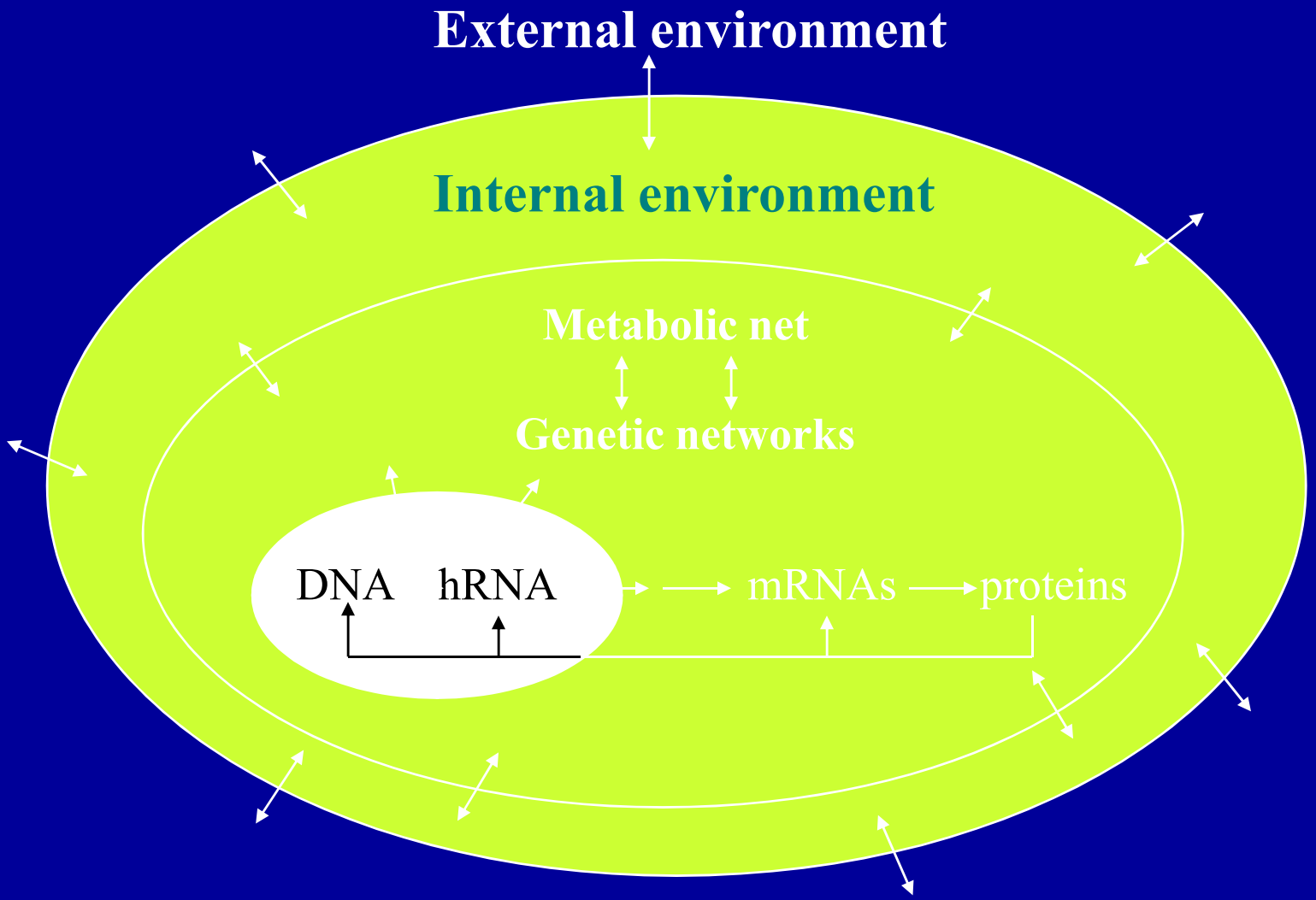


•How does the spatial and temporal organisation of living matter give rise to biological processes?

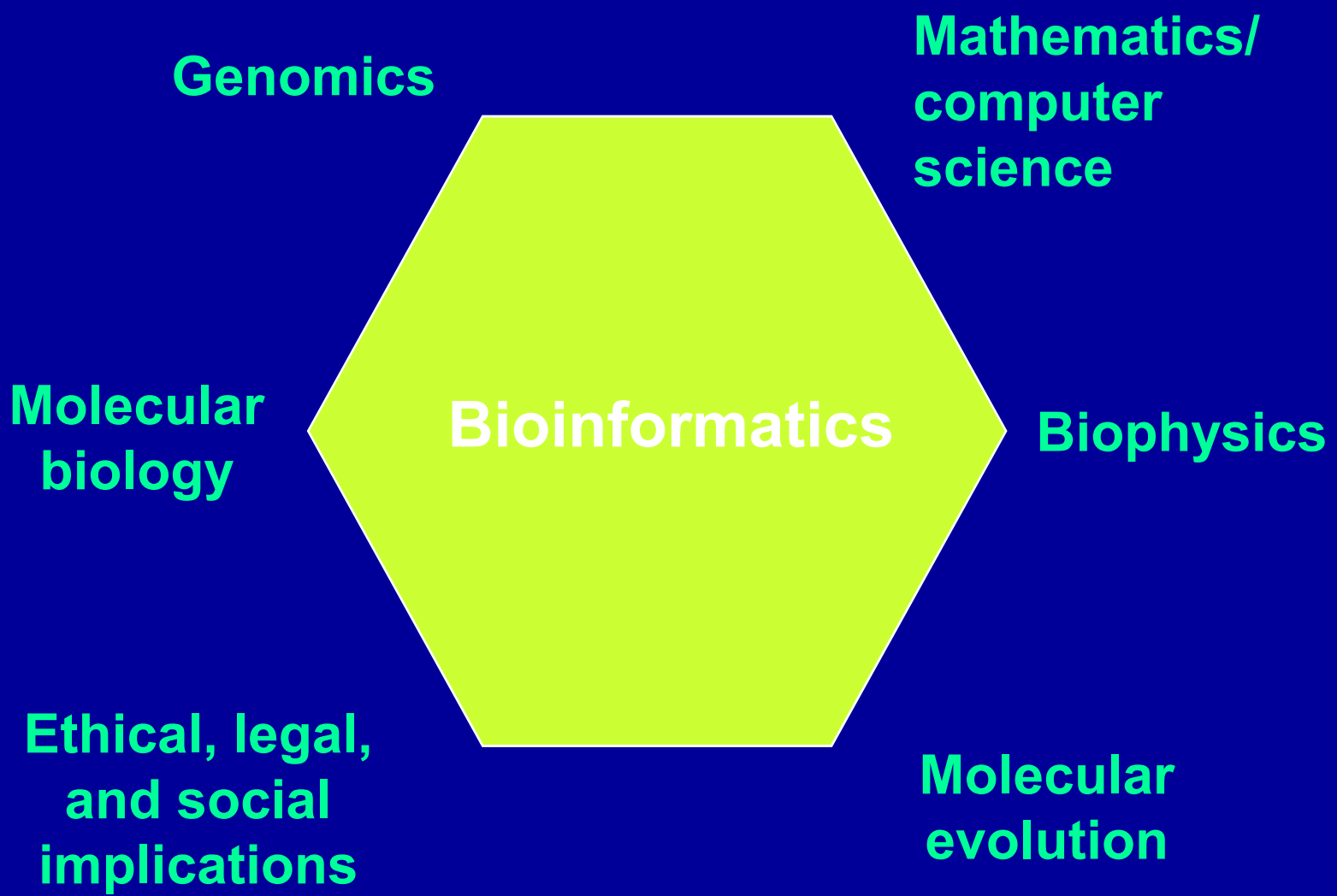
Global approaches: Toward a new Systems Biology



We do not know yet whether the information in the genome is sufficient to reconstruct an entire biological system. Information on building blocks not enough, information on their interactions is essential.



Bioinformatics in context



Current challenges to users

- **Potential hurdles:**

  Methods are in flux and not fully developed-scattered and heterogeneous resources

- **Remedies: Web resources**

  navigation guides
  integration of tools and databanks

http://www.biochem.ucl.ac.uk/~nagl/bioinformatics.html

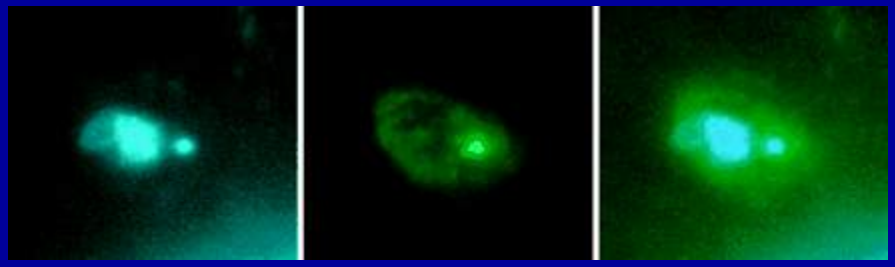
Sequence homology search of the genome of *Plasmodium falciparum*

Target identification for antimalerial drugs

The search for new antimalarial drugs

- Malaria is one of the leading causes of morbidity and mortality in the tropics.
- 300 to 500 million estimated clinical cases and 1.5 million to 2.7 million deaths per year.
- Nearly all fatal cases are caused by *Plasmodium falciparum.*
- The parasite's resistance to conventional antimalarial drugs such as chloroquine is growing at an alarming rate.

- ***P. falciparum*** **has a plastidlike organelle, called the apicoplast, acquired by endosymbiosis of an alga.**



Jomaa et al. (1999)

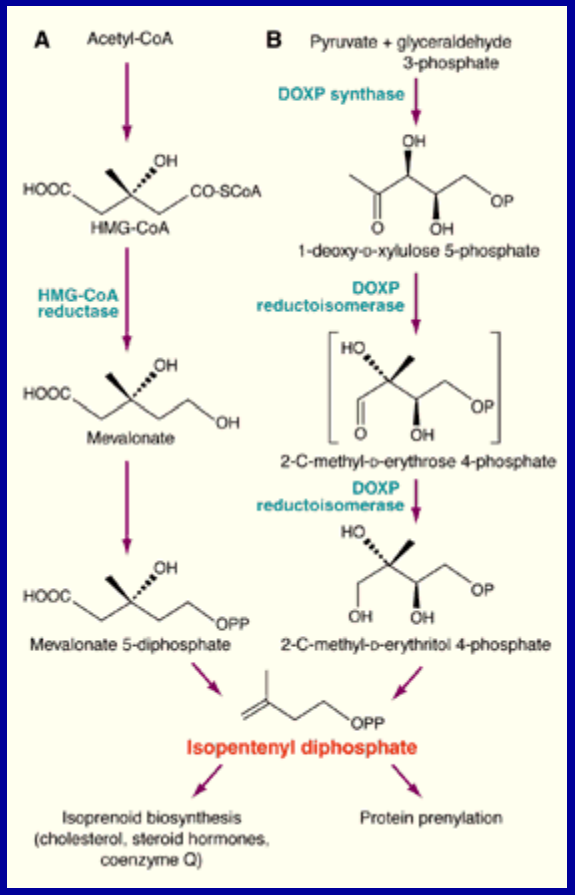
- **Self-replicating, maternally inherited (35kb, circular DNA).**
- **Comparative genome analysis: Search for orthologs.**

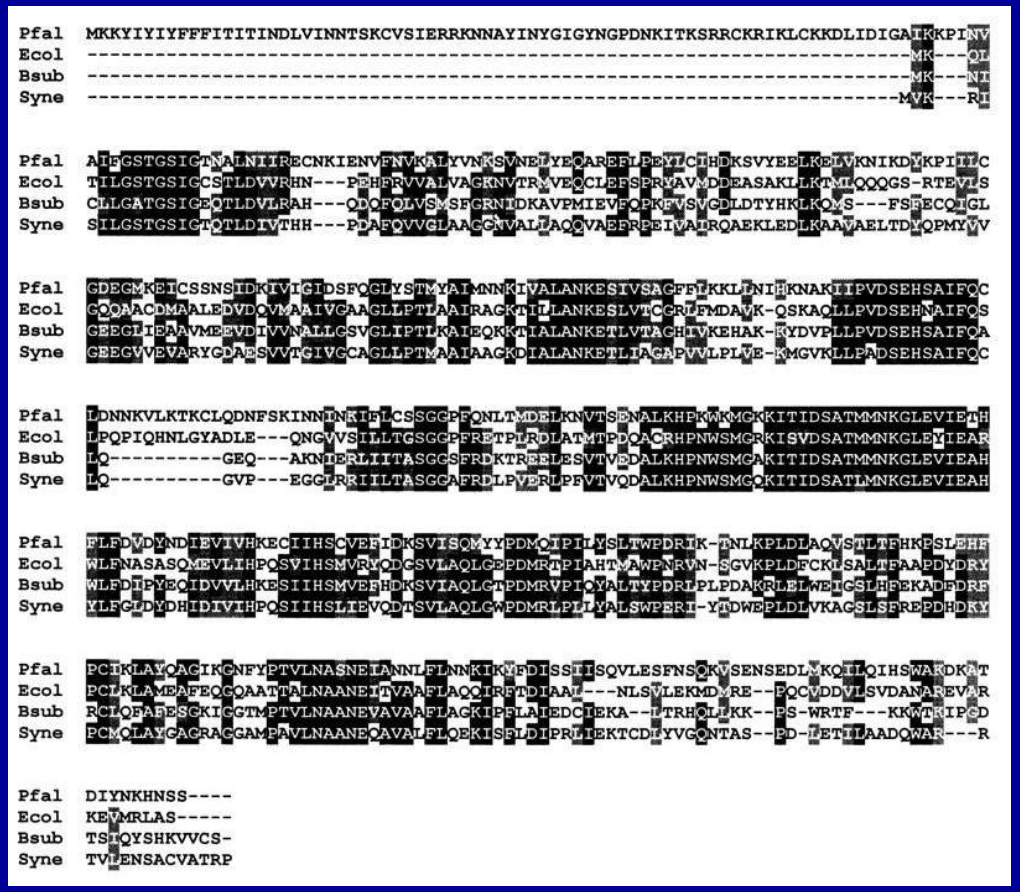
Apicoplast contains enzymes found in plant and bacterial, but not animal metabolic pathways.

- **Potential target for antimalerial drugs:**

DOXP reductoisomerase

Jomaa *et al*. (1999) Science 285: 1573-1576:





An Introduction to Linux Operating System

Zihui Han

Content

- What is Linux
- Why Linux
- The Linux system
- Security
- OS Comparison

What is Linux

- Linux is a true 32 bit UNIX-like OS developed originally for home PCs, but now it runs on a variety of platforms including PowerPC, Macintosh, Amiga, DEC Alpha, Sun Sparc, ARM, and many others. The source code for Linux is freely available to everyone. Linux was created by Linus Torvalds in 1991, and it has been developed with the help of many programmers across the Internet. Now it has evolved into a very functional, powerful and usable clone of Unix which has at least 10 million users worldwide.

Why Linux

- A Linux Distribution has thousands of dollars worth of software for no cost.
- Linux is a complete operating system:
  - stable - the crash of an application is much less likely to bring down the OS under Linux.
  - Reliable - Linux servers are often up for hundreds of days compared with the regular reboots required with a Windows system.
  - extremely powerful
- Linux provides a complete development environment.

Why Linux (continued)

- Excellent networking facilities
- Ideal environment to run servers such as a web server, or an ftp server.
- A wide variety of commercial software is available if not satisfied by the free software
- Easily upgradeable.
- Supports multiple processors.
- True multi-tasking, multi-user OS.
- An excellent window system called X, the equivalent of Windows but much more flexible.
- Full source code is provided and free.

The Linux System

- The Linux system excel in many area, ranging from end user concerns such as stability, speed, ease of use, to serious concerns such as development and networking.
  - Linux kernel
  - Linux networking
  - Linux file system

Linux Kernel

- The kernel is the central nervous system of Linux, include OS code which runs the whole computer. It provides resources to all other programs that you run under Linux, and manages all other programs as they run.
  - The kernel includes the code that performs certain specialized tasks, including TCP/IP networking.
  - The kernel design is modular, so that the actual OS code is very small to be able to load when it needs, and then free the memory afterwards, thus the kernel remains small and fast and highly extensible.

Linux Networking

- Networking comes naturally to Linux. In a real sense, Linux is a product of the Internet or World Wide Web (www).
- Linux is made for networking. Probably all networking protocols in use on the Internet are native to Unix and/or Linux. A large part of the Web is running on Linux boxes, e.g. : AOL

Networking protocols

- The Linux kernel supports several networking protocols:
  - TCP/IP - Transport Control Protocol/Internet Protocol
    - IP is the primary network protocol supported by Linux
  - IPX - Internetwork Packet Exchange
  - Applettalk DDP
  - Amateur Radio AX.25 Level 2

Supported Features

- Forwarding
- Firewall operations
- Proxy and Masquerading
- Accounting
- Tunneling and Intranets
- Aliasing

Linux File System

- Linux has an hierarchical, unified file system
- Supports 256-character filenames.
- All command line entries are case sensitive.
- Use the slash(/) rather than the backslash(\) you have been using in DOS.

Types of File

- Ordinary files
  - text files
  - data files
  - command text files
  - executable files
- directories
- links
  - rather than having multiple copies of a file, Linux uses linking to one file to save disk space.
- special device files

Security

- Encryption
- Secure shell(ssh)
- Principles of security

Encryption

- Encryption commonly used to secure data. It is the ancient technique of hiding information in plain sight. Include:
  - strong encryption - is stronger than the 40-bit encryption maximum that can be exported from the United States under U.S. law.
  - Public-key Encryption - is a type of asymmetric encryption, which is a system that you encrypt your message with one key, and the recipient decrypts it with a mathematically related, but different key.

The Secure Shell(ssh)

- The ssh and its tools use strong encryption to allow remotely located systems to exchange data securely.
- By using strong encryption, ssh significantly enhances the security of both the authentication process and the session itself.

Principles of security

- Two broad categories of attack exist:
  - unauthorized access
  - denial of service
- Defense against the attacks:
  - enforce the use of password
  - use TCP wrappers to limit which resources are made available to which categories of users.
  - monitor internal users, protect your organization against unauthorized or inappropriate use of the computer facilities to harass personnel.

Linux vs. Unix

- Linux is free, but Unix is not.
- Unix is compatible with Linux at the system call level, meaning most programs written for either Unix or Linux can be recompiled to run on the other system with a minimum of work. But Linux will run faster than Unix on the same hardware.

Linux vs. Microsoft Windows

- Both offer some of the graphics capabilities and include some networking capabilities. But Linux networking is excellent.
- Linux is multi-user, multi-tasking, but Microsoft Windows doesn’t support it.

Linux vs. Windows NT

- Linux needs 2MB RAM to try out, while NT needs 12 MB
- Linux needs at least 15 MB disk space, while NT needs 70 MB at least.
- Both system support multitasking
- Both system support multiprocessing.
- Both system support dynamic cache.
- Linux has full multi user support. Local users, modem users, and network users can all simultaneously run text and graphics programs. This is a powerful feature for business environments that is unmatched by NT.

Linux vs. Windows NT(continued)

- The issue of size is a great strength for Linux. It was designed to be as small and efficient as possible. NT's most important criterion was portability.
- Linux was built on the Internet, and hence has better support for networking then NT.
- Most software packages that run on Linux have their source code available, security problems are found and solved many times quicker then with NT.

Reference

- http://www.Linux.org
- http://www.croftj.net/~jam
- http://metalab.unc.edu/LDP/HOWTO/NET-3-HOWTO-4.html#ss4.1

Learning Unix/Linux

Bioinformatics Orientation 2008

Eric Bishop

Introduction: What is Unix?

- An operating system
- Developed at AT&T Bell Labs in the 1960's
- Command Line Interpreter
- GUIs (Window systems) are now available

Introduction: Unix vs. Linux

- Unix was the predecessor of Linux
- Linux is a variant of Unix
  - So is Mac OS X, so much of this tutorial applies to Macs as well
- Linux is open source
- Most of the machines you'll use in the Bioinformatics program are running the Linux OS

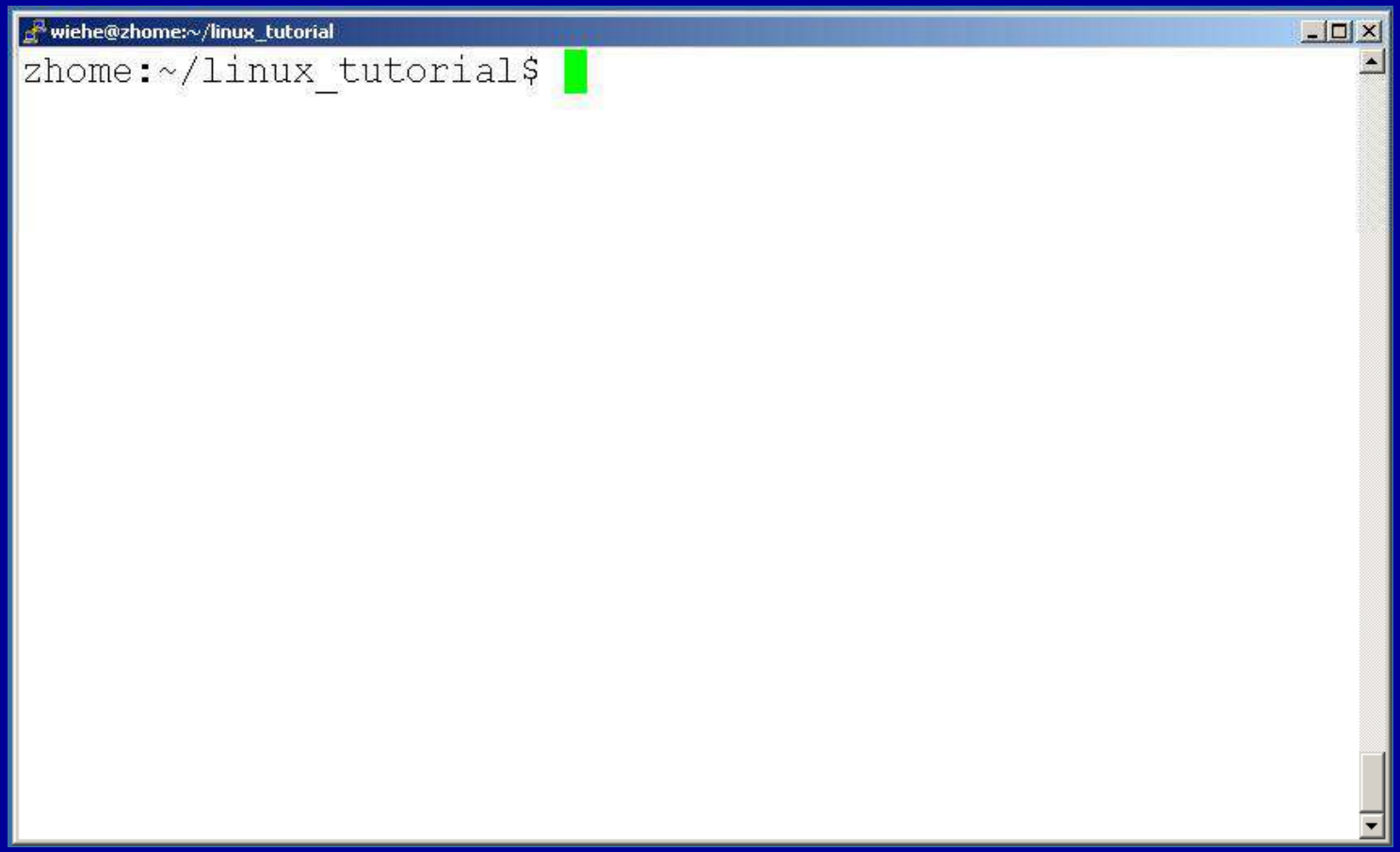
Introduction: Why Unix/Linux?

- Linux is free
- It's fully customizable
- It's stable (i.e. it almost never crashes)

- These characteristics make it an ideal OS for programmers and scientists

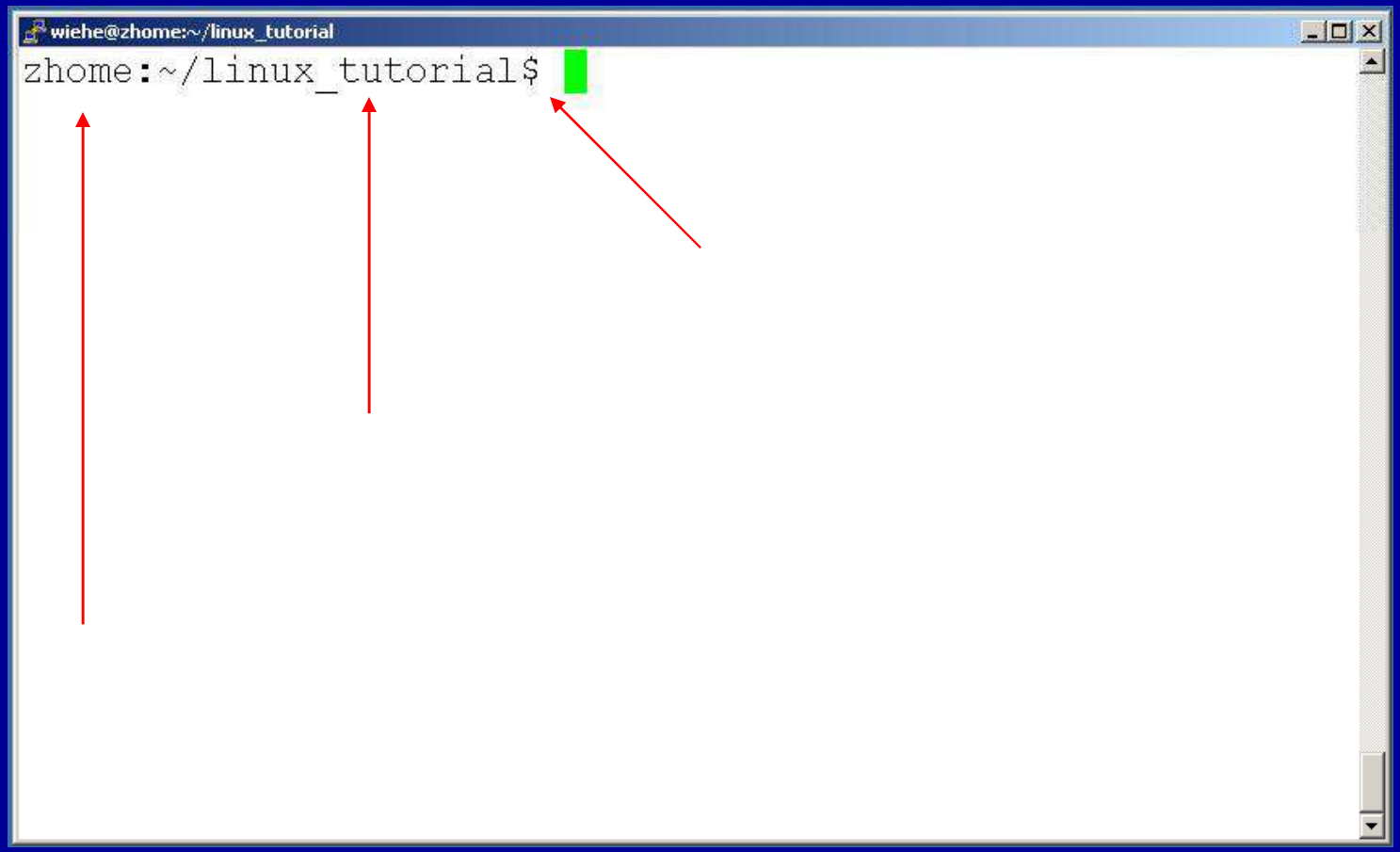
Connecting to a Unix/Linux system

- Open up a terminal:



Connecting to a Unix/Linux system

- Open up a terminal:



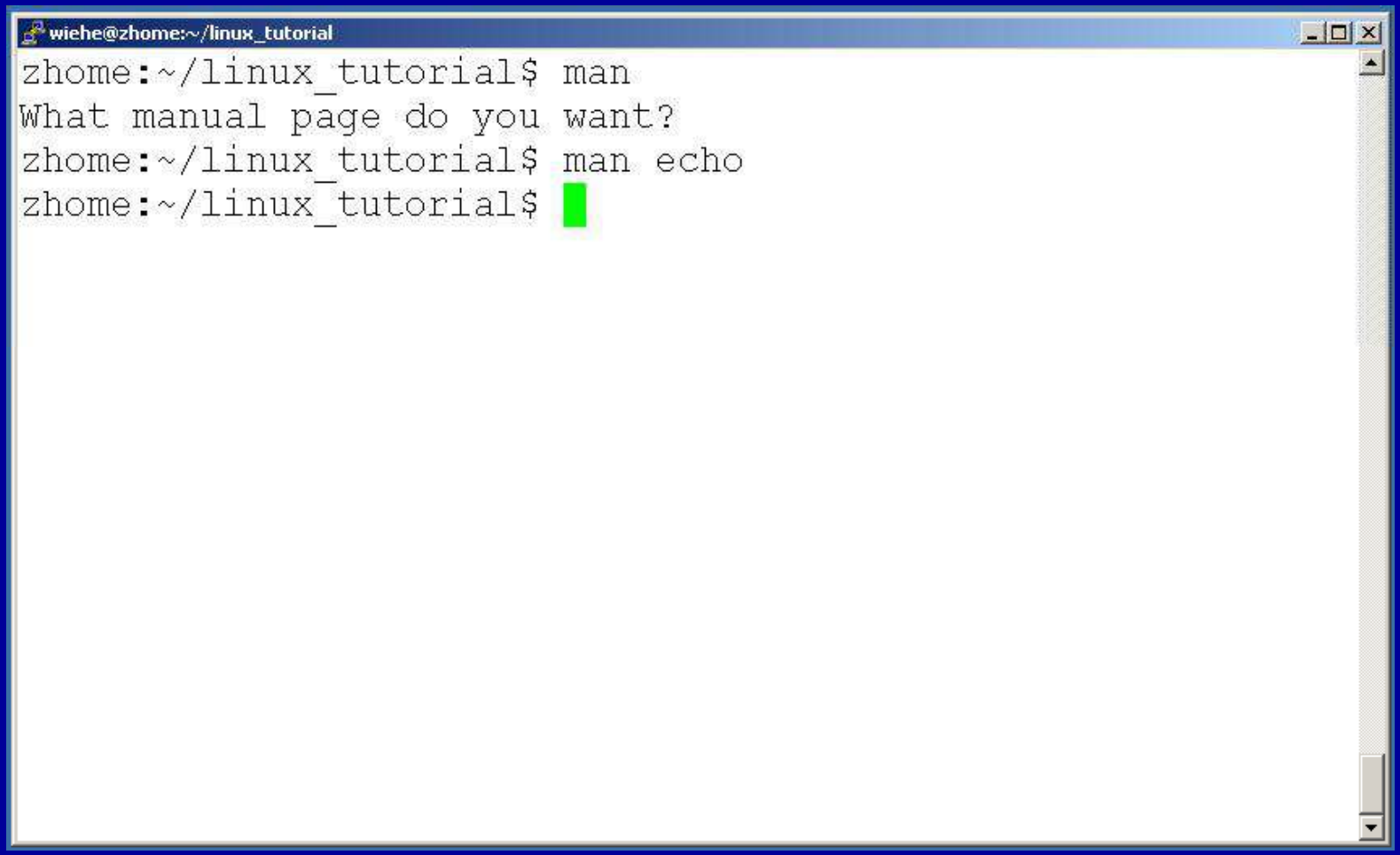
What exactly is a “shell”?

- After logging in, Linux/Unix starts another program called the **shell**
- The shell interprets commands the user types and manages their execution
  - The shell communicates with the internal part of the operating system called the **kernel**
  - The most popular shells are: tcsh, csh, korn, and bash
  - The differences are most times subtle
  - For this tutorial, we are using bash
- Shell commands are **CASE SENSITIVE!**

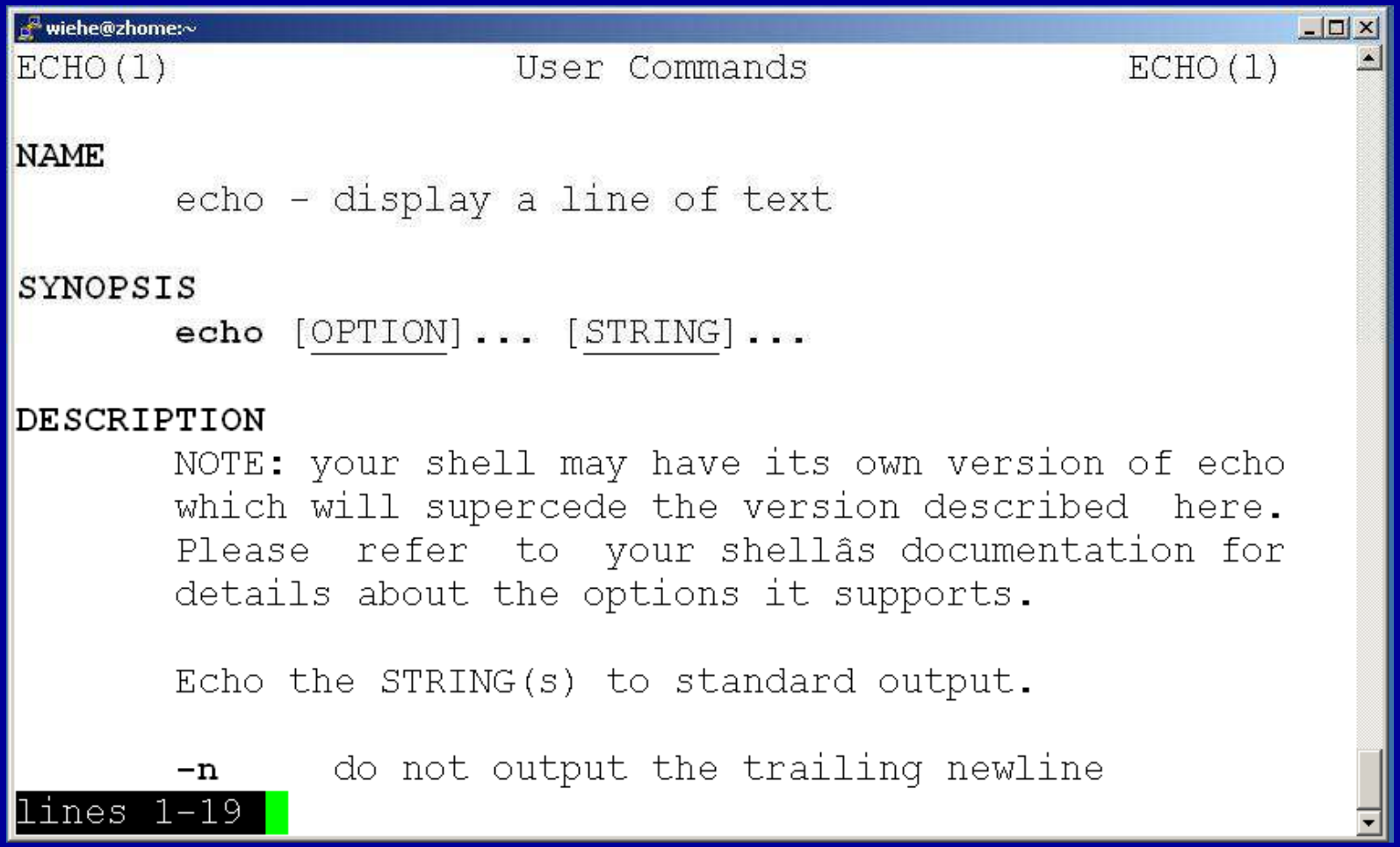
Help!

- Whenever you need help with a command type “man” and the command name

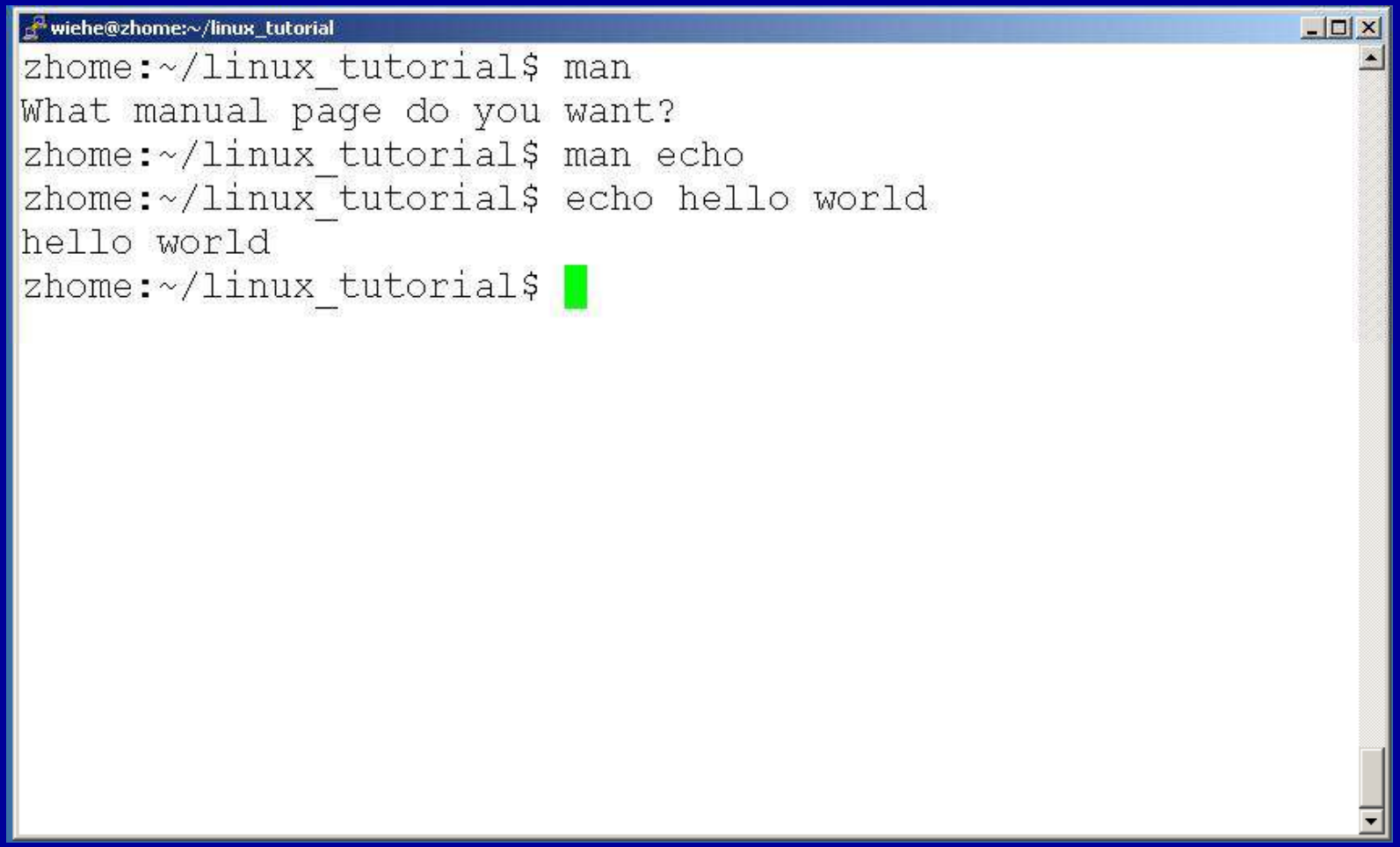
Help!


wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ man
What manual page do you want?
zhome:~/linux_tutorial$ man echo
zhome:~/linux_tutorial$

Help!

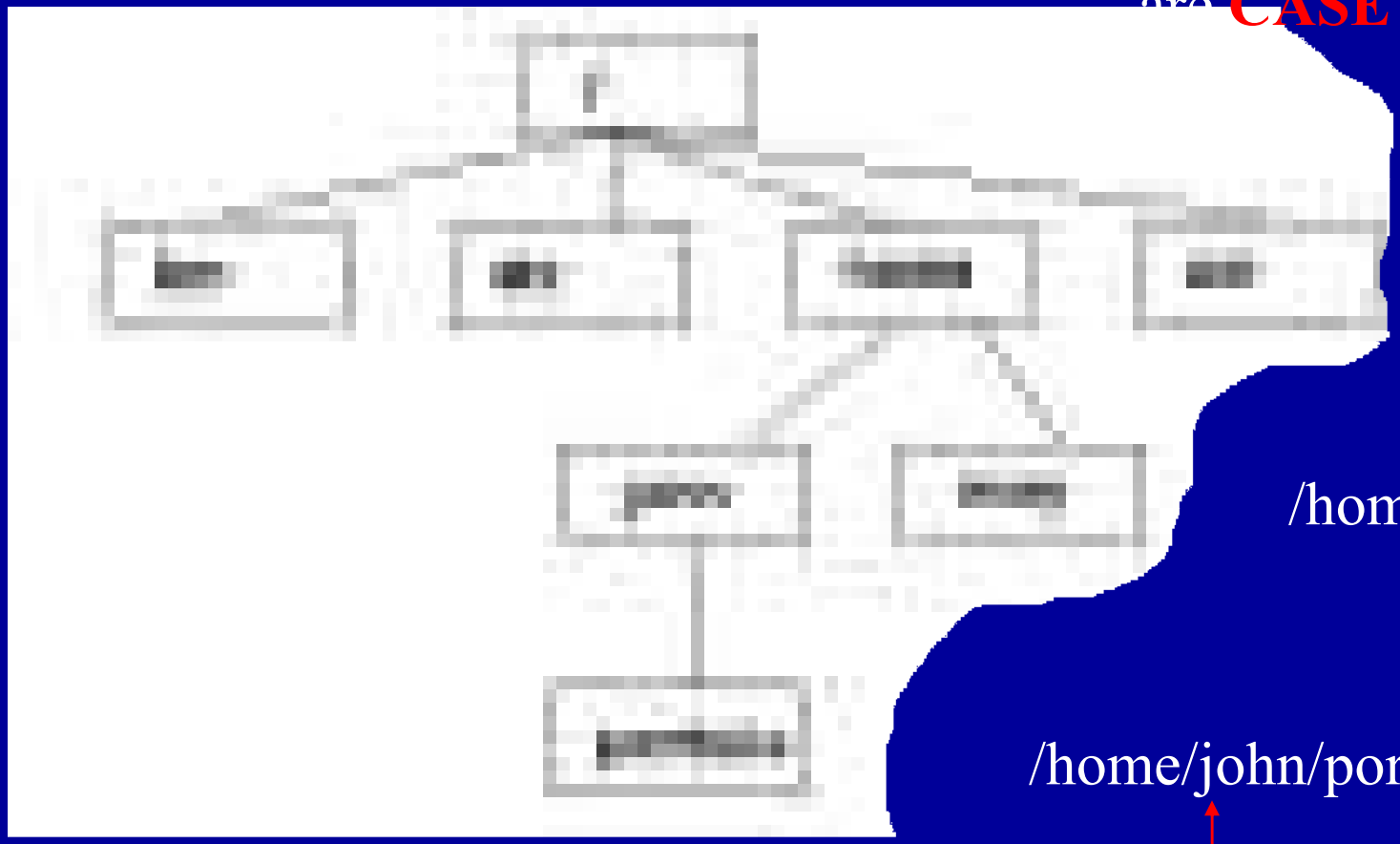

wiehe@zhome:~
ECHO(1) User Commands ECHO(1)
NAME
echo - display a line of text
SYNOPSIS
echo [OPTION]... [STRING]...
DESCRIPTION
NOTE: your shell may have its own version of echo which will supercede the version described here. Please refer to your shellâs documentation for details about the options it supports.
Echo the STRING(s) to standard output.
-n do not output the trailing newline
lines 1-19

Help!


wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ man
What manual page do you want?
zhome:~/linux_tutorial$ man echo
zhome:~/linux_tutorial$ echo hello world
hello world
zhome:~/linux_tutorial$

Unix/Linux File System

NOTE: Unix file names are **CASE SENSITIVE!**

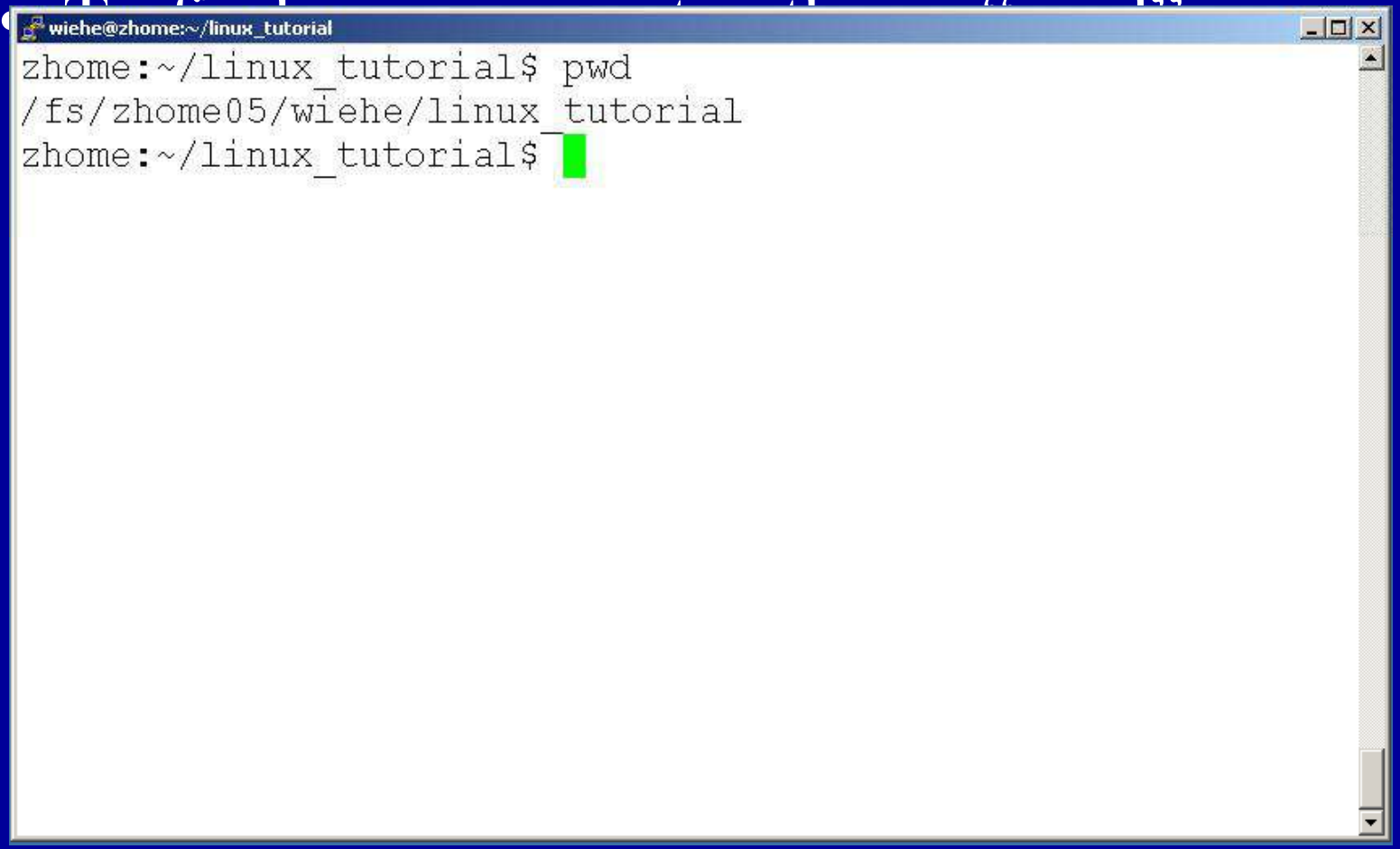


/home/mary/

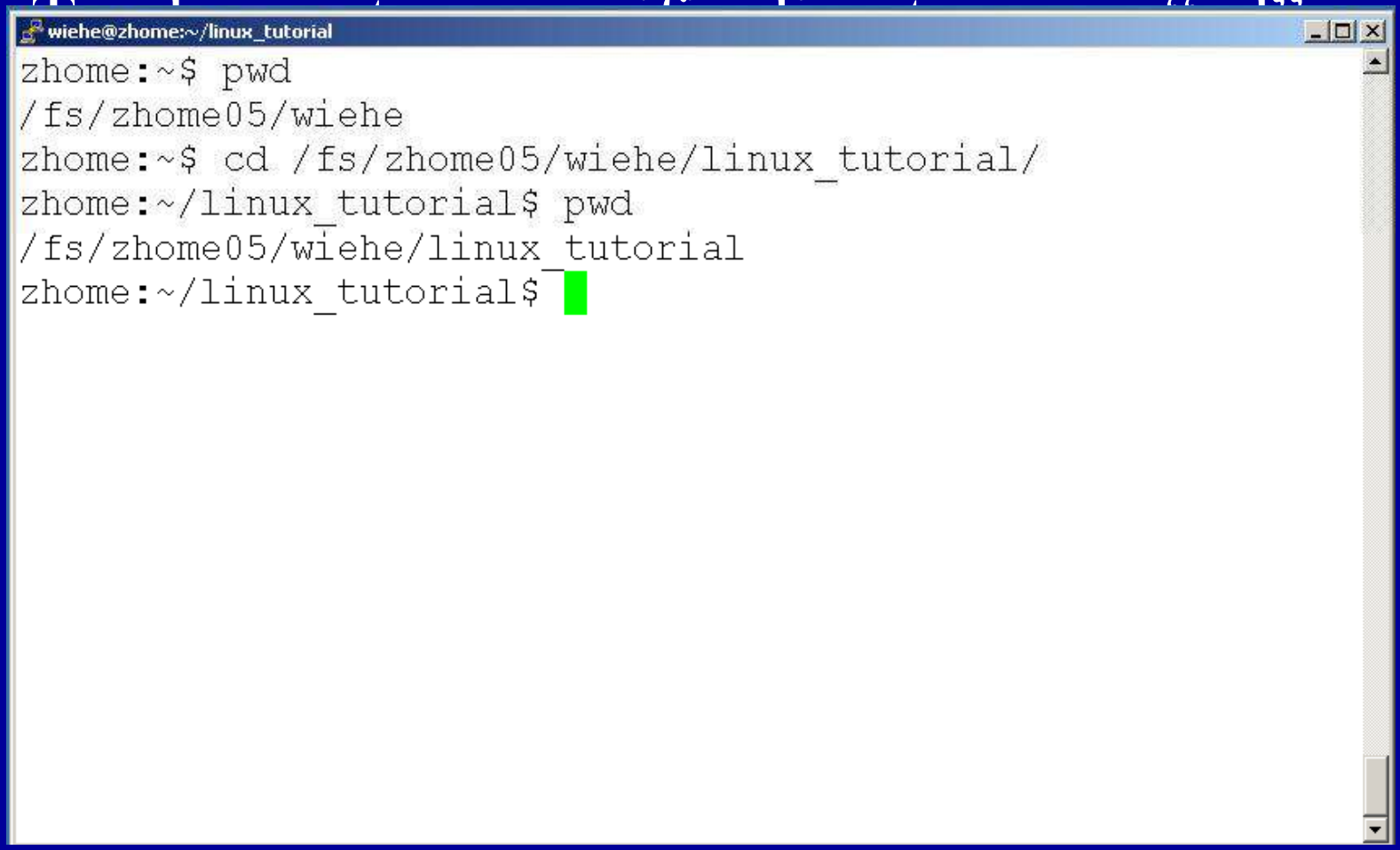
/home/john/portfolio/

The Path

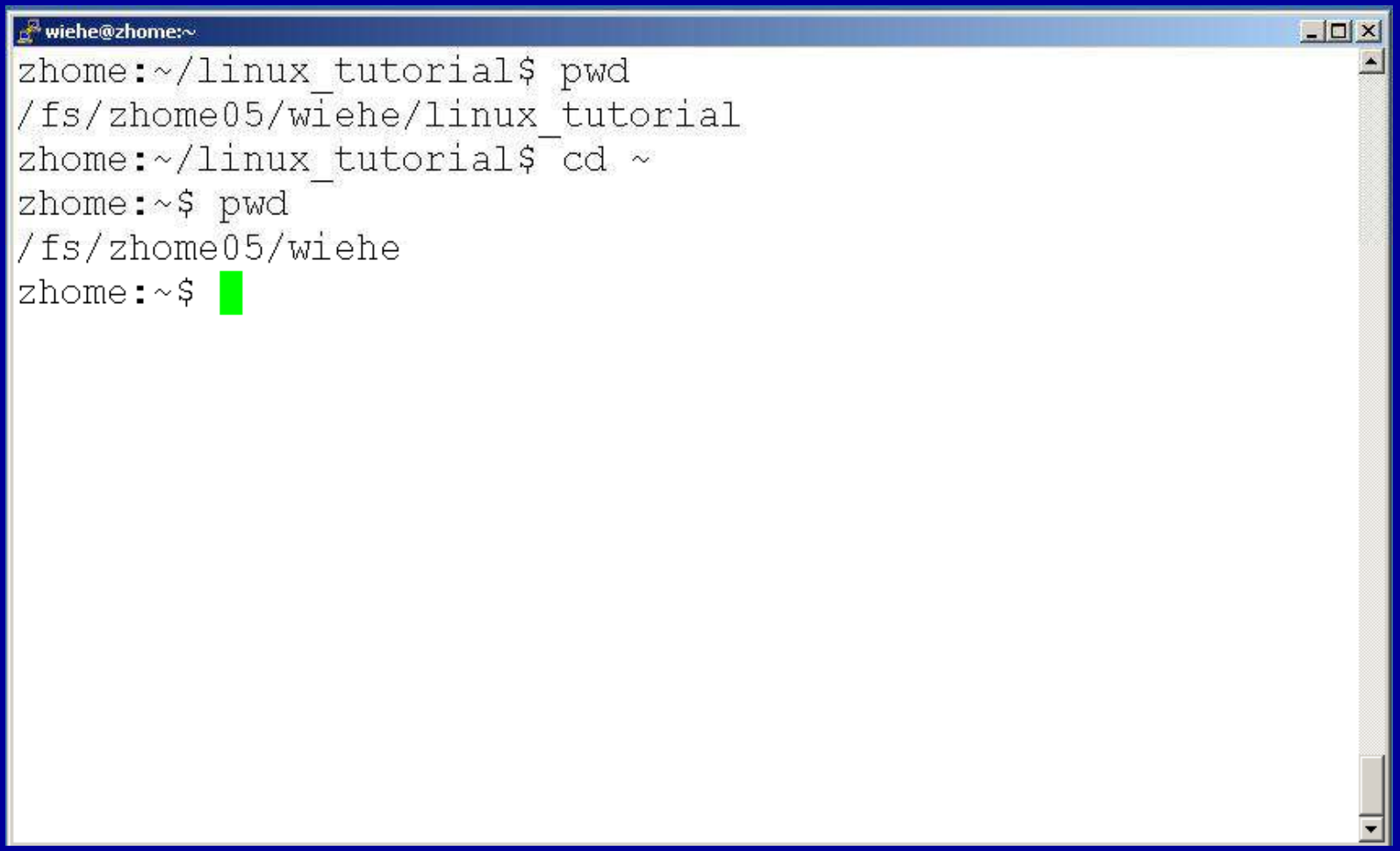
Command: pwd


wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ pwd
/fs/zhome05/wiehe/linux_tutorial
zhome:~/linux_tutorial$

Command: cd

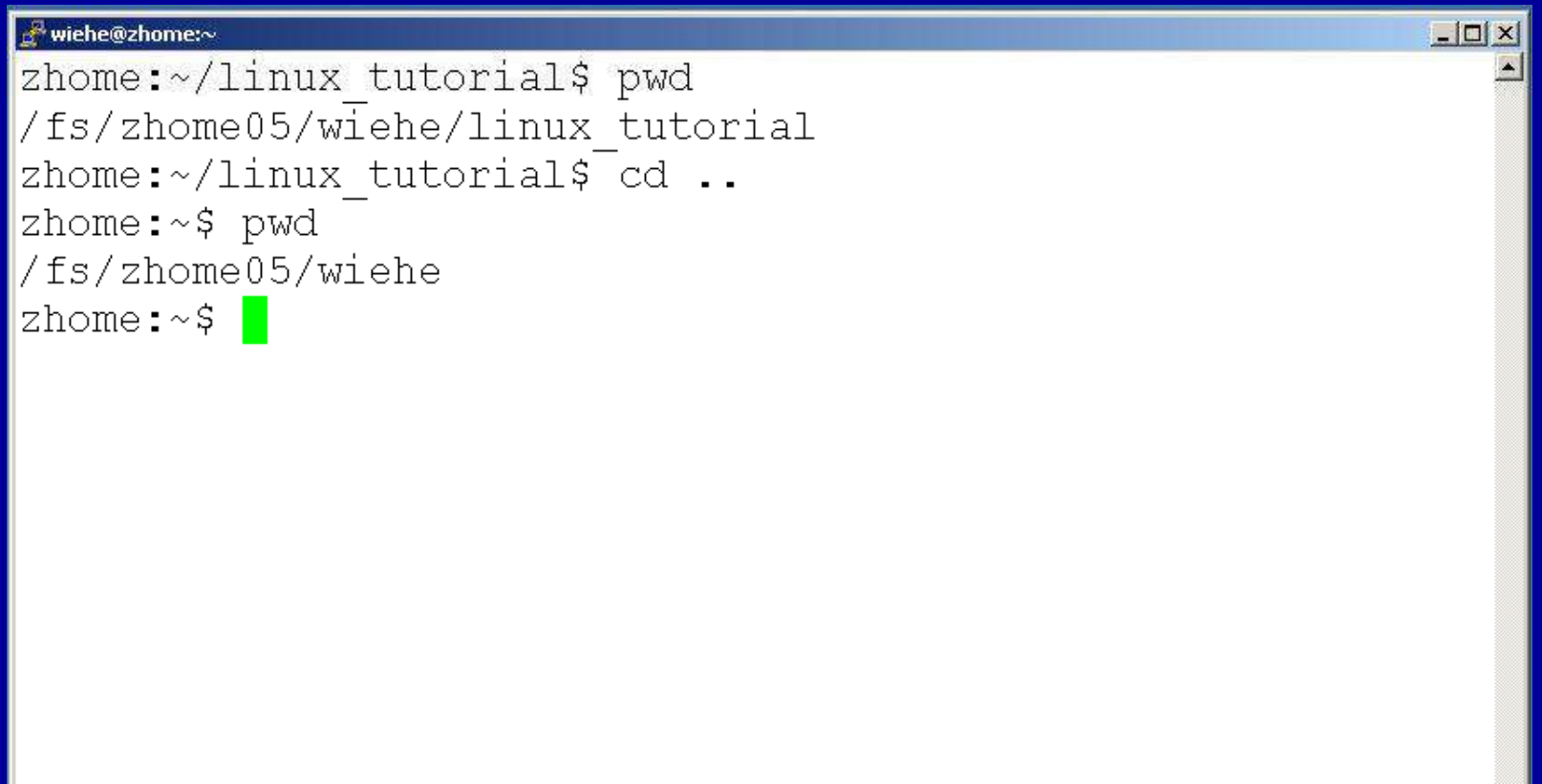
- 
wiehe@zhome:~/linux_tutorial
zhome:~$ pwd
/fs/zhome05/wiehe
zhome:~$ cd /fs/zhome05/wiehe/linux_tutorial/
zhome:~/linux_tutorial$ pwd
/fs/zhome05/wiehe/linux_tutorial
zhome:~/linux_tutorial$

Command: cd

- 
wiehe@zhome:~
zhome:~/linux_tutorial$ pwd
/fs/zhome05/wiehe/linux_tutorial
zhome:~/linux_tutorial$ cd ~
zhome:~$ pwd
/fs/zhome05/wiehe
zhome:~$

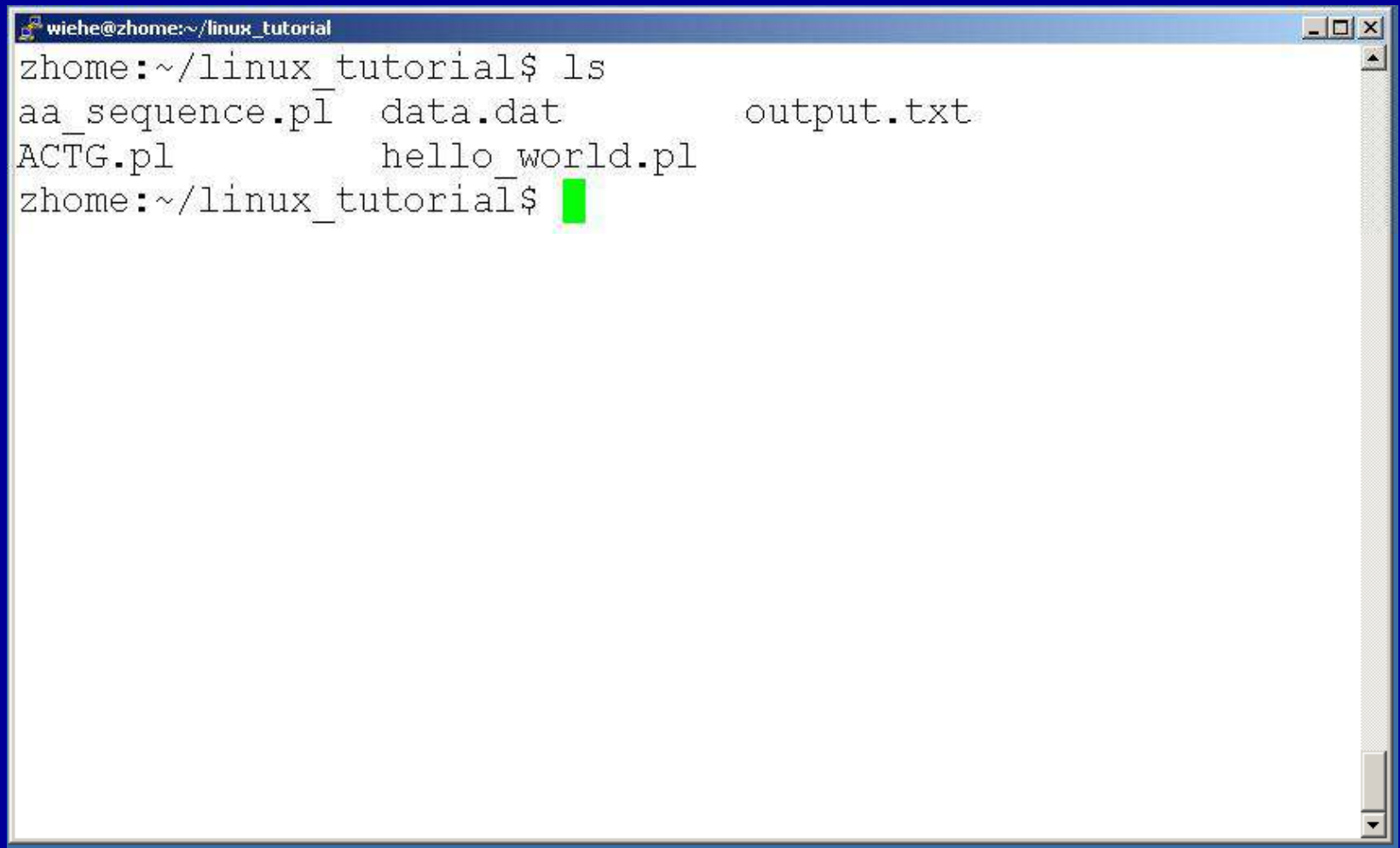
Command: cd

- ".." is the location of the directory below current one



Command: ls

- To list the files in the current directory use “ls”

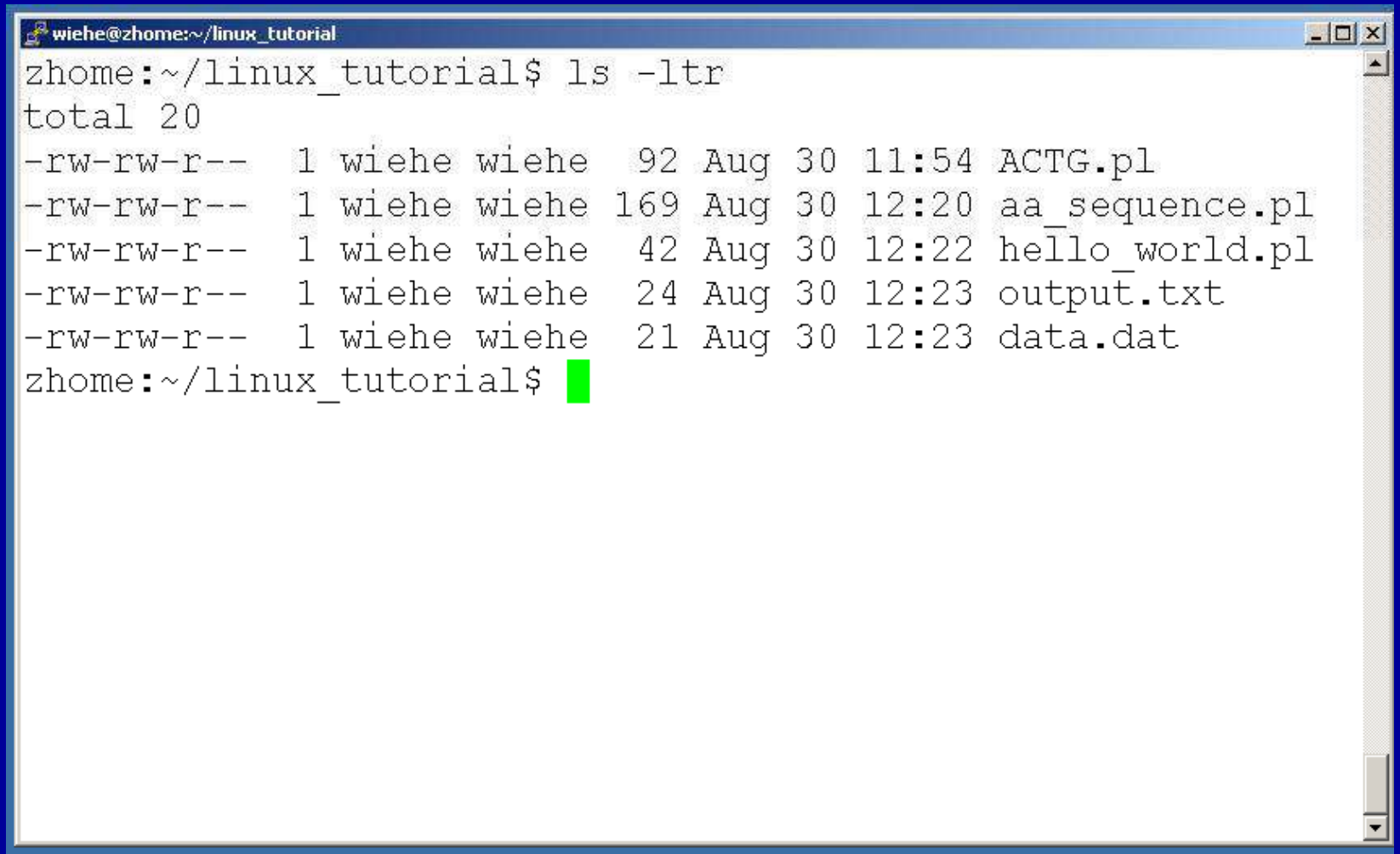


Command: ls

- ls has many options
  - -l long list (displays lots of info)
  - -t sort by modification time
  - -S sort by size
  - -h list file sizes in human readable format
  - -r reverse the order
- “man ls” for more options
- Options can be combined: “ls -ltr”

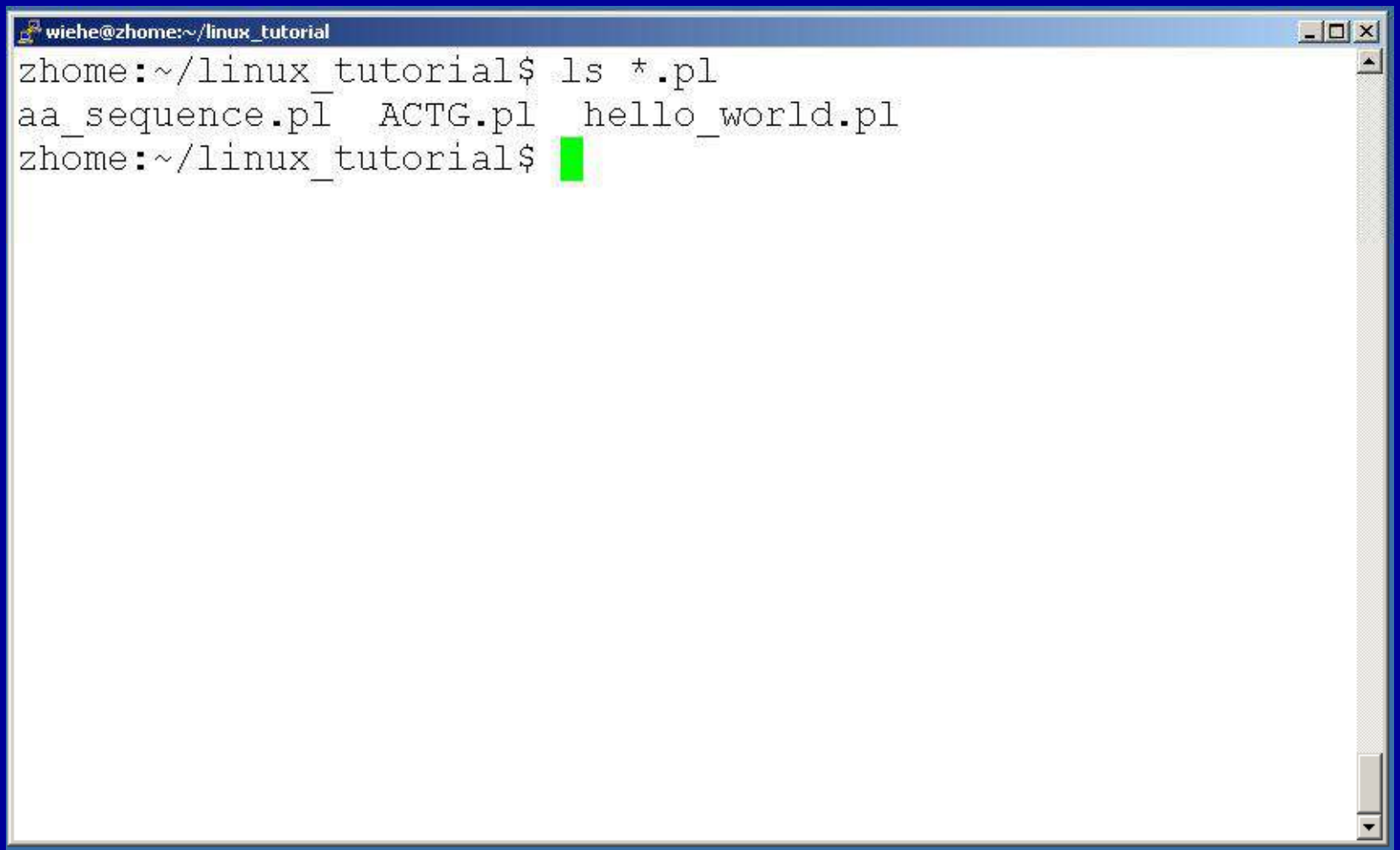
Command: ls -ltr

- List files by time in reverse order with long listing



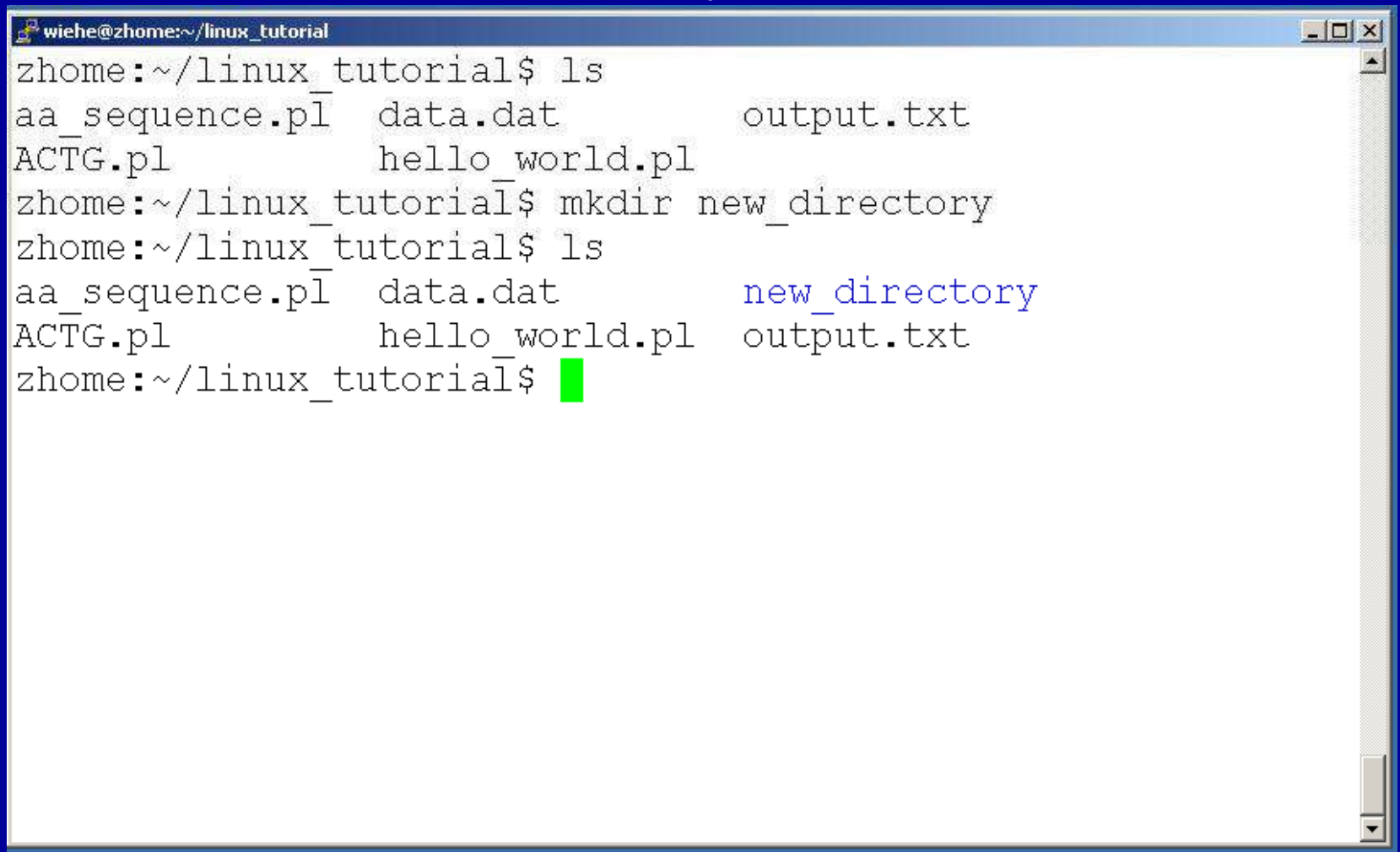
General Syntax: *

- "*" can be used as a wildcard in unix/linux



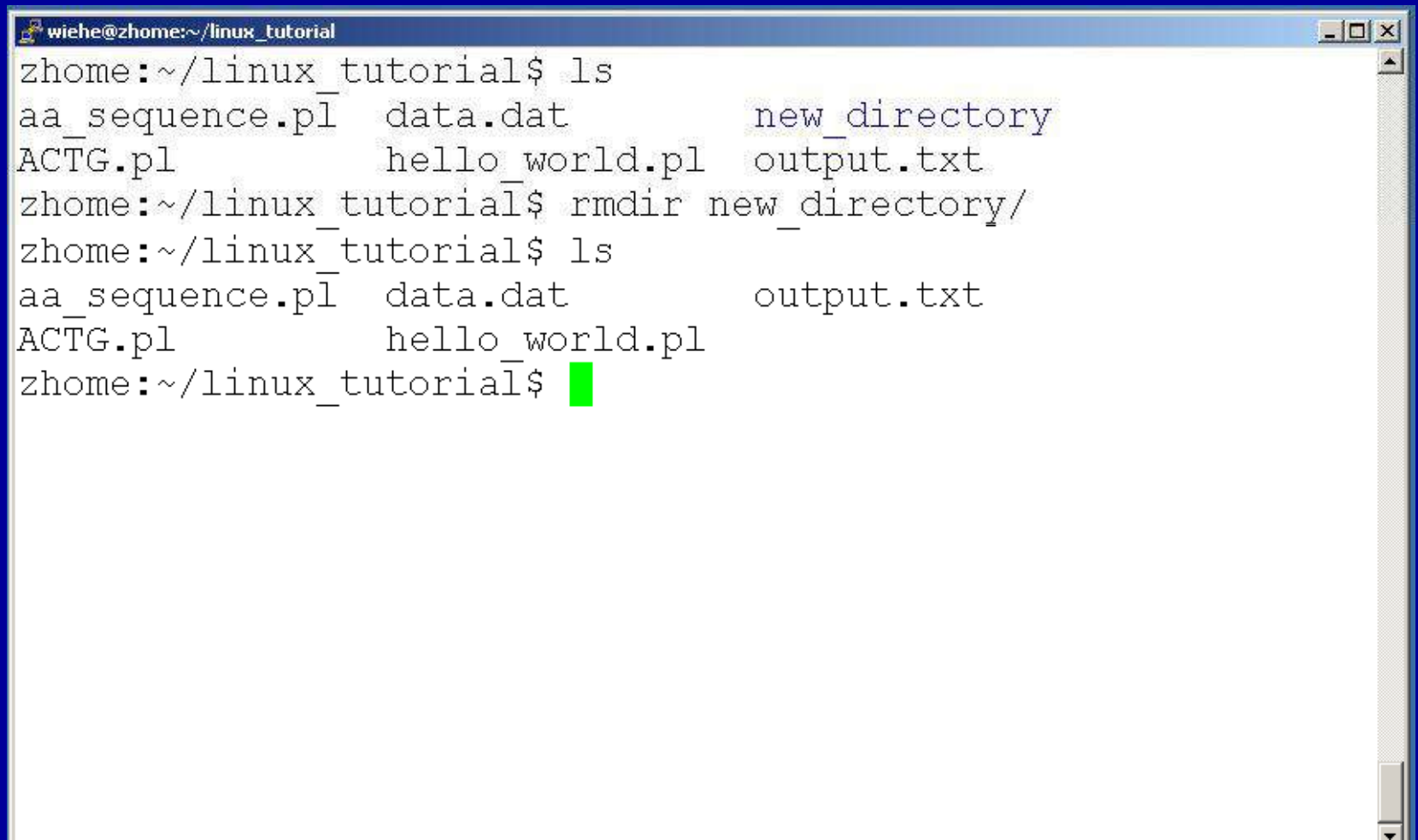
Command: mkdir

- To create a new directory use “mkdir”



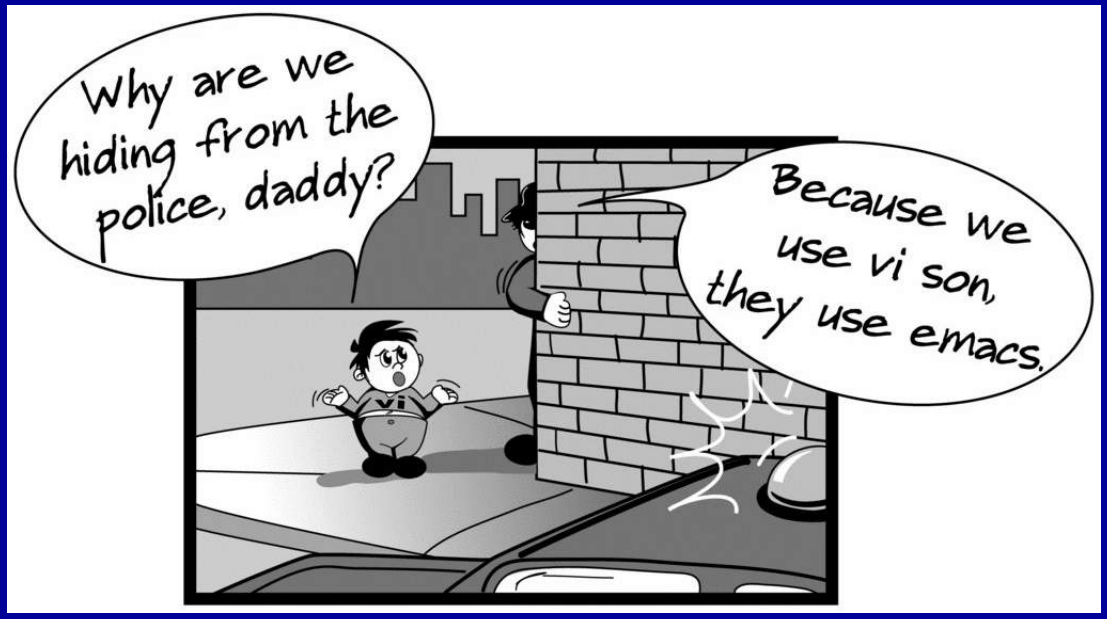
Command: rmdir

- To remove and empty directory use “rmdir”



Creating files in Unix/Linux

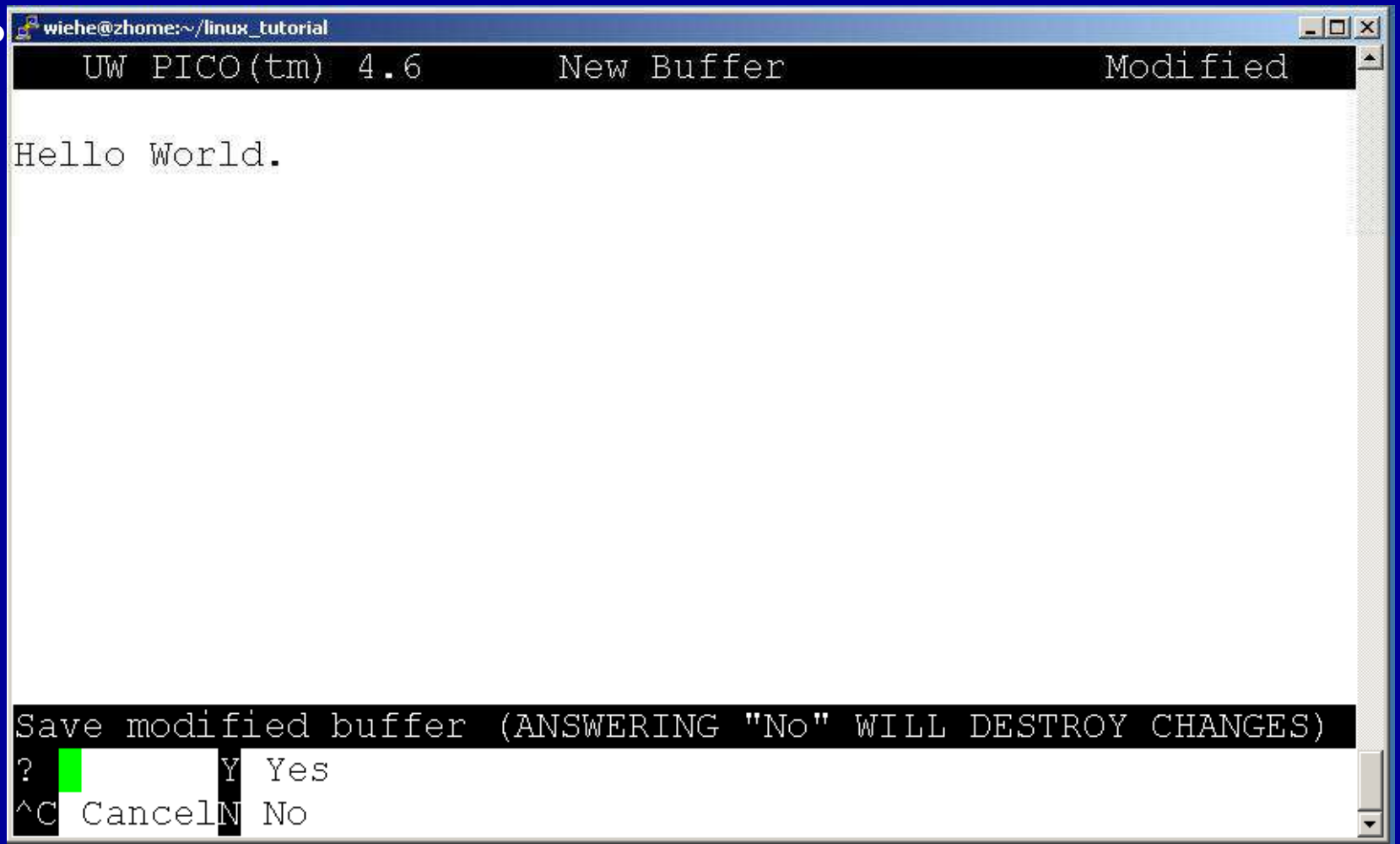
- Requires the use of an Editor
- Various Editors:
  1) nano / pico
  2) vi
  3) emacs



Editing a file using pico or nano

- 

Editing a file using pico



Displaying a file

- Various ways to display a file in Unix
  - cat
  - less
  - head
  - tail

Command: cat

- Dumps an entire file to standard output
- Good for displaying short, simple files

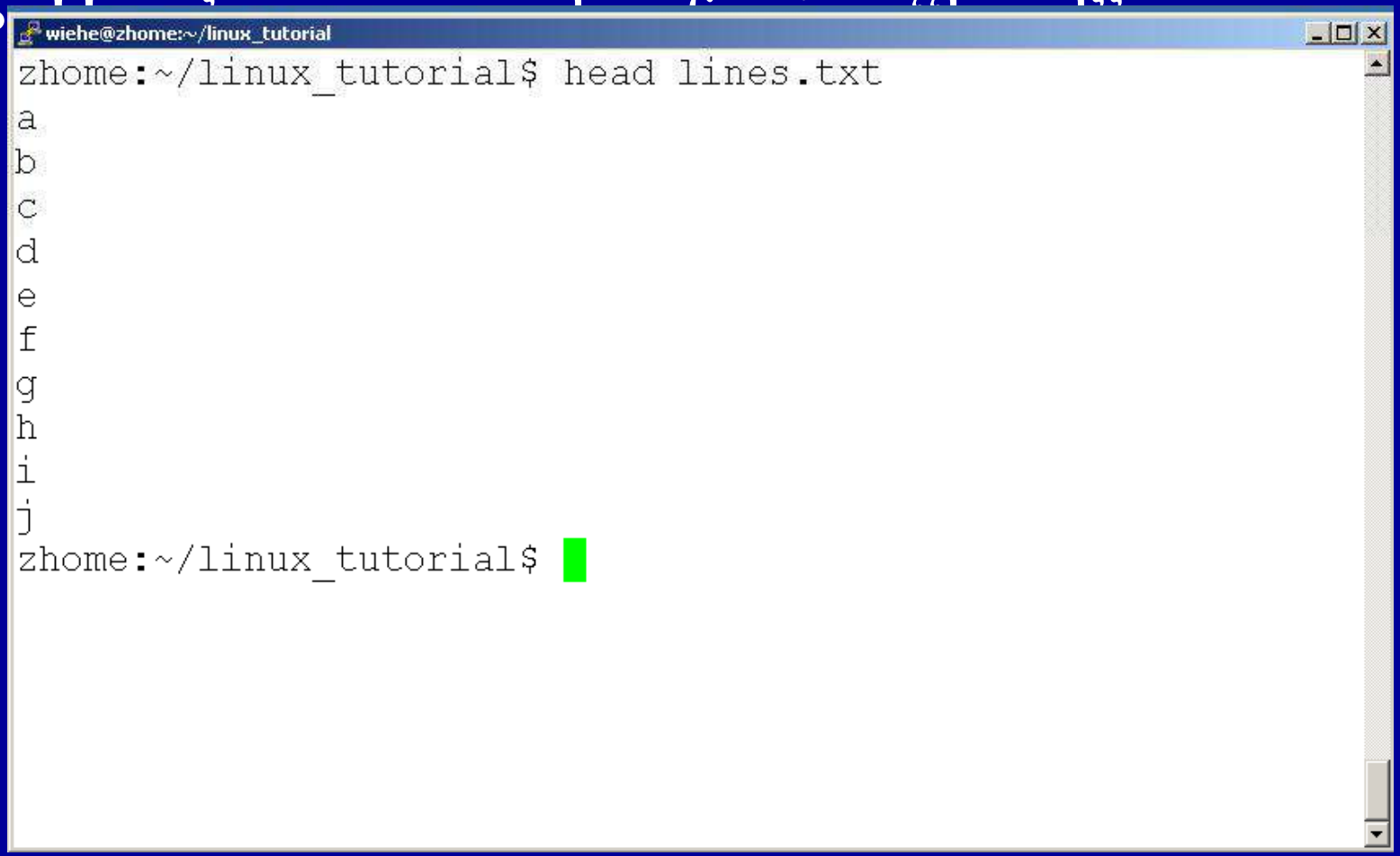
Command: less

- "less" displays a file, allowing forward/backward movement within it
  - return scrolls forward one line, space one page
  - y scrolls back one line, b one page
- use "/" to search for a string
- Press q to quit

Command: head

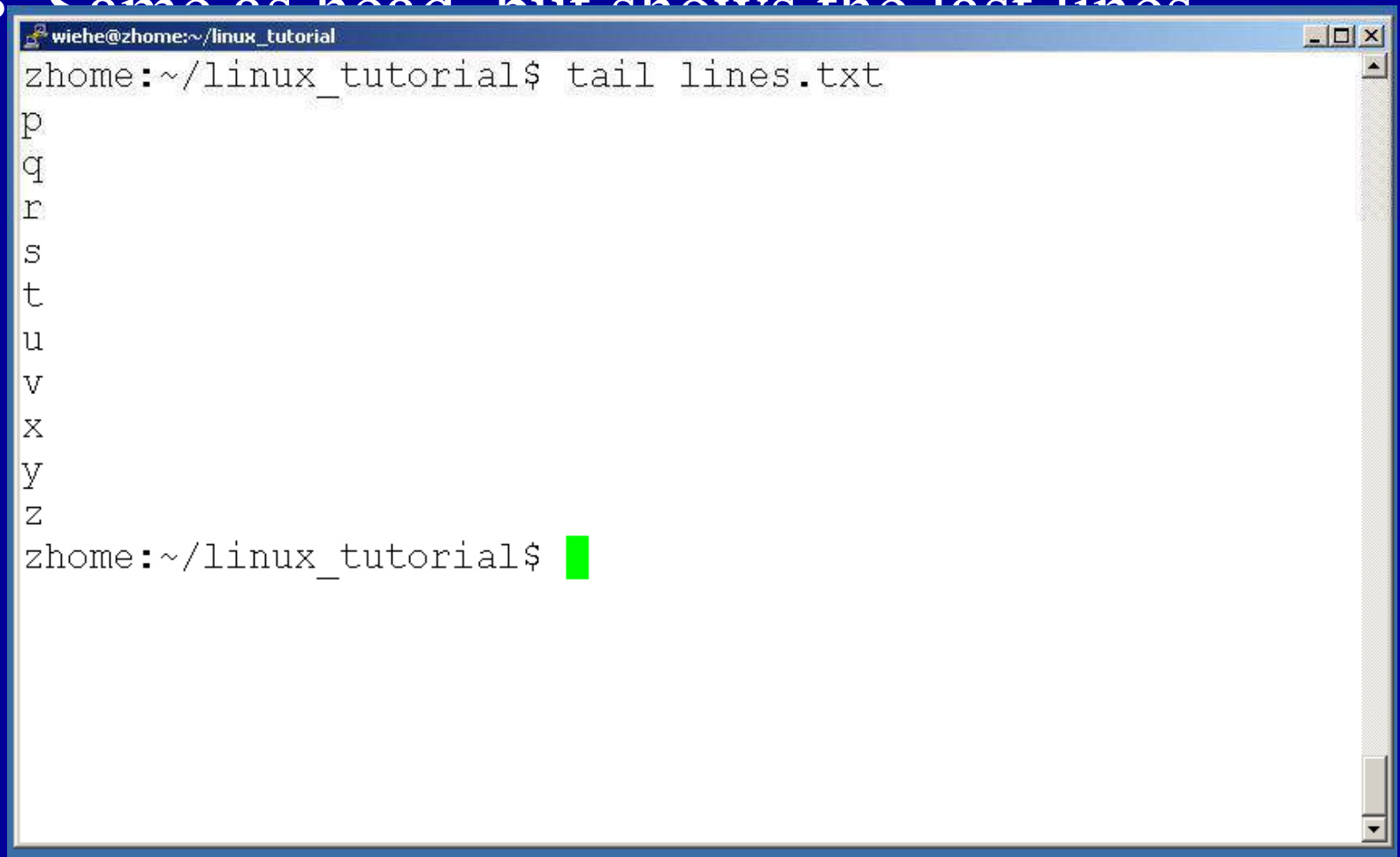
- “head” displays the top part of a file
- By default it shows the first 10 lines
- -n option allows you to change that
- “head -n50 file.txt” displays the first 50 lines of file.txt

Command: head

- 

Command: tail

- Same as head, but shows the last lines

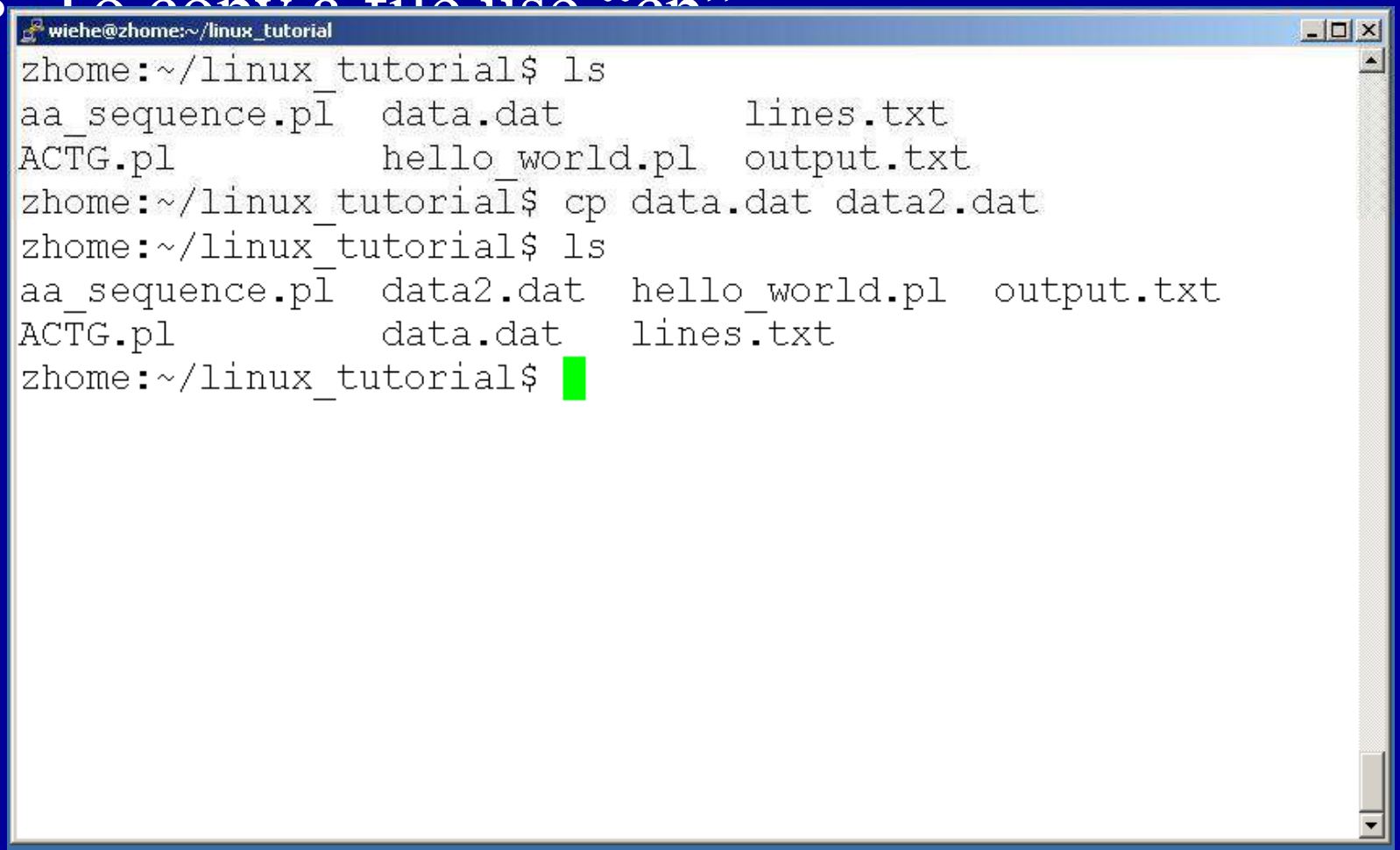


File Commands

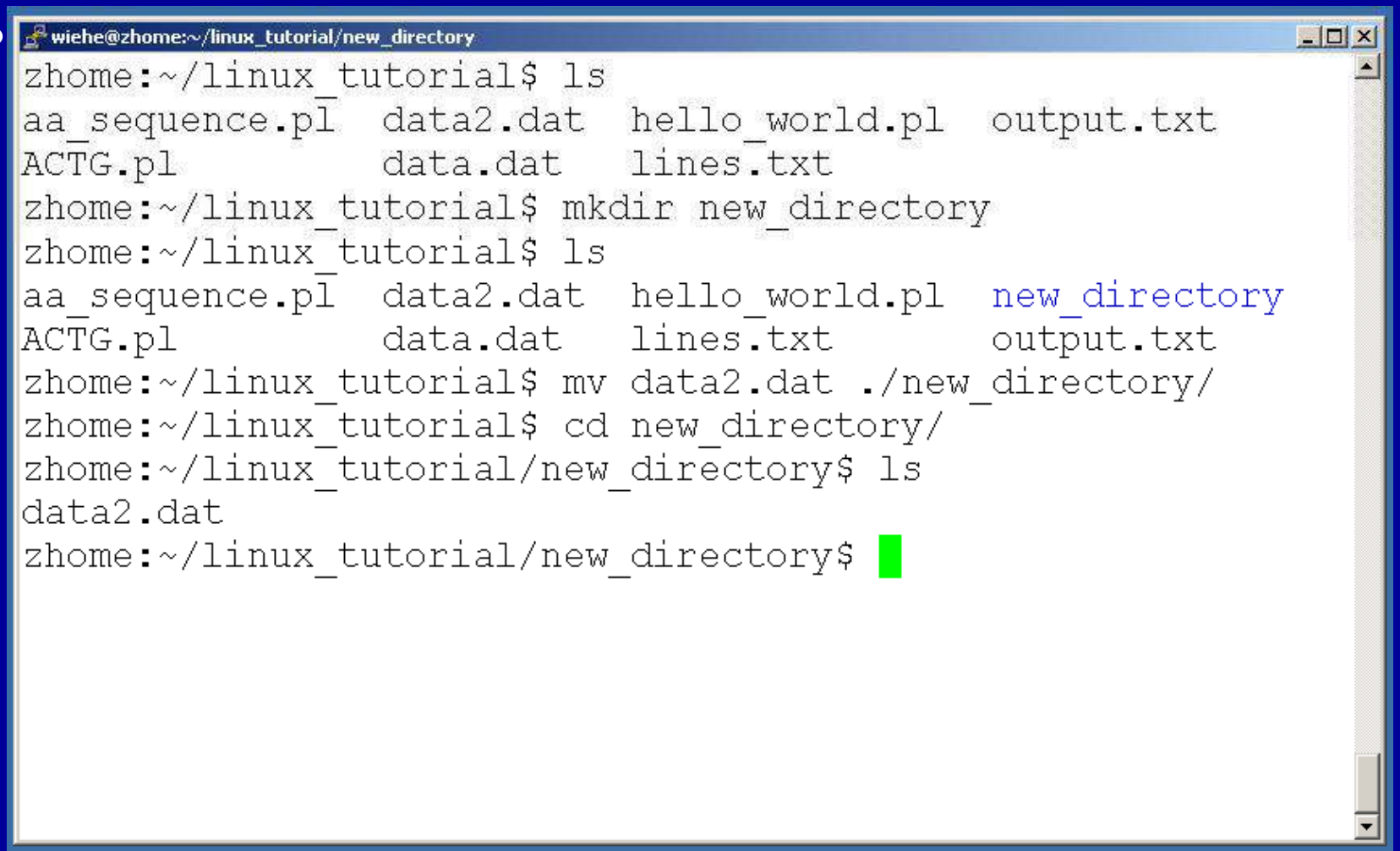
- Copying a file: cp
- Move or rename a file: mv
- Remove a file: rm

Command: cp

- To copy a file use "cp"

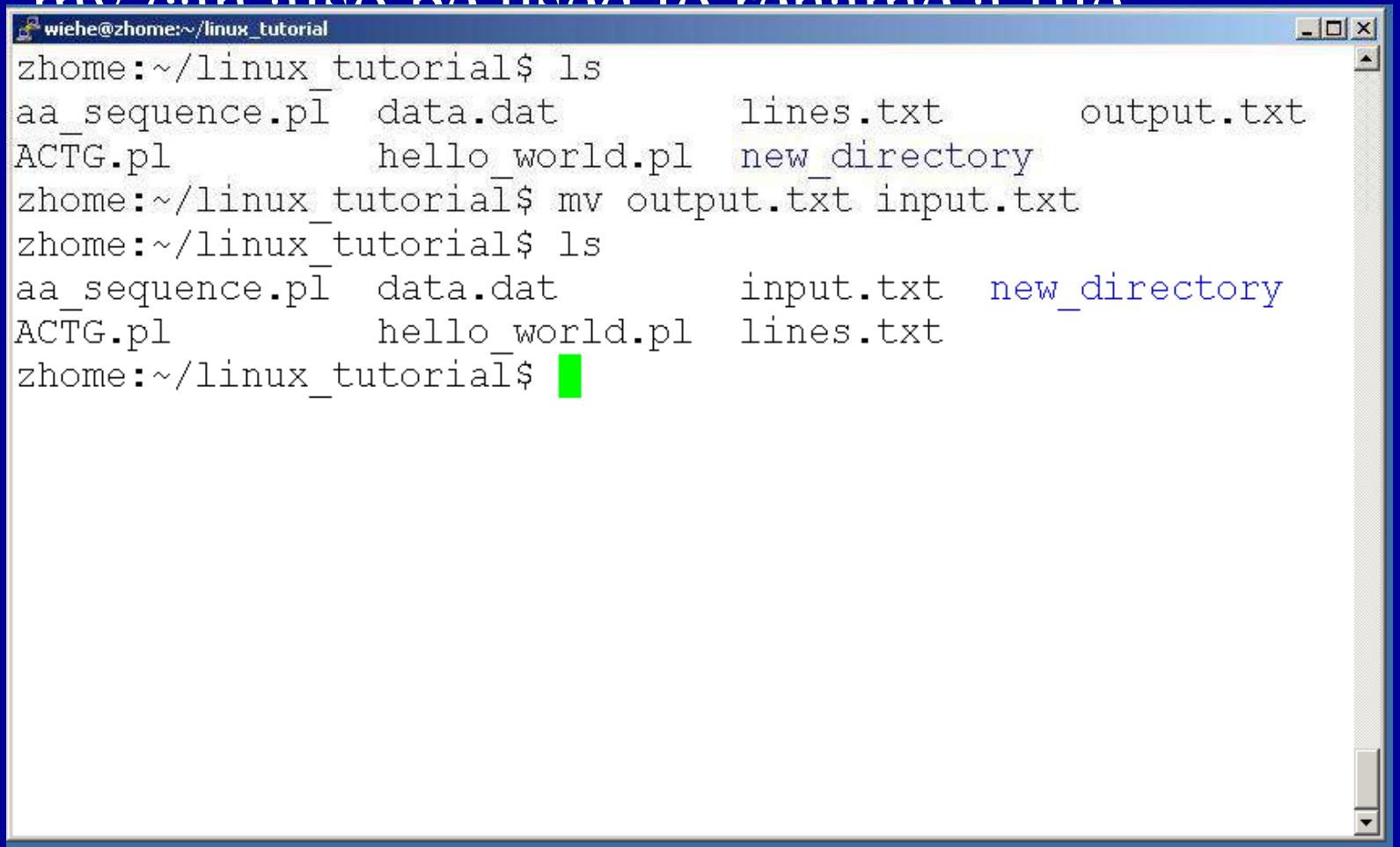


Command: mv

- 
wiehe@zhome:~/linux_tutorial/new_directory
zhome:~/linux_tutorial$ ls
aa_sequence.pl data2.dat hello_world.pl output.txt
ACTG.pl   data.dat   lines.txt
zhome:~/linux_tutorial$ mkdir new_directory
zhome:~/linux_tutorial$ ls
aa_sequence.pl data2.dat hello_world.pl new_directory
ACTG.pl   data.dat   lines.txt   output.txt
zhome:~/linux_tutorial$ mv data2.dat ./new_directory/
zhome:~/linux_tutorial$ cd new_directory/
zhome:~/linux_tutorial/new_directory$ ls
data2.dat
zhome:~/linux_tutorial/new_directory$

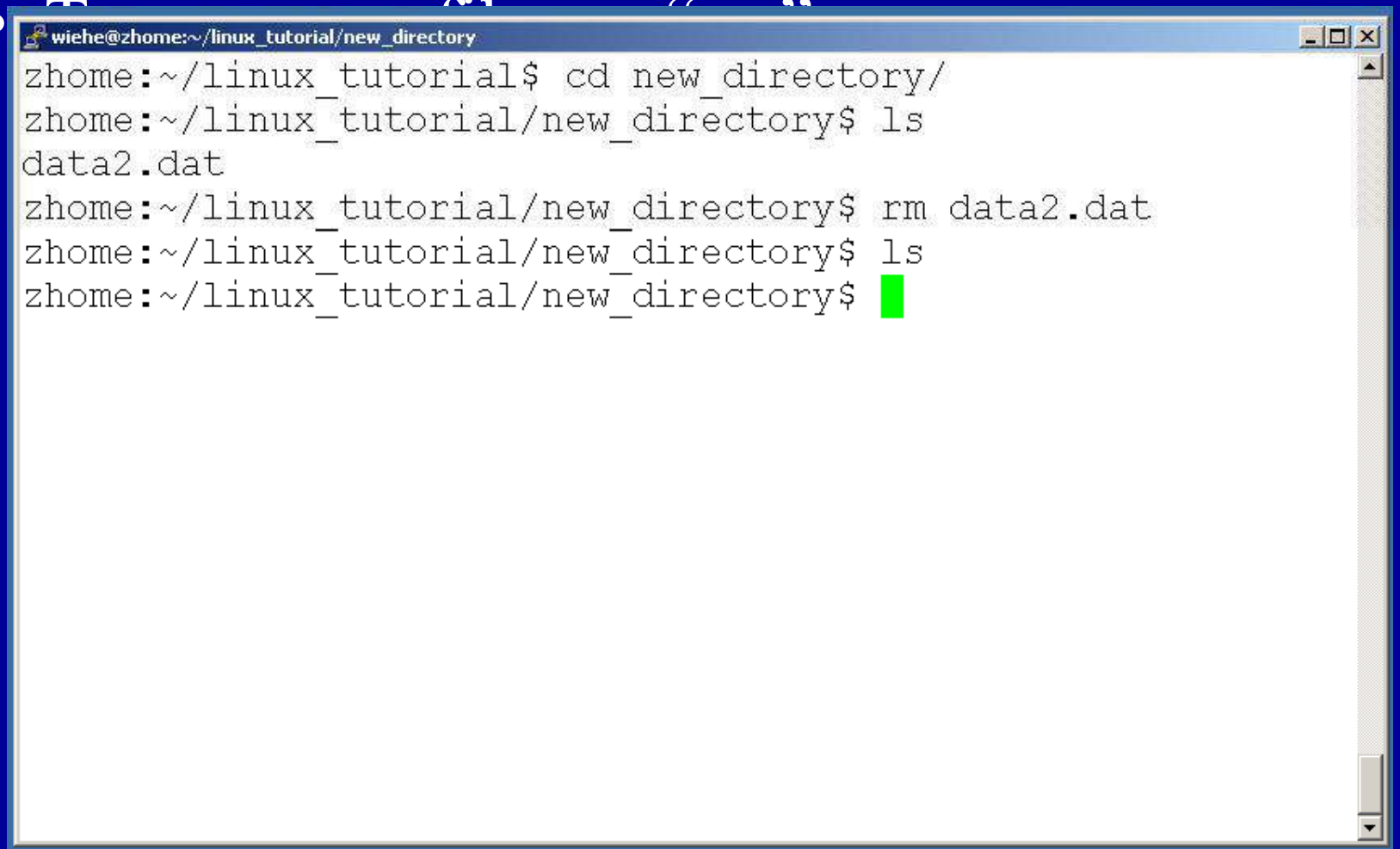
Command: mv

- mv can also be used to rename a file



Command: rm

- To remove a file, use the command "rm"



Command: rm

- To remove a file “recursively”: rm –r
- Used to remove all files and directories
- Be very careful, deletions are permanent in Unix/Linux

File permissions

- Each file in Unix/Linux has an associated permission level
- This allows the user to prevent others from reading/writing/executing their files or directories
- Use “ls -l *filename*” to find the permission level of that file

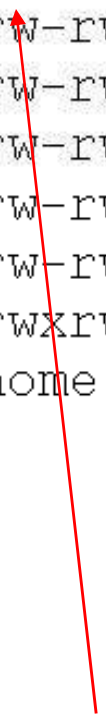
Permission levels

- “r” means “read only” permission
- “w” means “write” permission
- “x” means “execute” permission
  - In case of directory, “**x**” grants permission to list directory contents

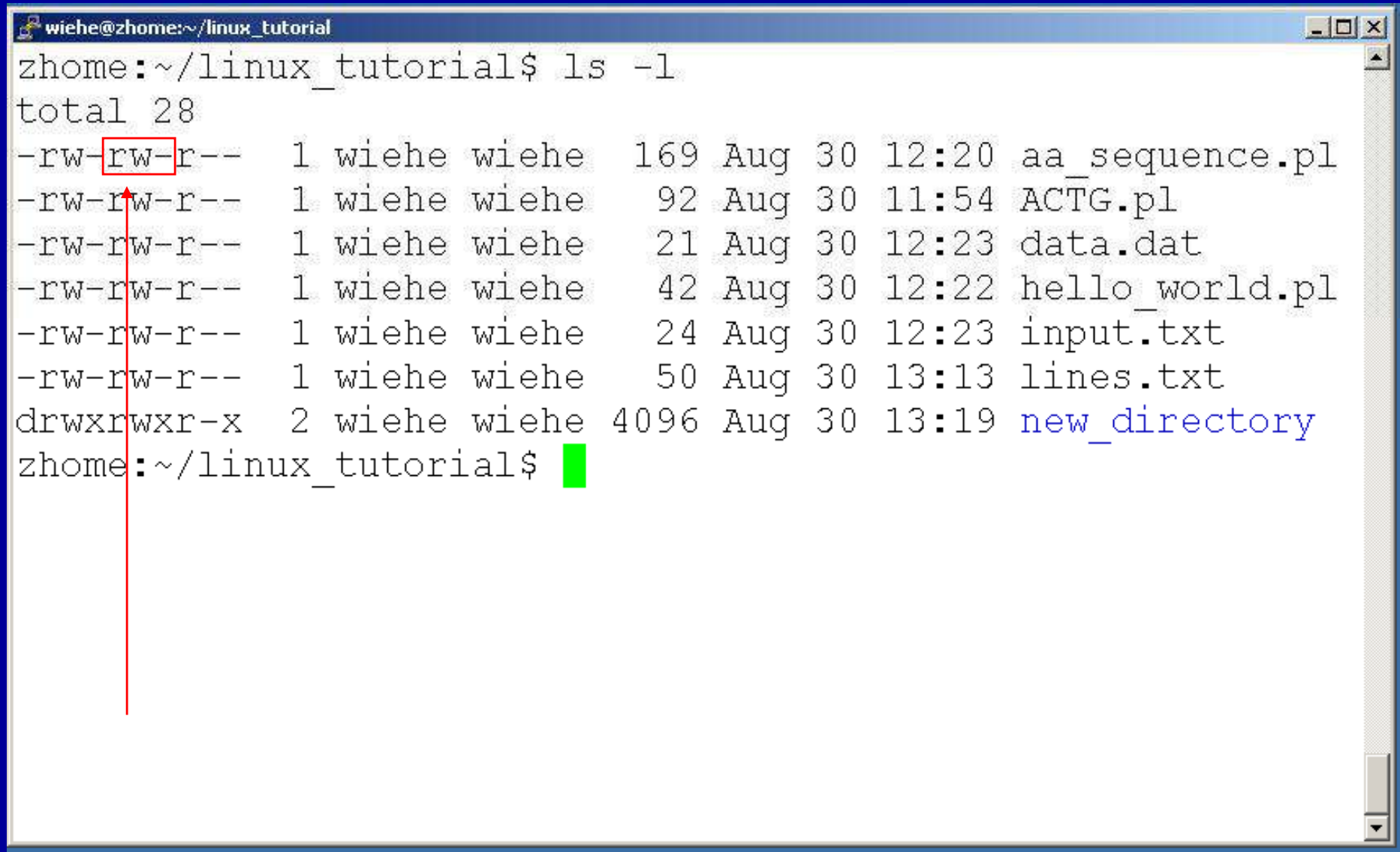
File Permissions

wiehe@zhome:~/linux_tutorial

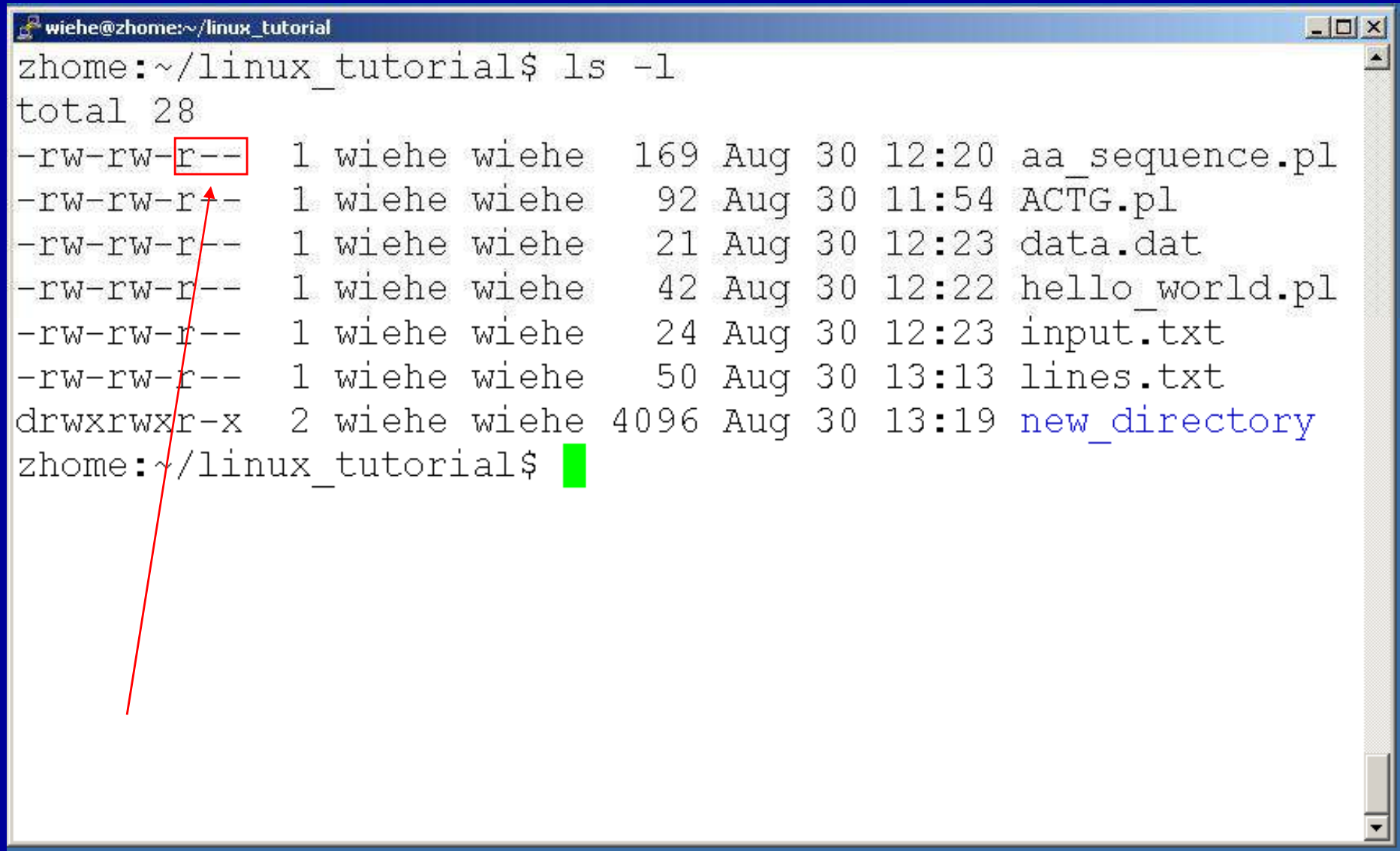
```
zhome:~/linux_tutorial$ ls -l
total 28
-rw-rw-r--  1 wiehe wiehe  169 Aug 30 12:20 aa_sequence.pl
-rw-rw-r--  1 wiehe wiehe   92 Aug 30 11:54 ACTG.pl
-rw-rw-r--  1 wiehe wiehe   21 Aug 30 12:23 data.dat
-rw-rw-r--  1 wiehe wiehe   42 Aug 30 12:22 hello_world.pl
-rw-rw-r--  1 wiehe wiehe   24 Aug 30 12:23 input.txt
-rw-rw-r--  1 wiehe wiehe   50 Aug 30 13:13 lines.txt
drwxrwxr-x  2 wiehe wiehe 4096 Aug 30 13:19 new_directory
zhome:~/linux_tutorial$
```



File Permissions

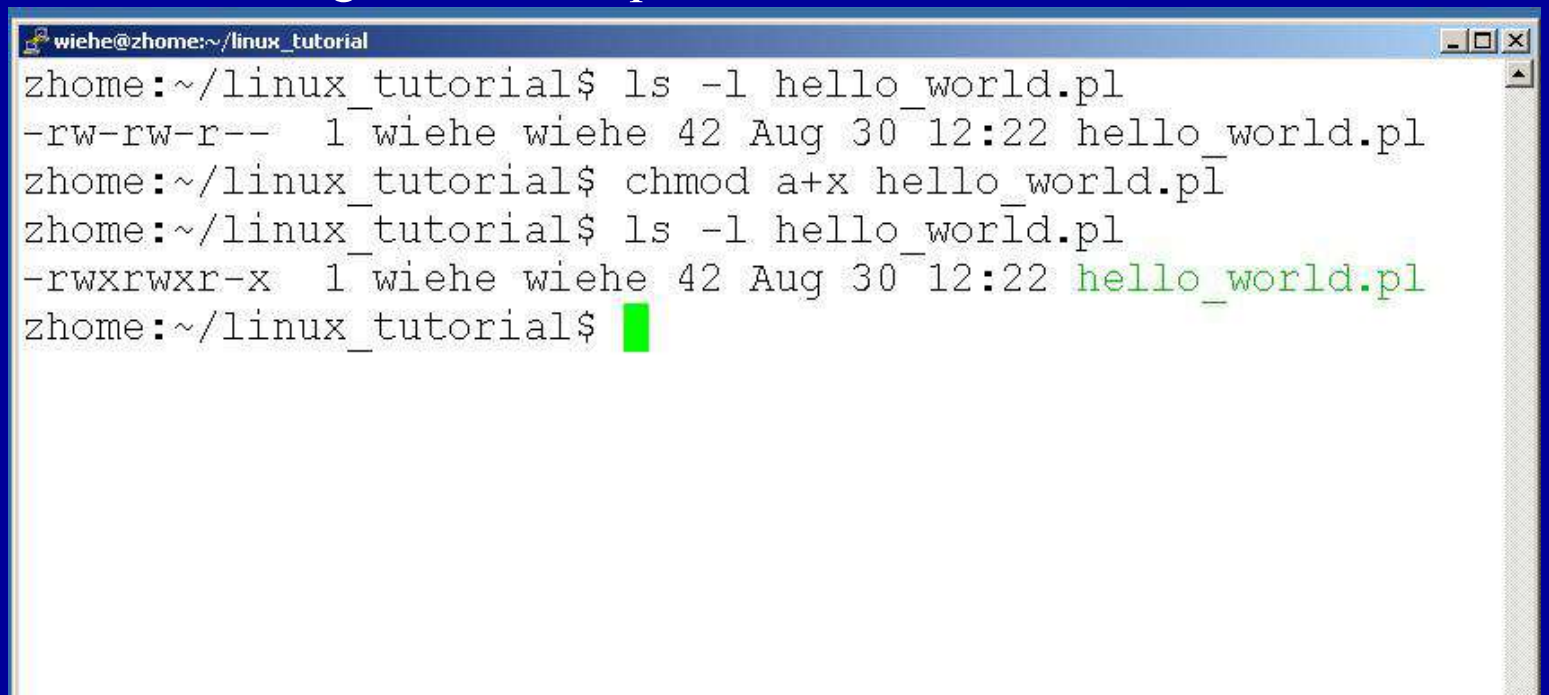


File Permissions



Command: chmod

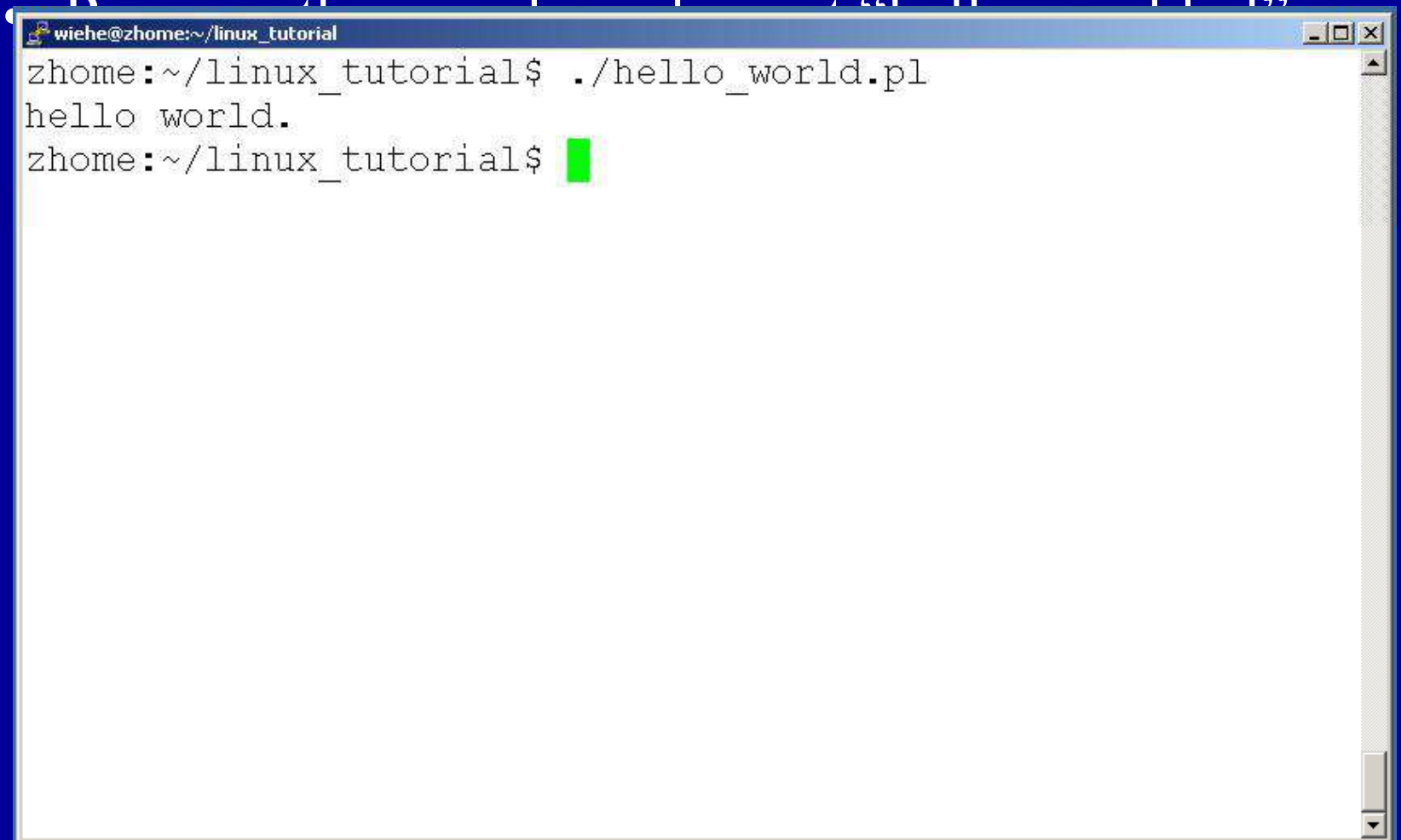
- If you own the file, you can change it's permissions with "chmod"
  - Syntax: chmod [**u**ser/**g**roup/**o**thers/**a**ll]+[permission] [file(s)]
  - Below we grant execute permission to all:



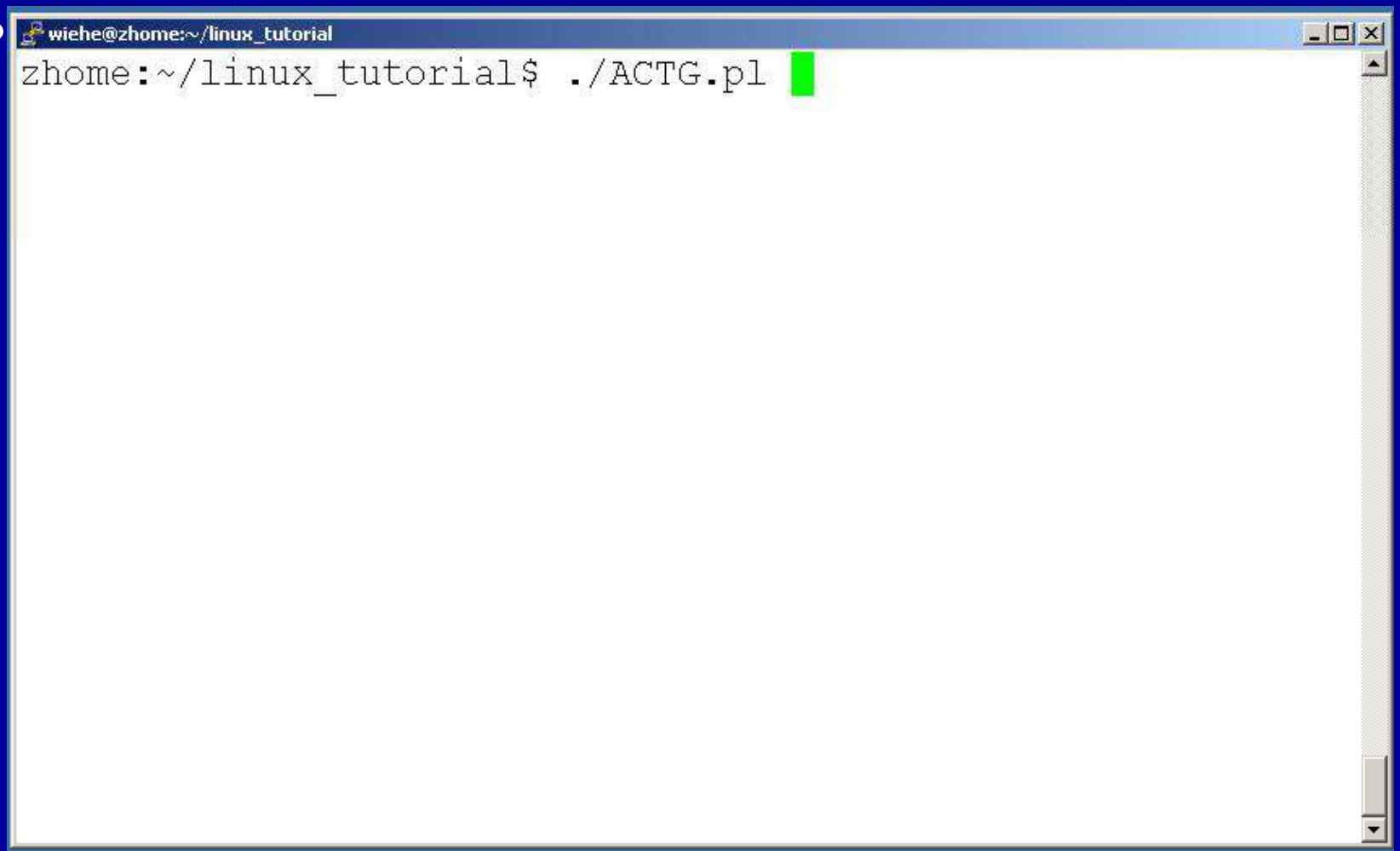
Running a program (a.k.a. a job)

- Make sure the program has executable permissions
- Use “./” to run the program

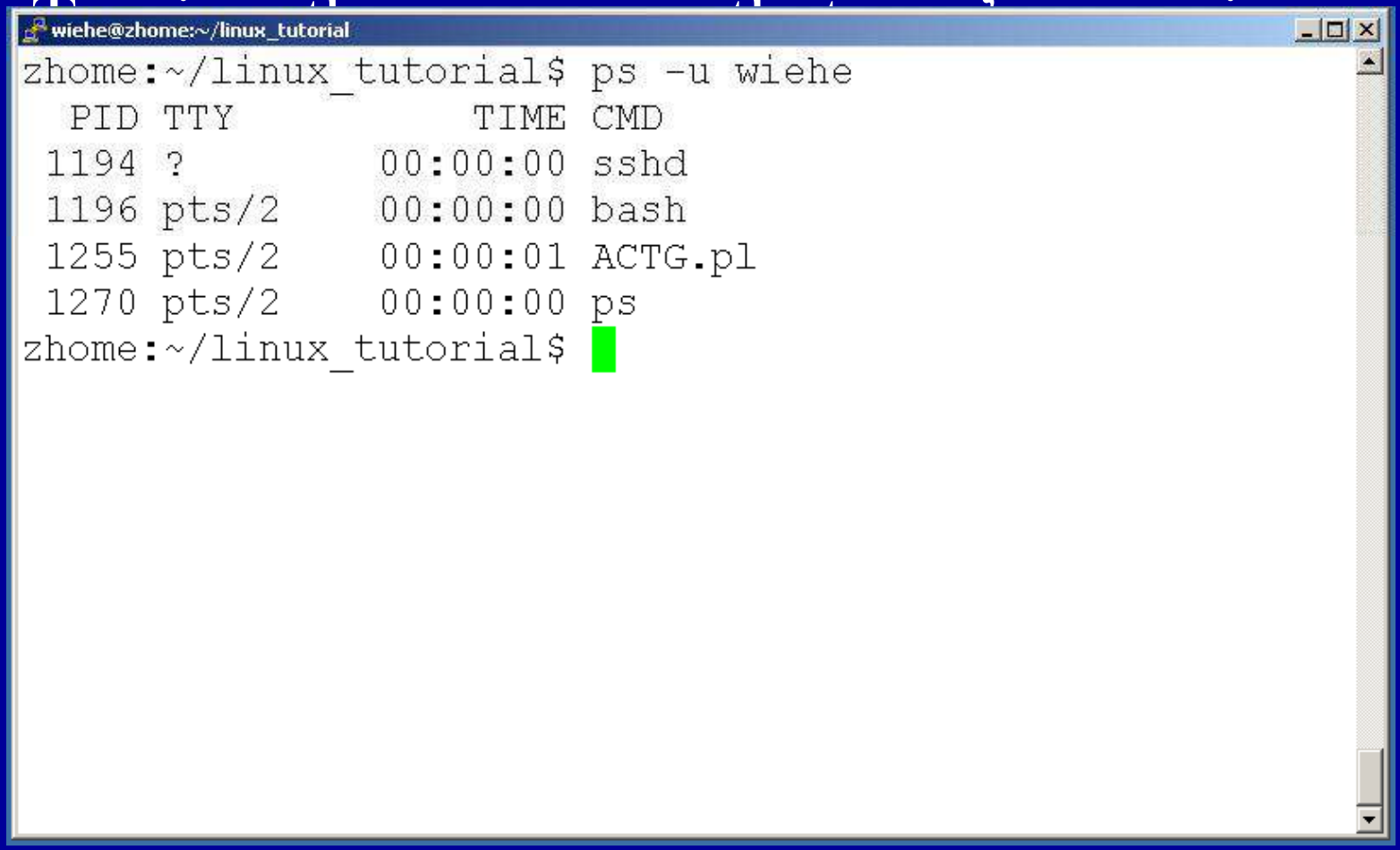
Running a program: an example



Ending a program

- 

Command: ps

- 
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ps -u wiehe
PID TTY TIME CMD
1194 ? 00:00:00 sshd
1196 pts/2 00:00:00 bash
1255 pts/2 00:00:01 ACTG.pl
1270 pts/2 00:00:00 ps
zhome:~/linux_tutorial$

Command: top

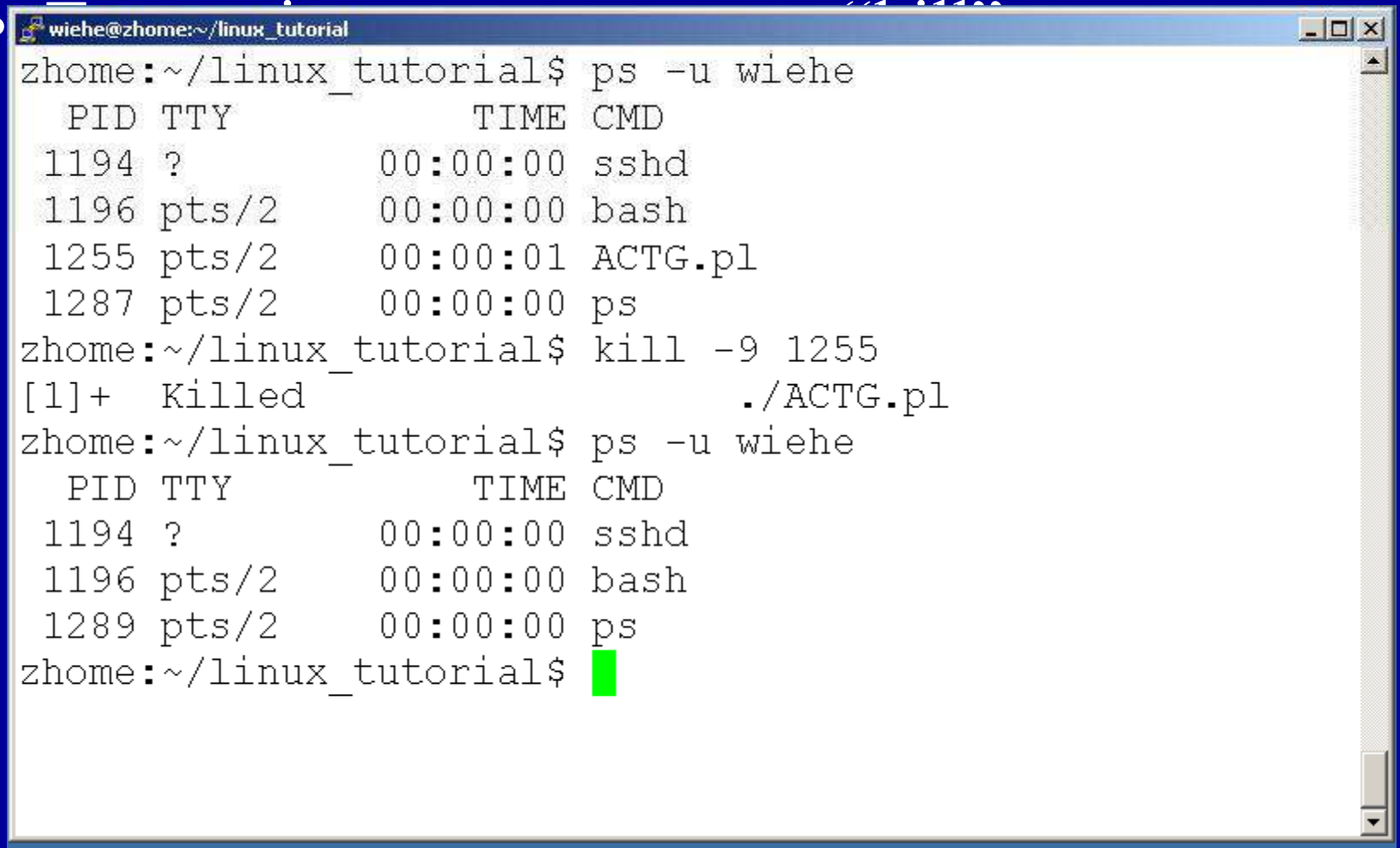
- Top is a tool for full

wiehe@zhome:~/linux_tutorial

```
top - 13:46:33 up 50 days,  4:26,  2 users,  load avera
Tasks:     total,     running,     sleeping,     stoppe
Cpu(s):       us,        sy,        ni,        id,        w
Mem:            total,            used,            free,
Swap:           total,            used,            free,
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM
3403	root	15	0	0	0	0	S	0.7	0.0
1	root	16	0	1604	324	292	S	0.0	0.0
2	root	RT	0	0	0	0	S	0.0	0.0
3	root	34	19	0	0	0	S	0.0	0.0
4	root	RT	0	0	0	0	S	0.0	0.0
5	root	34	19	0	0	0	S	0.0	0.0
6	root	RT	0	0	0	0	S	0.0	0.0
7	root	34	19	0	0	0	S	0.0	0.0
8	root	RT	0	0	0	0	S	0.0	0.0
9	root	34	19	0	0	0	S	0.0	0.0

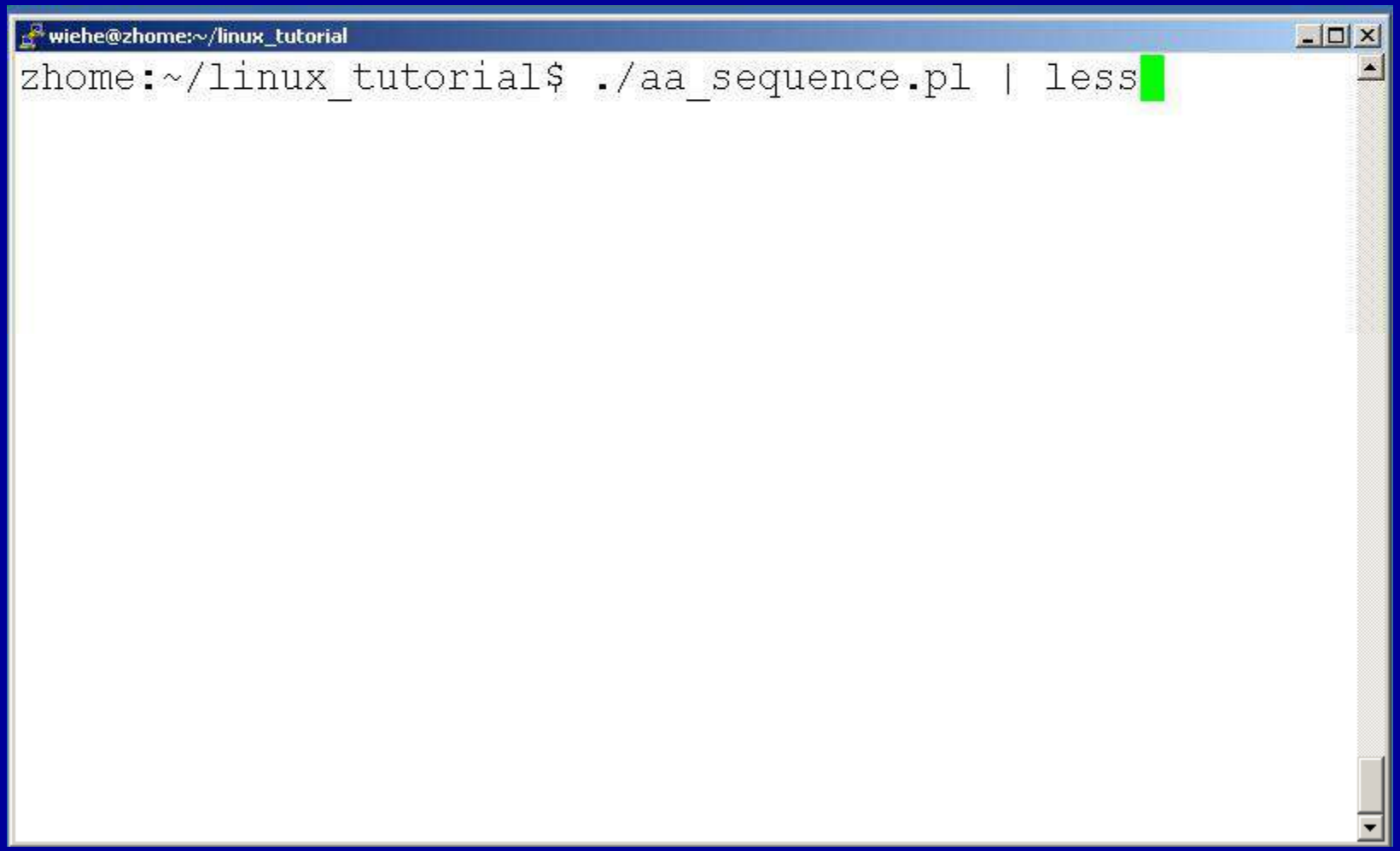
Command: kill

- 

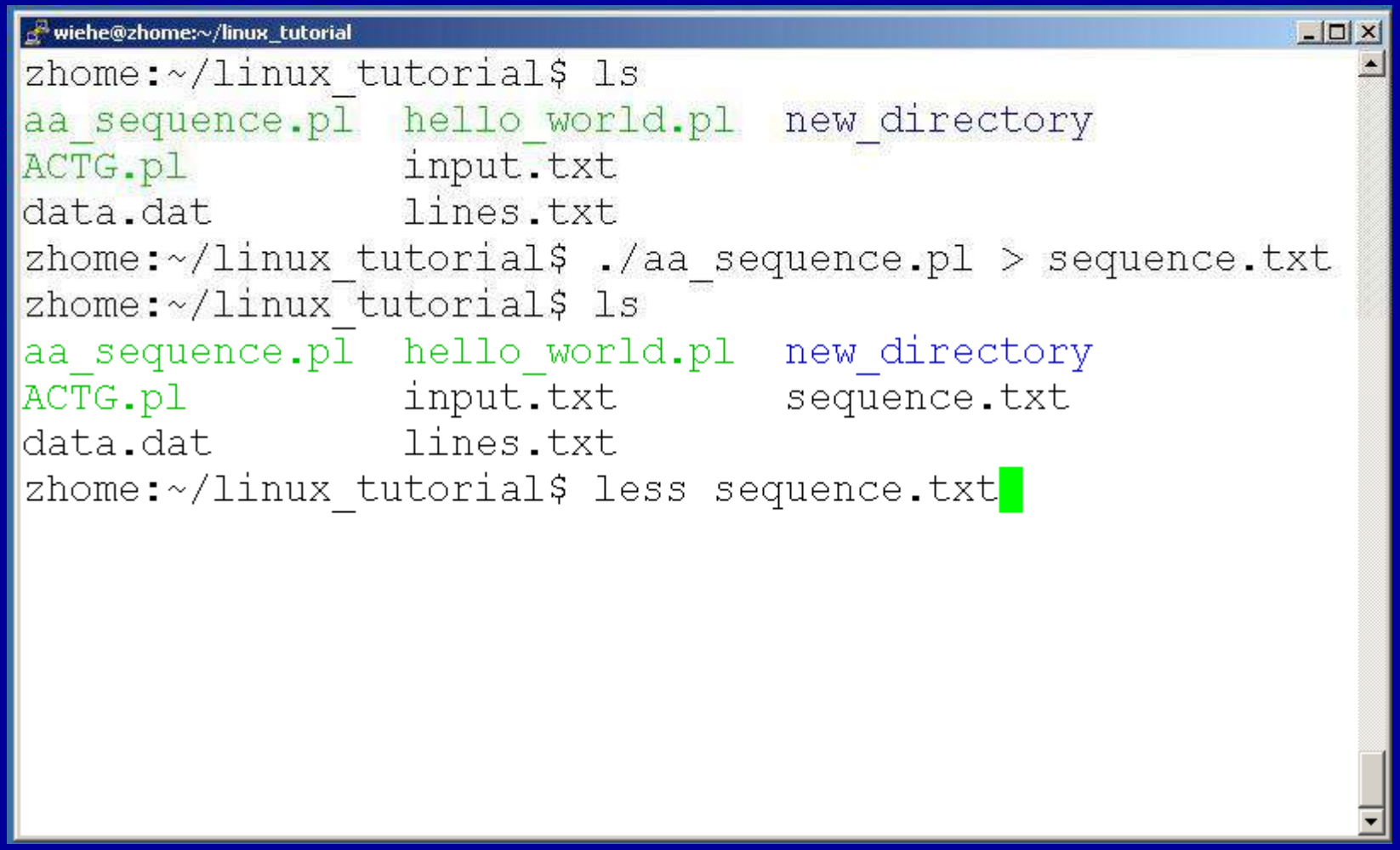
Input/Output Redirection ("piping")

- Programs can output to other programs
- Called "piping"
- "program_a | program_b"
  - program_a's output becomes program_b's input
- "program_a > file.txt"
  - program_a's output is written to a file called "file.txt"
- "program_a < input.txt"
  - program_a gets its input from a file called "input.txt"

A few examples of piping



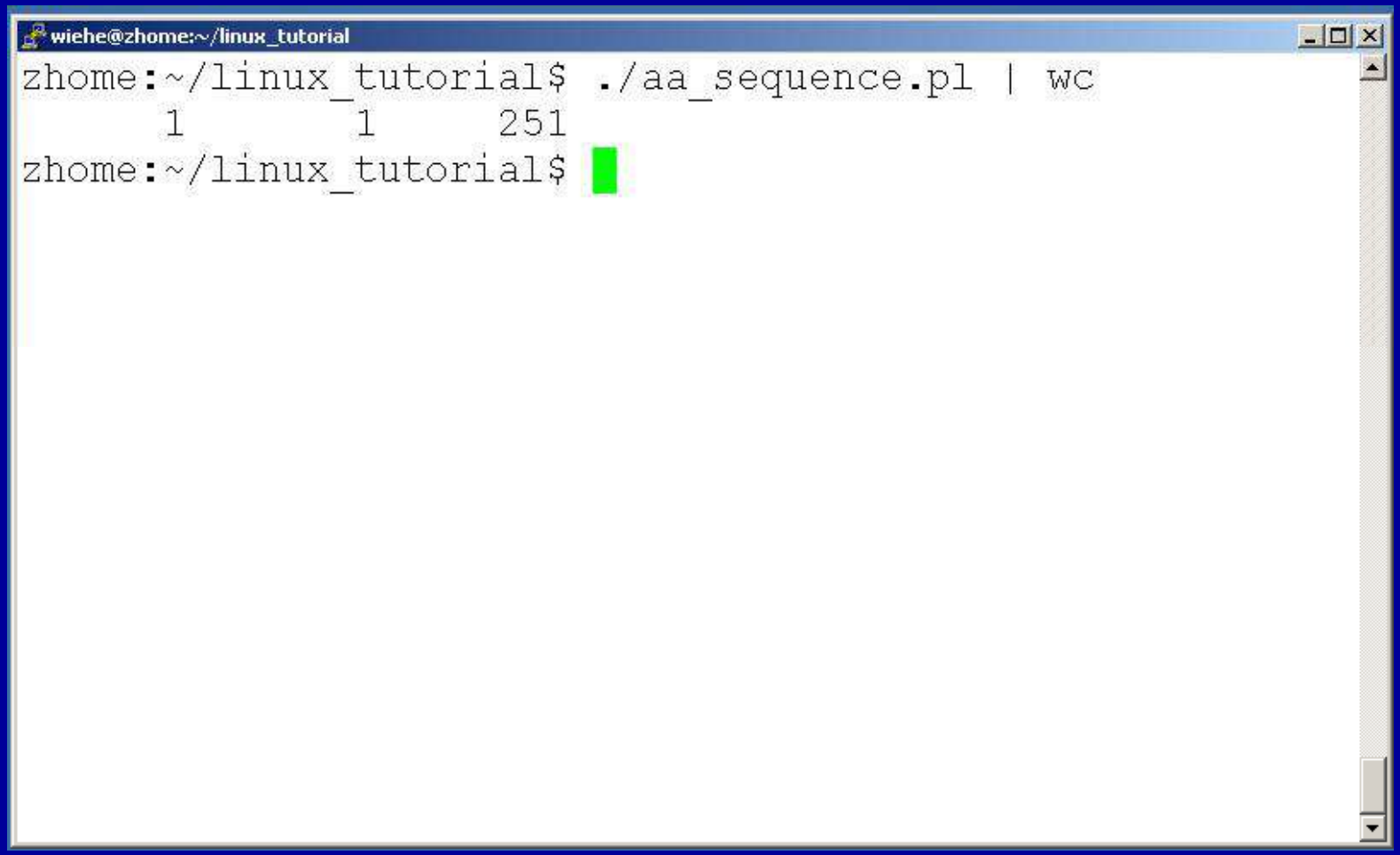
A few examples of piping



Command: wc

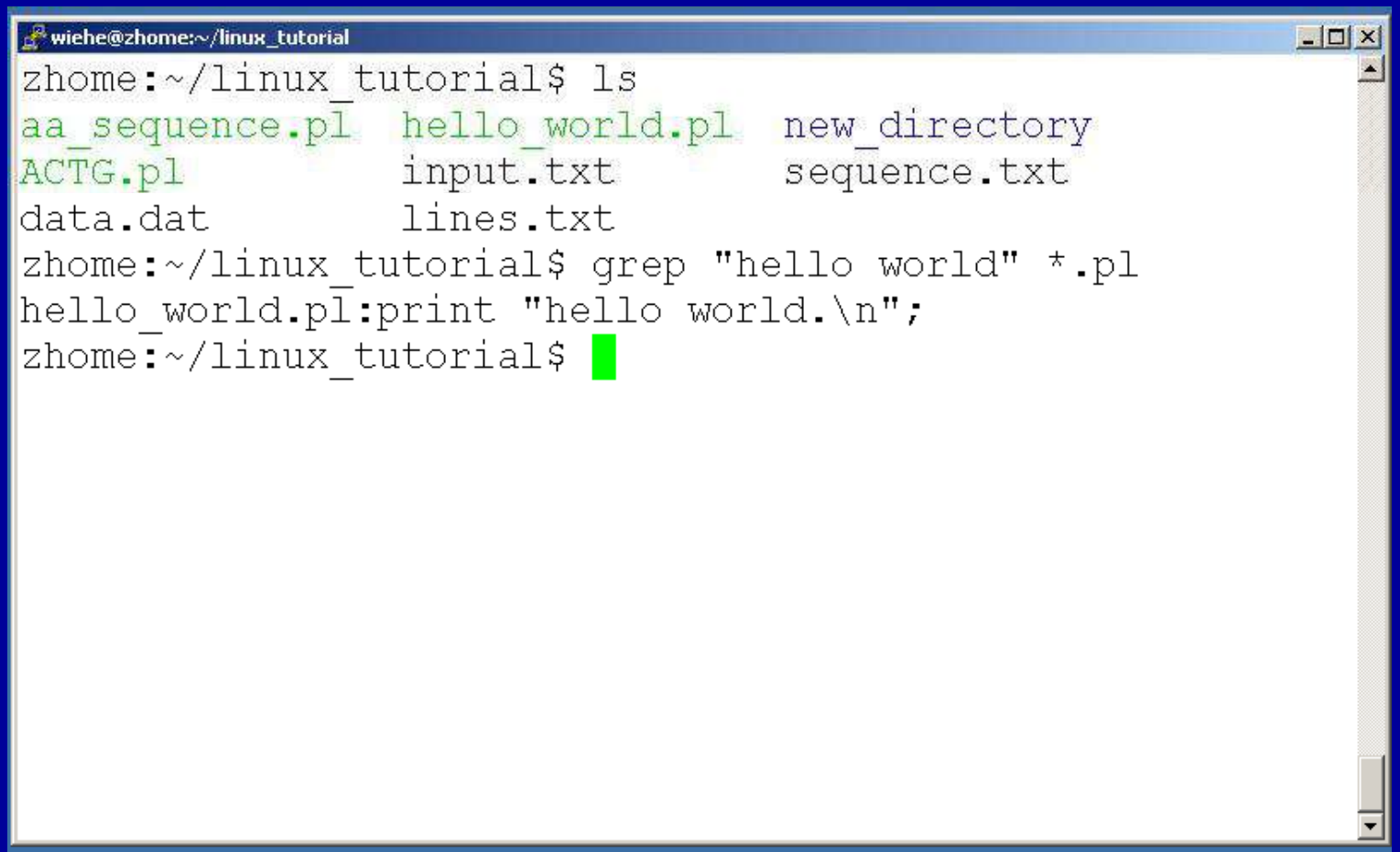
- To count the characters, words, and lines in a file use “wc”
- The first column in the output is lines, the second is words, and the last is characters

A few examples of piping



Command: grep

- To search files in a directory for a specific string use “grep”



Command: diff

- To compare to files for differences use “diff”
  - Try: diff /dev/null hello.txt
  - /dev/null is a special address -- it is always empty, and anything moved there is deleted

ssh, scp

- ssh is used to securely log in to remote systems, successor to telnet
- ssh [username]@[hostname]
- Try:

  ssh yourusername@localhost

  Type “exit” to log out of session

- Scp is used to copy files to/from remote systems, syntax is similar to cp:
  - scp [local path] [usernme]@[hostname]:[remote file path]
- Try:
  - **scp hello.txt yourusername@localhost:scp-test.txt**

Unix Web Resources

- http://www.ee.surrey.ac.uk/Teaching/Unix/
- http://www.ugu.com/sui/ugu/show?help.beginners
- http://en.wikipedia.org/wiki/Unix

UNIT - II

Sequence analysis

What is Sequence?

- A sequence is an ordered list of objects
- Biological sequence is a single, continuous molecule of nucleic acid or protein.

What is Sequence analysis?

- Comparing one or more sequences to identify the similarity.

Biological background…?

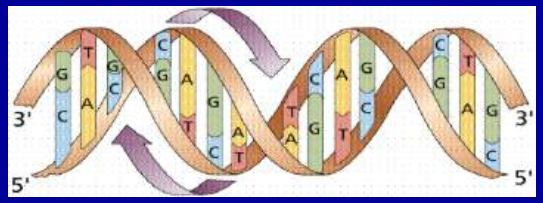
Biomolecules

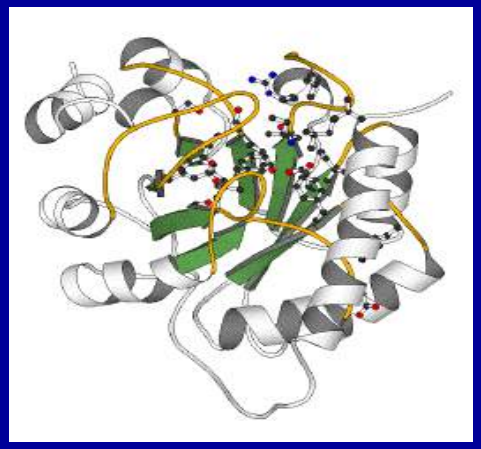
- Nucleic acids
- Proteins
- Enzymes
- Carbohydrates
- Lipids

} Macromolecules

- A **macromolecule** is a very large molecule commonly created by polymerization of smaller subunits (monomers)

The central dogma





DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

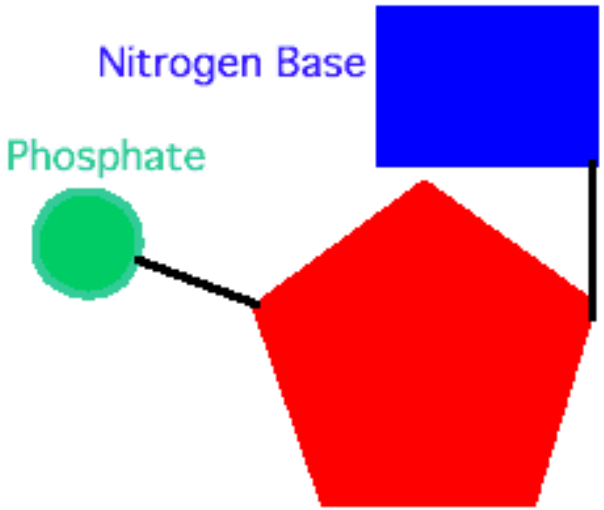
CCUGAGCCAACUAUUGAUGAA

PEPTIDE

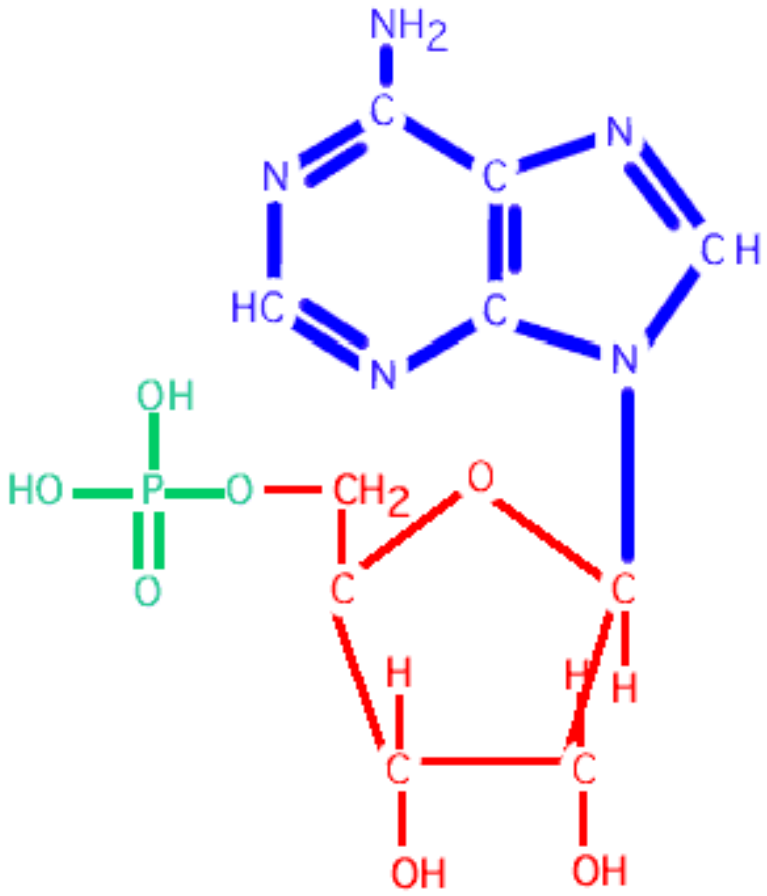
- **DNA**: units are the nucleotide residues A, C, G and T
- **RNA**: units are the nucleotide residues A, C, G and U
- **Proteins**: units are the amino acid residues A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y.

Structure of nucleotide

Basic Nucleotide Structure

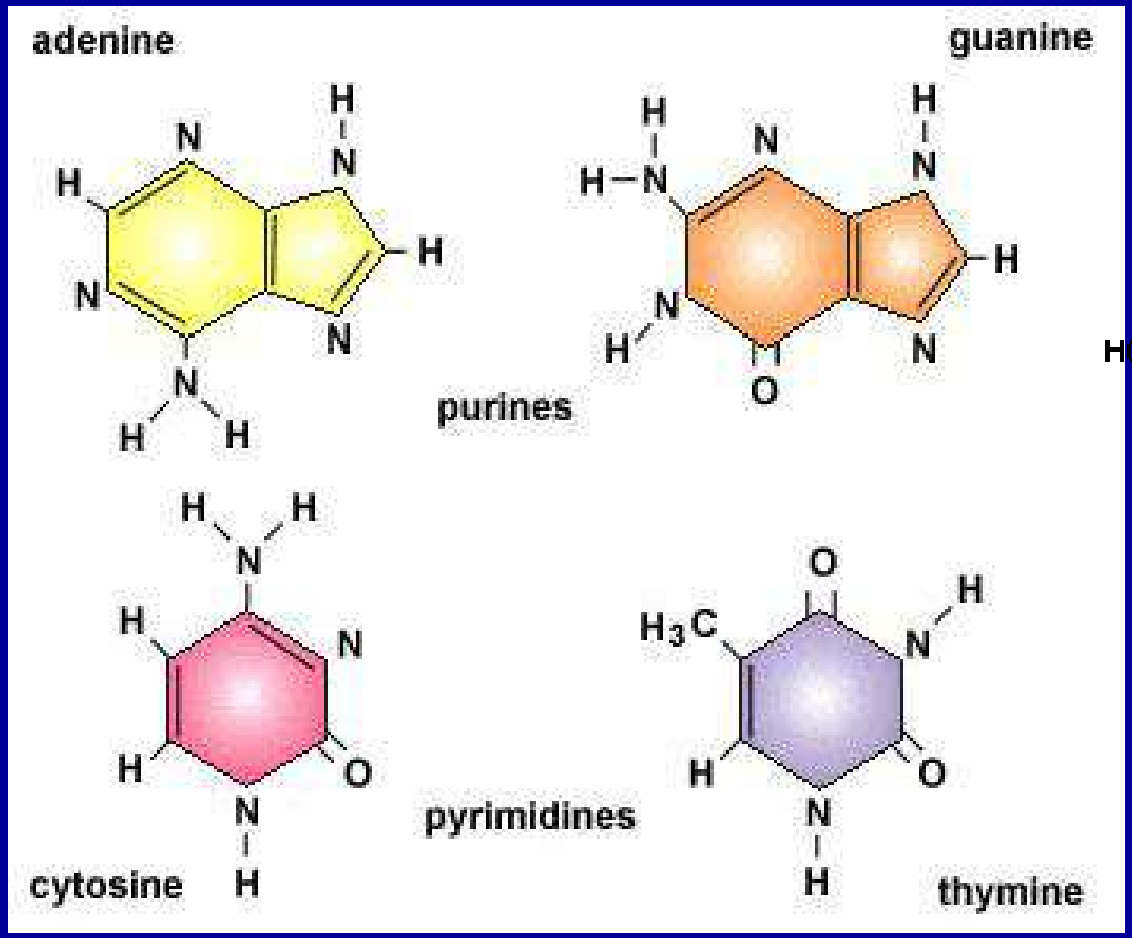


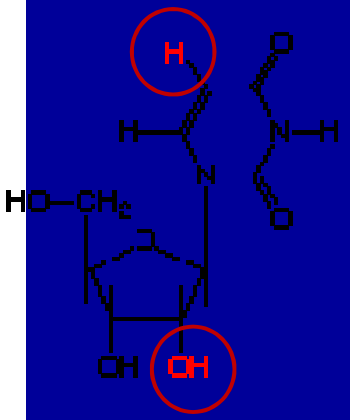
Example



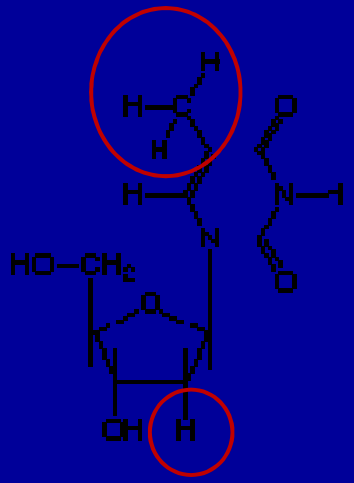
Adenosine 5' phosphoric acid

Bases



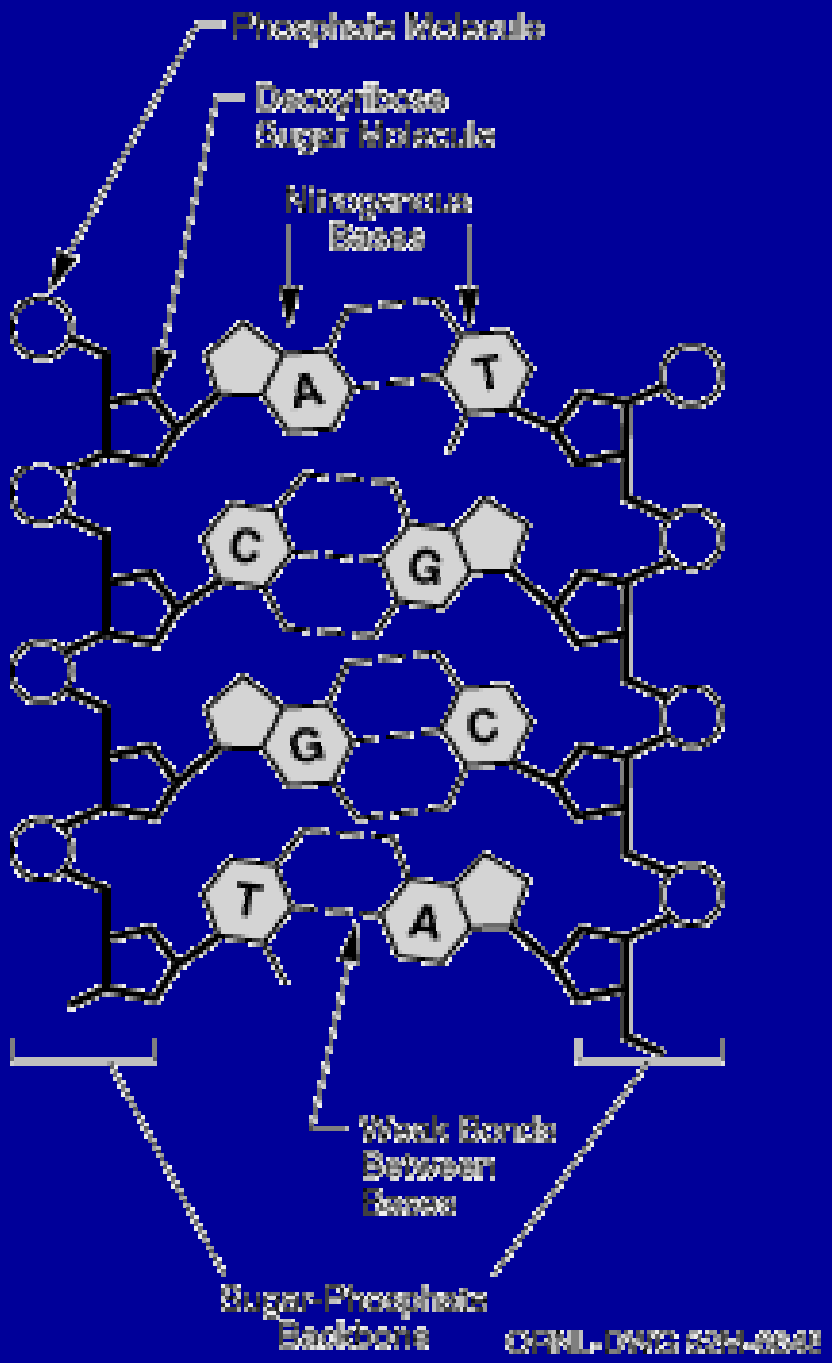


uracil (RNA)

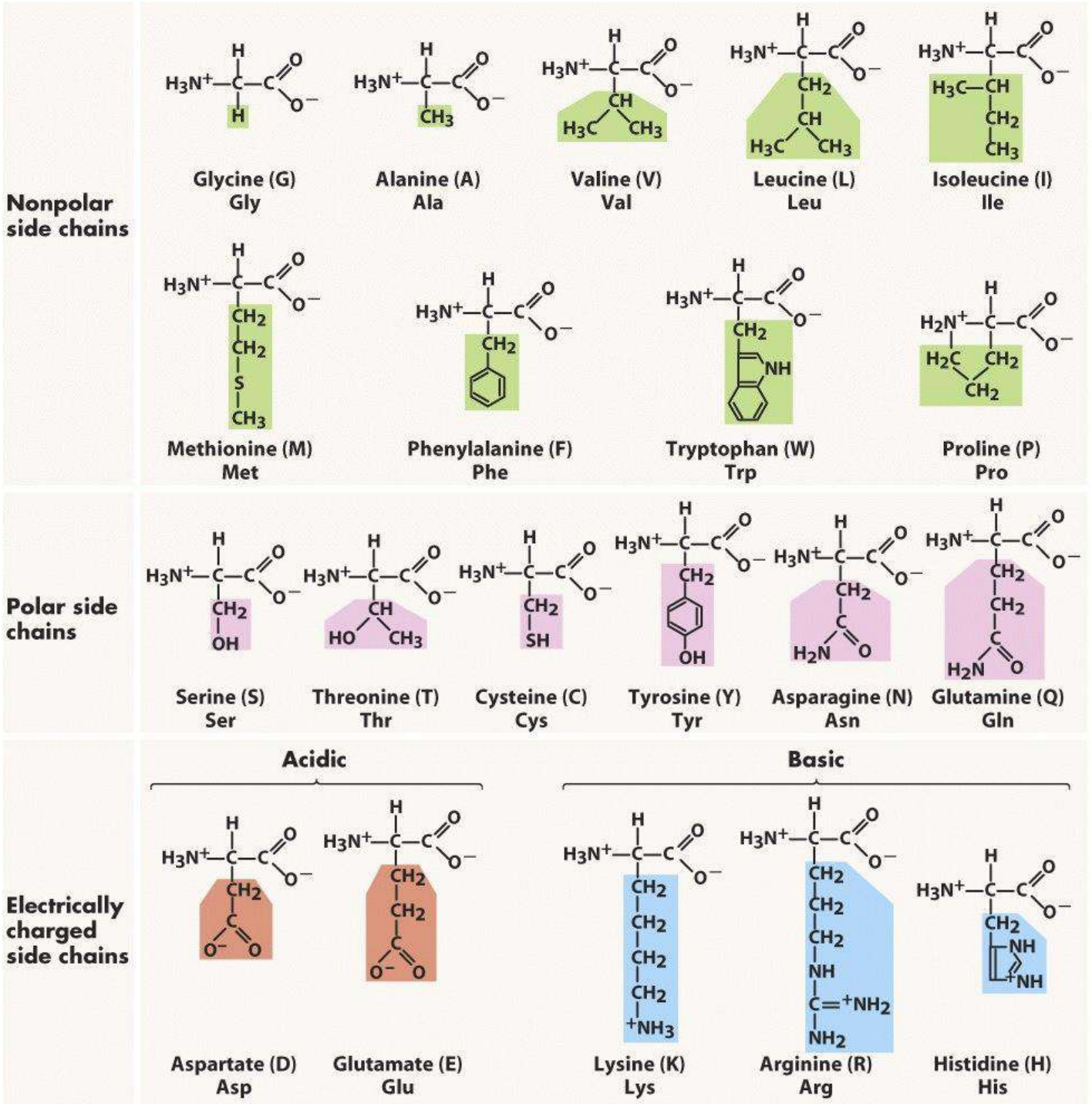


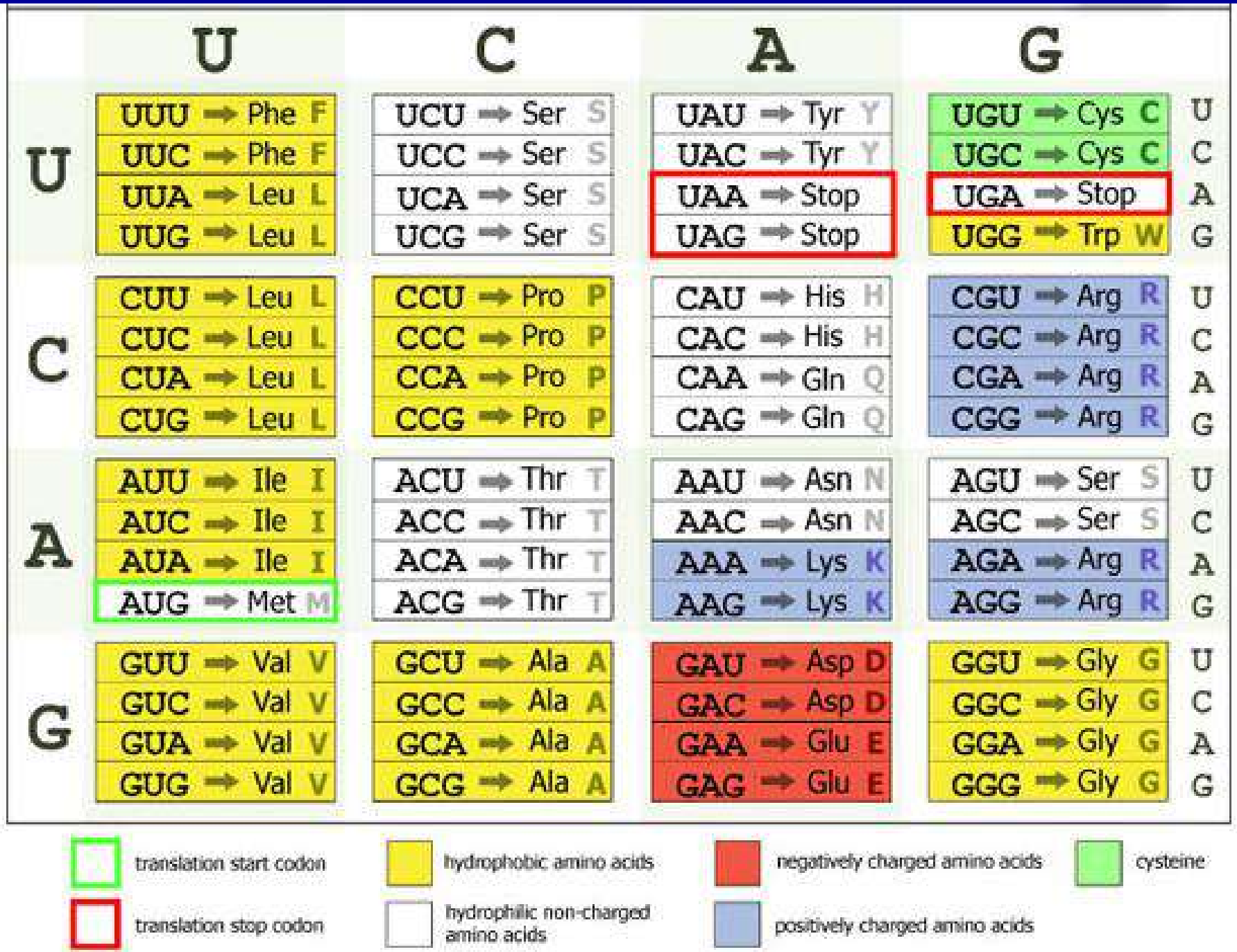
thymine (DNA)

Complimentary base pairing



Structure of Amino acids




U
C
A
G
UUU → Phe F
UUC → Phe F
UUA → Leu L
UUG → Leu L
UCU → Ser S
UCC → Ser S
UCA → Ser S
UCG → Ser S
UAU → Tyr Y
UAC → Tyr Y
UAA → Stop
UAG → Stop
UGU → Cys C
UGC → Cys C
UGA → Stop
UGG → Trp W
CUU → Leu L
CUC → Leu L
CUA → Leu L
CUG → Leu L
CCU → Pro P
CCC → Pro P
CCA → Pro P
CCG → Pro P
CAU → His H
CAC → His H
CAA → Gln Q
CAG → Gln Q
CGU → Arg R
CGC → Arg R
CGA → Arg R
CGG → Arg R
AUU → Ile I
AUC → Ile I
AUA → Ile I
AUG → Met M
ACU → Thr T
ACC → Thr T
ACA → Thr T
ACG → Thr T
AAU → Asn N
AAC → Asn N
AAA → Lys K
AAG → Lys K
AGU → Ser S
AGC → Ser S
AGA → Arg R
AGG → Arg R
GUU → Val V
GUC → Val V
GUA → Val V
GUG → Val V
GCU → Ala A
GCC → Ala A
GCA → Ala A
GCG → Ala A
GAU → Asp D
GAC → Asp D
GAA → Glu E
GAG → Glu E
GGU → Gly G
GGC → Gly G
GGA → Gly G
GGG → Gly G
translation start codon
translation stop codon
hydrophobic amino acids
hydrophilic non-charged amino acids
negatively charged amino acids
positively charged amino acids
cysteine

Sequence Alignment

- The identification of residue-residue correspondences
- The basic tool in bioinformatics

WHY Sequence Alignment ?

- For discovering functional, structural and evolutionary information in biological sequences
- Eases further tasks like:
  - Annotation of new sequences
  - Modeling of protein structures
  - Design and analysis of gene expression experiments

The Concept

- Sequence determines structure and structure determines function.

- An alignment is a mutual arrangement of two sequences
- Exhibits where two sequences are similar, and where they differ
- An ‘optimal’ alignment – most correspondences and the least differences

Terms of sequence comparison

Sequence identity

- Exactly same Nucleotide/Amino Acid in same position

Sequence similarity

- Substitutions with similar chemical properties

Sequence homology

- General term that indicates evolutionary relatedness among sequences
- Sequences are homologous if they are derived from a common ancestral sequence.

Things to consider

- To find the best alignment one needs to examine all possible alignment
- To reflect the quality of the possible alignments one needs to score them
- There can be different alignments with the same highest score
- Variations in the scoring scheme may change the ranking of alignments

Manual alignment

- When there are few gaps and the two sequences are not too different from each other, a reasonable alignment can be obtained by visual inspection.
- Advantages:
  (1) use of a powerful and trainable tool (the brain, well… some brains).
  (2) ability to integrate additional data

Disadvantage : The method is subjective and unscalable.

Types of Alignment

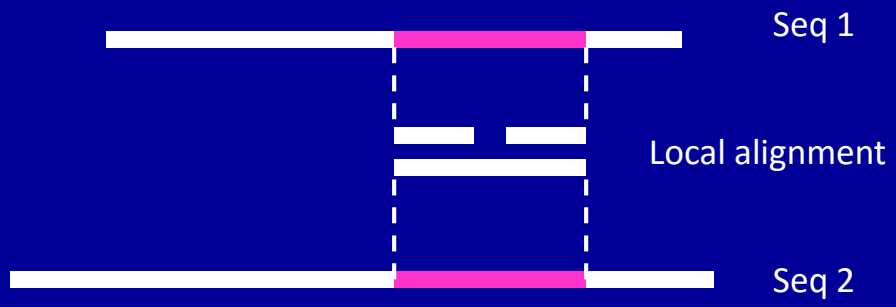
- Pairwise alignment
- Multiple alignment

- Local alignment
- Global alignment

Global versus local alignments

Global alignment: align full length of both sequences. (The “Needleman-Wunsch” algorithm).

Local alignment: find best partial alignment of two sequences (the “Smith-Waterman” algorithm).



Why Gap Penalties?

•The optimal alignment of two similar sequences is usually that which

- maximizes the number of matches and
- minimizes the number of gaps.
- There is a tradeoff between these two
  - adding gaps reduces mismatches

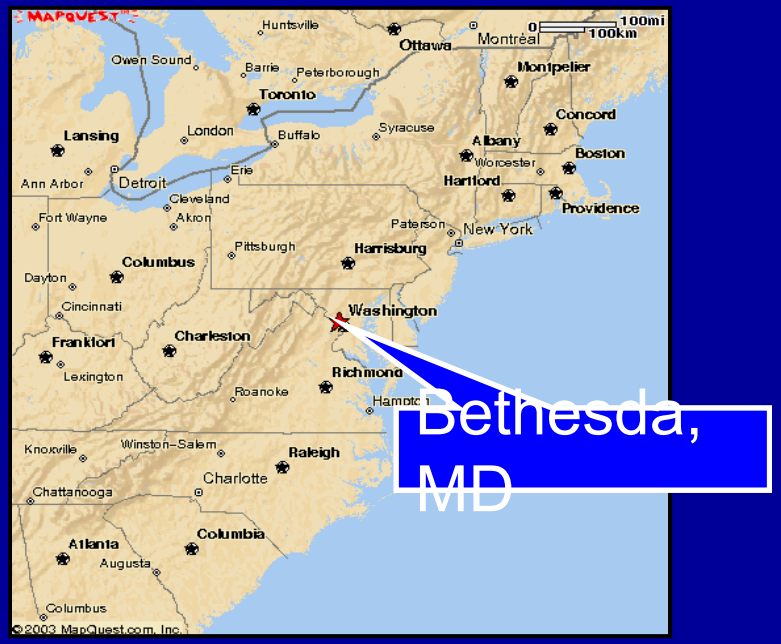
•Permitting the insertion of arbitrarily many gaps can lead to high scoring alignments of non-homologous sequences.

• Penalizing gaps forces alignments to have relatively few gaps.

NCBI Molecular Biology Resources

NCBI Databases

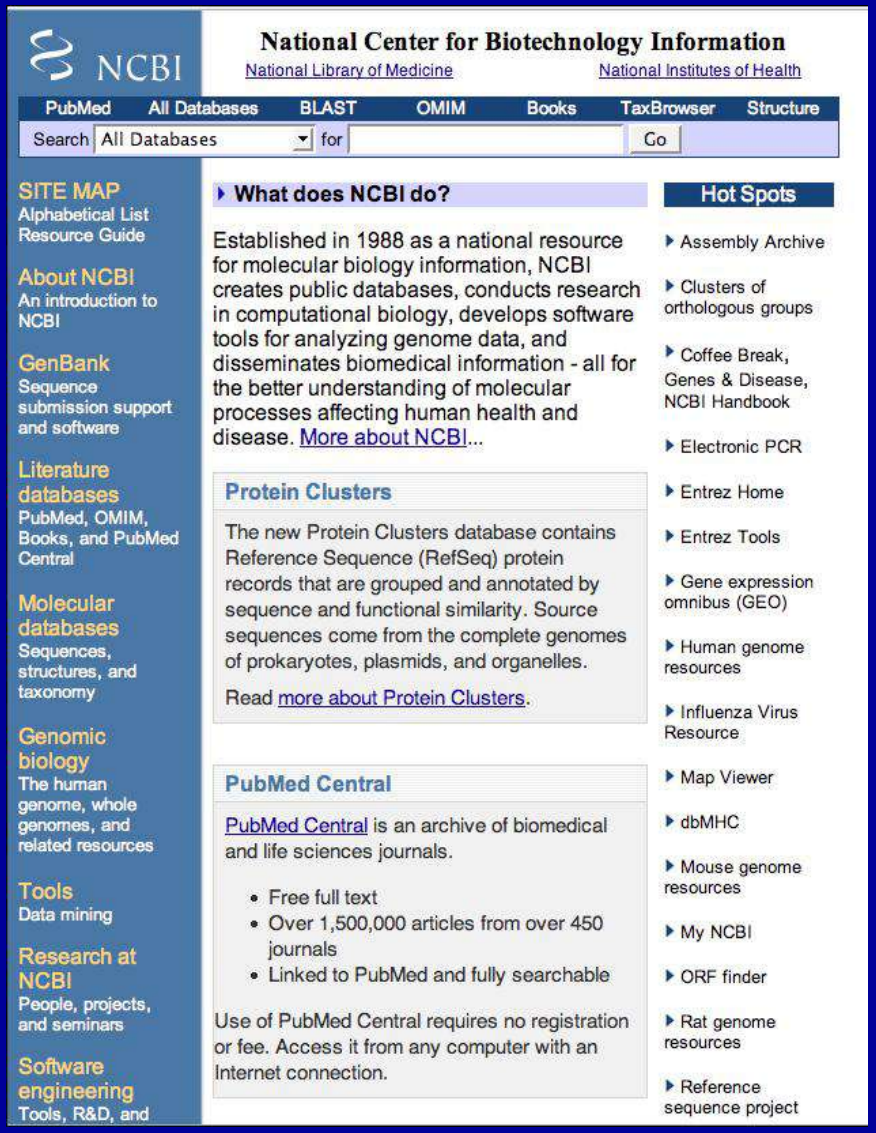
The National Center for Biotechnology Information



Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Web Access: www.ncbi.nlm.nih.gov



NCBI Databases and Services

- GenBank primary sequence database
- Free public access to biomedical literature
  - PubMed free Medline (3 million searches per day)
  - PubMed Central full text online access
- Entrez integrated molecular and literature databases
- BLAST highest volume sequence search service
  (100 – 200 K searches per day)
- VAST structure similarity searches
- Software and Databases

Types of Databases

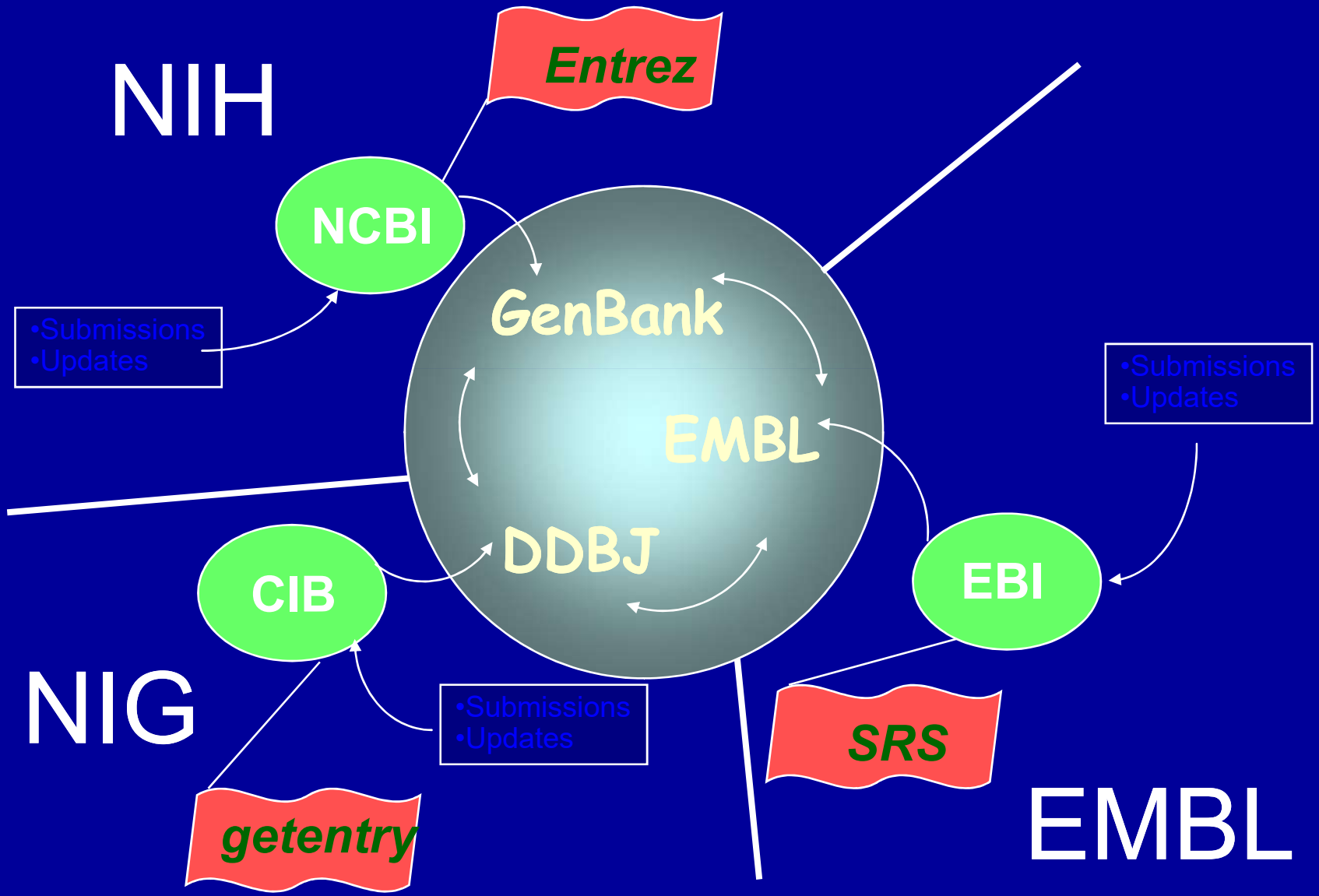
- Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the submitter
    - Examples: GenBank, SNP, GEO
- Derivative Databases
  - Built from primary data
  - Content controlled by third party (NCBI)
    - Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, Conserved Domain

What is GenBank?

NCBI's Primary Sequence Database

- Nucleotide only sequence database
- Archival in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - Redundant
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database

International Sequence Database Collaboration



Organization of GenBank: Traditional Divisions

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

Traditional Divisions:

- Direct Submissions
  (Sequin and BankIt)
- Accurate
- Well characterized

PRI Primate
PLN Plant and Fungal
BCT Bacterial and Archeal
INV Invertebrate
ROD Rodent
VRL Viral
VRT Other Vertebrate
MAM Mammalian
PHG Phage
SYN Synthetic (cloning vectors)
ENV Environmental Samples
UNA Unannotated

Entrez query: `gbdiv_xxx[Properties]`

Organization of GenBank: Bulk Divisions

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

BULK Divisions:

- Batch Submission
  (Email and FTP)
- Inaccurate
- Poorly characterized

EST Expressed Sequence Tag
GSS Genome Survey Sequence
HTG High Throughput Genomic
STS Sequence Tagged Site
HTC High Throughput cDNA
PAT Patent

Entrez query: `gbdiv_xxx[Properties]`

A Traditional GenBank Record

```
LOCUS       AF124527               2540 bp    mRNA    linear   PLN 29-JAN-2004
DEFINITION  Prunus persica ethylene receptor (ETR1) mRNA, complete cds.
ACCESSION   AF124527
VERSION     AF124527.1  GI:6841074
KEYWORDS    .
SOURCE      Prunus persica (peach)
  ORGANISM  Prunus persica
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Rosales; Rosaceae; Amygdaloideae; Prunus.
REFERENCE   1  (bases 1 to 2540)
  AUTHORS   Bassett,C.L., Artlip,T.S. and Callahan,A.M.
  TITLE     Characterization of the peach homologue of the ethylene receptor,
            PpETR1, reveals some unusual features regarding transcript
            processing
  JOURNAL   Planta 215 (4), 679-688 (2002)
   PUBMED   12172852
REFERENCE   2  (bases 1 to 2540)
  AUTHORS   Bassett,C.B., Artlip,T.S. and Nickerson,M.L.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-JAN-1999) Appalachian Fruit Research Station,
            USDA-ARS, 45 Wiltshire Road, Kearneysville, WV 25430, USA
FEATURES             Location/Qualifiers
     source          1..2540
                     /organism="Prunus persica"
                     /mol_type="mRNA"
                     /cultivar="Loring"
                     /db_xref="taxon:3760"
                     /dev_stage="III B/C fruit"
     gene            1..2540
                     /gene="ETR1"
     CDS             269..2485
                     /gene="ETR1"
                     /codon_start=1
                     /product="ethylene receptor"
                     /protein_id="AAF28893.1"
                     /db_xref="GI:6841075"
                     /translation="MEACNCIEPQWPADELLMKYQYISDFFIALAYFSIPLELIYFVK
                     KSAVFPYRWVLVQFGAFIVLCGATHLINLWTFSMHSRTVAIVMTTAKVLTAVVSCATA
                     LMLVHIIPDLLSVKTRELFLKNKAAELDREMGLIRTQEETGRHVRMLTHEIRSTLDRH
                     TILKTTLVELGRTLALEECALWMPTRTGLELQLSYTLRQQNPVGYTVPIHLPVINQVF
                     SSNRALKISPNSPVARMRPLAGKHMPGEVVAVRVPLLHLSNFQINDWPELSTKRYALM
                     VLMLPSDSARQWHVHELELVEVVADQVAVALSHAAILEESMRARDLLMEQNIALDLAR
                     REAETAIRARNDFLAVMNHEMRTPMHAIIALSSLLQETELTPEQRLMVETILKSSHLL
                     ATLINDVLDLSRLEDGSLQLEIATFNLHSVFREVHNLIKPVASVKKLSVSLNLAADLP
                     VQAVGDEKRLMQIVLNVVGNAVKFSKEGSISITAFVAKSESLRDFRAPEFFPAQSDNH
                     FYLRVQVKDSGSGINPQDIPKLFTKFAQTQSLATRNSGGSGLGLAICKRFVNLMEGHI
                     WIESEGPGKGCTAIFIVKLGFAERSNESKLPFLTKVQANHVQTNFPGLKVLVMDDNGS
                     VTKGLLVHLGCDVTTVSSIDEFLHVISQEHKVVFMDVCMPGIDGYELAVRIHEKFTKR
                     HERPVLVALTGNIDKMTKENCMRVGMDGVILKPVSVDKMRSVLSELLEHRVLFEAM"
ORIGIN
        1 gcacgagggc tcaccgagcg agctagctct tcaggagtca aggcttctgg gtgaggggaa
       61 gaagaagaag cttctttgat gtgttggggt gccaatctaa agaggaagaa gaaggcctct
      121 aatgtattga ggtcggctgt ctgggctgcc gatctgtgtt gaatggatag tttggtagag
      181 atgcttcaac gacatagggt ggctgaaaag ggtttgaaga aagtgaagga ggaaaccaag
                                        ...
     2401 tatactgaaa cctgtctcag ttgataaaat gaggagtgtt ttatcagaac tgttggagca
     2461 tcgagtttta tttgaggcta tgtaagatat aggaaaattg ttctagtgaa ggaaagattt
     2521 aaatggaaaa aaaaaaaaaa
//
```

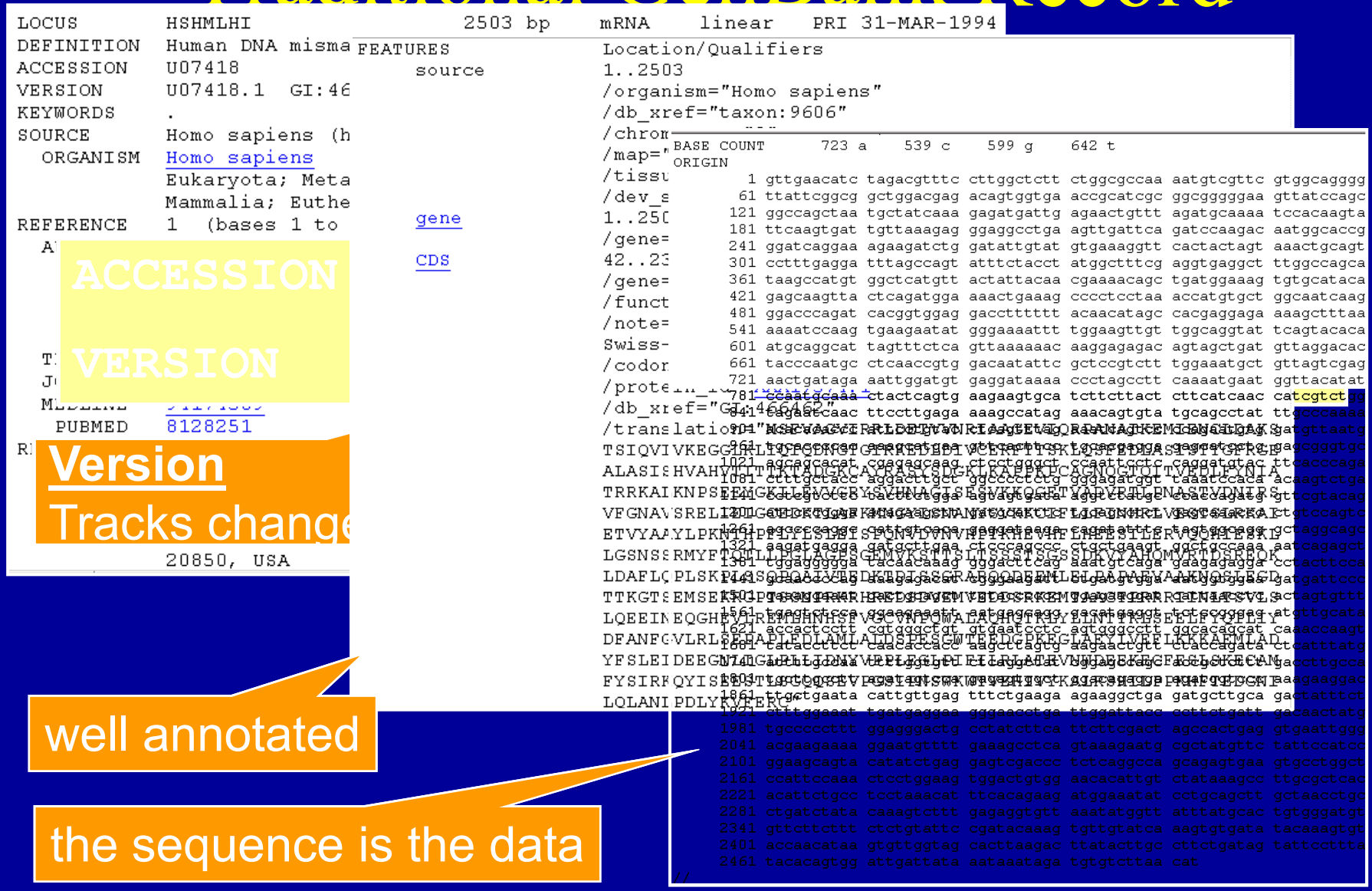
Header

The Flatfile Format

Feature Table

Sequence

Traditional GenBank Record



Version

Tracks change

well annotated

the sequence is the data

Accessing information on molecular sequences

Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences.
You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.

What is an accession number?

An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**protein**
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

Four ways to access DNA and protein sequences

[1] Entrez Gene with RefSeq
[2] UniGene
[3] European Bioinformatics Institute (EBI) and Ensembl (separate from NCBI)
[4] ExPASy Sequence Retrieval System (separate from NCBI)

Note: LocusLink at NCBI was recently retired. The third printing of the book has updated these sections (pages 27-31).

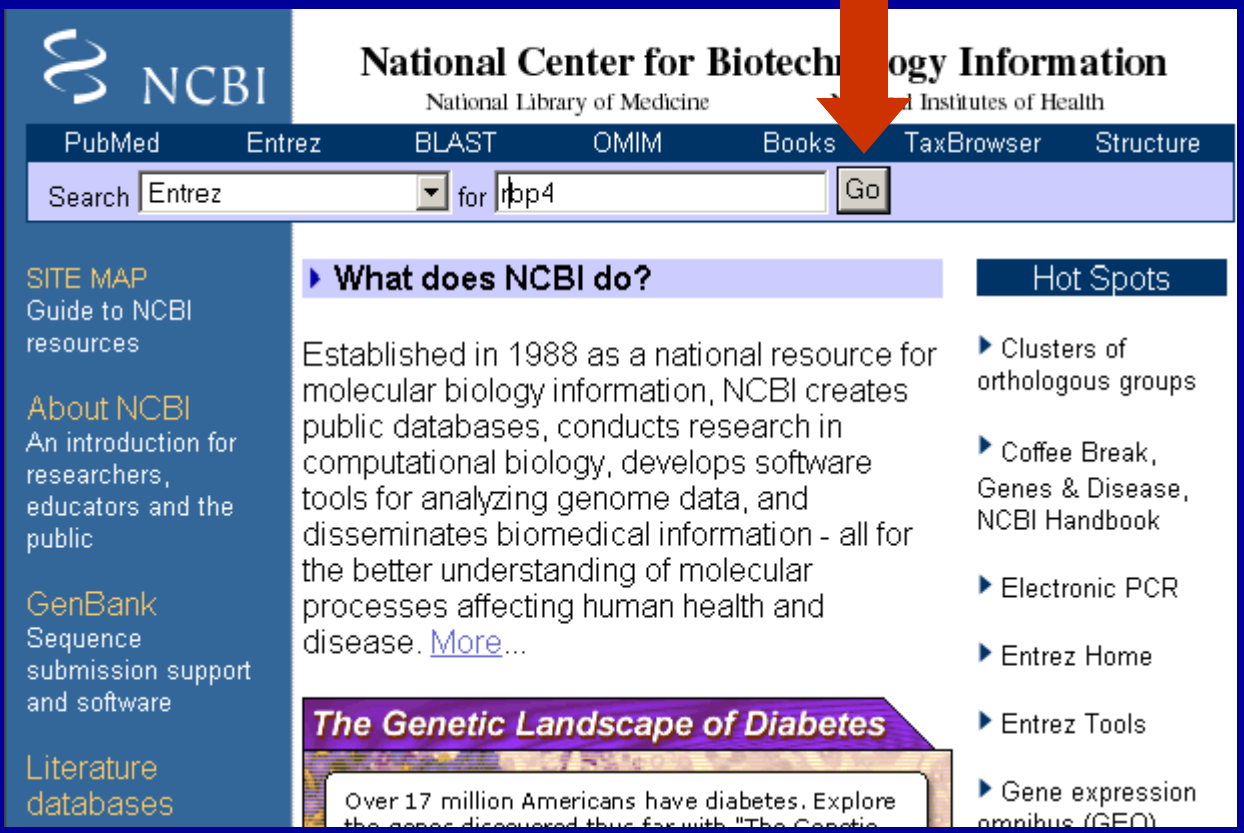
4 ways to access protein and DNA sequences

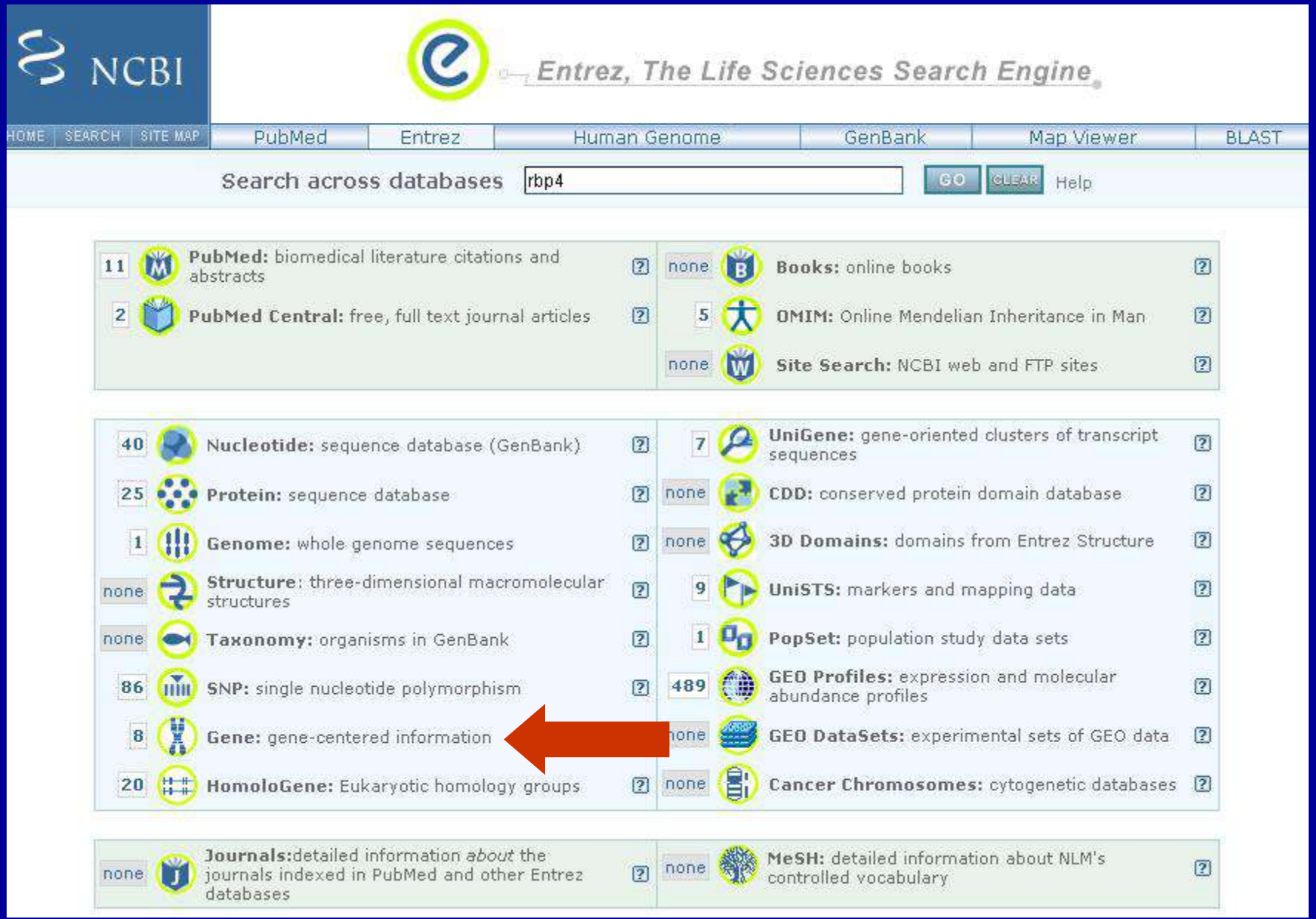
[1] Entrez Gene with RefSeq

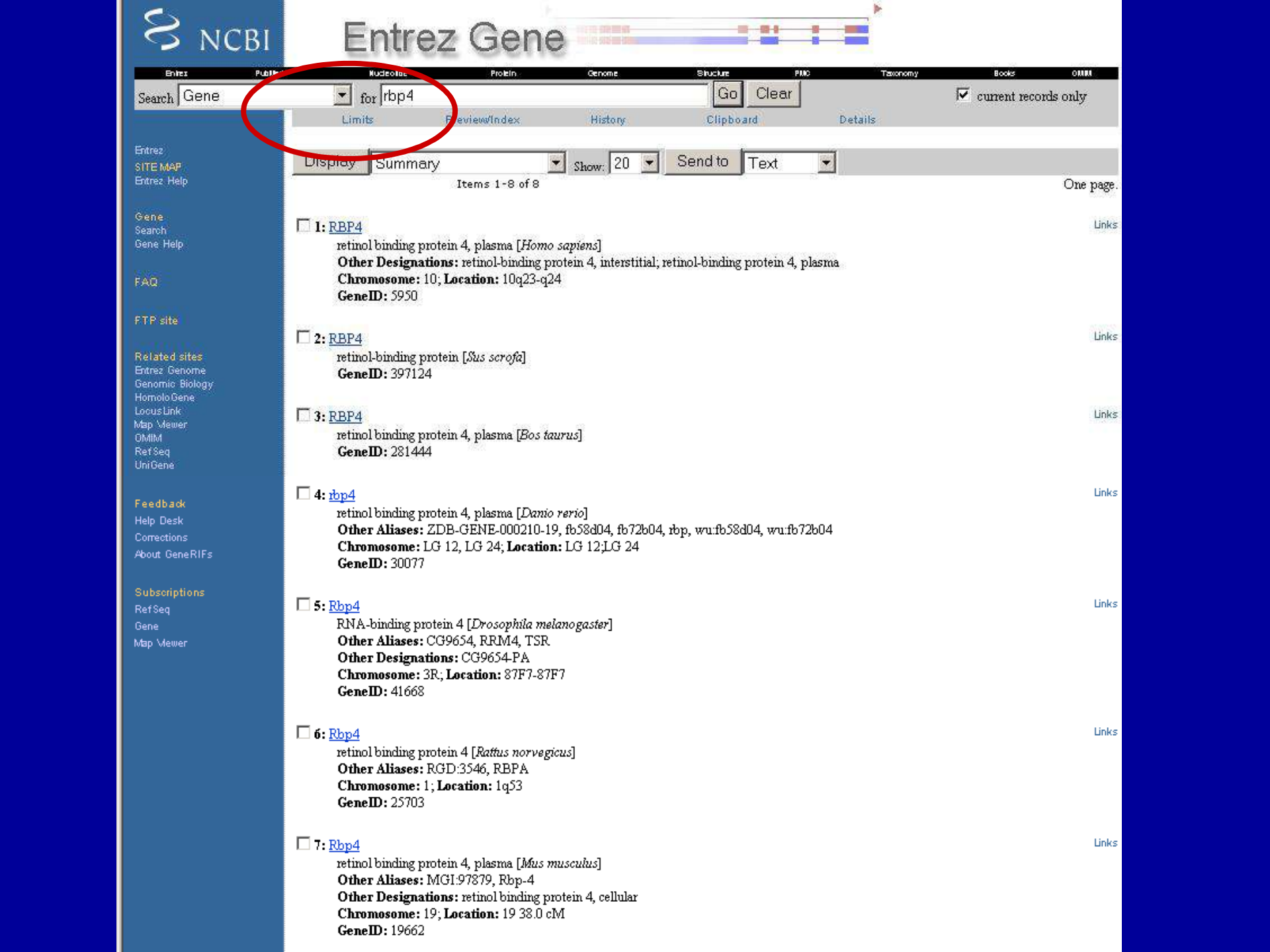
Entrez Gene is a great starting point: it collects key information on each gene/protein from major databases. It covers all major organisms.

RefSeq provides a curated, optimal accession number for each DNA (NM_006744)
or protein (NP_007635)

From the NCBI home page, type “rbp4” and hit “Go”




NCBI
Entrez, The Life Sciences Search Engine
HOME
SEARCH
SITE MAP
PubMed
Entrez
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
rbp4
GO
CLEAR
Help
11 PubMed: biomedical literature citations and abstracts
2 PubMed Central: free, full text journal articles
none Books: online books
5 OMIM: Online Mendelian Inheritance in Man
none Site Search: NCBI web and FTP sites
40 Nucleotide: sequence database (GenBank)
25 Protein: sequence database
1 Genome: whole genome sequences
none Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
86 SNP: single nucleotide polymorphism
8 Gene: gene-centered information
20 HomoloGene: Eukaryotic homology groups
7 UniGene: gene-oriented clusters of transcript sequences
none CDD: conserved protein domain database
none 3D Domains: domains from Entrez Structure
9 UniSTS: markers and mapping data
1 PopSet: population study data sets
489 GEO Profiles: expression and molecular abundance profiles
none GEO DataSets: experimental sets of GEO data
none Cancer Chromosomes: cytogenetic databases
none Journals: detailed information about the journals indexed in PubMed and other Entrez databases
none MeSH: detailed information about NLM's controlled vocabulary


NCBI
Entrez Gene
Search Gene for rbp4 Go Clear current records only
Limits Preview/Index History Clipboard Details
Entrez
SITE MAP
Entrez Help
Gene
Search
Gene Help
FAQ
FTP site
Related sites
Entrez Genome
Genomic Biology
HomoloGene
LocusLink
Map Viewer
OMIM
RefSeq
UniGene
Feedback
Help Desk
Corrections
About GeneRIFs
Subscriptions
RefSeq
Gene
Map Viewer
Display Summary Show: 20 Send to Text
Items 1-8 of 8
One page.
1: RBP4
retinol binding protein 4, plasma [Homo sapiens]
Other Designations: retinol-binding protein 4, interstitial; retinol-binding protein 4, plasma
Chromosome: 10; Location: 10q23-q24
GeneID: 5950
2: RBP4
retinol-binding protein [Sus scrofa]
GeneID: 397124
3: RBP4
retinol binding protein 4, plasma [Bos taurus]
GeneID: 281444
4: rbp4
retinol binding protein 4, plasma [Danio rerio]
Other Aliases: ZDB-GENE-000210-19, fb58d04, fb72b04, rbp, wu:fb58d04, wu:fb72b04
Chromosome: LG 12, LG 24; Location: LG 12;LG 24
GeneID: 30077
5: Rbp4
RNA-binding protein 4 [Drosophila melanogaster]
Other Aliases: CG9654, RRM4, TSR
Other Designations: CG9654-PA
Chromosome: 3R; Location: 87F7-87F7
GeneID: 41668
6: Rbp4
retinol binding protein 4 [Rattus norvegicus]
Other Aliases: RGD:3546, RBPA
Chromosome: 1; Location: 1q53
GeneID: 25703
7: Rbp4
retinol binding protein 4, plasma [Mus musculus]
Other Aliases: MGI:97879, Rbp-4
Other Designations: retinol binding protein 4, cellular
Chromosome: 19; Location: 19 38.0 cM
GeneID: 19662
Links

NCBI

Entrez Gene

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books | OMIM

Search Gene for rbp4 Go Clear current records only

Limits | Preview/Index | History | Clipboard | Details

- To apply your query to all fields, leave this page unchanged.
- Boolean operators AND, OR, NOT must be in upper case.
- To limit your search to a specific field, select:
  All Fields
  or use search field tags enclosed in square brackets, e.g. aaa[title].
- You may also limit your search by checking boxes below.

Exclude: clear

☐ Mitochondria ☐ Plasmids ☐ Plastids ☐ Pseudogenes ☐ RefSeqs ☐ NEWENTRY

Include Only: clear

☐ Genomic ☐ Mitochondria ☐ Plasmids ☐ Plastids ☐ RefSeqs ☐ NEWENTRY

Limit by RefSeq Status: clear

☐ Inferred ☐ Known ☐ Model ☐ Predicted ☐ Provisional ☐ Reviewed ☐ Validated

Limit by Taxonomy: clear

☐ Mammalia
- ☐ Bos taurus
- ☐ Canis familiaris
- ☑ Homo sapiens
- ☐ Mus musculus
- ☐ Pan troglodytes
- ☐ Rattus norvegicus
- ☐ Sus scrofa

☐ Non-mammalian vertebrates
- ☐ Danio rerio
- ☐ Gallus gallus
- ☐ Xenopus laevis
- ☐ Xenopus tropicalis

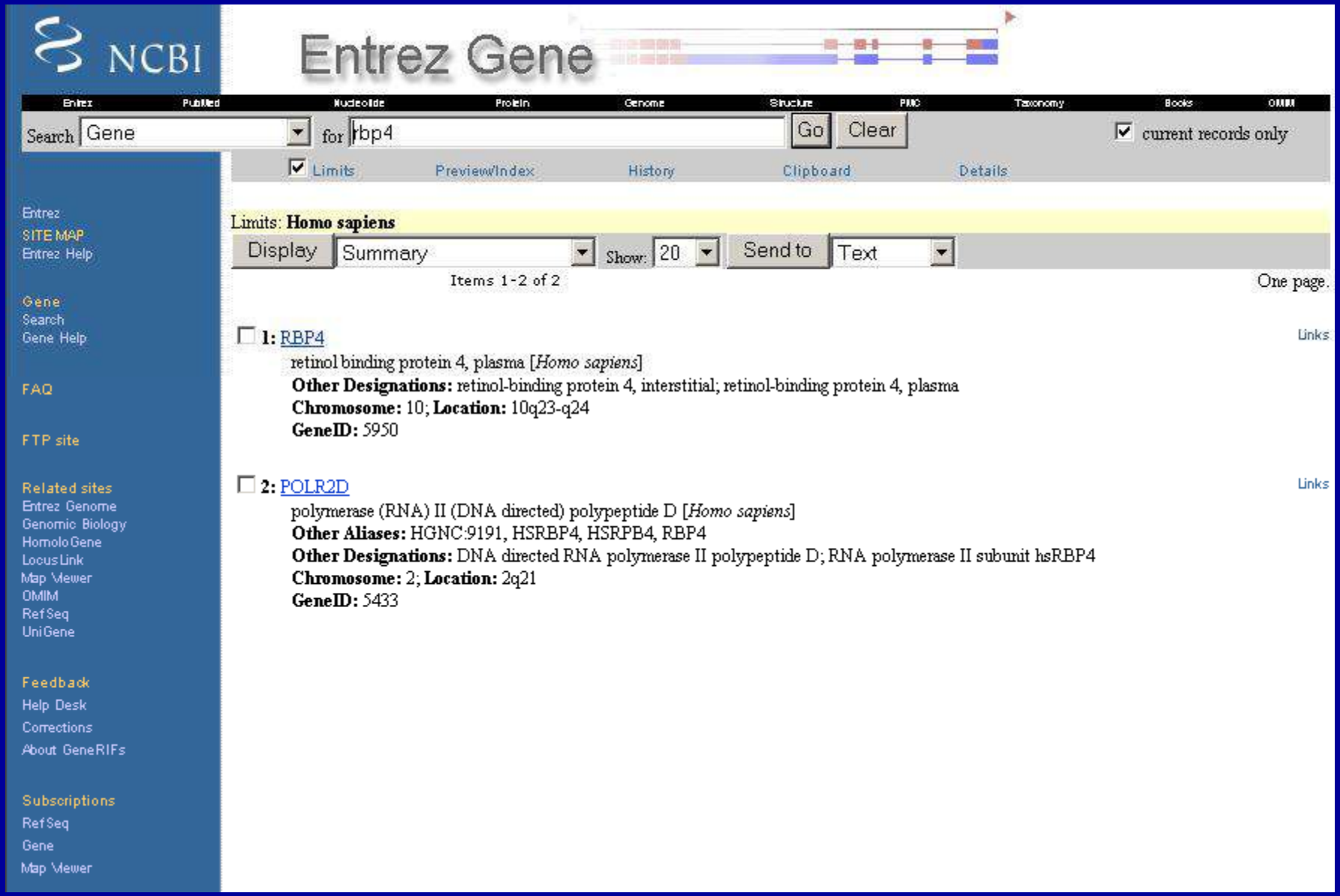
☐ Invertebrates
- ☐ Anopheles gambiae
- ☐ Caenorhabditis elegans
- ☐ Drosophila melanogaster
- ☐ Plasmodium falciparum

☐ Plants
- ☐ Arabidopsis thaliana
- ☐ Oryza sativa

☐ Fungi
- ☐ Magnaporthe grisea
- ☐ Neurospora crassa
- ☐ Saccharomyces cerevisiae
- ☐ Schizosaccharomyces pombe

☐ Bacteria
- ☐ Actinobacteria
- ☐ Bacillus
- ☐ Chlamydia
- ☐ Cyanobacteria
- ☐ Escherichia coli
- ☐ Mycobacterium tuberculosis
- ☐ Mycoplasmatales
- ☐ Proteobacteria
- ☐ Pseudomonas
- ☐ Salmonella
- ☐ Spirochaetes
- ☐ Streptococcus
- ☐ Yersinia

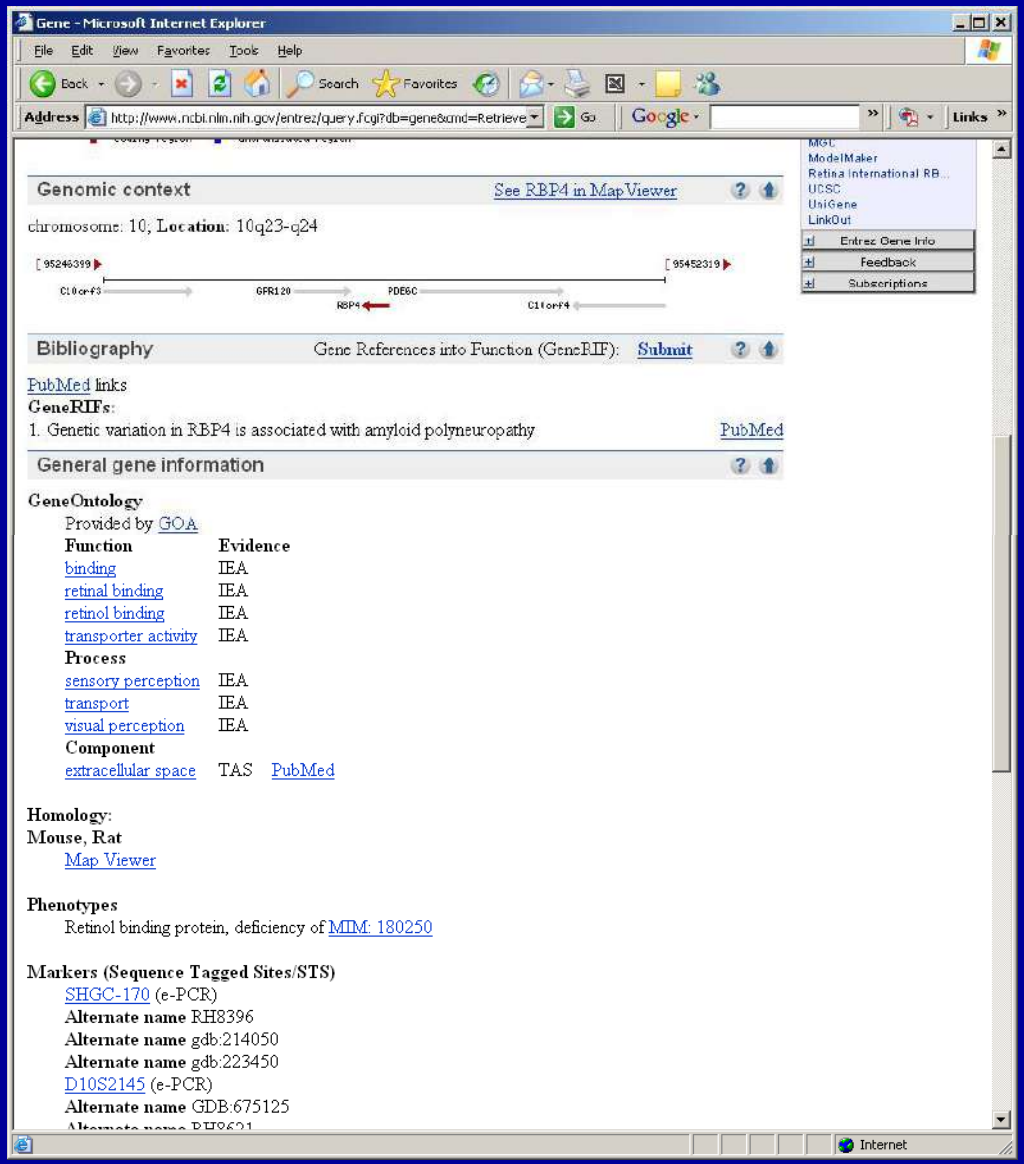
By applying limits, there are now just two entries



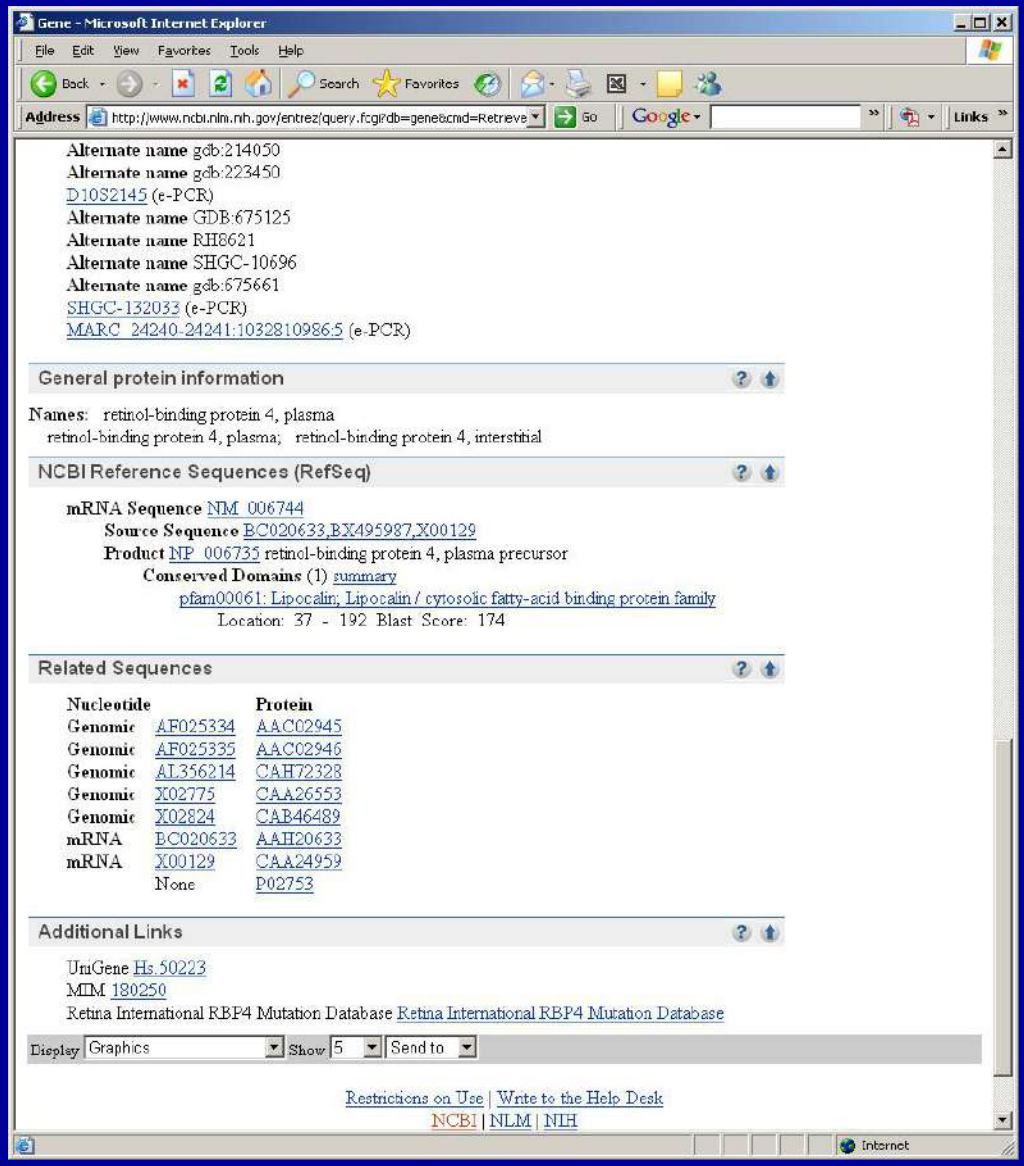
Entrez Gene (top of page)

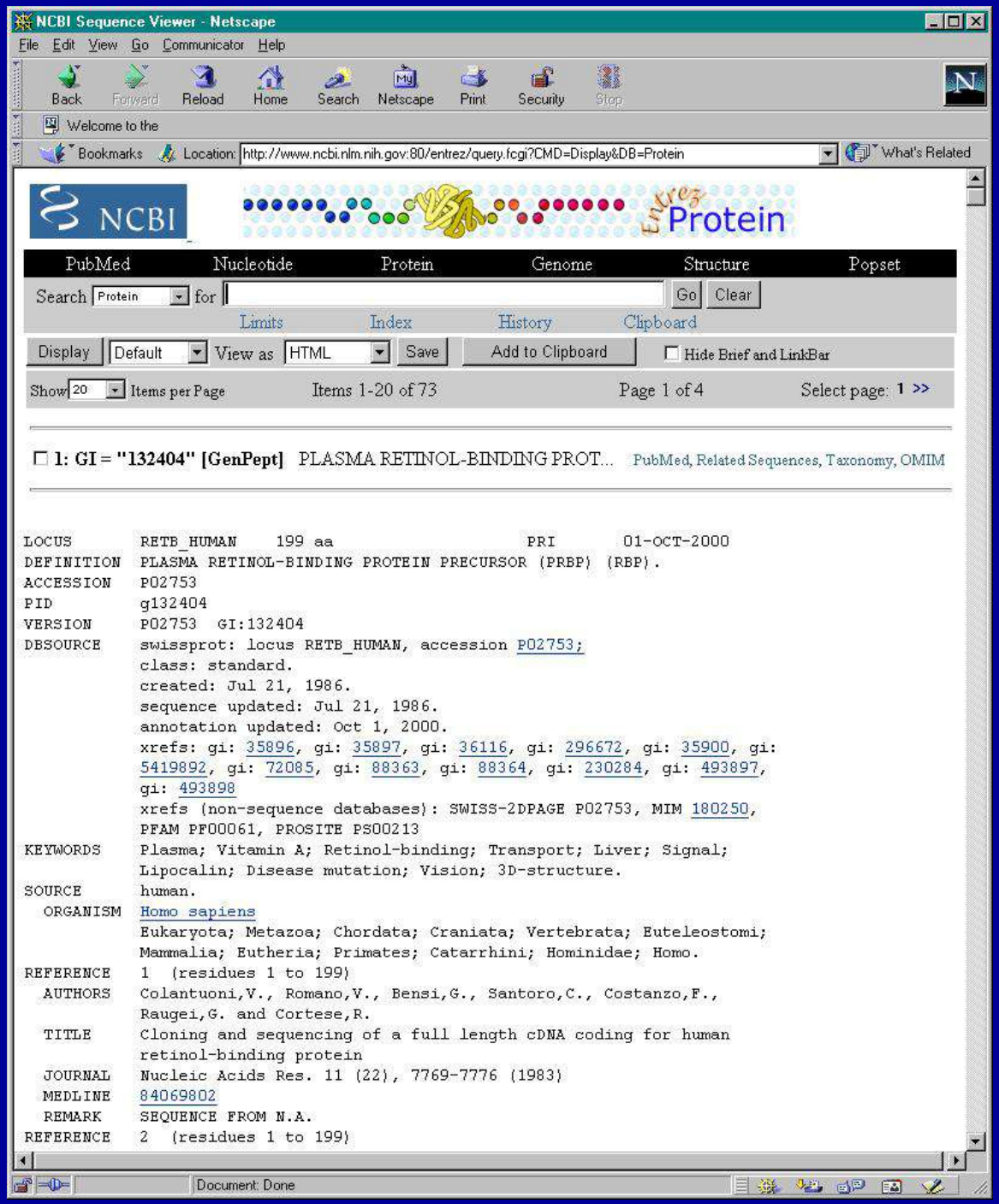
Note that links to many other RBP4 database entries are available

Entrez Gene (middle of page)



Entrez Gene (bottom of page)




NCBI Sequence Viewer - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Welcome to the
Bookmarks Location: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=Display&DB=Protein What's Related
NCBI
Entrez Protein
PubMed Nucleotide Protein Genome Structure Popset
Search Protein for Go Clear
Limits Index History Clipboard
Display Default View as HTML Save Add to Clipboard Hide Brief and LinkBar
Show 20 Items per Page Items 1-20 of 73 Page 1 of 4 Select page: 1 >>
1: GI = "132404" [GenPept] PLASMA RETINOL-BINDING PROT... PubMed, Related Sequences, Taxonomy, OMIM
LOCUS RETB_HUMAN 199 aa PRI 01-OCT-2000
DEFINITION PLASMA RETINOL-BINDING PROTEIN PRECURSOR (PRBP) (RBP).
ACCESSION P02753
PID g132404
VERSION P02753 GI:132404
DBSOURCE swissprot: locus RETB_HUMAN, accession P02753;
class: standard.
created: Jul 21, 1986.
sequence updated: Jul 21, 1986.
annotation updated: Oct 1, 2000.
xrefs: gi: 35896, gi: 35897, gi: 36116, gi: 296672, gi: 35900, gi:
5419892, gi: 72085, gi: 88363, gi: 88364, gi: 230284, gi: 493897,
gi: 493898
xrefs (non-sequence databases): SWISS-2DPAGE P02753, MIM 180250,
PFAM PF00061, PROSITE PS00213
KEYWORDS Plasma; Vitamin A; Retinol-binding; Transport; Liver; Signal;
Lipocalin; Disease mutation; Vision; 3D-structure.
SOURCE human.
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (residues 1 to 199)
AUTHORS Colantuoni,V., Romano,V., Bensi,G., Santoro,C., Costanzo,F.,
Raugei,G. and Cortese,R.
TITLE Cloning and sequencing of a full length cDNA coding for human
retinol-binding protein
JOURNAL Nucleic Acids Res. 11 (22), 7769-7776 (1983)
MEDLINE 84069802
REMARK SEQUENCE FROM N.A.
REFERENCE 2 (residues 1 to 199)
Document: Done

```
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
            Eutheria; Lagomorpha; Leporidae; Oryctolagus.
REFERENCE   1  (residues 1 to 201)
  AUTHORS   Sundelin,J., Laurent,B.C., Anundi,H., Tragardh,L., Larhammar,D.,
            Bjorck,L., Eriksson,U., Akerstrom,B., Jones,A., Newcomer,M.,
            Peterson,P.A. and Rask,L.
  TITLE     Amino acid sequence homologies between rabbit, rat, and human serum
            retinol-binding proteins
  JOURNAL   J. Biol. Chem. 260 (10), 6472-6480 (1985)
  MEDLINE   85207643
REFERENCE   2  (residues 1 to 201)
  AUTHORS   Lee,S.Y., Ubels,J.L. and Soprano,D.R.
  TITLE     The lacrimal gland synthesizes retinol-binding protein
  JOURNAL   Exp. Eye Res. 55 (1), 163-171 (1992)
  MEDLINE   93011736
COMMENT     On May 31, 1997 this sequence version replaced gi:72086.
            For general comments see the entry for the human protein
            (PIR:VAHU).
FEATURES             Location/Qualifiers
     source          1..201
                     /organism="Oryctolagus cuniculus"
                     /db_xref="taxon:9986"
     Protein         1..201
                     /product="retinol-binding protein precursor"
     Region          1..18
                     /region_name="domain"
                     /note="signal sequence"
     Region          19..201
                     /region_name="product"
                     /note="retinol-binding protein"
     Bond            bond(22,178)
                     /bond_type="disulfide"
     Region          33..192
                     /region_name="domain"
                     /note="lipocalin homology #label LIP"
     Bond            bond(88,192)
                     /bond_type="disulfide"
     Bond            bond(138,147)
                     /bond_type="disulfide"
ORIGIN
        1 mewvwalvll aalgsgrger dcrvssfrvk enfdkarfag twyamakkdp eglflqdniv
       61 aefsvdengh msatakgrvr llnnwdvcad mvgtftdted pakfkmkywg vasflqrgnd
      121 dhwiidtdyd tfavqyscrl lnfdgtcads ysfvfsrdph glppdvqklv rqrqeelcls
      181 rqyrlivhng ycddksvrnl l
//
```

NCBI Sequence Viewer - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Stop

Welcome to the

Bookmarks Location: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=Display&DB=Protein What's Related

NCBI Entrez Protein

PubMed Nucleotide Protein Genome Structure Popset

Search Protein for Go Clear

Limits Index History Clipboard

Display Default View as HTML Save Add to Clipboard Hide Brief and LinkBar

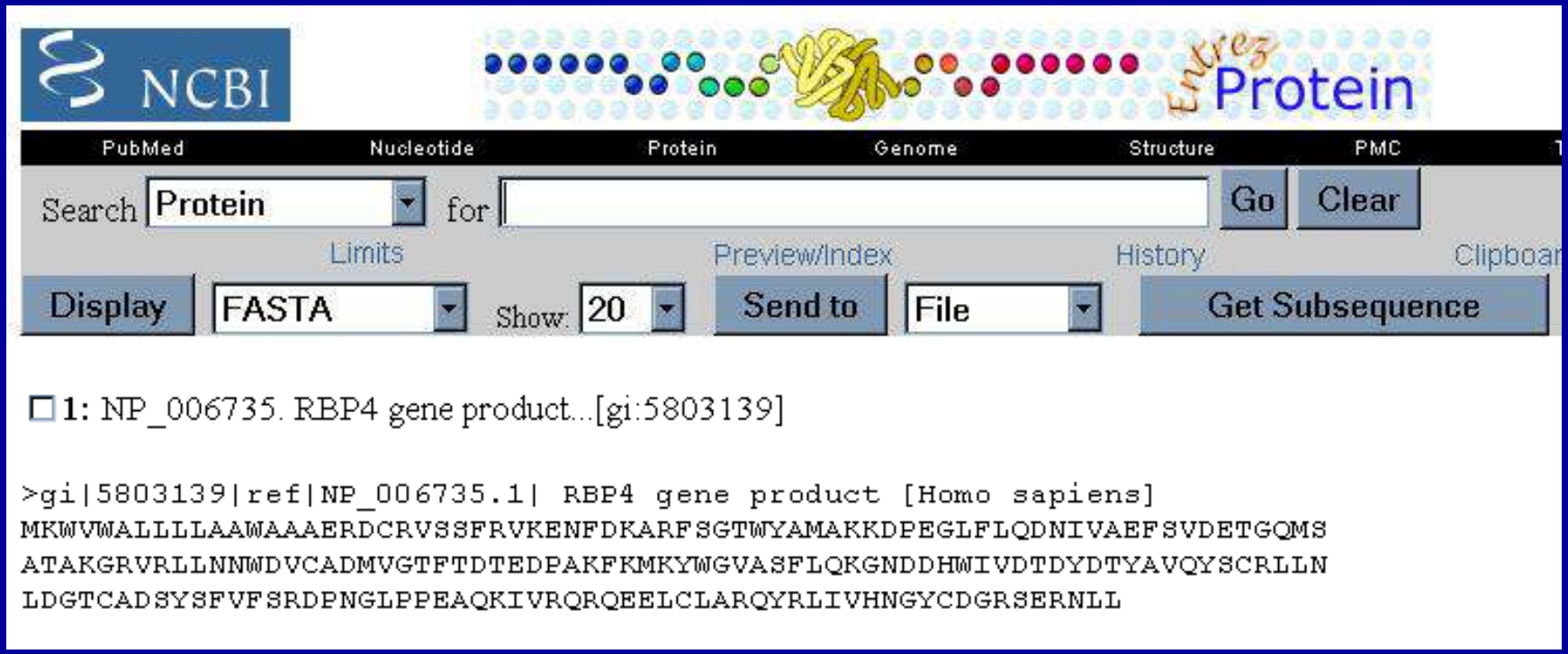
Show 20 Items per Page Items 1-20 of 73 Page 1 of 4 Select page: 1 >>

1: GI = "132404" [GenPept] PLASMA RETINOL-BINDING PROT... PubMed, Related Sequences, Taxonomy, OMIM

```
LOCUS       RETB_HUMAN    199 aa                    PRI       01-OCT-2000
DEFINITION  PLASMA RETINOL-BINDING PROTEIN PRECURSOR (PRBP) (RBP).
ACCESSION   P02753
PID         g132404
VERSION     P02753  GI:132404
DBSOURCE    swissprot: locus RETB_HUMAN, accession P02753;
            class: standard.
            created: Jul 21, 1986.
            sequence updated: Jul 21, 1986.
            annotation updated: Oct 1, 2000.
            xrefs: gi: 35896, gi: 35897, gi: 36116, gi: 296672, gi: 35900, gi:
            5419892, gi: 72085, gi: 88363, gi: 88364, gi: 230284, gi: 493897,
            gi: 493898
            xrefs (non-sequence databases): SWISS-2DPAGE P02753, MIM 180250,
            PFAM PF00061, PROSITE PS00213
KEYWORDS    Plasma; Vitamin A; Retinol-binding; Transport; Liver; Signal;
            Lipocalin; Disease mutation; Vision; 3D-structure.
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 199)
  AUTHORS   Colantuoni,V., Romano,V., Bensi,G., Santoro,C., Costanzo,F.,
            Raugei,G. and Cortese,R.
  TITLE     Cloning and sequencing of a full length cDNA coding for human
            retinol-binding protein
  JOURNAL   Nucleic Acids Res. 11 (22), 7769-7776 (1983)
  MEDLINE   84069802
  REMARK    SEQUENCE FROM N.A.
REFERENCE   2  (residues 1 to 199)
```

Document: Done

FASTA format



UNIT - III

Introduction and Biological Databases

Overview

- Introduction
  - What Is Bioinformatics?
  - Goal
  - Scope
  - Applications
  - Limitations

- Introduction to biological Databases
- What Is a Database?
- Types of Databases
- Biological Databases
- Pitfalls of Biological databases
- Information Retrieval from Biological databases

What is Bioinformatics?

- Bioinformatics involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules
- Bioinformatics & Computational Biology

Goal

- Better understand a living cell and how it functions at the molecular level

Scope

- The development of computational tools and databases
- The application of these tools and databases in generating biological knowledge

Scope

- Tools development:
  - Writing software for sequence, structural, and functional analysis
  - Construction and curating of biological databases

Tools: Used in three areas

- Molecular Sequence Analysis
- Molecular Structural Analysis
- Molecular Functional Analysis

Sequence Analysis

- Sequence Alignment
- Sequence Database Searching
- Motif and Pattern Discovery
- Gene and Promoter Finding
- Reconstruction of Evolutionary Relationships
- ...

Structural Analysis

- Protein and nucleic acid structure
  - Analysis
  - Comparison
  - Classification
  - Prediction

Functional Analysis

- Gene Expression Profiling
- Protein– Protein Interaction Prediction
- Protein Sub cellular Localization Prediction
- Metabolic Pathway Reconstruction
- ...

Applications

- Drug design
- Agricultural biotechnology
- Forensic DNA analysis

Limitations

- Fighting a battle without intelligence is inefficient and dangerous

Introduction to Biological Databases

What is a Database?

- Type of Databases:
  - Relational Databases
  - Object-Oriented Databases

Biological Databases

- Primary Databases
- Secondary Databases

Databases in Bioinformatics

- Sequence databases
- Sequence analysis
- Functional genomics
- Literature databases
- Structural databases
- Metabolic pathway databases
- Specialized databases

Pitfalls of Biological Databases

- Errors in Sequence Databases
- Redundancy in the Primary Sequence Databases
- False or Incomplete Genes Annotations

Errors in Nucleotide Sequences

- sequencing errors
- frame-shifts
- Contaminated with sequences from cloning vectors
  - Exceptional Care for sequences produced before the 1990s

Redundancy

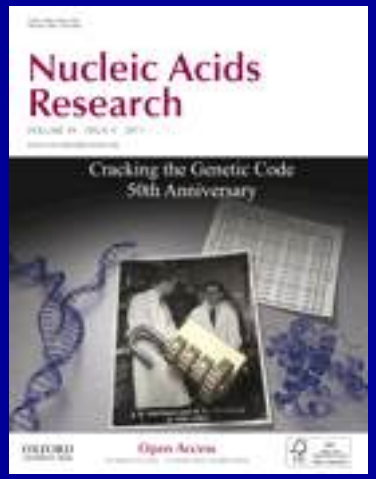
- repeated submission
  - identical or overlapping sequences by the same or different authors
  - revision of annotations
  - dumping of expressed sequence tags (EST) data
  - poor database management

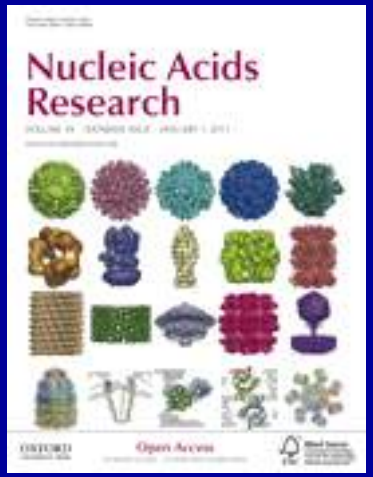
Bioinformatics Databases

- Growing steadily in number
  - Growing amazingly in size
- Specialization
  - Which genome they contain (mouse, human, all of them)
  - Which types of information about the genome they contain
- Contain information such as
  - Sequences: of bases and of residues
  - Structure: 3d conformations of known proteins
  - Families: Which sets of genes are known to be homologous
  - Annotations: which processes each gene is involved in
    - And lots of other information

The definitive source….

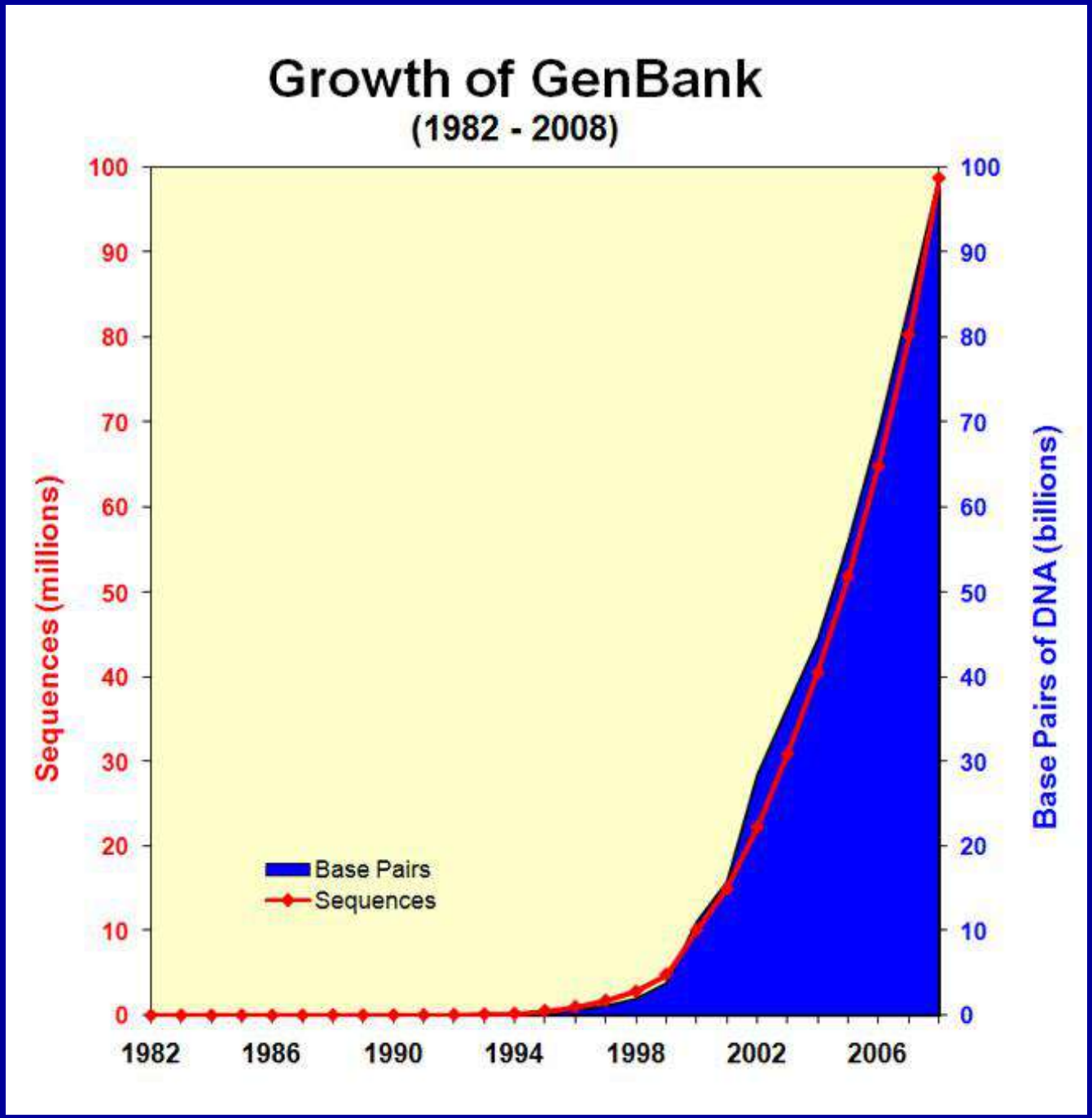
- More than 1300 DB
- http://nar.oxfordjournals.org/content/39/suppl_1.toc





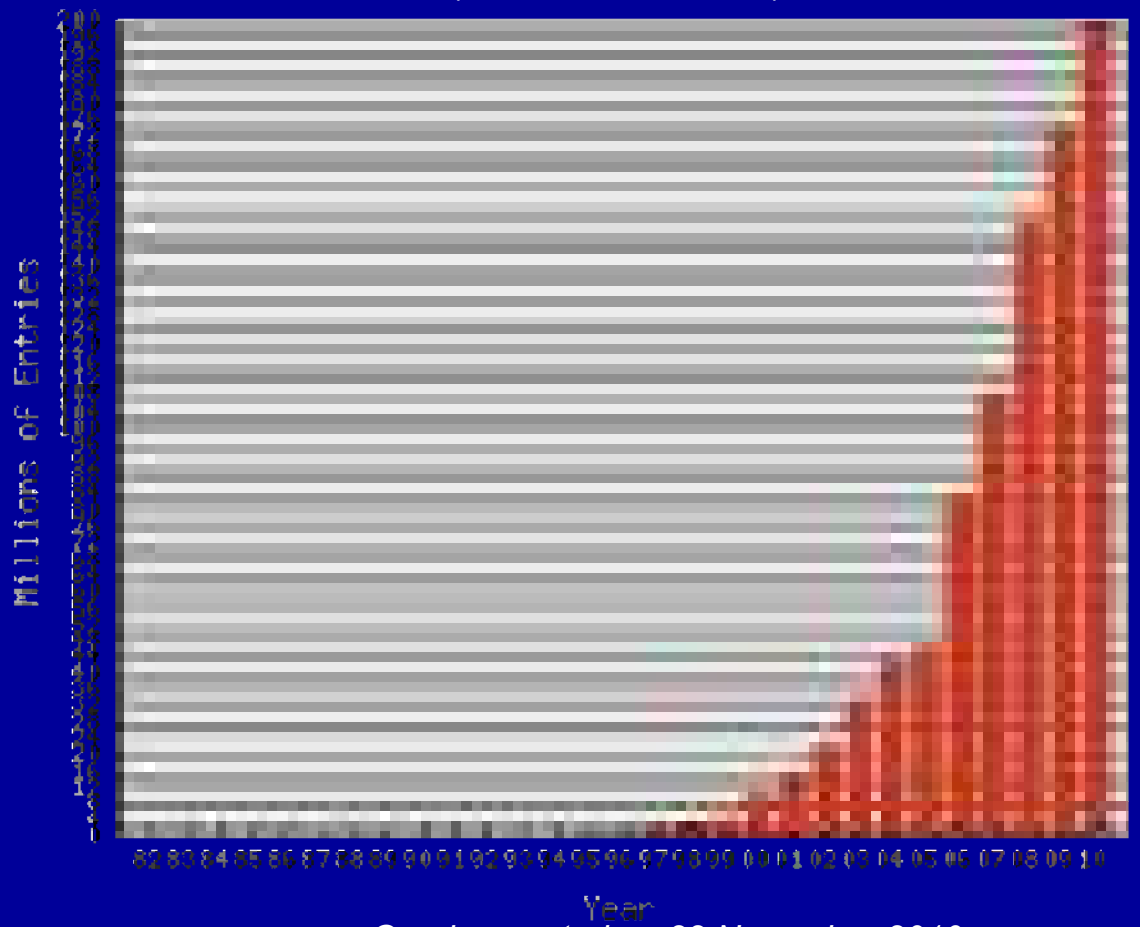
DNA Sequence databases

- Main repositories:
  - GenBank (US)
    - (http://www.ncbi.nlm.nih.gov/Genbank/index.html)
  - EMBL (Europe)
    - (http://www.ebi.ac.uk/embl/)
  - DDBJ (Japan)
    - (http://www.ddbj.nig.ac.jp/)
- Primary databases
  - DNA sequences are identical


Growth of GenBank
(1982 - 2008)
Sequences (millions)
Base Pairs of DNA (billions)
100
90
80
70
60
50
40
30
20
10
0
1982
1986
1990
1994
1998
2002
2006
Base Pairs
Sequences

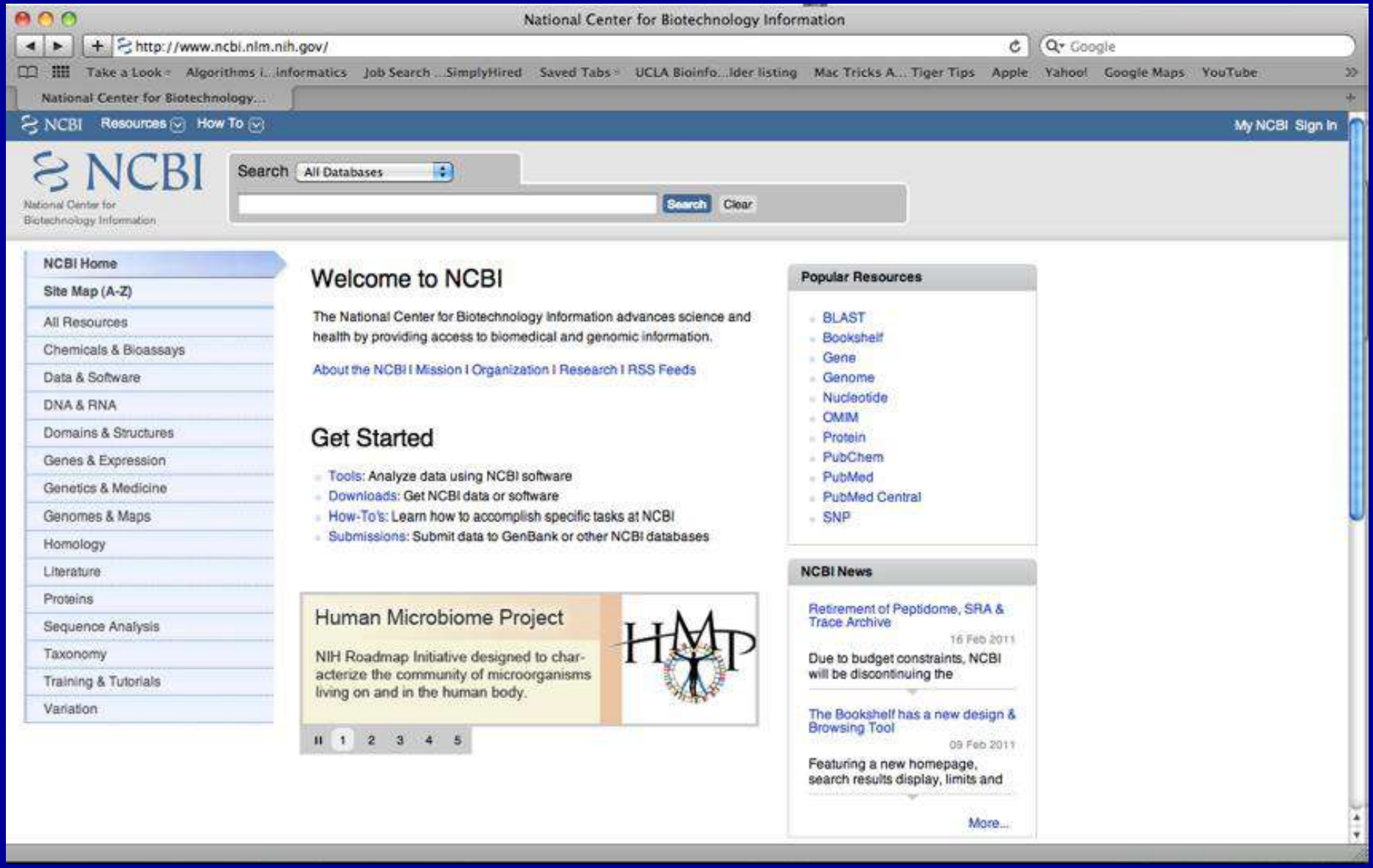
EMBL Database

Number of entries
(current 199,575,971)



Graphs created on 22 November 2010

http://www.ebi.ac.uk/embl/Services/DBStats/

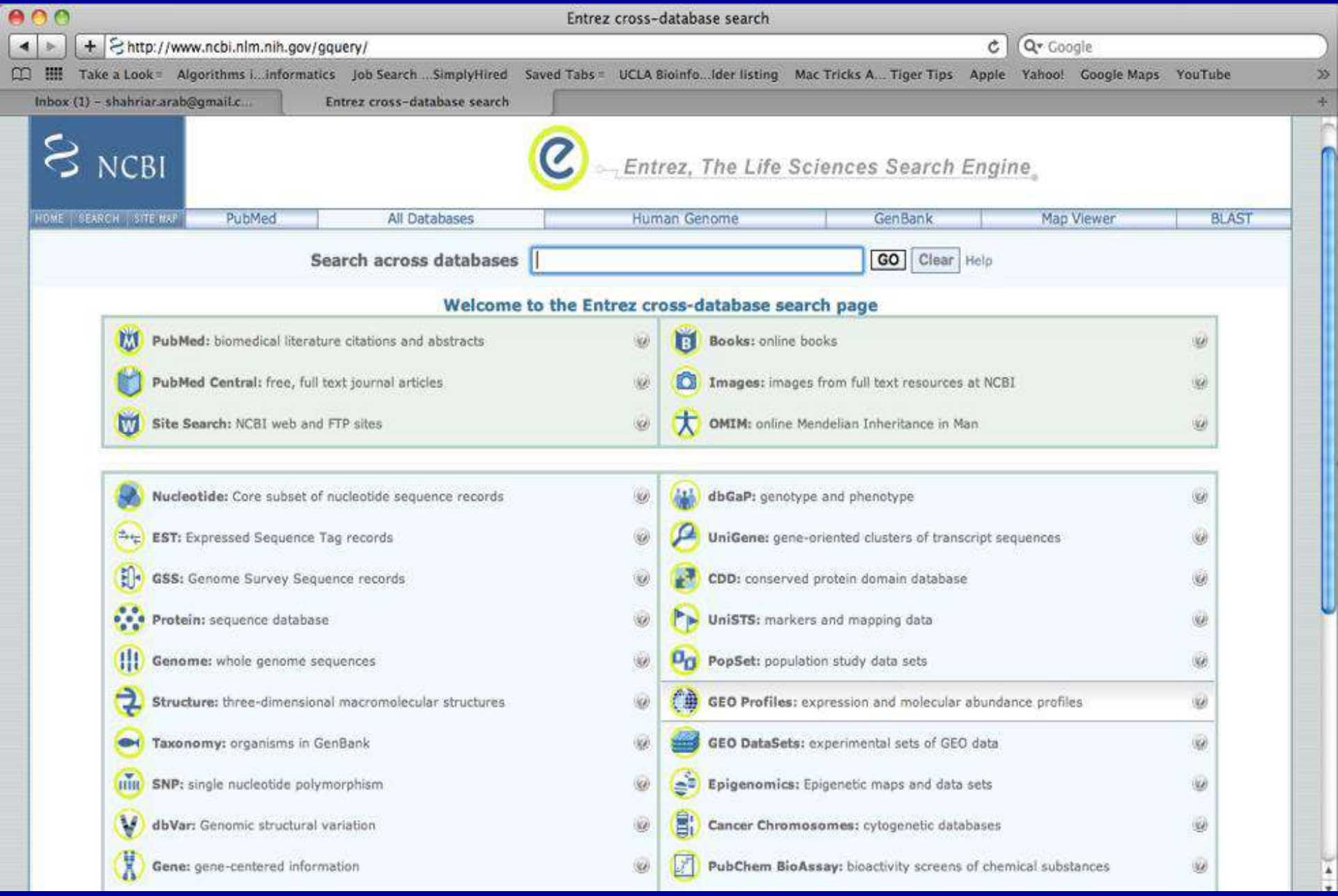


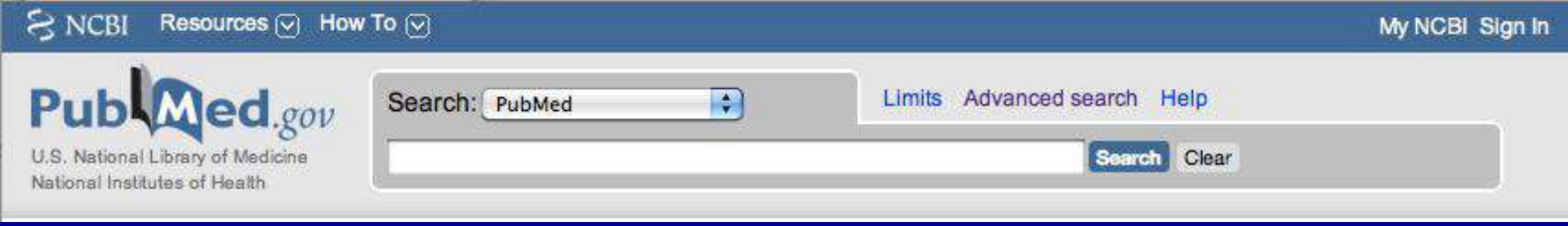
www.ncbi.nlm.nih.gov

ENTREZ

- NCBI (USA) National Center for Biotechnology Information
  - PubMed: The biomedical literature (PubMed)
  - Nucleotide sequence database (Genbank)
  - Protein sequence database
  - Structure: three-dimensional macromolecular structures
  - Genome: complete genome assemblies
  - PopSet: population study data sets
  - OMIM: Online Mendelian Inheritance in Man
  - Taxonomy: organisms in GenBank
  - Books: online books
  - ProbeSet: gene expression and microarray datasets
  - 3D Domains: domains from Entrez Structure
  - UniSTS: markers and mapping data
  - SNP: single nucleotide polymorphisms
  - CDD: conserved domains
  - Journals: journals in Entrez
  - UniGene: gene-oriented clusters of transcript sequences
  - PMC: full-text digital archive of life sciences journal literature

http://www.ncbi.nlm.nih.gov/Entrez/


Entrez cross-database search
http://www.ncbi.nlm.nih.gov/gquery/
Google
Take a Look
Algorithms i...informatics
Job Search ...SimplyHired
Saved Tabs
UCLA Bioinfo...lder listing
Mac Tricks A...
Tiger Tips
Apple
Yahoo!
Google Maps
YouTube
Inbox (1) - shahriar.arab@gmail.c...
Entrez cross-database search
NCBI
Entrez, The Life Sciences Search Engine
HOME
SEARCH
SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
GO
Clear
Help
Welcome to the Entrez cross-database search page
PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Site Search: NCBI web and FTP sites
Books: online books
Images: images from full text resources at NCBI
OMIM: online Mendelian Inheritance in Man
Nucleotide: Core subset of nucleotide sequence records
EST: Expressed Sequence Tag records
GSS: Genome Survey Sequence records
Protein: sequence database
Genome: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNP: single nucleotide polymorphism
dbVar: Genomic structural variation
Gene: gene-centered information
dbGaP: genotype and phenotype
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
UniSTS: markers and mapping data
PopSet: population study data sets
GEO Profiles: expression and molecular abundance profiles
GEO DataSets: experimental sets of GEO data
Epigenomics: Epigenetic maps and data sets
Cancer Chromosomes: cytogenetic databases
PubChem BioAssay: bioactivity screens of chemical substances



PubMed is…

- National Library of Medicine's search service
- >20 million citations in MEDLINE
- links to participating online journals
- PubMed tutorial (via side bar)

Entrez integrates…

- the scientific literature;
- DNA and protein sequence databases;
- 3D protein structure data;
- population study data sets;
- assemblies of complete genomes

Entrez is a search and retrieval system

Sequence Databases

- Annotated sequence databases
  - SWISS-PROT, GenBank etc…
  - Usage: identifying function, retrieving information
- Low-annotation sequence databases
  - EST databases, high-throughput genome sequences
  - Usage: discovery of new genes

General Protein Databases

- SWISS-PROT
  - Manually curated
  - high-quality annotations, less data
- GenPept/TREMBL
  - Translated coding sequences from GenBank/EMBL
  - Few annotations, more up to date
- PIR
  - Phylogenetic-based annotations
- All 3 now combining efforts to form UniProt (http://www.uniprot.org)

Low-annotation Databases

- ESTs (Expressed Sequence Tags)
  - Low quality sequences generated by high - volume sequencing the 3’ or 5’ end of cDNAs
- High-throughput genome sequences
  - Produced by mass-sequencing of genomic DNA

Non-redundant Databases

- Sequence data only: cannot be browsed, can only be searched using a sequence
- Combine sequences from more than one database
- Examples:
  - NR Nucleic (genbank+EMBL+DDBJ+PDB DNA)
  - NR Protein (SWISS-PROT+TrEMBL+GenPept+PDB protein)

Sequence & Structure Databases

- PDB (Protein Databank)
  - Stores 3-dimensional atomic coordinates for biological molecules including protein and nucleic acids
  - Data obtained by X-ray crystallography, NMR, or computer modelling
  - http://www.rcsb.org/pdb/
- MMDB (Molecular Modelling database)
  - Over 28,000 3D macromolecular structures, including proteins and polynucleotides
  - (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure)
- SCOP (Structural Classification of Proteins)
  - Classification of proteins according to structural and evolutionary relationships

File Formats

- GenBank/GB, genbank flatfile format
- NBRF format
- EMBL, EMBL flatfile format
- Swissprot
- GCG, single sequence format of GCG software
- DNAStrider, for common Mac program
- Pearson/Fasta, a common format used by Fasta programs and others
- Phylip3.2, sequential format for Phylip programs
- Phylip, interleaved format for Phylip programs (v3.3, v3.4)
- Plain/Raw, sequence data only (no name, document, numbering)
- MSF multi sequence format used by GCG software
- PAUP"s multiple sequence (NEXUS) format
- ASN.1 format used by NCBI

EMBL Format

```
ID TRBG361 standard; mRNA; PLN; 1859 BP.
XX
AC X56734; S46826;
XX
SV X56734.1
XX
DT 12-SEP-1991 (Rel. 29, Created)
DT 15-MAR-1999 (Rel. 59, Last updated, Version 9)
XX
DE Trifolium repens mRNA for non-cyanogenic beta-
glucosidase
XX
KW beta-glucosidase.
XX
OS Trifolium repens (white clover)
OC Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
Tracheophyta;
OC Spermatophyta; Magnoliophyta; eudicotyledons; core
eudicots; rosids;
OC eurosids I; Fabales; Fabaceae; Papilionoideae;
Trifolieae; Trifolium.
XX
RN [5]
RP 1-1859
RX MEDLINE; 91322517.
RX PUBMED; 1907511.
RA Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT "Nucleotide and derived amino acid sequence of the
cyanogenic
RT beta-glucosidase (linamarase) from white clover
(Trifolium repens L.).";
RL Plant Mol. Biol. 17(2):209-219(1991).
XX
RN [6]
RP 1-1859
RA Hughes M.A.;
RT ;
RL Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ
databases.
RL M.A. Hughes, UNIVERSITY OF NEWCASTLE UPON TYNE,
MEDICAL SCHOOL, NEW CASTLE
RL UPON TYNE, NE2 4HH, UK
XX
DR GOA; P26204.
DR MENDEL; 11000; Trirp;1162;11000.
DR SWISS-PROT; P26204; BGLS_TRIRP.
XX
FH Key Location/Qualifiers
FH
FT source 1..1859
FT /db_xref="taxon:3899"
FT /mol_type="mRNA"
FT /organism="Trifolium repens"
FT /tissue_type="leaves"
FT /clone_lib="lambda gt10"
FT /clone="TRE361"
FT CDS 14..1495
FT /db_xref="GOA:P26204"
FT /db_xref="SWISS-PROT:P26204"
FT /note="non-cyanogenic"
FT /EC_number="3.2.1.21"
FT /product="beta-glucosidase"
FT /protein_id="CAA40058.1"
FT /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT mRNA 1..1859
FT /evidence=EXPERIMENTAL
XX
SQ Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt 60
cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag 120
tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga 180
aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata 240
tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta 300
caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc 360
ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaa
```

Swissprot format

Swiss-Prot: Q823P0

NiceProt - a user-friendly view of this Swiss-Prot entry

```
ID   DNA1_CHLCV     STANDARD;      PRT;   450 AA.
AC   Q823P0;
DT   10-OCT-2003 (Rel. 42, Created)
DT   10-OCT-2003 (Rel. 42, Last sequence update)
DT   25-OCT-2004 (Rel. 45, Last annotation update)
DE   Chromosomal replication initiator protein dnaA 1.
GN   Name=dnaA1; Synonyms=dnaA-1; OrderedLocusNames=CCA00368;
OS   Chlamydophila caviae.
OC   Bacteria; Chlamydiae; Chlamydiales; Chlamydiaceae; Chlamydophila.
OX   NCBI_TaxID=83557;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=GPIC;
RX   MEDLINE=22569155; PubMed=12682364 [NCBI, ExPASy, EBI, Israel, Japan]; DOI=10.1093/n
RA   Read T.D., Myers G.S.A., Brunham R.C., Nelson W.C., Paulsen I.T.,
RA   Heidelberg J.F., Holtzapple E.K., Khouri H.M., Federova N.B.,
RA   Carty H.A., Umayam L.A., Haft D.H., Peterson J.D., Beanan M.J.,
RA   White O., Salzberg S.L., Hsia R.-C., McClarty G., Rank R.G.,
RA   Bavoil P.M., Fraser C.M.;
RT   "Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC):
RT   examining the role of niche-specific genes in the evolution of the
RT   Chlamydiaceae.";
RL   Nucleic Acids Res. 31:2134-2147(2003).
CC   -!- FUNCTION: Plays an important role in the initiation and regulation
CC       of chromosomal replication. Binds to the origin of replication; it
CC       binds specifically double-stranded DNA at a 9 bp consensus (dnaA
CC       box): 5'-TTATC(C/A)A(C/A)A-3'. DnaA binds to ATP and to acidic
CC       phospholipids (By similarity).
CC   -!- SIMILARITY: Belongs to the dnaA family.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AE016995; AAP05115.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   HSSP; P03004; 1J1V. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
DR   TIGR; CCA00368; -.
DR   HAMAP; MF_00377; -; 1.
DR   InterPro; IPR003593; AAA ATPase.
DR   InterPro; IPR001957; Bac DnaA.
DR   InterPro; IPR010921; Trp repress rep.
DR   InterPro; Graphical view of domain structure.
DR   Pfam; PF00308; Bac DnaA; 1.
DR   Pfam; Graphical view of domain structure.
DR   PRINTS; PR00051; DNAA.
DR   SMART; SM00382; AAA; 1.
DR   TIGRFAMs; TIGR00362; DnaA; 1.
DR   PROSITE; PS01008; DNAA; 1.
DR   ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR   HOBACGEN [Family / Alignment / Tree]
DR   BLOCKS; Q823P0.
DR   ProtoNet; Q823P0.
DR   ProtoMap; Q823P0.
DR   PRESAGE; Q823P0.
DR   DIP; Q823P0.
DR   ModBase; Q823P0.
DR   SMR; Q823P0.
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   ATP-binding; Complete proteome; DNA replication; DNA-binding.
FT   NP_BIND     156    163       ATP (Potential).
SQ   SEQUENCE   450 AA;  51099 MW;  CF440A7B30021OD8 CRC64;
     MLTCSDCSTW EQFVNYVKTR CSKTAFENWI SPIQIIEETQ EKIRLEVPNI FVQNYLLDNY
     KQDLCSFVPL DAQGEPALEF VVAEIKKAPA QPIAPREPQE SPAETFEESK DFELKLNAAY
     RFDNFIEGPS NQFVKSAAVG IAGRPGRSYN PLFIHGGVGL GKTHLLHAVG HYVREHHKNL
     RVHCITTEAF INDLVQHLRL KSIDKMKNFY RSLDLLLVDD IQFLQNRQNF EEEFCNTFET
     LINLNKQIVI TSDKPPGQLK LSERIIARME WGLVAHVGIP DLETRVAILQ HKAEQKGLHI
     PNEIAFYIAD HIYGNVRQLE GAINKLTAYC RLFGKTLTES IVRDTLRELF RSPSKQKVSV
     ESILKSVATV FQVKLQDLKG TSRSKELVLA RQVAMYLAKT LITDSLVAIG SAFGKTHSTV
     LYACKTIEQK IEKDETLTRQ ISLCKNHIVG
//
```

Specialized Sequence Databases

- Focus on a specific type of sequences
- Sequences are often modified or specially annotated
- Usage depends on the database
- Examples:
  - Ribosomal RNA databases
  - Immunology databases

Protein domain databases

- Pfam (http://www.sanger.ac.uk/Software/Pfam/)
  - Collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families
- SMART **(a Simple Modular Architecture Research Tool)**
  - Identification and annotation of genetically mobile domains and the analysis of domain architectures
  - (http://smart.embl-heidelberg.de/help/smart_about.shtml
- CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)
  - Combines SMART and Pfam databases
  - Easier and quicker search

Sequence Motif Databases

- Scan Prosite (http://www.expassy.org/prosite) and PRINTS (http://bioinf.man.ac.uk/dbbrowser/PRINTS/)
  - Store conserved motifs occurring in nucleic acid or protein sequences
  - Motifs can be stored as consensus sequences, alignments, or using statistical representations such as residue frequency tables

Ribosomal RNA Databases

- RDP (Michigan State University, USA)
  - http://rdp.cme.msu.edu/html/
- rRNA database (University of Antwerp, Belgium)
  - http://rrna.uia.ac.be/
- ribosomal RNA sequences are pre-aligned according to their secondary structure
- Usage: creating data sets for molecular phylogeny, especially for microbial taxonomy and identification

Immunological Sequence Databases

- The Kabat Database of Sequences of Proteins of Immunological Interest
  - www.hgmp.mrc.ac.uk/Bioinformatics/Databases/kabatp-help.html
  - Sequences are classified according to antigen specificity, and available in pre-aligned format
- The Immunogenetics database (IMGT)
  - http://imgt.cnusc.fr:8104/
  - Focuses on immunoglobulins, T-cell receptors and MHC genes

Genome Databases

- Focus on one organism or group of organisms:
  - Colibase (E. coli and related species) http://colibase.bham.ac.uk/
  - GDB (human) http://www.gdb.org/
  - Flybase (Drosophila) http://flybase.bio.indiana.edu/
  - WormBase (C. elegans) http://wormbase.org
  - AtDB (Arabidopsis) http://www.arabidopsis.org
  - SGD (S. cerevisiae) http://genome-www.stanford.edu/Saccharomyces/

Expression Databases

- RNA expression
  - Results of microarray experiments measuring the change in specific mRNA content under certain conditions
  - Array Express (EBI) and Geo (NCBI)
  - Not user friendly
- Proteome databases
  - 2D gel electrophoresis images representing the protein content of a cell or tissue under specific conditions
  - SWISS 2D PAGE at http://us.expasy.org/ch2d/

Other Database Types

- Literature
  - MEDLINE (http://ncbi.nlm.nih.gov/PubMed/)
  - HighWire (http://www.highwire.org)
- Variation
  - dbSNP (http://ncbi.nlm.nih.gov/SNP/)
  - HGBase (http://hgbase/interactiva/de)
- Metabolic pathways
  - KEGG (http://kegg.genome.ad.jp/kegg/)
  - WIT (http://wit.mcs/anl.gov/WIT2)
- Organisms and nomenclature
  - Taxonomies (e.g.: http://ncbi.nlm.nih.gov/Taxonomy/)
  - Mendel (http://mbclserver.rutgers.edu/CPGN)

Methods for Accessing Data

- local installation
- screen scraping
- BioPerl
- FTP sites

Local Installations

- SRS
  - Need to obtain license from Lion Biosceinces
- Download data from FTP sites
- Ensembl
  - "framework to organize biology around the sequences of large genomes"
  - www.ensembl.org

Screen Scraping

- URL spoofing
  - construction of URLs that replicate the query
- html parsing
  - extraction of results from html pages returned by query
- Requirements
  - html module
  - knowledge of query mechanism
- Method NOT advocated by most data providers

BioPerl

- BioPerl is a collection of modules that facilitates the development of Perl scripts for bioinformatics applications.

- www.bioperl.org

SWISSPROT

- European/Swiss Bioinformatics Institute 1986
- Highly accurate, hand curated resource
- Aims:
  - Have a high level of annotation
    - Often by the people who have been working with the gene
  - Have a low level of redundancy

http://www.ebi.ac.uk/trembl/

TrEMBL

- SWISSPROT's Big Brother
  - All genes which have been left out of SWISSPROT
  - Computer annotated rather than human annotated

PROSITE

http://ca.expasy.org/prosite/

- Families of proteins
- Can search using regular expressions
  - Similar to unix commands using wildcards, etc.
  - E.g., [AC]-x-V-x(4)-{ED}
  - Interpreted as:
  - [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}
- Families exhibit these patterns
  - So we can search over families
- 1574 documents about 1308 different patterns

http://pfam.sanger.ac.uk/

PFAM

- Maintained by the Sanger Centre (Cambridge)
- Protein families aligned using HMMs
  - Hidden Markov Models (see later lecture)
- Given a new sequence
  - Find families which the sequence might fit into
- Sequence Coverage
  - 11912 families
  - Split into Pfam-A (high quality) and Pfam-B (low quality)

SCOP and CATH

http://scop.mrc-lmb.cam.ac.uk/scop/
http://www.cathdb.info/

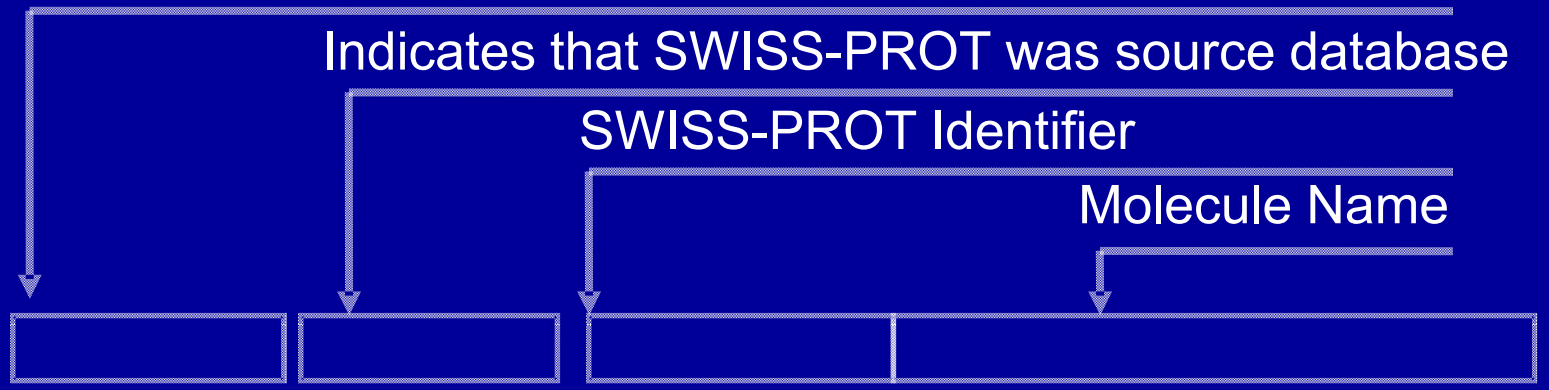
- SCOP
  - Structural Classification of Proteins
  - Hierarchically ordered and manually curated
  - 38221 PDB Entries
  - 110800 Domains
- CATH
  - Classification of protein domain structures
  - 124 folds
  - 226 Superfamily
  - 1148 Sequence family
  - 14473 Domain

Using Databases with the FASTA Format

- May need to know the FASTA format
  - For residue sequences
- First line must start with a > sign
  - First line contains identification information for gene
- Other lines contain the residue sequence
  - OK to have a ragged right format

Example FASTA Format

Geninfo num, assigned by the NCBI



- > gi|121664|sp|P00435|GSHC_BOVIN GLUTATHIONE PEROXIDASE
- mcaaqrsaaalaaaaprtvyafsarplaggepfnlsslrgkvllienvak
- slcgttvrdytqmndlqrrlgprglvvlgfpcnqfghqenakneeilncl
- yvrpgggfepnfmlfekcevngekahplfaflrevlptpsddatalmtdp
- kfitwspvcrndvswnfekflvgpdgvpvrrysrrfltidiepdietlls
- qgasa

Analyzing Results Using PERL Scripts

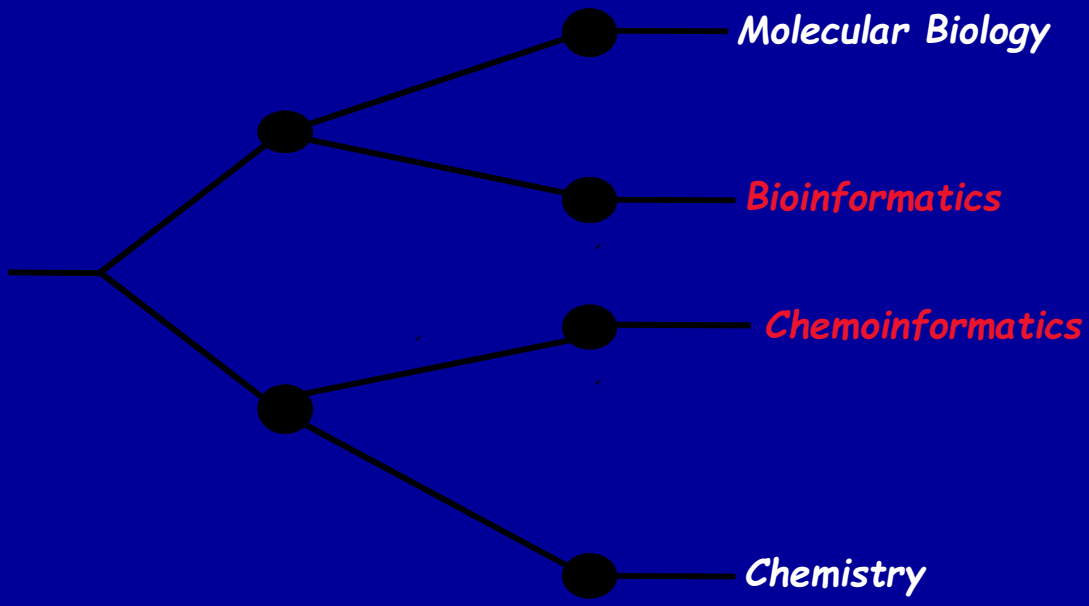
- Database servers now do:
  - Increasingly specific analysis of your results
- But you will eventually need to do analysis
- Ideal programming language is PERL
  - Designed to manipulate text and files
  - Can use it to play around with (manipulate) strings
- Will be using it in the coursework
  - PERL Tutorial

PDB Format

- The PDB format consists of a collection of fixed format records that describe :
  - Atomic coordinates,
  - Chemical and biochemical features
  - Experimental details of the structure determination
  - Some structural features such as
    - Secondary structure assignments,
    - Hydrogen bonding
    - Biological assemblies
    - Active sites

UNIT - IV

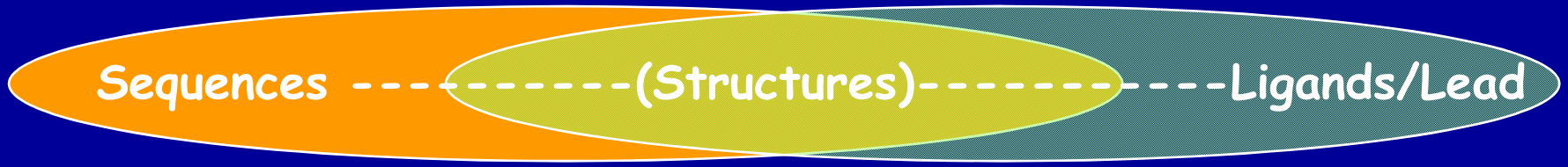
Chemoinformatics And Bioinformatics



Bioinformatics and Chemoinformatics Scale

Bioinformatics

Chemoinformatics



Genes
Genomes

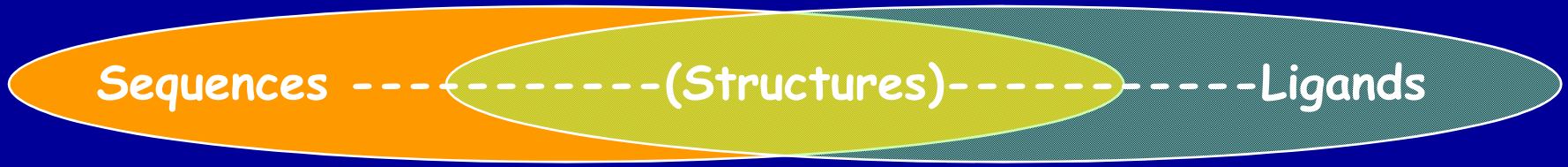
Proteins

Drugs

Bioinformatics and Chemoinformatics
Describing the DATA

Bioinformatics

Chemoinformatics



Chain → Chain Atom → Each Atom →

AGCTGTCGAGGGATAGGACA
TATACATAAATTAATATAAT

Strings

3D-Structure

SMILES, Sybyl,
Matrices…
Structure Descriptors

Bioinformatics and Chemoinformatics
Storing the DATA

Bioinformatics

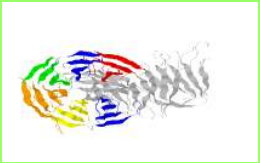
Chemoinformatics

Sequences ----------(Structures)----------Ligands

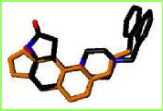
Chain ⋯⋯ Chain Atom ⋯⋯ Each Atom ⋯⋯▶

AGCTGTCGAGGGATAGGACA
TATACATAAATTAATATAAT

Strings



3D-Structure



SMILES, Sybyl, Matrices, descriptors

Sequence ⋯⋯ Coordinates ⋯⋯ Chemical Structure ⋯⋯▶

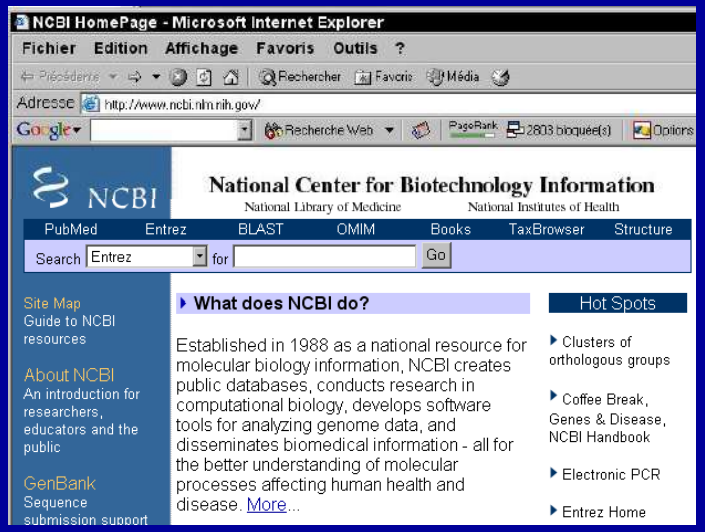
GenBank: Genomes
Genpep/NR: proteins
SwissProt

PDB

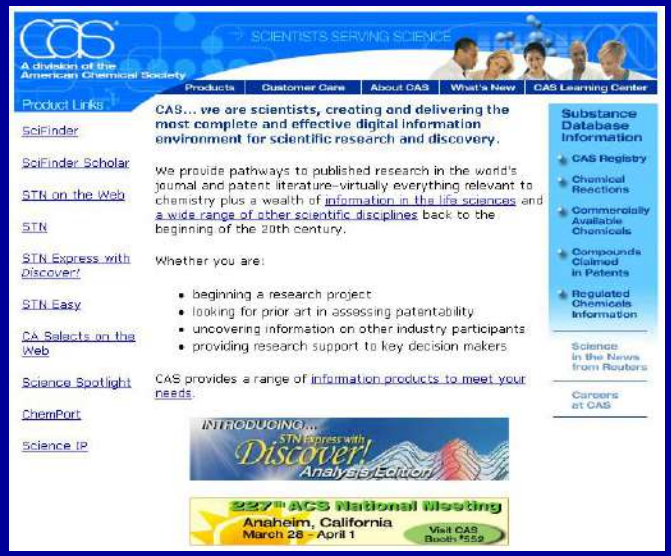
CAPLUS
REGISTRY
…

Related Techniques Databases

Bioinformatics: NCBI/EBI



Chemoinformatics CAS/Beilstein/MDL

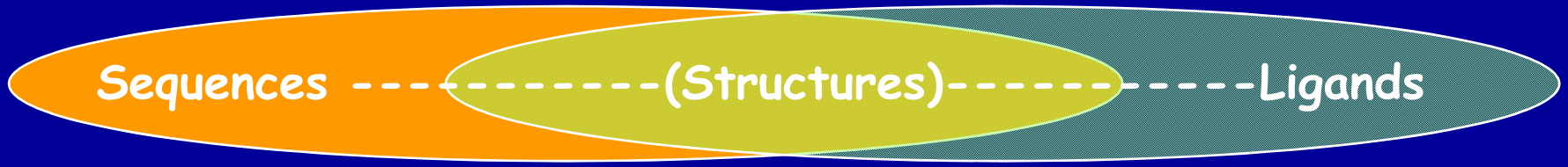


Finding the right Database is NOT an issue as most data are public.

Bioinformatics and Chemoinformatics
Comparing the DATA

Bioinformatics

Chemoinformatics

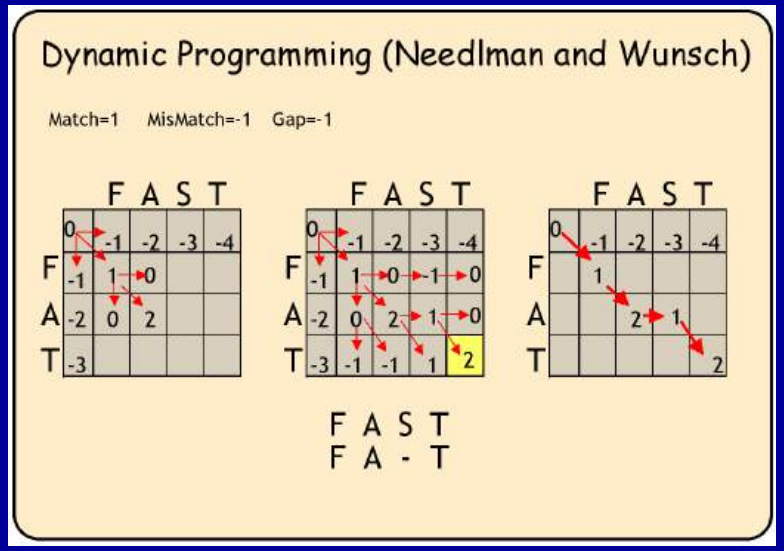


Similarity ..... **Fold** ..... **Descriptor** →

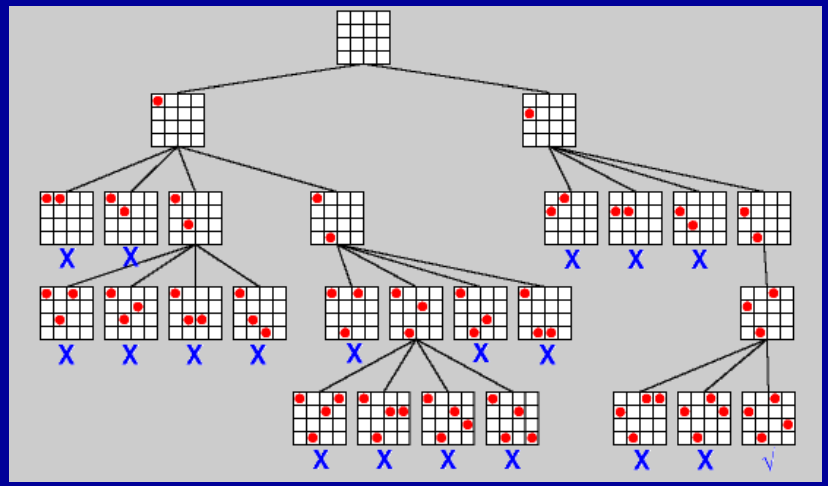
Evolutionary model	3D-Model	Descriptor Similarity
Dynamic Programming	Structure/Structure Comparison	Backtracking Algorithm
BLAST	LSQman DALI SAP	CACTUS (Cactus.nci.nih.gov)

Related Techniques
Searching Databases

Bioinformatics: Dynamic Programming



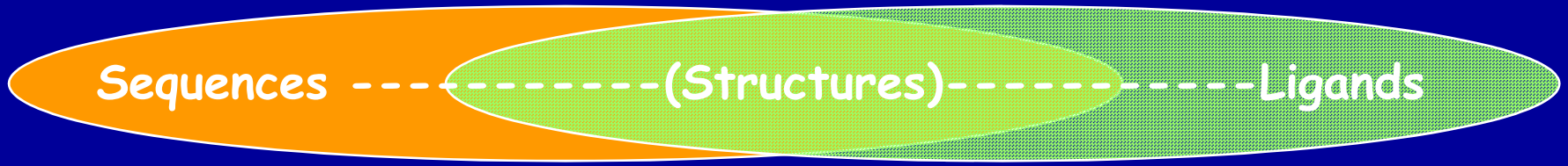
Chemoinformatics Backtracking



Bioinformatics and Chemoinformatics
Building Models

Bioinformatics

Chemoinformatics

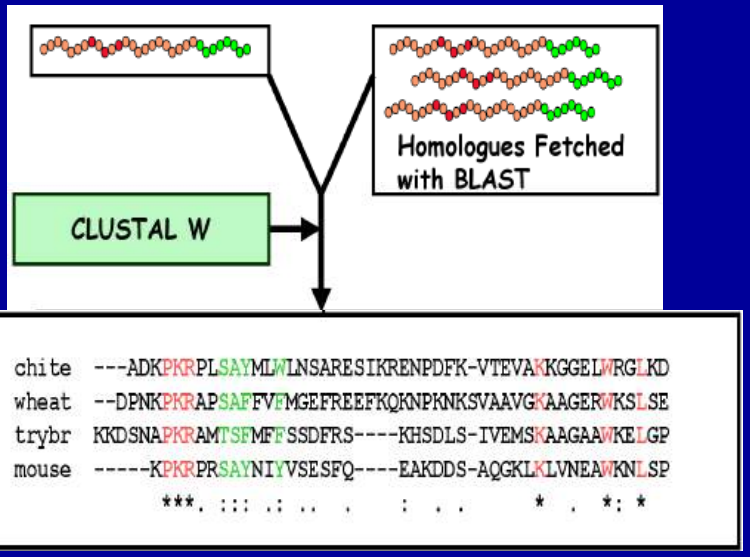


MSA ..... **Fold** ..... **Descriptor** ➔

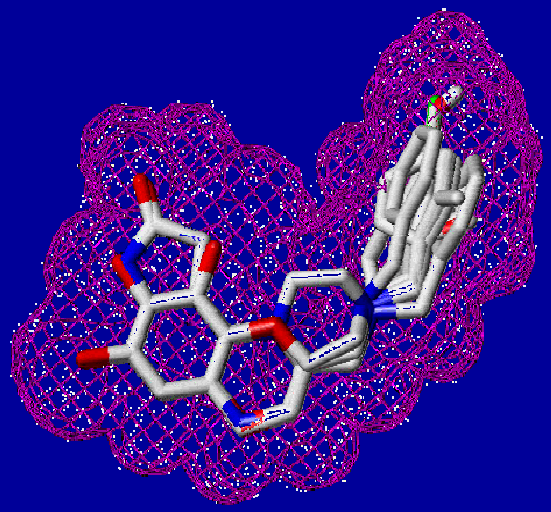
Multiple Sequence Alignments ClustalW, TCoffee	3D-Model Classifications DALI SCOP CAT	Descriptor Computation

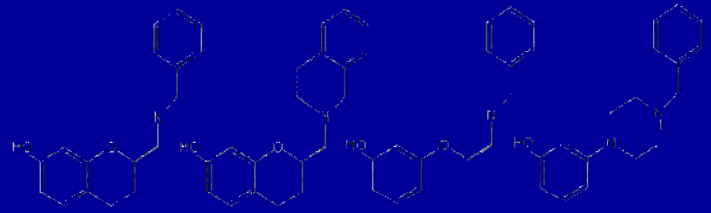
Related Techniques
Profiles

Bioinformatics:



Chemoinformatics

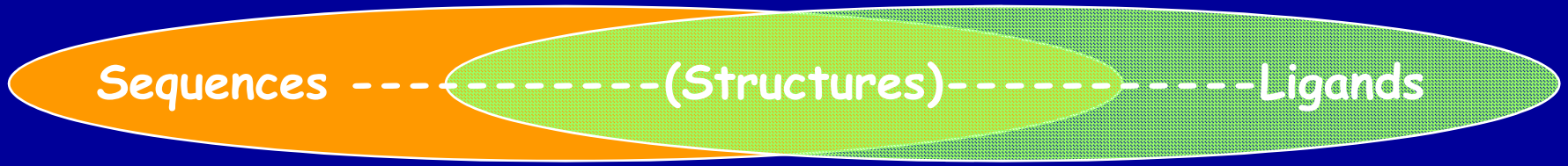




Bioinformatics and Chemoinformatics
Predicting Properties

Bioinformatics

Chemoinformatics



Function ..... **Binding** ..... **Descriptor** →

Multiple Comparisons

Profiles Pfam
Function
Structure
Phylogeny

Docking/Mecanisms

Molecular Dynamic
Docking Predictions

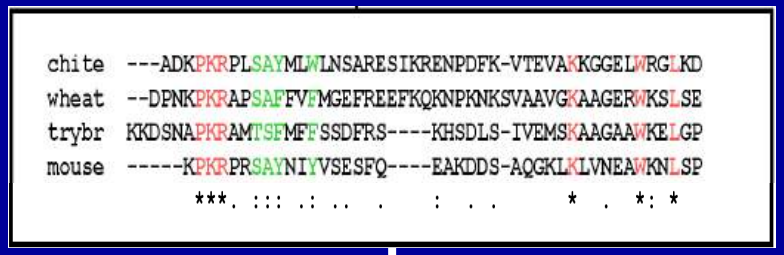
QSAR

Relate descriptors to

Activity/Toxicity

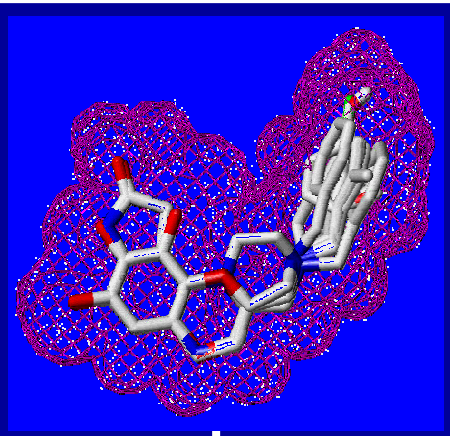
Related Techniques
Profiles and Descriptors

Bioinformatics:



Database Search
With a Profile

Chemoinformatics



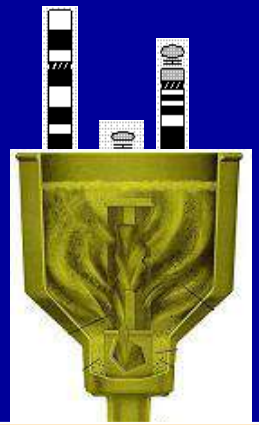
Pharmacophore

Database Search
With a Profile

Bioinformatics and Chemoinformatics
The Questions

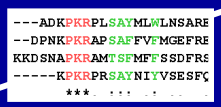
Bioinformatics: Target

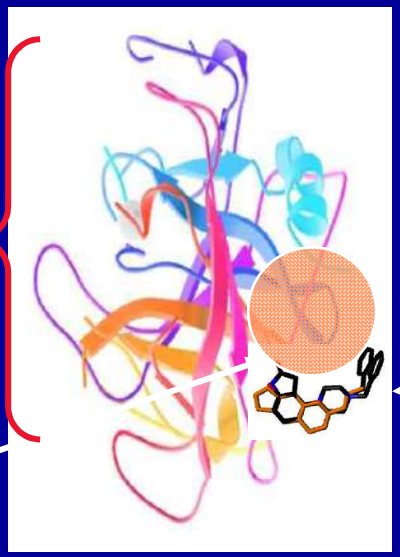
Chemoinformatics Lead



Target

Site



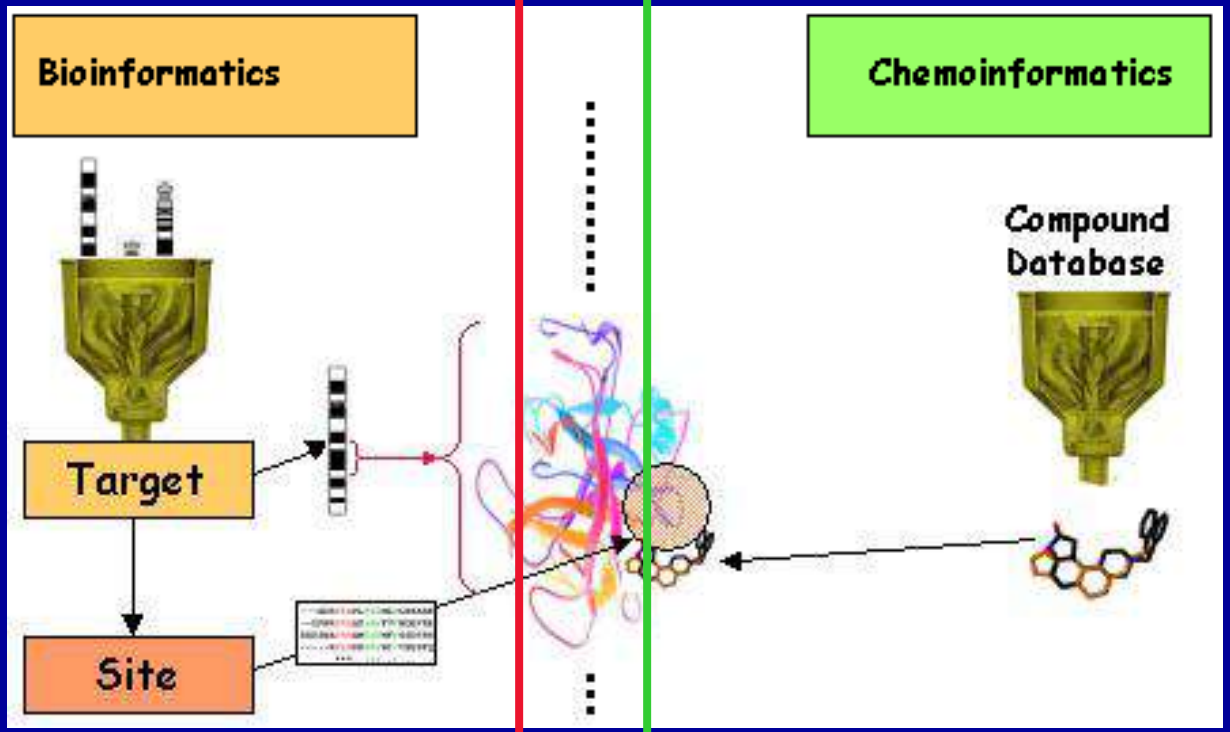


Compound Database

Bioinformatics and Chemoinformatics
The Questions

Evolutionary Trace

Chemical Modelling



Bioinformatics and Chemoinformatics
The Questions

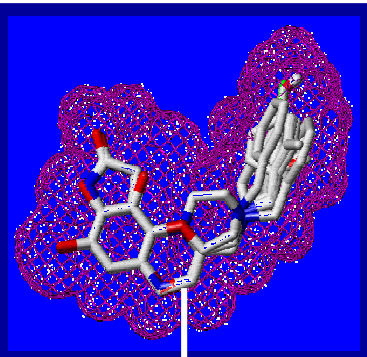
Bioinformatics: Function

Comparative Genomics

↓

Function

Chemoinformatics QSAR



Quantitative Structure-Activity Relationship

↓

Activity/Toxicity

Bioinformatics and Chemoinformatics
The Questions

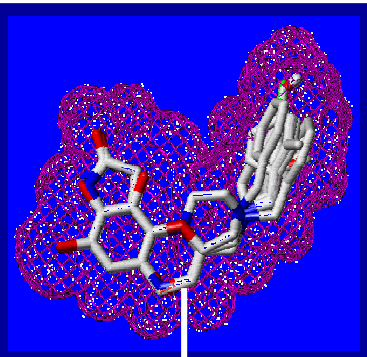
Bioinformatics: Target

Comparative Genomics
Molecular Dynamic

↓

Docking

Chemoinformatics Lead



2D/3D QSAR
Molecular Dynamic

↓

In Silico Screening

Bioinformatics and Chemoinformatics
The Algorithms

Bioinformatics: Target

Chemoinformatics Lead

Graph Theory

Neural Networks

Genetic Algorithms

Dynamic Programming

Hidden Markov Models

Backtrace

Bioinformatics/ Chemoinformatics

Where Do They Meet

Homology based SAR predictions

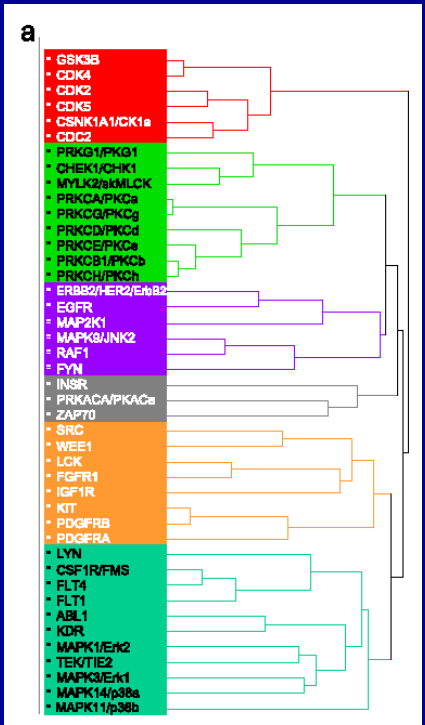
Homology based SAR predictions

- SAR Matrices contain affinity Data

Targets / Compounds	T1	T2	T3	T4	T5	T6	T7
C1	8	4	>20	2			
C2	5	7		>20		>20	
C3	1.1	3		>20		>20	
C4	>20	>20		>20			
C5			2	0.9			
C6	>20	>20	0.1	0.09			
C7	0.1	0.2	0.1	0.1	0.3		
C8	0.77	0.2	>20	>20	>20		
C9	0.57	0.27	>20	>20			
...	0.2	>20	>20	>20	>20	>20	

Homology based SAR predictions

- Targets can be clustered from the SAR Data

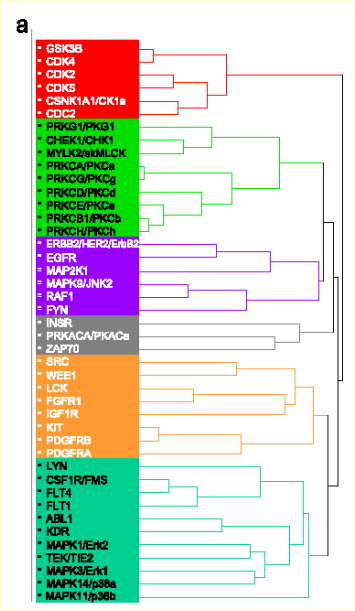


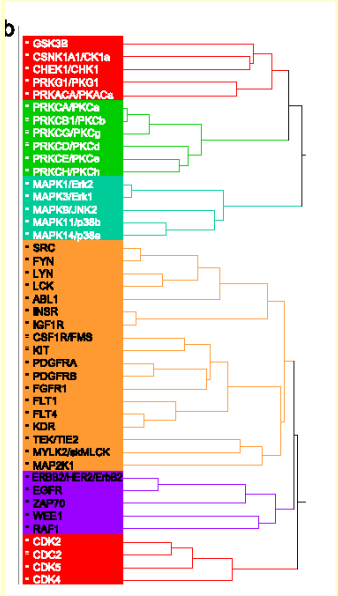
Targets / Compounds	T1	T2	T3	T4	T5	T6	T7
C1	6	4	>20	2			
C2	5	7		>20		>20	
C3	1.1	3		>20		>20	
C4	>20	>20		>20			
C5			2	0.9			
C6	>20	>20	0.1	0.09			
C7	0.1	0.2	0.1	0.1	0.3		
C8	0.77	0.2	>20	>20	>20		
C9	0.57	0.27	>20	>20			
...	0.2	>20	>20	>20	>20	>20	

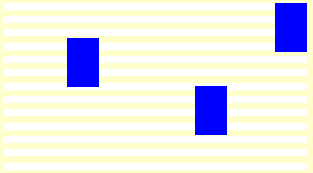
Homology based SAR predictions

- Targets can be clustered using similarity

Targets / Compounds	T1	T2	T3	T4	T5	T6	T7
C1	6	4	>20	2			
C2	5	7		>20		>20	
C3	1.1	3		>20		>20	
C4	>20	>20		>20			
C5			2	0.9			
C6	>20	>20	0.1	0.09			
C7	0.1	0.2	0.1	0.1	0.3		
C8	0.77	0.2	>20	>20	>20		
C9	0.57	0.27	>20	>20			
...	0.2	>20	>20	>20	>20	>20	

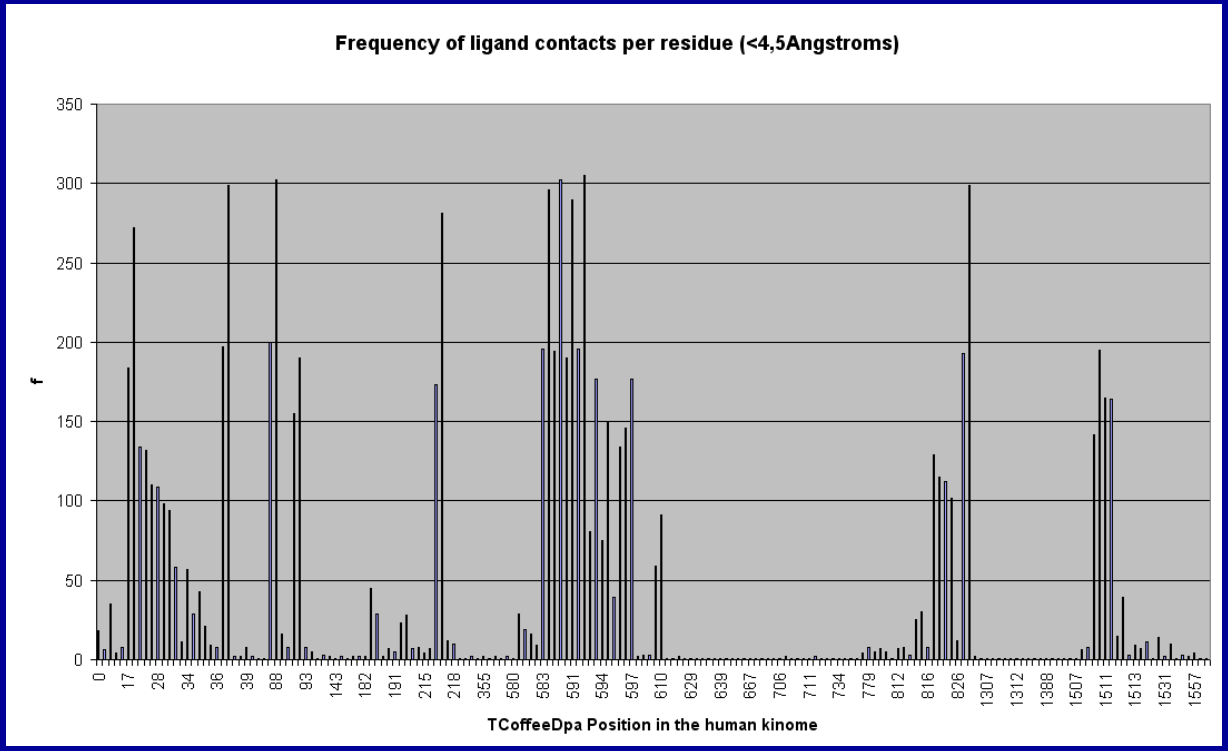






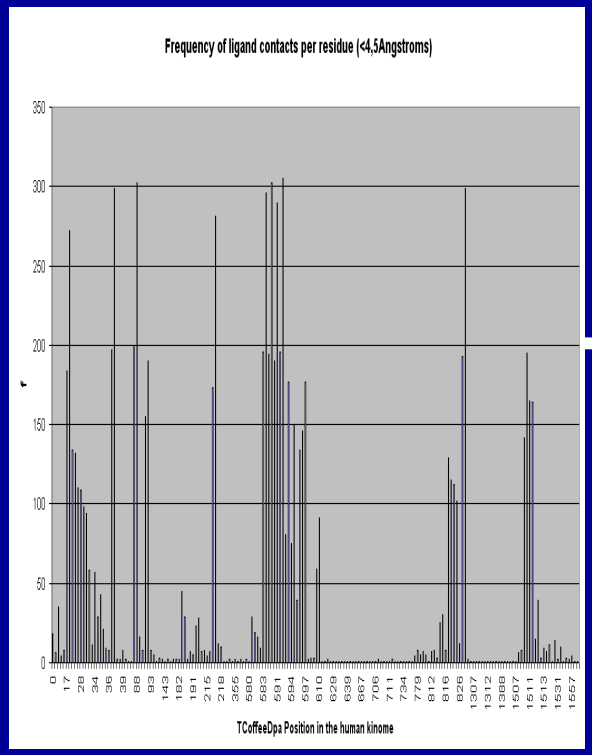
Homology based SAR predictions

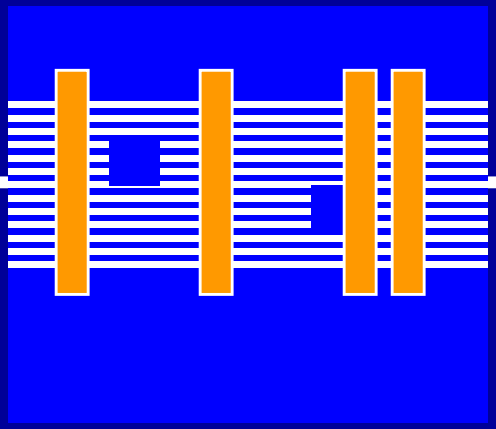
- Sequence Clustering Can be done using Key positions

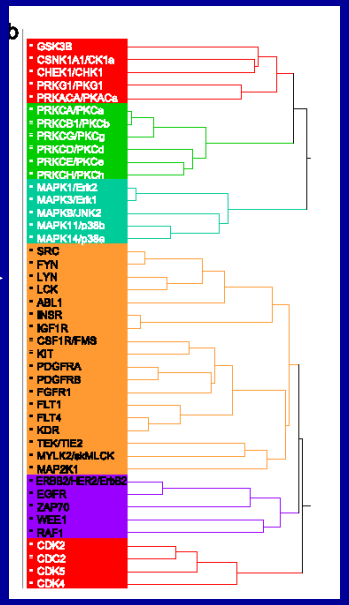


Homology based SAR predictions

- Sequence Clustering Can be done using Key positions







Homology based SAR predictions

Clustering Based on Active Site Residues

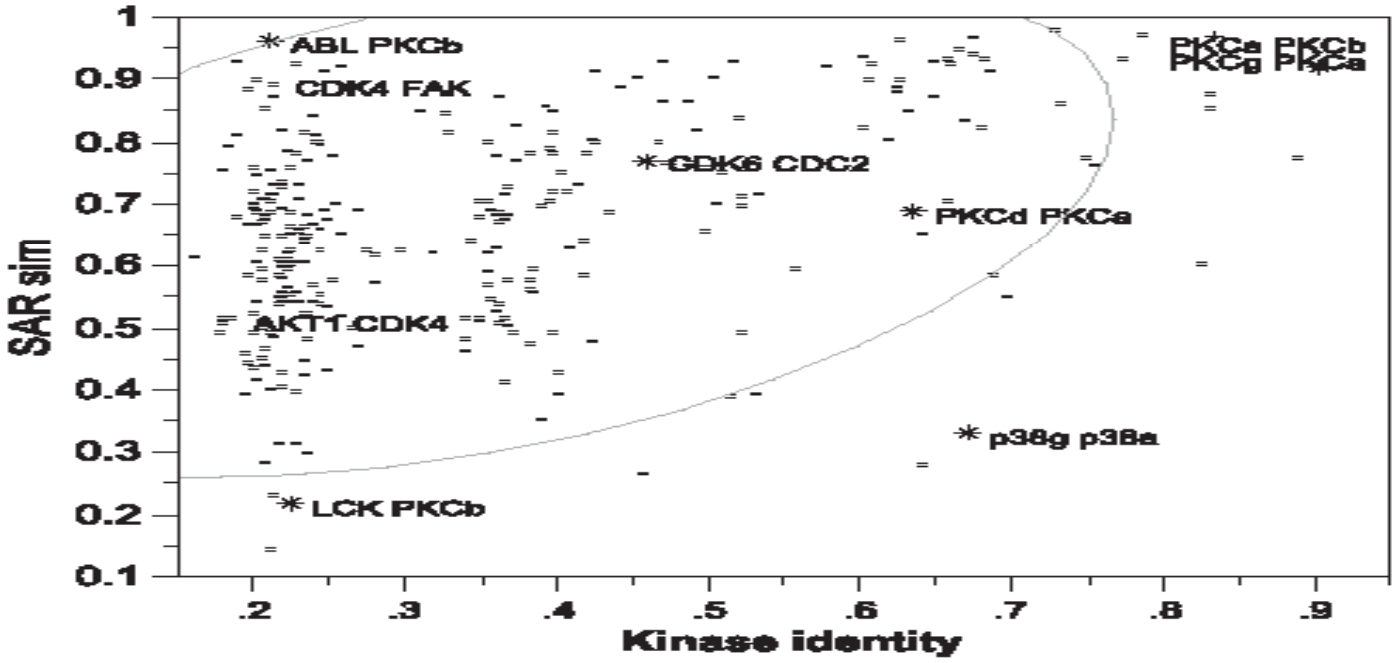


Fig. 4. Relationship between catalytic domain sequence identity and SAR similarity between pairs of kinase targets for 292 pairs with selectivity data. For this plot, only kinase pairs are shown for which data on at least two active compounds were available. Note that the available selectivity data to define kinase space covers only 20% (292 of 1485 pairs for 55 targets) possible combinations. Each dot represents a pair of kinase targets; some examples of pairs are displayed.

Homology based SAR predictions

–Pairs of
Comparable Kinases

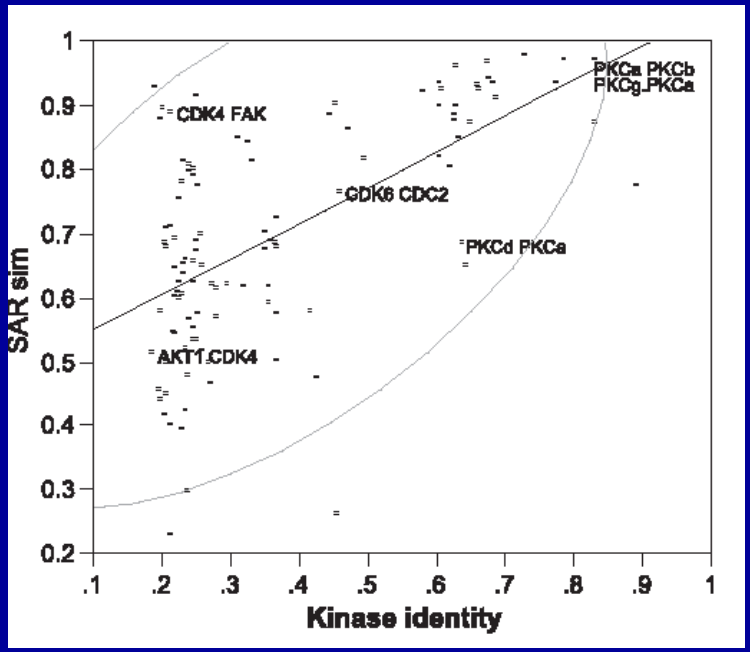
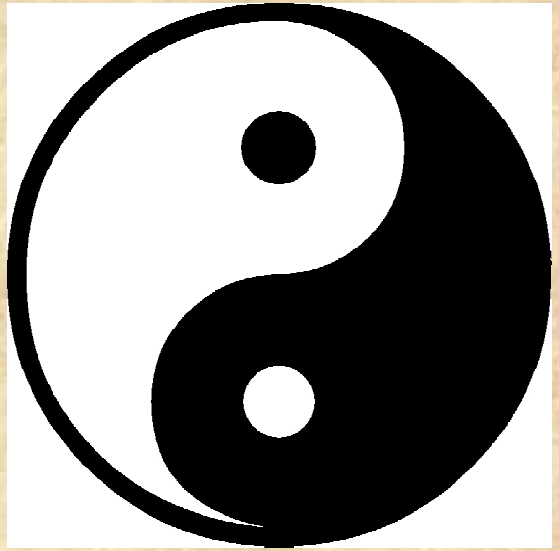


Fig. 5. Relationship between catalytic domain sequence identity and SAR similarity between 111 pairs of targets for kinases with large gatekeeper sizes. The R^2 value for the linear fitting is 0.37.

Bioinformatics/ Chemoinformatics

Main Differences ?



Bioinformatics and Chemoinformatics

Bioinformatics:

- Macro-molecules
- Evolutionary Signal

Chemoinformatics

- Small Molecules
- Chemical Modelling

UNIT - V

What is pharmacy informatics?

Topics

- What is Pharmacy Informatics?
- Who is an Informatics Pharmacist?
- What is the role of an Informatics Pharmacist?
- What are some current technologies?
- How can I pursue a career in Pharmacy Informatics?
- Where can I get more information?

Definition

- “a unique subset of medical informatics that focuses on the use of information technology and drug information to optimize medication use.”
- “the use and integration of data, information, knowledge, technology, and automation in the medication-use process for the purpose of improving health outcomes.”

ASHP statement on the pharmacist’s role in informatics

Simple Definition:

- The application of technology and information systems to the medication use process to improve outcomes and increase safety and efficiency
- Informatics pharmacists are involved in the design, implementation, customization and support of health information systems and technologies.

Who is an Informatics Pharmacist?

Who is an Informatics Pharmacist?

- An informatics pharmacist is a dual specialist
  - Knowledgeable about both pharmacy practice and informatics
    - Is able to analyze pharmacy practice from analytical design perspective
    - Is able to analyze health informatics technologies from a clinical/operational perspective
  - Has the ability to look at both the “big picture” and the individual details and processes.

Competencies/Characteristics

- Strong understanding of pharmacy practice
- Knowledgeable about the medication use process
- Knowledgeable about information systems, healthcare technology and automation
- Basic understanding of database design and function

Competencies/Characteristics

- Current with relevant standards, regulations and initiatives
- Ability to anticipate future needs and challenges
- Ability to think about the “end user”
- Ability to teach and guide others
- Communication skills

Competencies/Characteristics

- Project management skills
- Technology oriented
- Innovative
- Analytical

What is the role of an Informatics Pharmacist?

Role of an Informatics Pharmacist

- Ensure patient safety
- Provide guidance and leadership for all technology initiatives that support medication use
- Customize and tailor health information systems and technology to the needs of practice

Role of an Informatics Pharmacist

- Serve as a liaison between pharmacy and other departments
  - Information Technology
  - Nursing
  - Physicians
  - Vendors

Role of an Informatics Pharmacist

- Provide education to healthcare professionals and managers
- Serve as a resource for hospital staff
- Provide recommendations regarding vendor selection

Technology in Pharmacy Practice

- The field of pharmacy is rapidly evolving
  - pharmacy model is changing with the introduction of information technology and automation
- Technology is continuously being developed with the intention of increasing safety, reducing cost and increasing efficiency.

What are some current technologies relevant to medication use?

Current Technology

- Health Information Systems
- Electronic Medication Administration Records
- Computerized Provider Order Entry
- Clinical Decision Support
- Electronic Prescribing

Current Technology

- Automated Dispensing Cabinets
- Inventory Management Systems
- Bar Coding
- Radio-frequency identification
- Smart Pumps
- Robotics

How can I pursue a career in Pharmacy Informatics?

Becoming a Pharmacy Informaticist

- Pharmacy informaticists should have:
  - real world pharmacy practice experience
  - thorough understanding of the medication use process
  - extensive knowledge about the health information systems and all medication related technologies

Becoming a Pharmacy Informaticist

- There are a growing number of residency/training programs available
  - Typically 12 months in length
- Some individuals obtain formalized training while others have had extensive experience in utilizing health information systems and technology

Residency Training

- Four PGY2 specialty residencies (ASHP accredited status)
  - Vanderbilt University Medical Center
  - University of Utah Hospitals and Clinics
  - University of Michigan Hospitals and Health Centers
  - Oregon Health and Science University Hospitals and Clinics

Residency Training

- Five PGY2 specialty residencies (ASHP candidate/pre-candidate status)
  - Clarian Health Partners
  - James A. Haley Veterans Hospital
  - University of Louisville Health Care
  - VA San Diego Healthcare System
  - The Johns Hopkins Hospital

Non ASHP Accredited Training

- Brigham and Women's Hospital
- Cedar-Sinai Medical Center
- Ohio State Medical Center
- Penn State Hershey Medical Center
- Sentara Health System
- University of California San Francisco

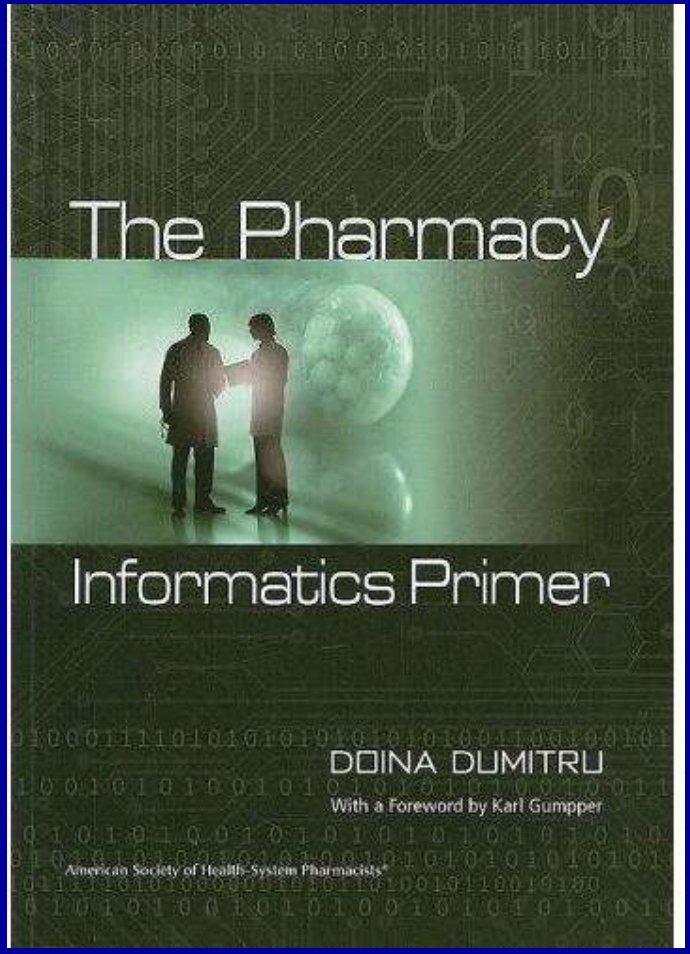
Where can I get more information?

Additional Resources

- Visit the Pharmacy Informatics and Technology section of the American Society of Health-System Pharmacists website

Additional Resources

- The Pharmacy Informatics Primer is a great resource for pharmacy managers, information technology project managers and pharmacy students.
  - Provides practical information regarding the most relevant topics in pharmacy informatics



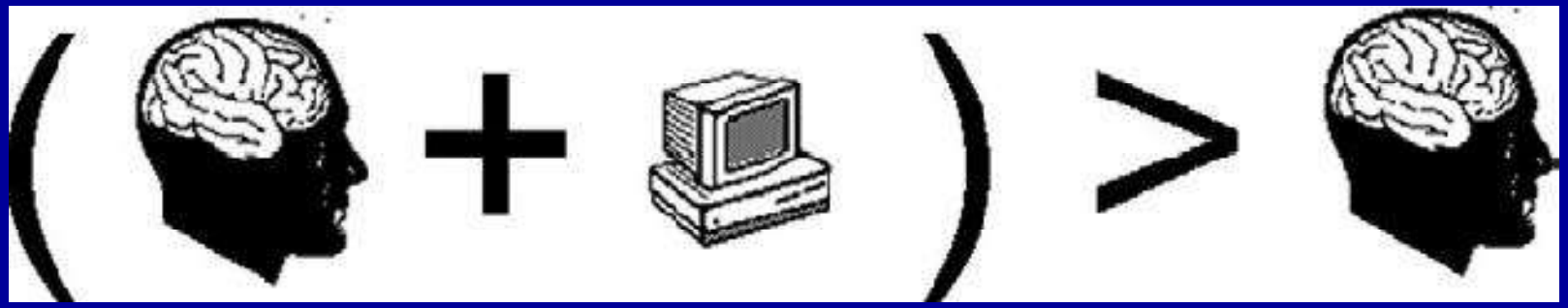
Health Informatics – Electronic Medical Record Advantages

Learning Objectives

- Describe Health Informatics as it relates to the electronic medical record.
- Recognize the utility of regional business intelligence tools related to health care metrics and patient outcomes.
- Identify different ways technology can enhance antimicrobial stewardship.
- Understand the general impacts of converting to electronic clinical documentation.
- Explain how electronic medical records contribute to quality measures and patient safety.

What is Health Informatics?

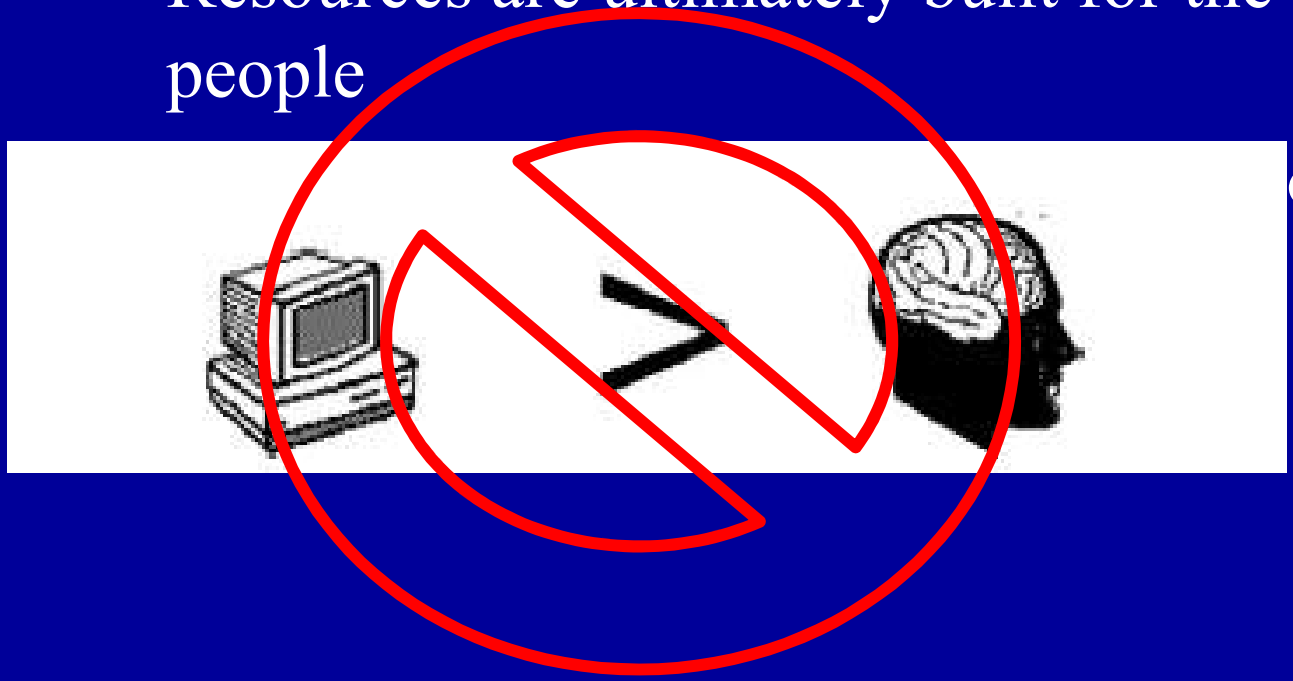
- The fundamental theorem:
  - "A person working in partnership with an information resource is 'better' than that same person unassisted."



Friedman CP. J Am Med Inform Assoc. 2009;16:169-170

What is Health Informatics?

- More about people than technology:
  - Resources are ultimately built for the benefit of people



of

What is Health Informatics?

- The resource must offer new information:
  - Most challenging aspect of designing effective information resources
- The interaction between person and resource cannot be predicted in advance:
  - Whether the theorem holds depends on this interaction
    - Poor resource design
    - Person lacks sufficient knowledge of domain

Friedman CP. J Am Med Inform Assoc. 2009;16:169-170

Assessment

The Fundamental Theorem of Health Informatics states that:

A. Information resources are more efficient that humans
B. All information resources have intuitive human interfaces
C. Working with an information resource elicits better results than working unassisted
D. Informatics is all about technology, people play an insignificant role

Learning Objectives

- Recognize the utility of regional business intelligence tools related to health care metrics and patient outcomes.

Business Intelligence Tools

- Process large amounts of raw data into meaningful information:
  - Create Dashboards –
    - Interactive collection of information to help make informed decisions
  - Monitor Performance –
    - Scorecards track progress towards goals
    - More easily identify areas requiring attention

What are some examples of BI tools at your site?

High Risk Opioid Registry

VDW. VISN 20 HRR. Accessed: 1/13/15

High Risk Opioid Registry

Clinic	PCP	Resident Name	MHTC	Patient Name	Last 4	Age	Sex	OEF

Upcoming Apointments	Date

	Risk Ranks		Recent Screens			SI			Active Flags		Admits/ED Visits	
Morph Equiv	Opioid Risk Percentile	CAN Score	PHQ2	PTSD	AUDIT-C	SI Flag Active	SI High Risk HF	SI Dx Visits	Drug Seeking	Behavior	Admits	ED Visits
40	98%	80					Yes	2			1	2

Other Risk Indicators														
Pain Dxs	Cancer Dx	Terminal	MDD	PTSD	OTP	PT	Benzo Rx Active	Tobacco Use	Apnea	Home-less	Naloxone Kit Rxs	OD Dxs	ADs Rxs	Pain Scores >= 7
5	Yes		19	55			Yes		2					2

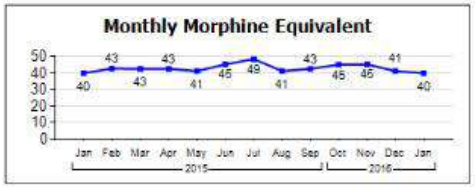
Measures	Last Opiate Agreement & iMed Consent		V20 Opioid Safety Case Review		Past Encounters					
	Agreement	iMed Consent	Last Case Review	Appropriate for Taper	Last UDS	Last PMP	Last OSRB	Last PC Visit	Last MH Visit	Last SATP Visit
		12/16/15			12/11/15	12/17/15		12/16/15	11/18/15	

Resources	Agreement Notes	iMed Consents		UDS Tests	PMP Entries	OSRB Entries	PC Visits	MH Visits	SUD Dxs
				2	2		15	16	1

Opioid Rx History (Past 12 Mos)									
Total Days	Early Fills	%	Avg Days Between Fills	Total Fills	Active Fills	Quantity	Days Supply	Morph Equiv	Prescribers
350	8	62%	-1	13				40	5

Opioid & Benzo Rxs in Past 60 Days						
Prescriber	Local Drug Name With Dose	Release Date	Rx Status	Qty	Days Supply	Equiv
	HYDROCODONE 10/ACETAMINOPHEN 325 MG TAB	11/13/15	DISCONTINUED	112	28	40
	CLONAZEPAM 2MG TAB (GEN)	12/31/15	ACTIVE	15	30	20
	HYDROCODONE 10/ACETAMINOPHEN 325 MG TAB	12/11/15	DISCONTINUED	112	28	40

	Most Recent Urinary Drug Screen(s)							
UDS Date	Amphetamine	Barbiturate	Benzo	Methadone	Opiate	Oxycodone	PCP	THC
12/11/15		NEG	POS	NEG	POS	NEG		NEG
1/16/14		NEG	NEG		NEG			NEG



VDW. VISN 20 HRR. Accessed: 1/13/15

BI Benefits

- Provider/Performance feedback has led to improved outcomes in all of the following practices:
  - Tobacco cessation
  - Anesthesia guidelines
  - Hypertension management
  - Treatment of cirrhosis

Bentz CJ, et al. Nicotine Tob Res. 2007 Mar;9(3):341-9
Loy V et al. Am J Med Qual. 2014 Oct 27 [ePub]
Luders S et al. Eur J Cardiovasc Prev Rehabil. 2010 Jun:17(3):271-9

Other VA Initiatives

Assessment

Business intelligence tools provide an efficient method for achieving which of the following:

A. Reporting performance indicators to providers to improve health care metrics
B. Providing real time clinical decision support to ensure guideline adherence
C. Enhancing usability of existing EMRs
D. Ensuring existing decision support reflects the most current evidence

Thanks