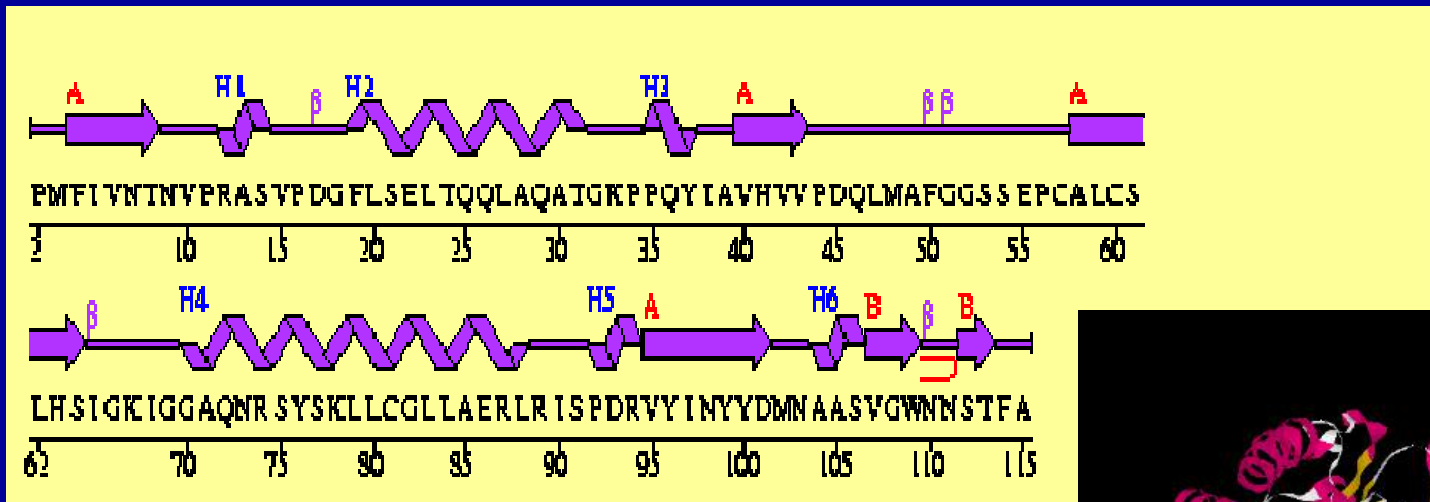


# **M.Sc. Bioinformatics**

# UNIT - I

# Introduction to bioinformatics



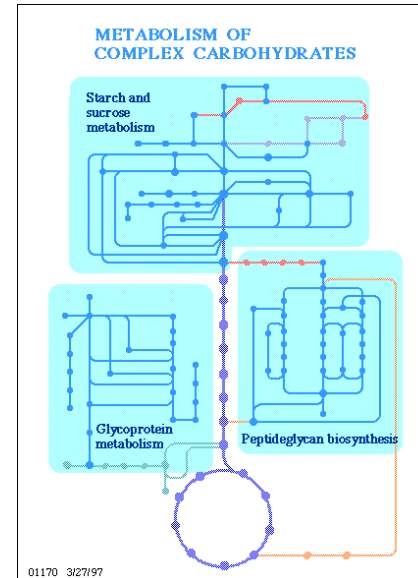
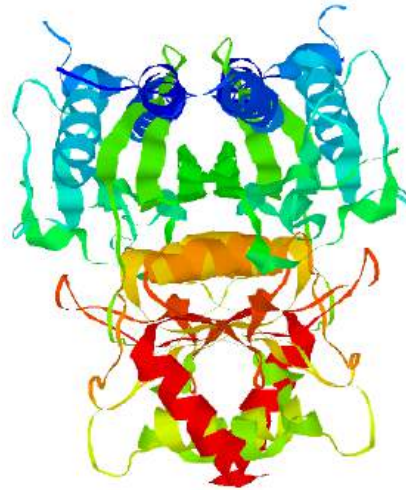
# What is bioinformatics?

- 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  
NAIIVEPEKRGKYVCFDPL  
DGSSNIDCLVSIPTIFGIYR  
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  
TCTGAGGAAGCCTGGGCATGTGCA  
AATGAAGTTAACCTAA...
```

**> 500, 000 genes  
sequenced to date**



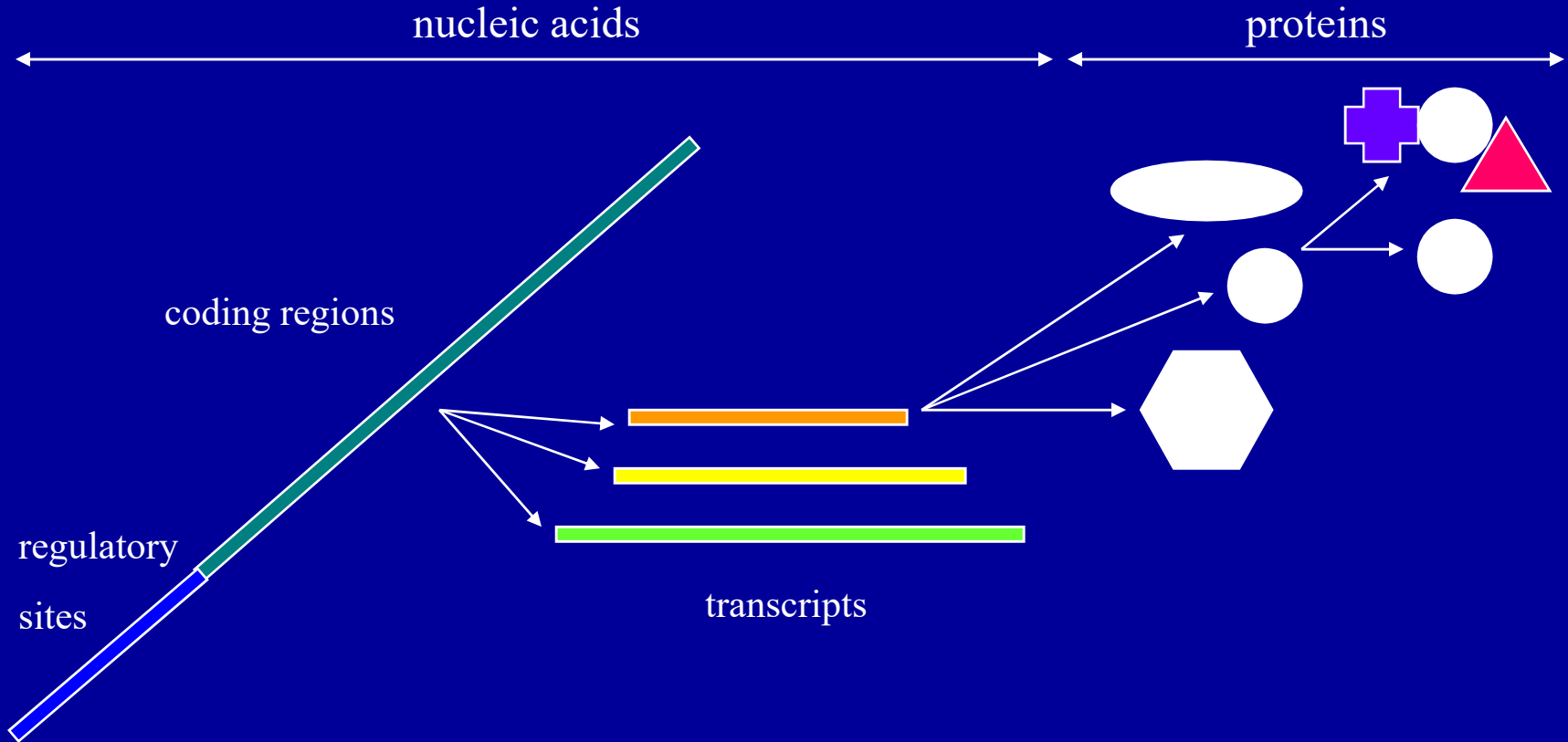
**Expected number of  
unique protein  
structures:**

**~ 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**

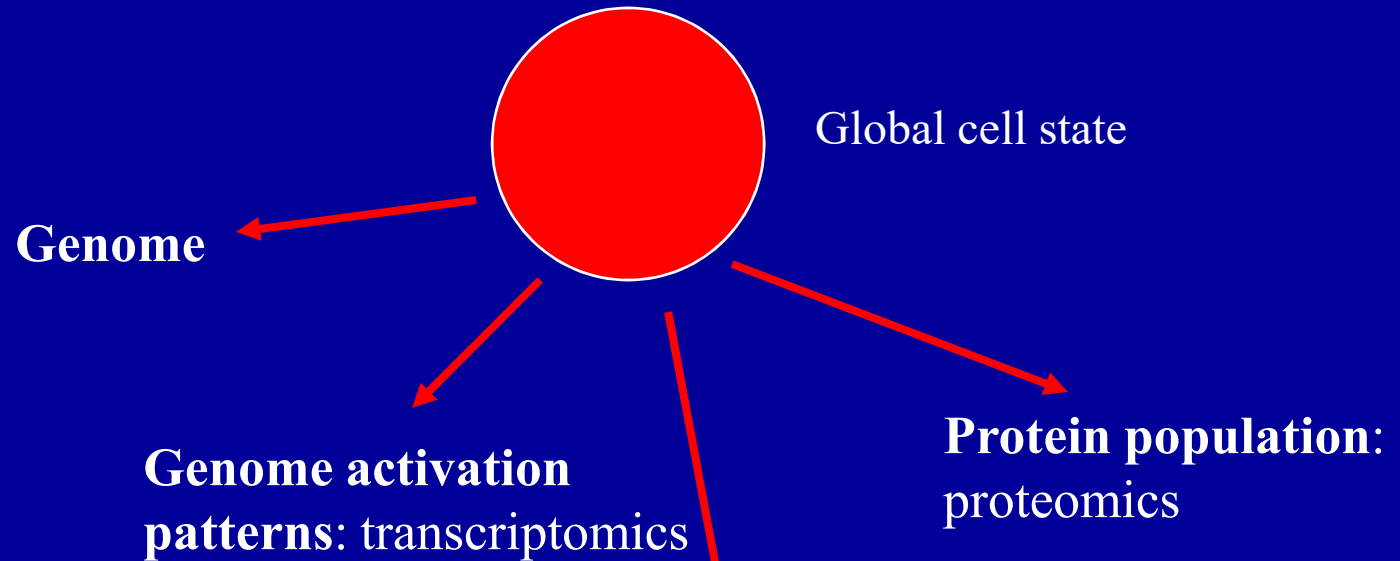
# Information processing in cells



**One-to-many mappings!**

**Context-dependence!**

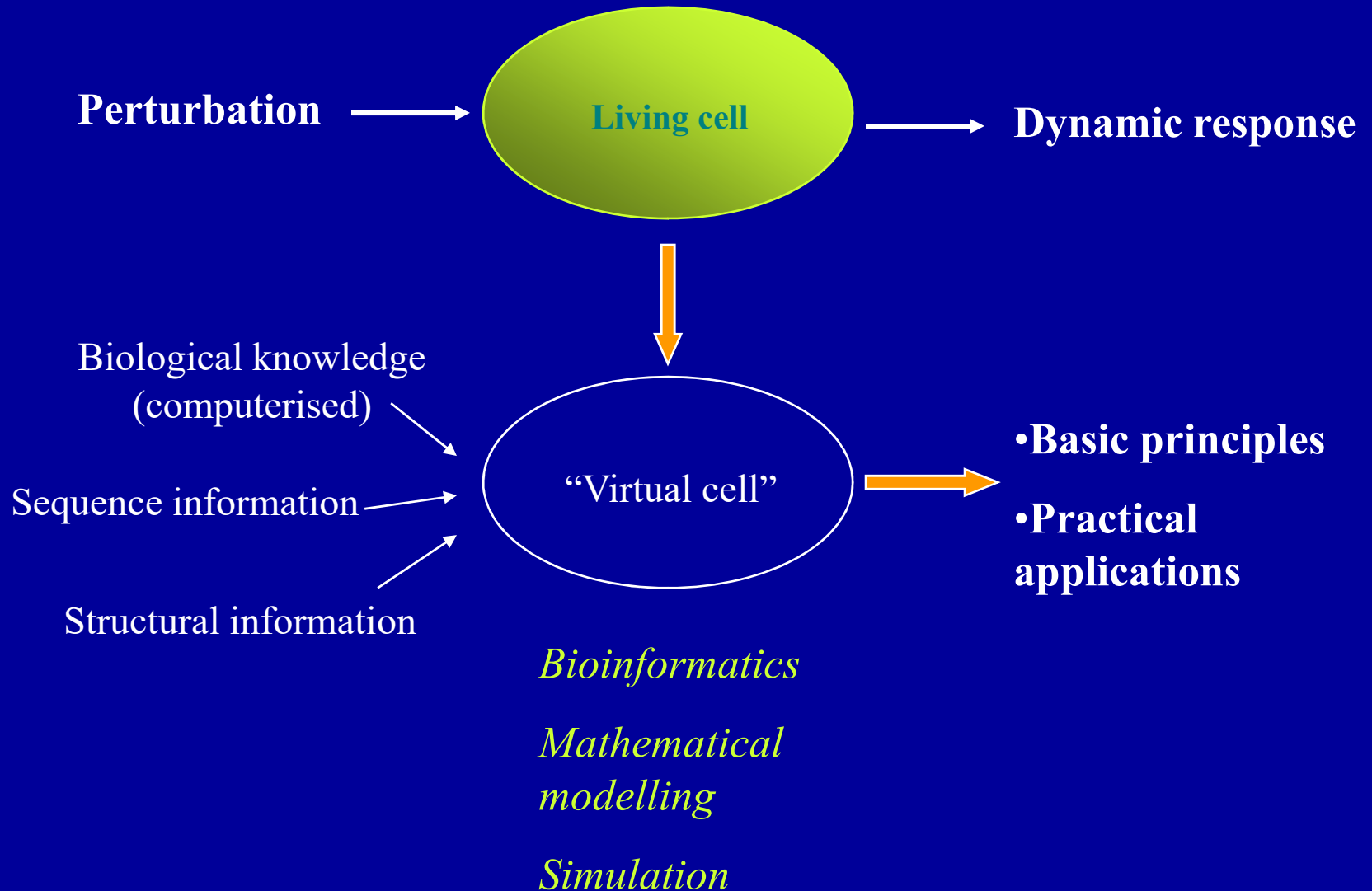
# Global approaches: Toward a new Systems Biology



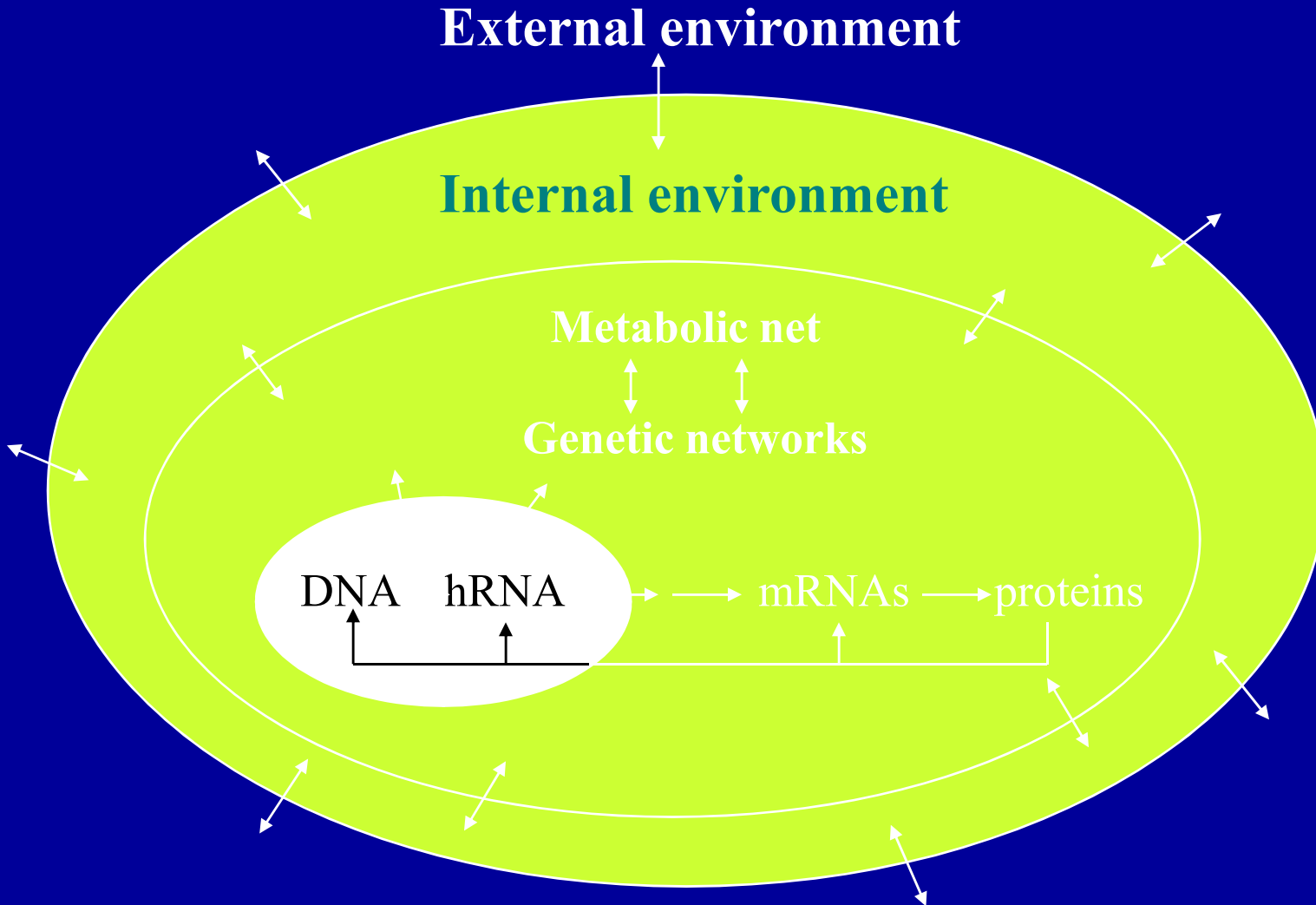
•How does the spatial and temporal organisation of living matter give rise to biological processes?

tissue imaging ↔ EM ↔ X-ray, NMR  
cells  
molecular complexes

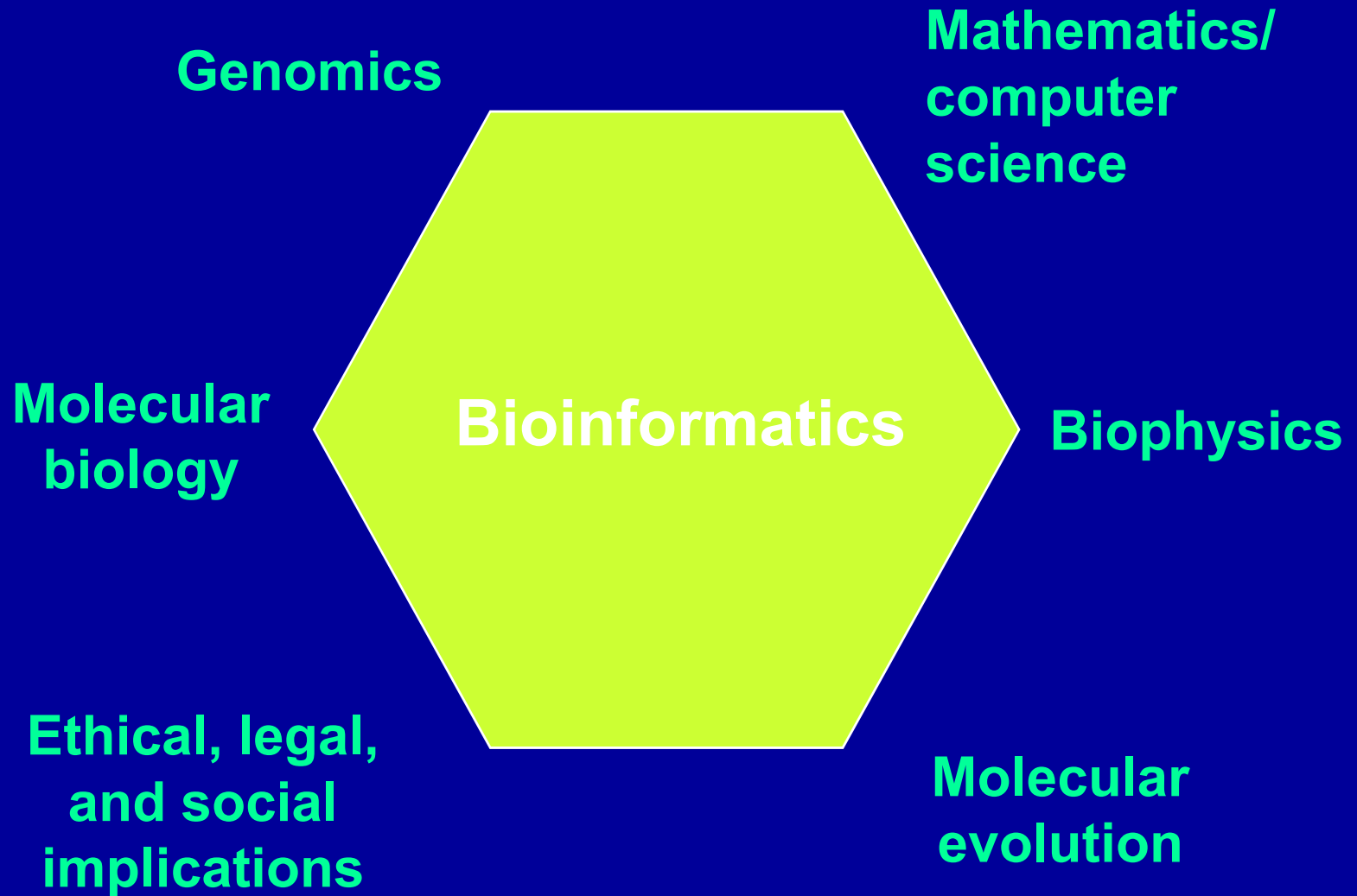
# Global approaches: Toward a new Systems Biology



**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*

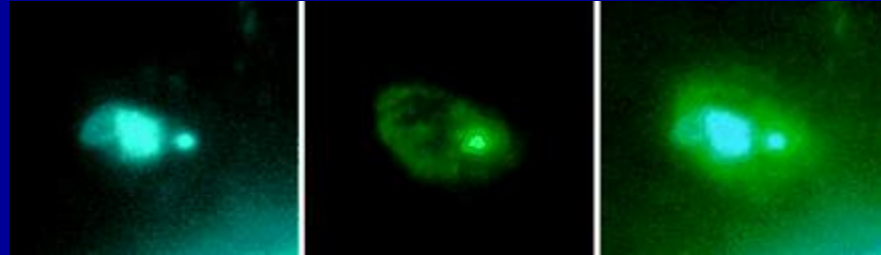


Target identification for antimalarial  
drugs

# 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).

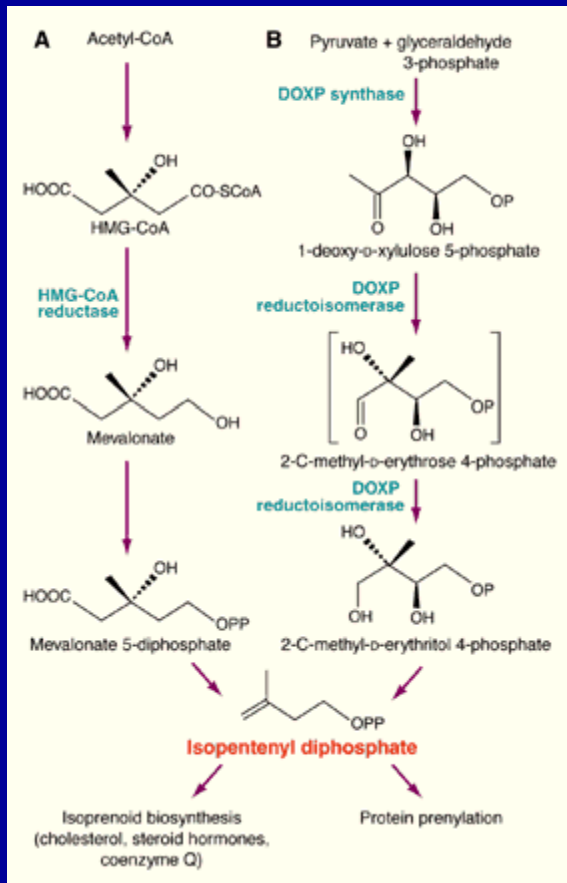
- Comparative genome analysis: Search for orthologs.

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

- Potential target for antimalarial drugs:

DOXP reductoisomerase

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



Pfal	MKKYIIYFFFITITINDLVINNTSKCVSIERRKNNAYINYGIGYNGPDNKITKSRRCRKRILCKCKDLIDIGATKKKPINV
Ecol	-----MK-----QL
Bsub	-----MK-----NI
Syne	-----MVK-----RI

Pfal	AIEGSGSICGALNIIIECNKLIENVNVKALVNNKSVNELYEQAREFLBEYICHHDKSVYEBLKEVLKNIKDKPKPIILC
Ecol	TLIGSTGSIGCSLTDVVRHN---PEHFRVVALVAGKRVTRAVEOCLEFSRRAVYDDEASAKLKKMIQQQS-RTEVLS
Bsub	CILGATGSIGEQTLDLVLRH---QDFQLVSMFGRNIDKAVEMIEVFOPEKVEVSGDLDTYHRLKQMS---FSIECQIGL
Syne	SILGSTGSICGTQLDLVTHH---PDAFOVVGAAACGNVALLAQOVAEERREIYALROAEKLEDLKAVALDLMQPMYVV

Pfal	GDEGKEICSSNSIDKIVLIGIDSFQGLYSTMYALMNNKIVALANKESIVSAGFFFKKLENIHKNAKLIIPVDSEHSAIFOC
Ecol	QQAACDMAALEDDVDQVAAIVGAAQLLPTLAAIRACKTILLANKESLVTGCRUFMDAVK-QSKAQLLPVDSEHNAIFOS
Bsub	GEEGLIEAAVMEVDIVVNALLGSVGLIPTLKAIEQKRTIALANKETLVTAGHIVKEHAK-RYDVP LLPVDSEHSAIFQA
Syne	GEEGVVEVARYDAESVVVIGVGCAGLLPTMAAIAAGKDFALANKETLIIAGAPVVLPLVE-KMGVKLLPADSEHSAIFOC

Pfal	LNNKVLKTKCLQDNFSKINNINKIELCSSGGFFONLTMDELKNVTSENALKHPNKKMGKKITIDSATMNNKGLVIEIHH
Ecol	LPQPIQHNLGYADLE---QNGVVSILLTSGGCFRETPTLRDLATMPDQACRHPNWSMGRKISVDSATMNNKGLVIEIAR
Bsub	LQ-----GEO---AKNIEERLIITASGGSFRDKTREELSVTVEDALKHPNWSMCAKITIDSATMNNKGLVIEIHH
Syne	LQ-----GVP---EGGFRIIILTASGGAFLPLVERLPFVTVQDALKHPNWSMCAKITIDSATMNNKGLVIEIHH

Pfal	FLFDVVDNDIEVIVHKECIHSCVEEIDKSVISQMYPPDMQIPILYSLTWPDRIK-TNLKPLDLAOVSTLTFHKPSELEHF
Ecol	WLFNASASQMEVLIHPQSVIHSMVVWQCSVLAQLGEPDMRTPIAHTNAWPNRVN-SGVKPLDECKLSALTEAAPPDMDRY
Bsub	WLEDFHPEQIDVVLHKEIHSMSVEEHDKSVIAQLGTPDMRVPIQALTYPPDRLPPLDANKRLELWEGSLHPEKADEDFRE
Syne	YLFGLDYDHDIDIVIHPCSIHSLIEVQDASVLAQLGWPDMRLEPLLYALSWPERI-YDWEPLDLVKAGSLSFREPDHDKY

Pfal	PCIKLAYQAGIKGNFYPTVLNASNEIANNLNNKIKKEDISSIISQVLESFNSSQVSENSEDLMKQILQIHSWAKDKAT
Ecol	PCLKLAPEAFEGQQAATTALNAANEITVAAFLAQQHREFDIAAL---NLSVLEKMDYRE--EQCVDDVLSVDANAREVAR
Bsub	RCLQFAEBSCKIGGTMPTVLNAANEVAVAAFLAGKIPFLAIEDCEKA--LTRHQLLKK--PS-WRTF---KNWTHIPED
Syne	PCMOLAYGACRAGGAMPVVLNAANECAVALFLOEKISFLDIPRLIEKTCDIYVQNTAS--PD-ETTLAADQAR---R

Pfal	DIYNKHNS-----
Ecol	KEVMRLAS-----
Bsub	TSIQYSHKVVCS-
Syne	TVIENSACVATRP

# 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
  - Appletalk 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 than NT.
- Most software packages that run on Linux have their source code available, security problems are found and solved many times quicker than 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

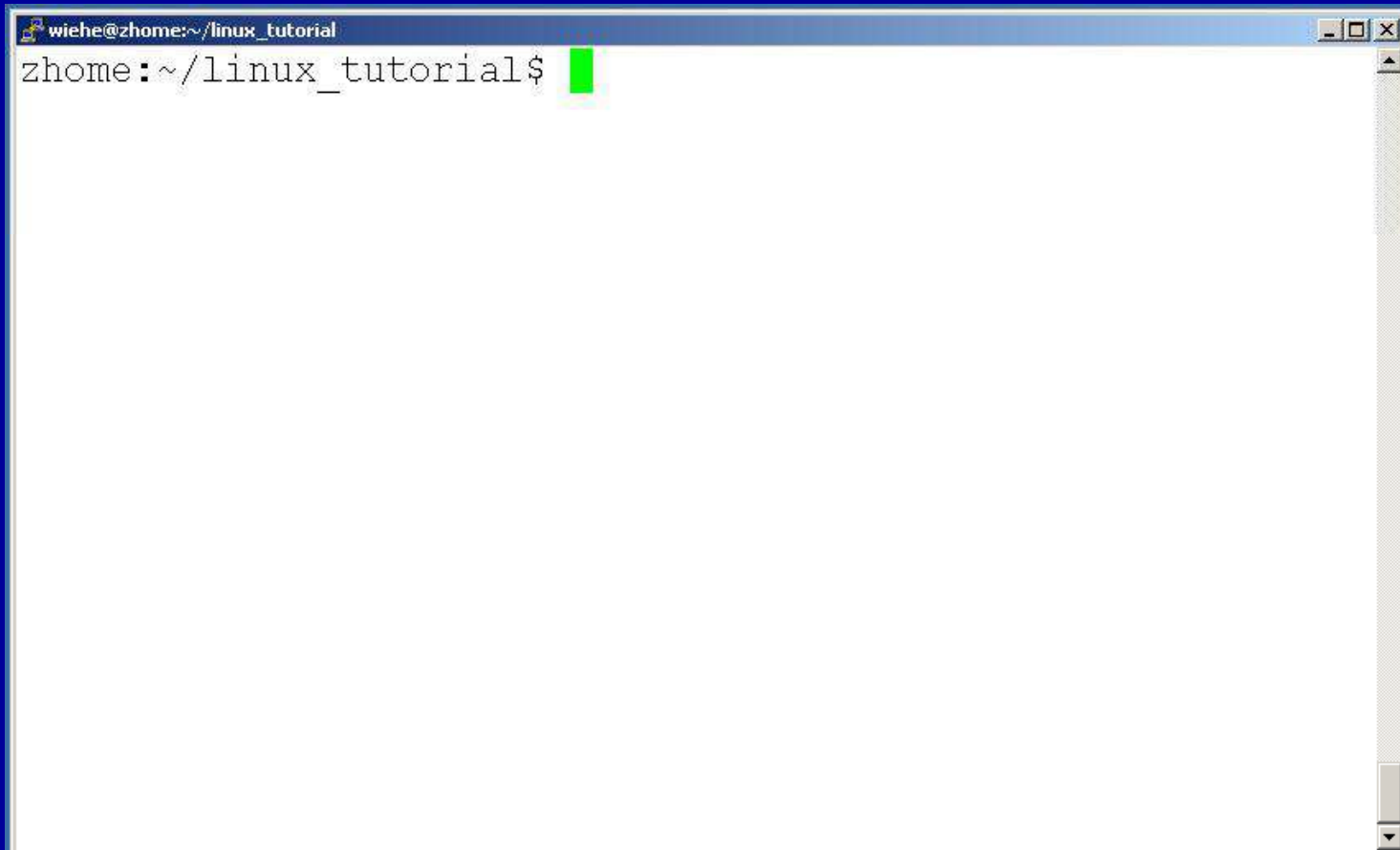
- Unix was the predecessor of Linux
- Linux is a variant of Unix
  - So is Mac OS X, so much of this tutorial applies to Macs as well
- Linux is open source
- Most of the machines you'll use in the Bioinformatics program are running the Linux OS

# 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

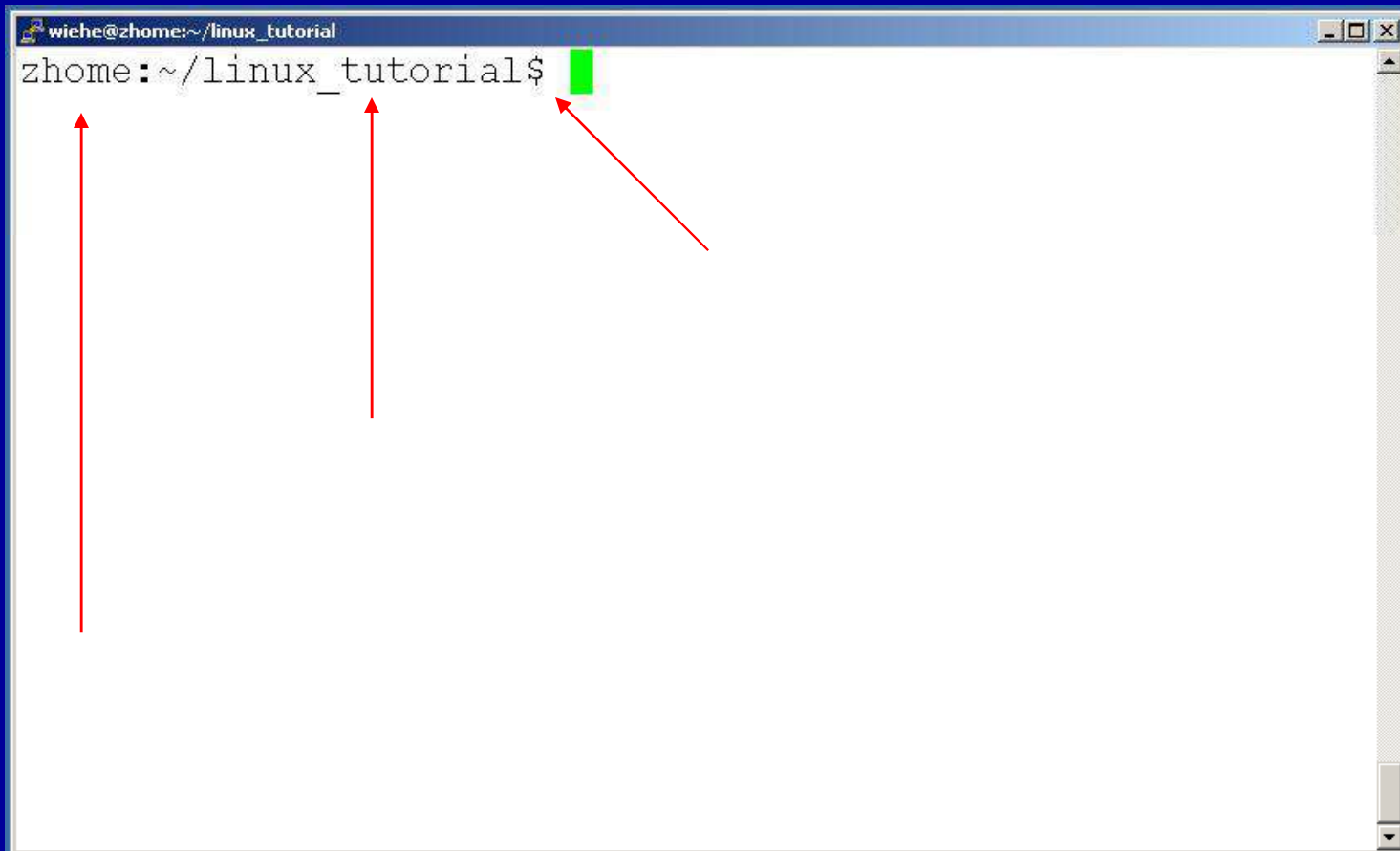
- Open up a terminal:

A screenshot of a terminal window. The title bar at the top reads "wiehe@zhome:~/linux\_tutorial". The main area of the terminal shows the prompt "zhome:~/linux\_tutorial\$" followed by a green cursor block. The window has standard Linux window controls (minimize, maximize, close) in the top right corner and a scrollbar on the right side.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$
```

# Connecting to a Unix/Linux system

- Open up a terminal:



A terminal window titled "wiehe@zhome:~/linux\_tutorial" is shown. The prompt is "zhome:~/linux\_tutorial\$". A green cursor is positioned at the end of the prompt. Three red arrows point to the prompt: one to the "zhome:" part, one to the "~/" part, and one to the "\$" part.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$
```

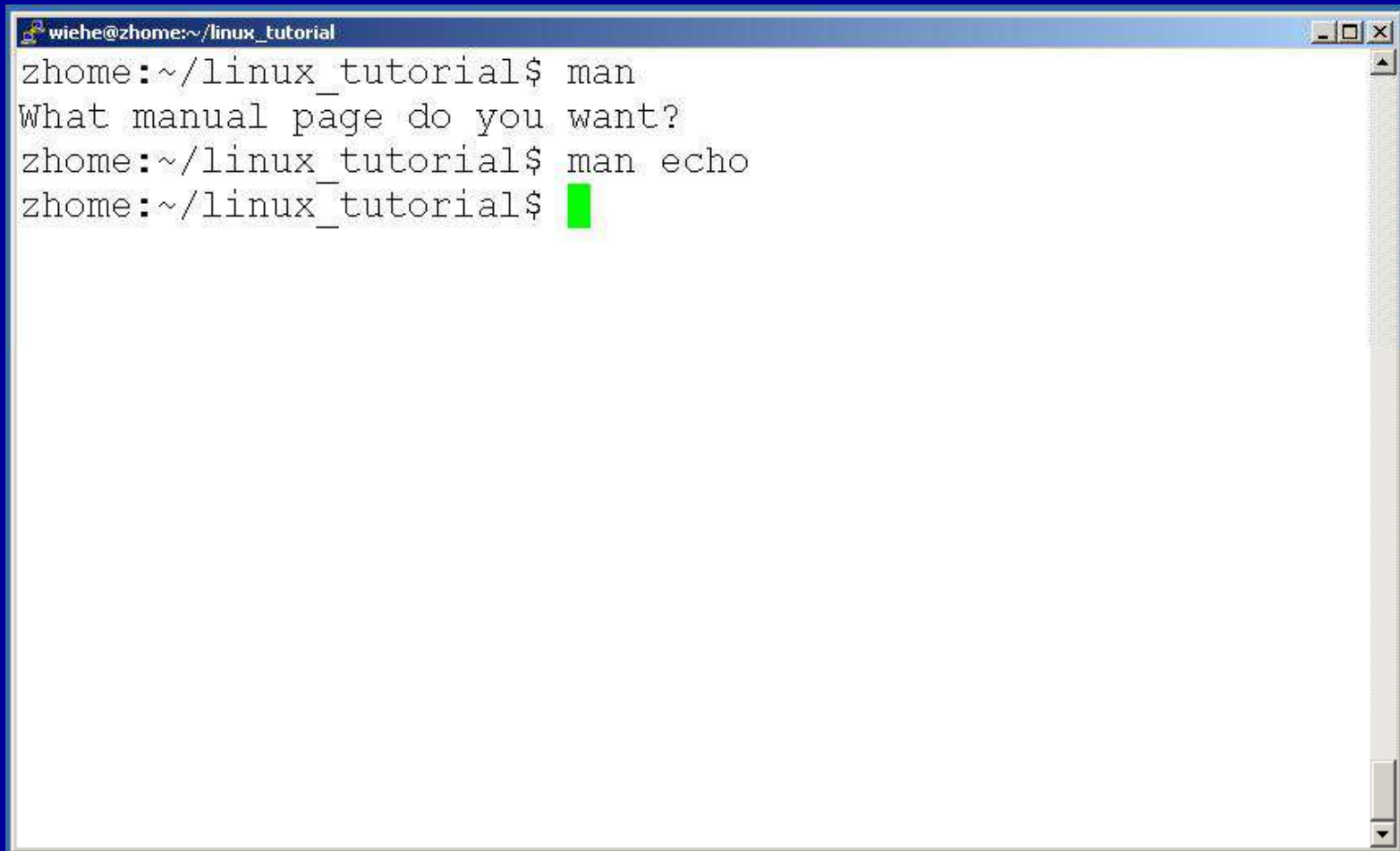
# What exactly is a “shell”?

- 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!

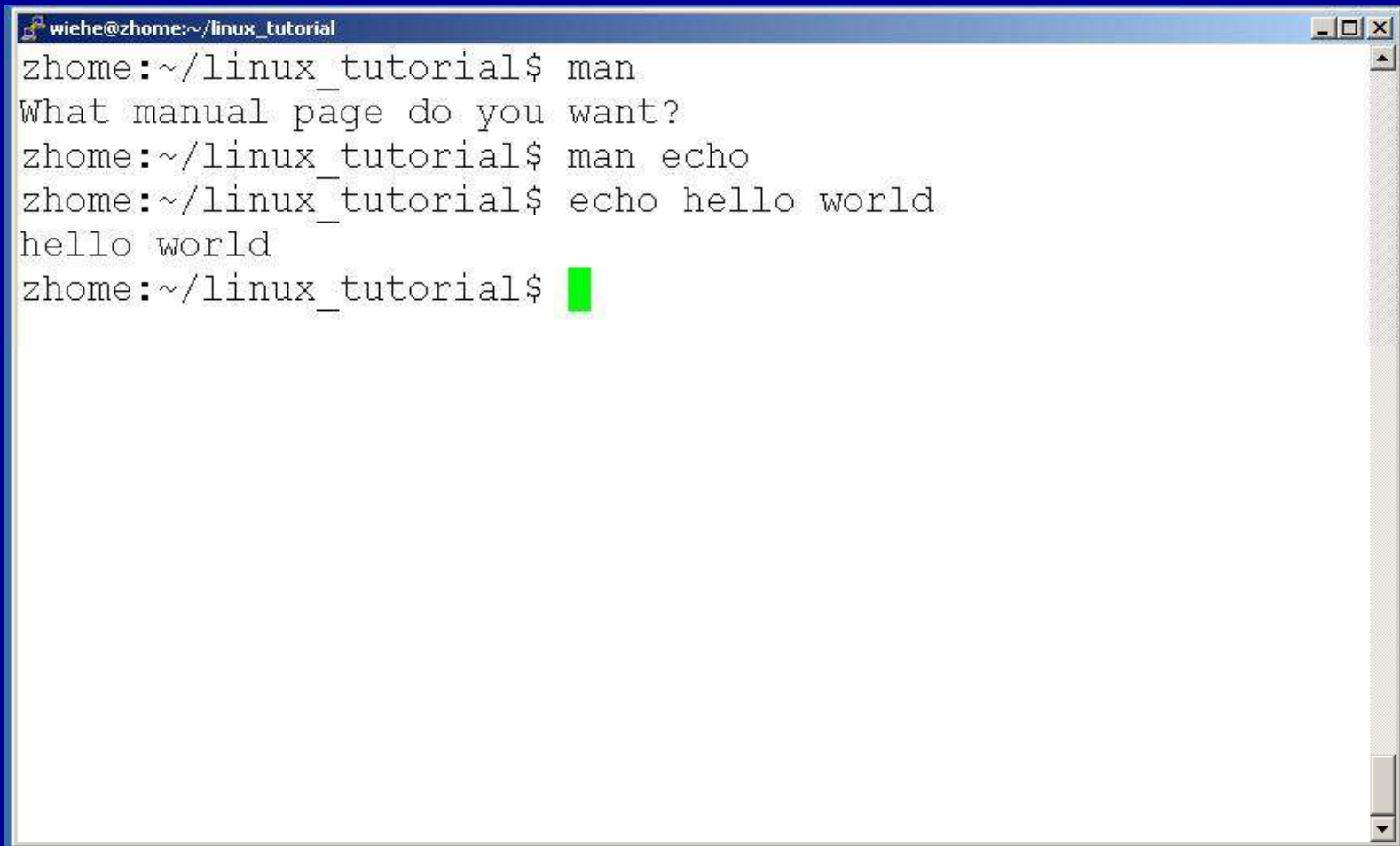
A terminal window with a blue title bar containing the text "wiehe@zhome:~/linux\_tutorial". The window has standard window controls (minimize, maximize, close) in the top right corner and a vertical scrollbar on the right side. The terminal content shows a sequence of commands and their outputs:

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

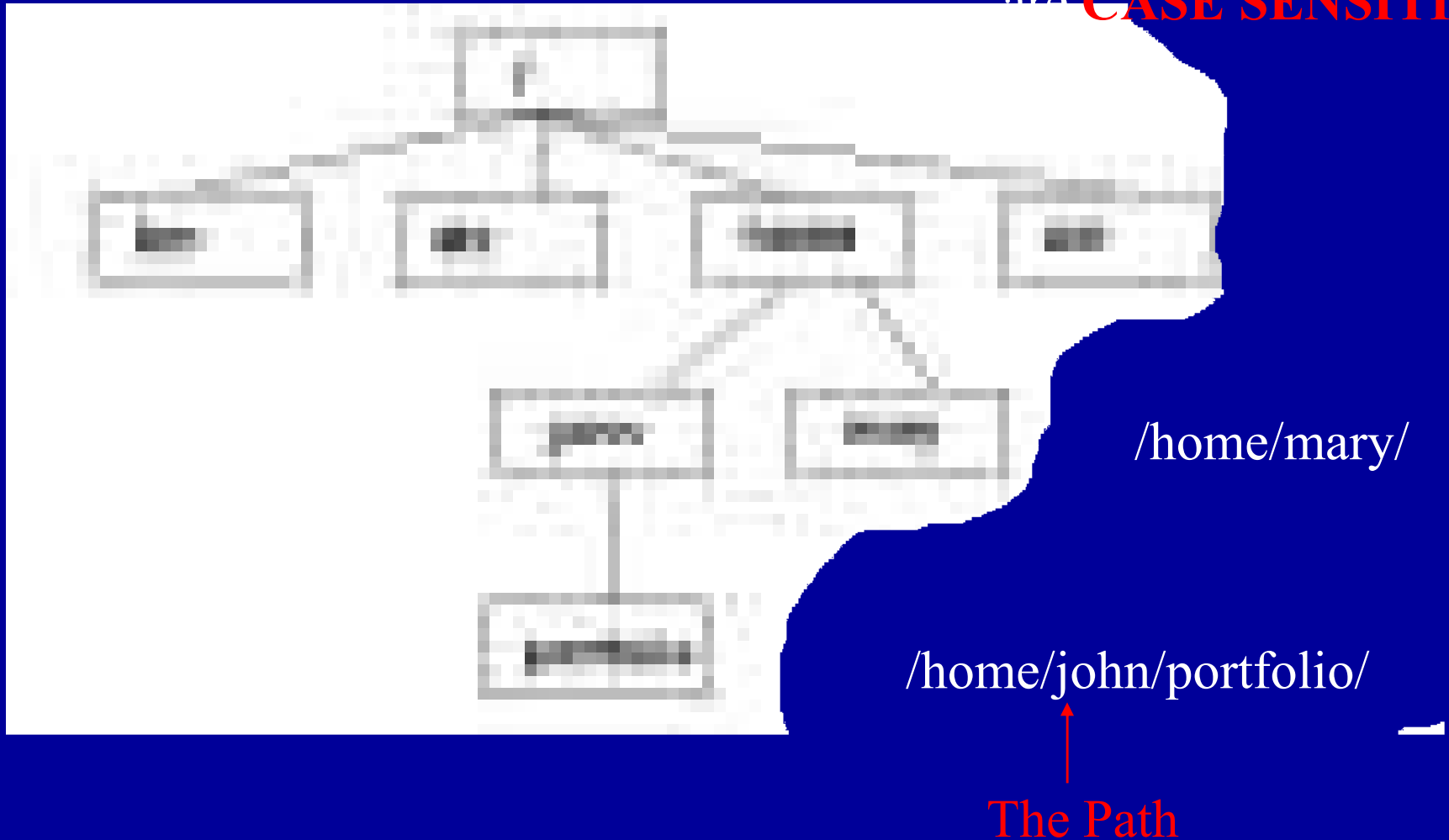
# Help!



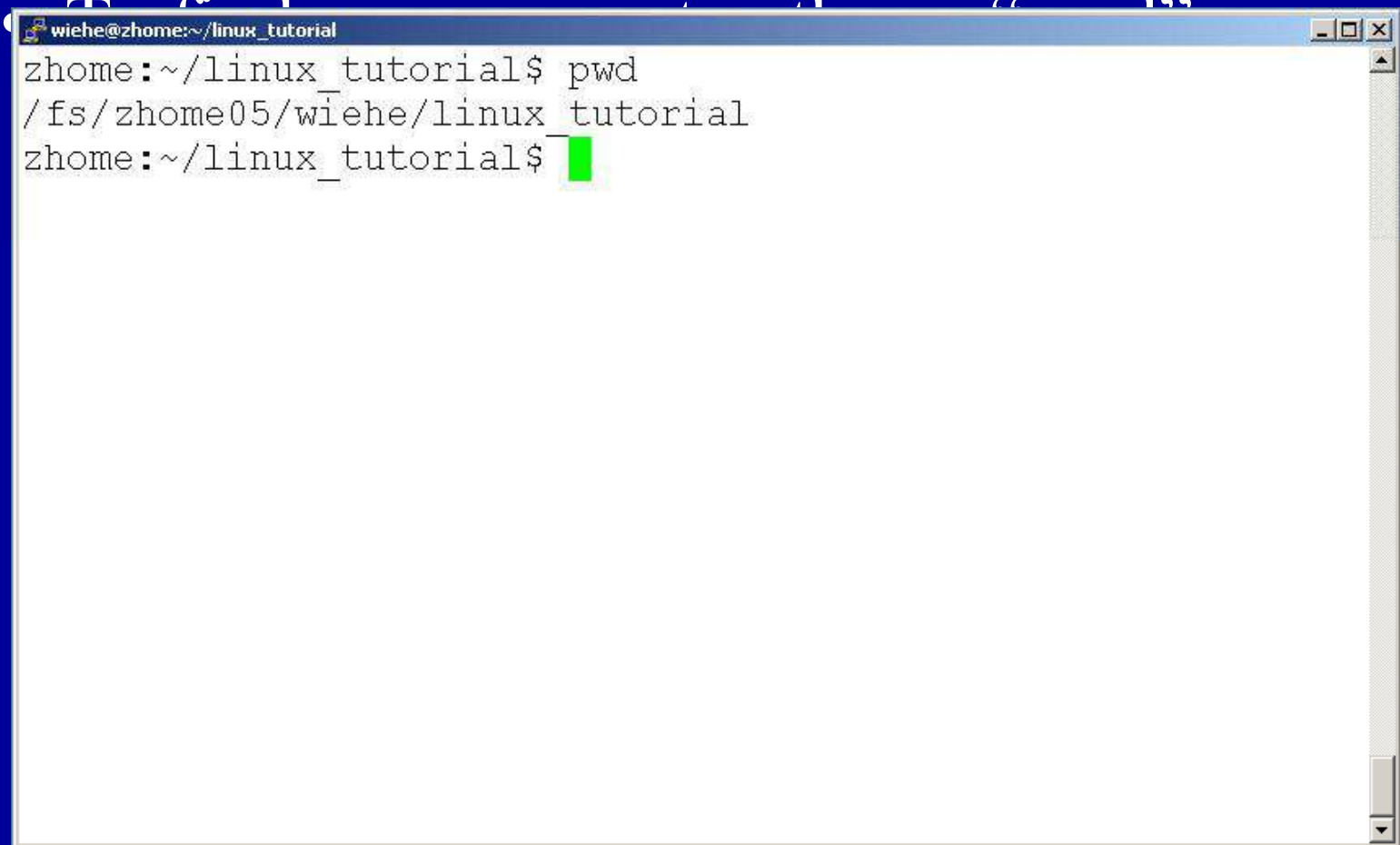
```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ man
What manual page do you want?
zhome:~/linux_tutorial$ man echo
zhome:~/linux_tutorial$ echo hello world
hello world
zhome:~/linux_tutorial$ █
```

# Unix/Linux File System

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



# Command: pwd



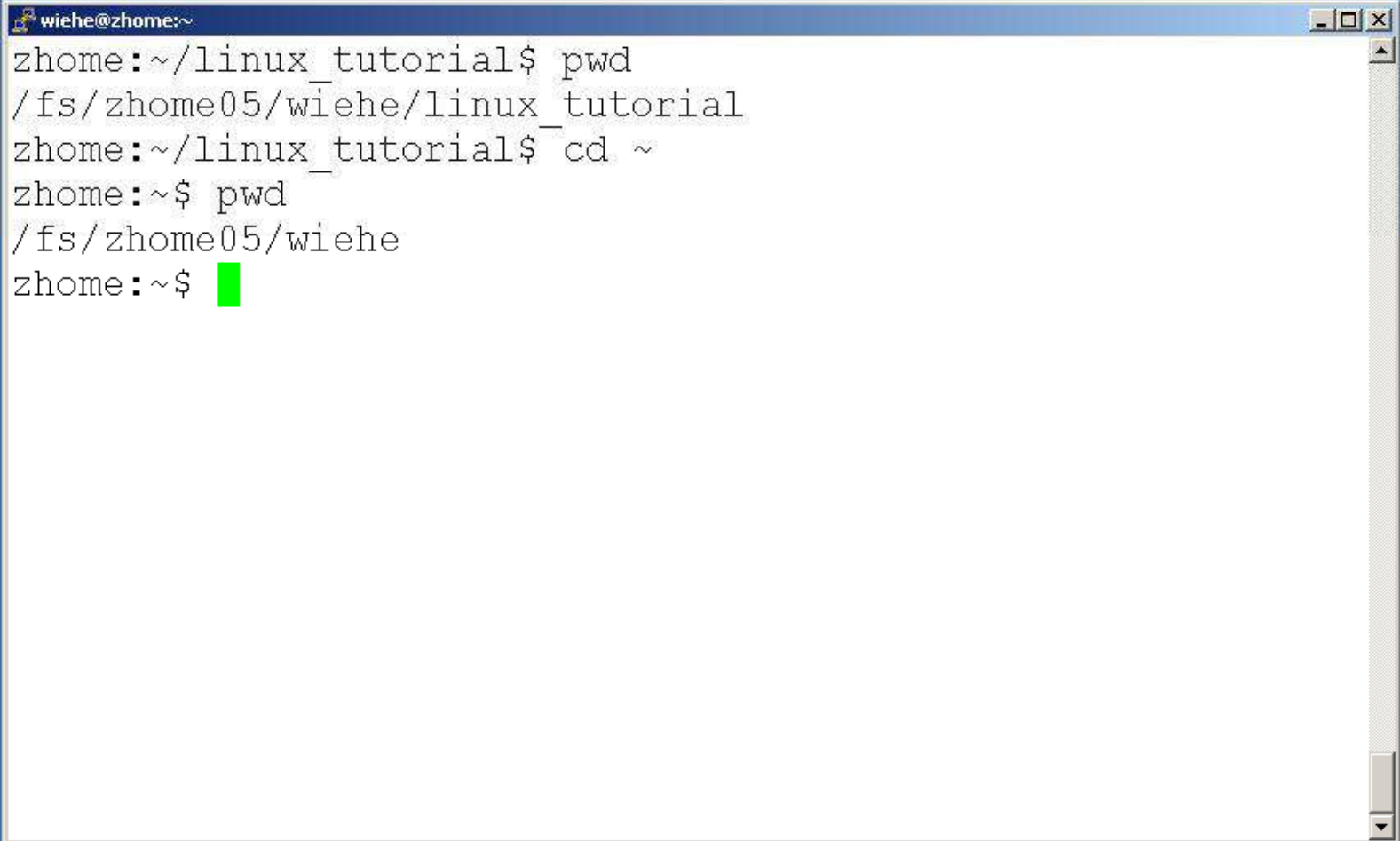
```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ pwd
/fs/zhome05/wiehe/linux_tutorial
zhome:~/linux_tutorial$
```

A terminal window titled "wiehe@zhome:~/linux\_tutorial" showing the execution of the "pwd" command. The output is "/fs/zhome05/wiehe/linux\_tutorial". The prompt "zhome:~/linux\_tutorial\$" is shown again on the next line with a green cursor.

# 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$ █
```

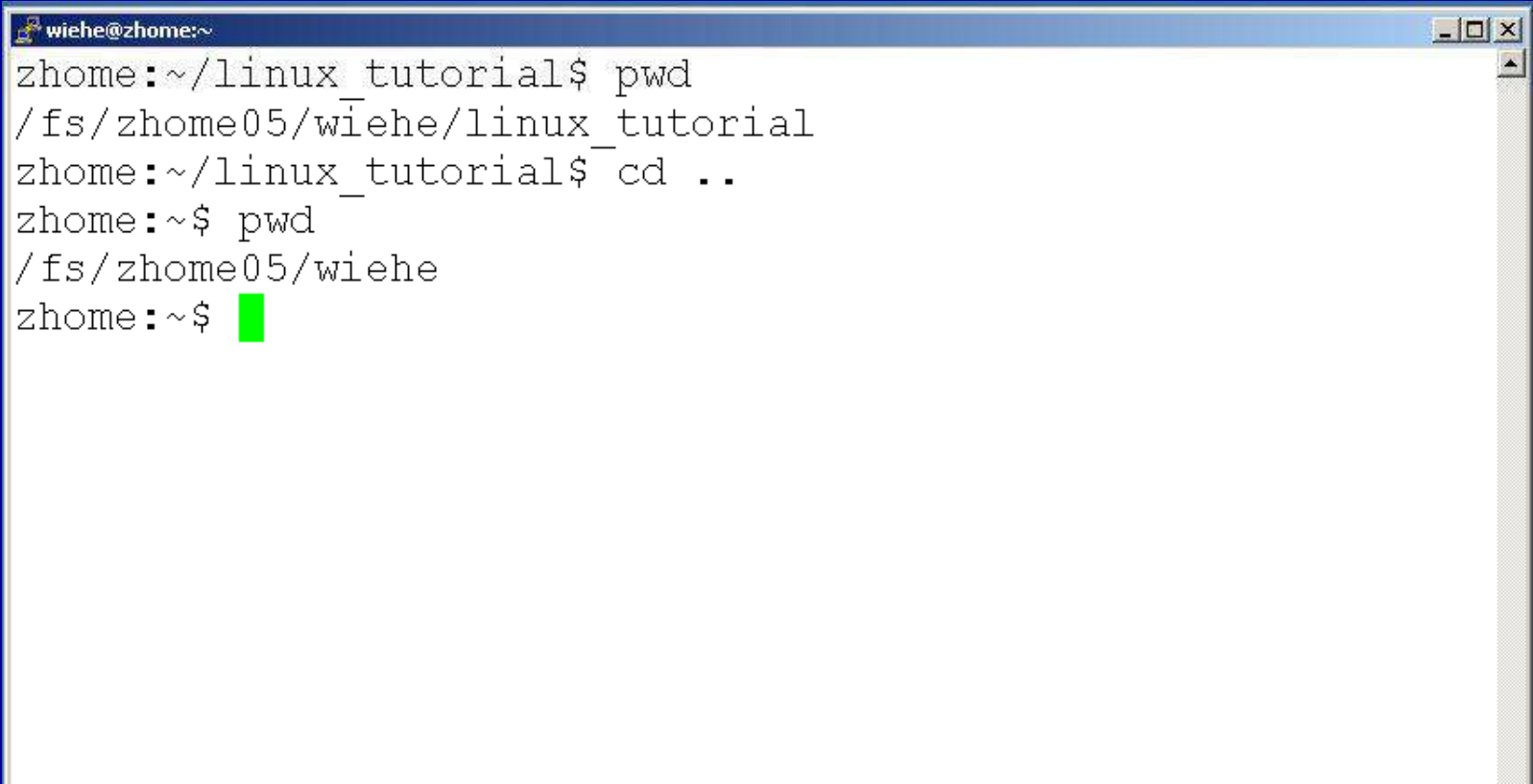
# Command: cd

- A terminal window titled "wiehe@zhome:~" showing the following commands and output:

```
zhome:~/linux_tutorial$ pwd
/fs/zhome05/wiehe/linux_tutorial
zhome:~/linux_tutorial$ cd ~
zhome:~$ pwd
/fs/zhome05/wiehe
zhome:~$ █
```

# Command: cd

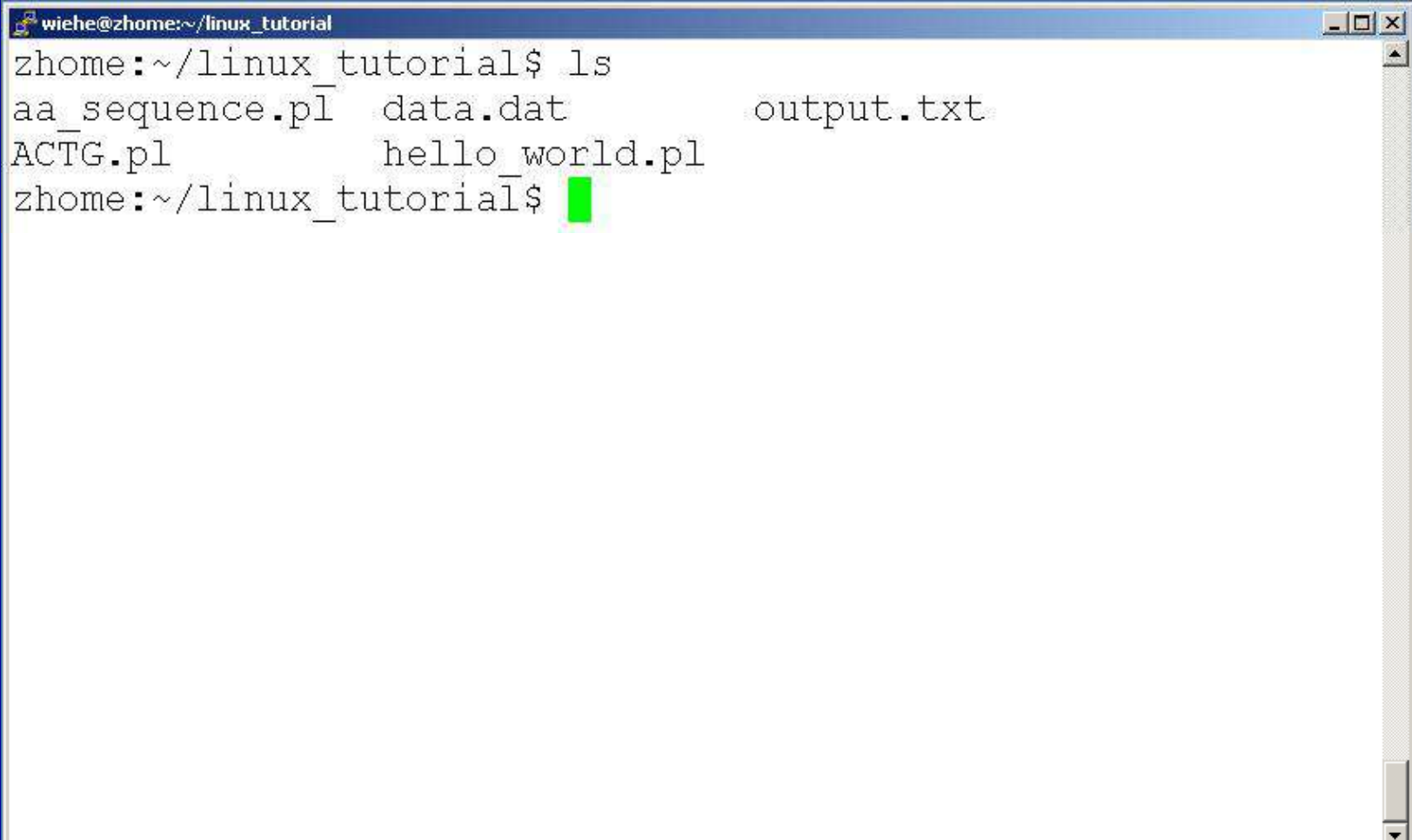
- “..” is the location of the directory below current one

A terminal window titled 'wiehe@zhome:~' showing a sequence of commands and their outputs. The user starts in the directory ~/linux\_tutorial, runs 'pwd' to get /fs/zhome05/wiehe/linux\_tutorial, then runs 'cd ..' to move to the parent directory, and finally runs 'pwd' to get /fs/zhome05/wiehe. A green cursor is visible at the end of the final prompt.

```
wiehe@zhome:~  
zhome:~/linux_tutorial$ pwd  
/fs/zhome05/wiehe/linux_tutorial  
zhome:~/linux_tutorial$ cd ..  
zhome:~$ pwd  
/fs/zhome05/wiehe  
zhome:~$ █
```

# Command: ls

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

A terminal window titled 'wiehe@zhome:~/linux\_tutorial' showing the execution of the 'ls' command. The output lists four files: 'aa\_sequence.pl', 'data.dat', 'output.txt', and 'hello\_world.pl'. The prompt 'zhome:~/linux\_tutorial\$' is shown again at the end of the output, followed by a green cursor.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat      output.txt
ACTG.pl        hello_world.pl
zhome:~/linux_tutorial$ █
```

# 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”

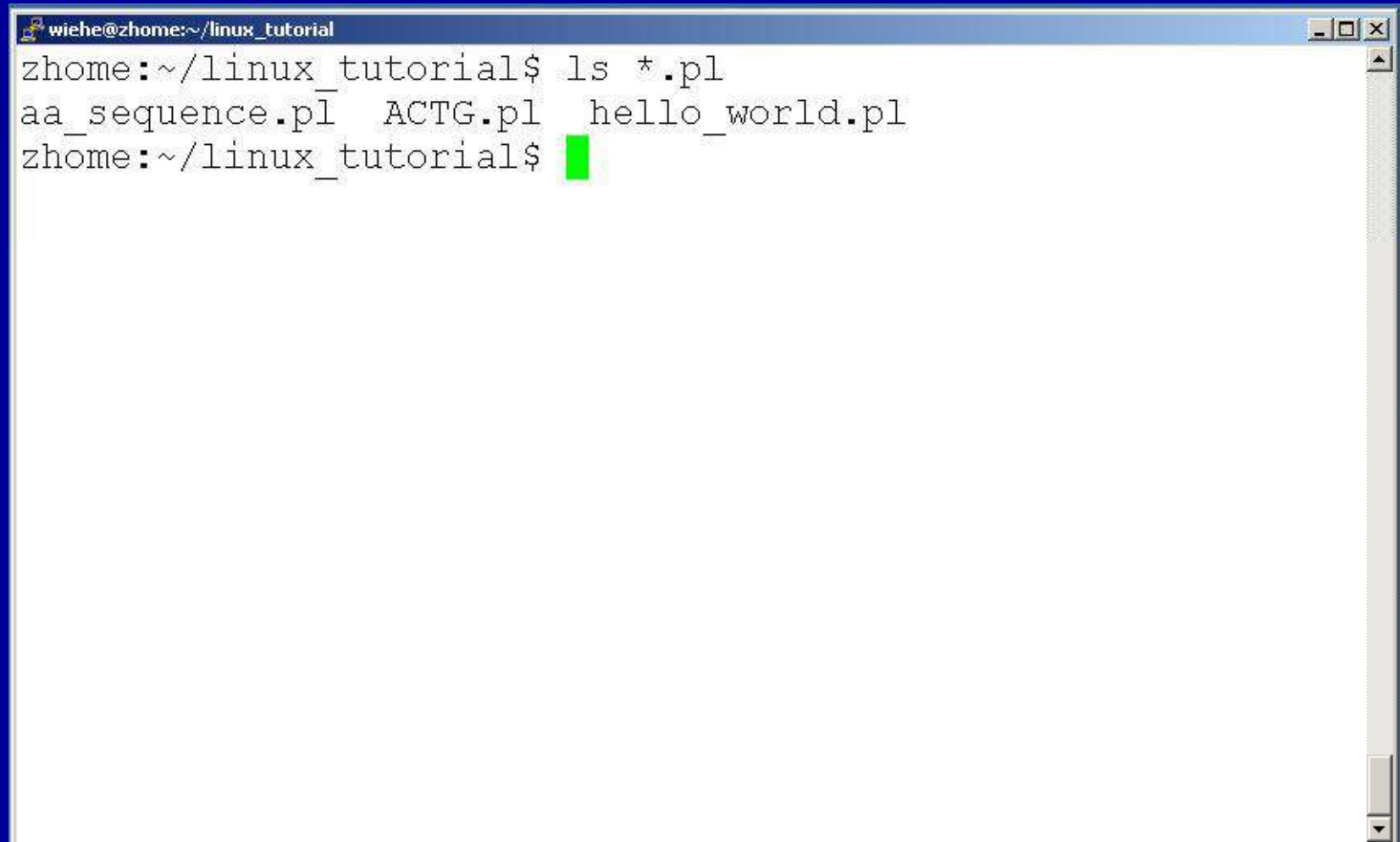
# Command: ls -ltr

- List files by time in reverse order with long listing

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls -ltr
total 20
-rw-rw-r-- 1 wiehe wiehe  92 Aug 30 11:54 ACTG.pl
-rw-rw-r-- 1 wiehe wiehe 169 Aug 30 12:20 aa_sequence.pl
-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 output.txt
-rw-rw-r-- 1 wiehe wiehe  21 Aug 30 12:23 data.dat
zhome:~/linux_tutorial$ █
```

# General Syntax: \*

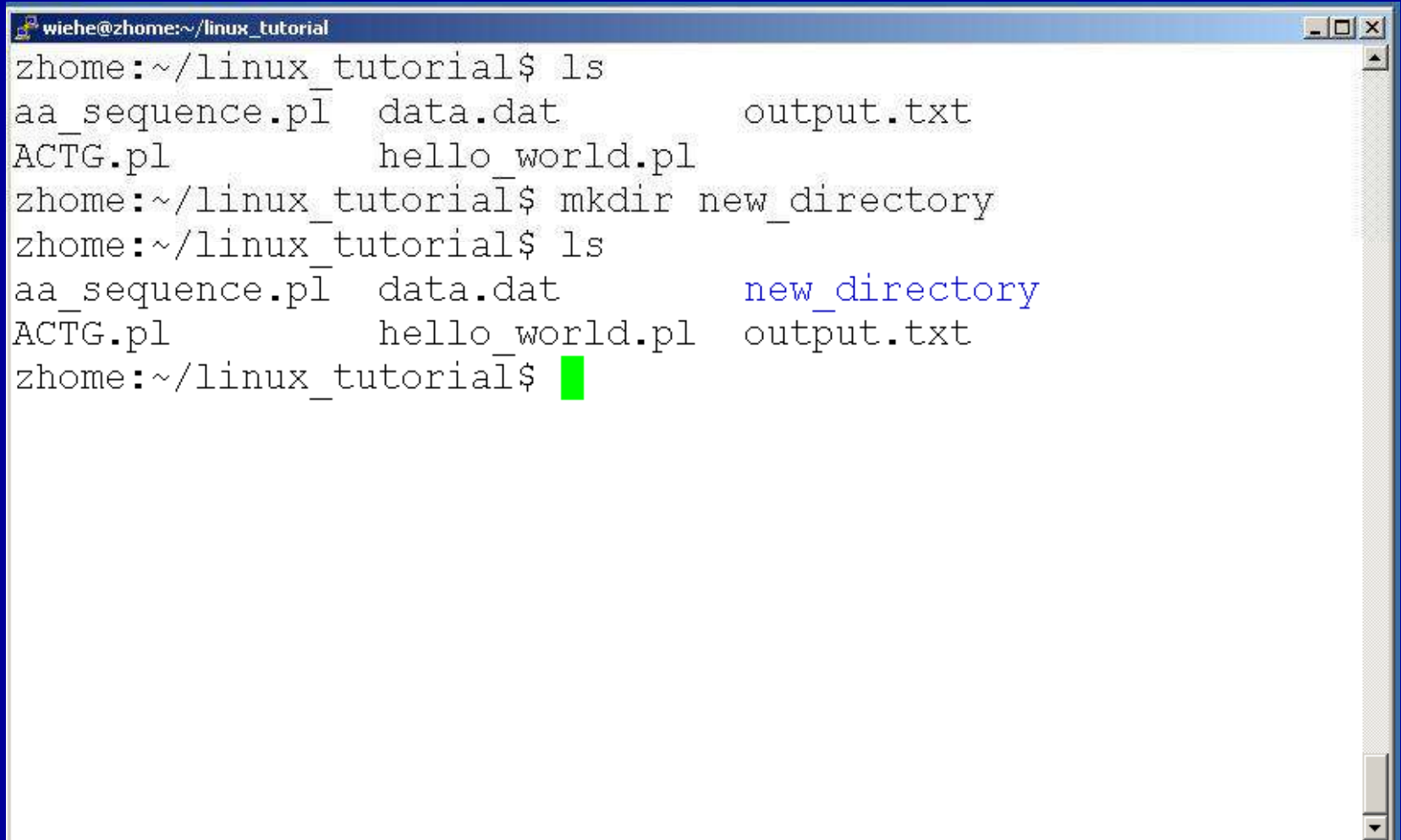
- “\*” can be used as a wildcard in unix/linux

A terminal window with a blue title bar containing the text 'wiehe@zhome:~/linux\_tutorial'. The terminal content shows a command 'ls \*.pl' being executed, resulting in the output 'aa\_sequence.pl ACTG.pl hello\_world.pl'. The prompt 'zhome:~/linux\_tutorial\$' is visible at the end of the line.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls *.pl
aa_sequence.pl  ACTG.pl  hello_world.pl
zhome:~/linux_tutorial$
```

# Command: mkdir

- To create a new directory use “mkdir”



```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat          output.txt
ACTG.pl        hello_world.pl
zhome:~/linux_tutorial$ mkdir new_directory
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat          new_directory
ACTG.pl        hello_world.pl   output.txt
zhome:~/linux_tutorial$ █
```

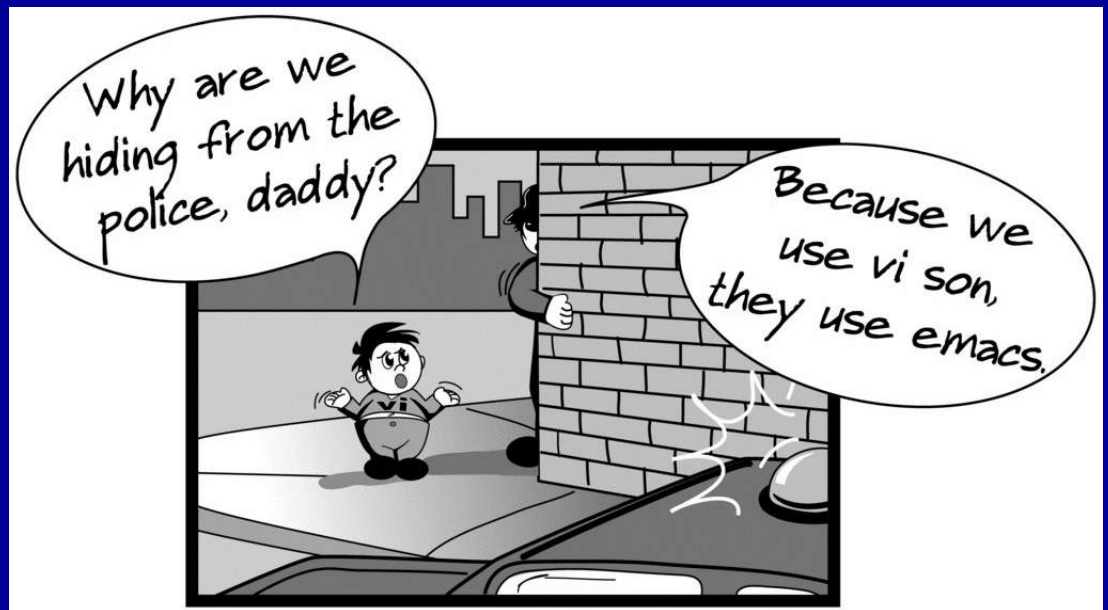
# Command: rmdir

- To remove an empty directory use “rmdir”

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat          new_directory
ACTG.pl        hello_world.pl   output.txt
zhome:~/linux_tutorial$ rmdir new_directory/
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat          output.txt
ACTG.pl        hello_world.pl
zhome:~/linux_tutorial$ █
```

# 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



The image shows a terminal window with the nano text editor. The window title is "wiehe@zhome:~/linux\_tutorial". The editor's status bar at the top displays "UW PICO (tm) 4.6" and "New Buffer". A green cursor is visible on the left side of the editor area. On the right side, there is a small tooltip that says "Learning UNI:". At the bottom of the window, a help menu is displayed with the following text:

```
^G Get He ^O WriteO ^R Read F ^Y Prev P ^K Cut Te ^C Cur Po
^X Exit  ^J Justif ^W Where ^V Next P ^U UnCut ^T To Spe
```

# Editing a file using pico



The screenshot shows a terminal window with the pico text editor. The window title is "wiehe@zhome:~/linux\_tutorial". The editor's status bar at the top displays "UW PICO(tm) 4.6", "New Buffer", and "Modified". The main editing area contains the text "Hello World.". At the bottom, a prompt asks to "Save modified buffer (ANSWERING 'No' WILL DESTROY CHANGES)". The user has entered "Y" for "Yes". The prompt also lists "^C Cancel" and "N No".

```
wiehe@zhome:~/linux_tutorial
UW PICO(tm) 4.6          New Buffer          Modified
Hello World.

Save modified buffer (ANSWERING "No" WILL DESTROY CHANGES)
? Y Yes
^C Cancel N No
```

# 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

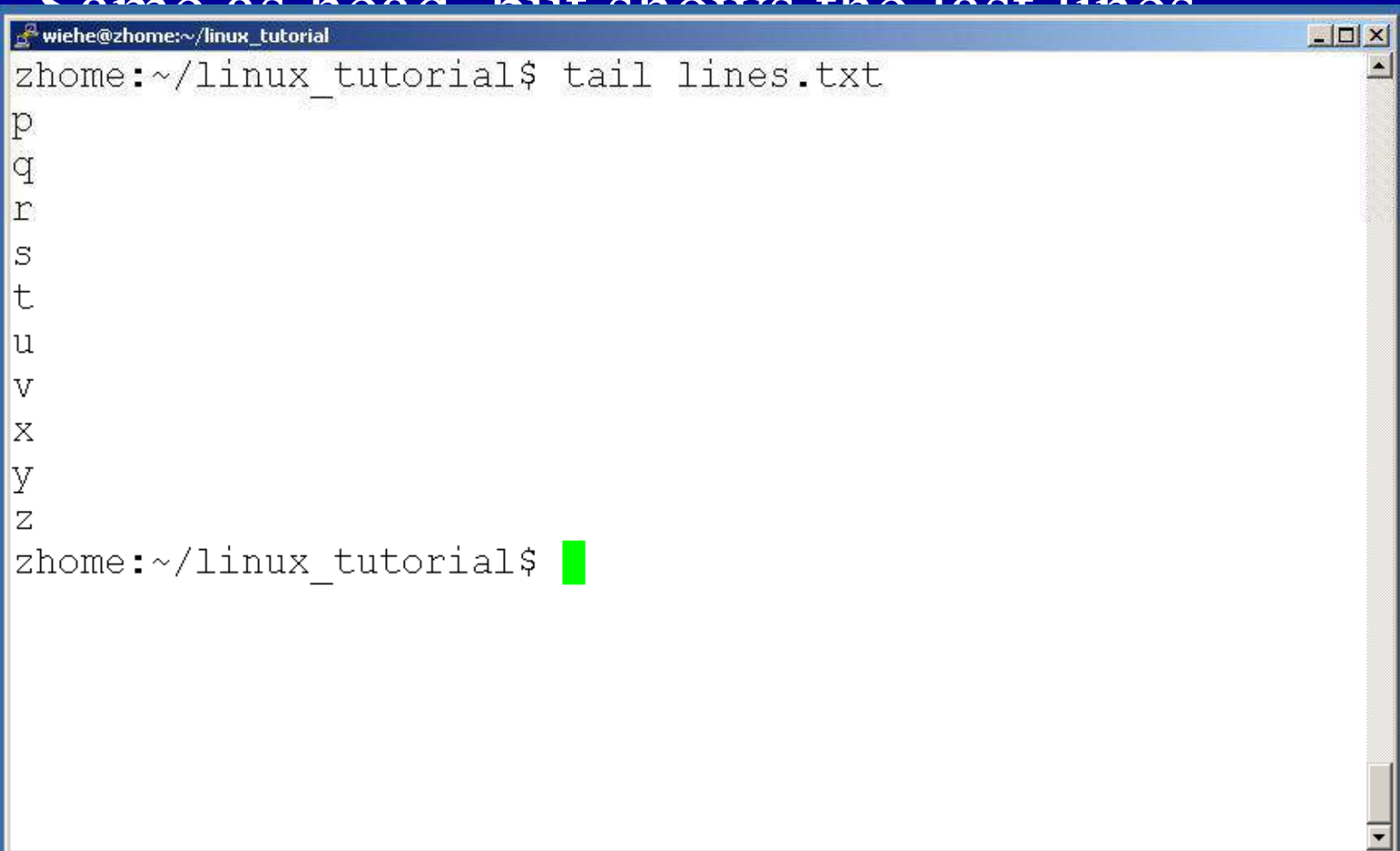
- “head” displays the top part of a file
- By default it shows the first 10 lines
- -n option allows you to change that
- “head -n50 file.txt” displays the first 50 lines of file.txt

# Command: head

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ head lines.txt
a
b
c
d
e
f
g
h
i
j
zhome:~/linux_tutorial$ █
```

# Command: tail

- Same as head, but shows the last lines

A terminal window with a blue title bar containing the text 'wiehe@zhome:~/linux\_tutorial'. The terminal content shows the command 'tail lines.txt' being executed, resulting in the output 'p', 'q', 'r', 's', 't', 'u', 'v', 'x', 'y', 'z'. The prompt 'zhome:~/linux\_tutorial\$' is followed by a green cursor block.

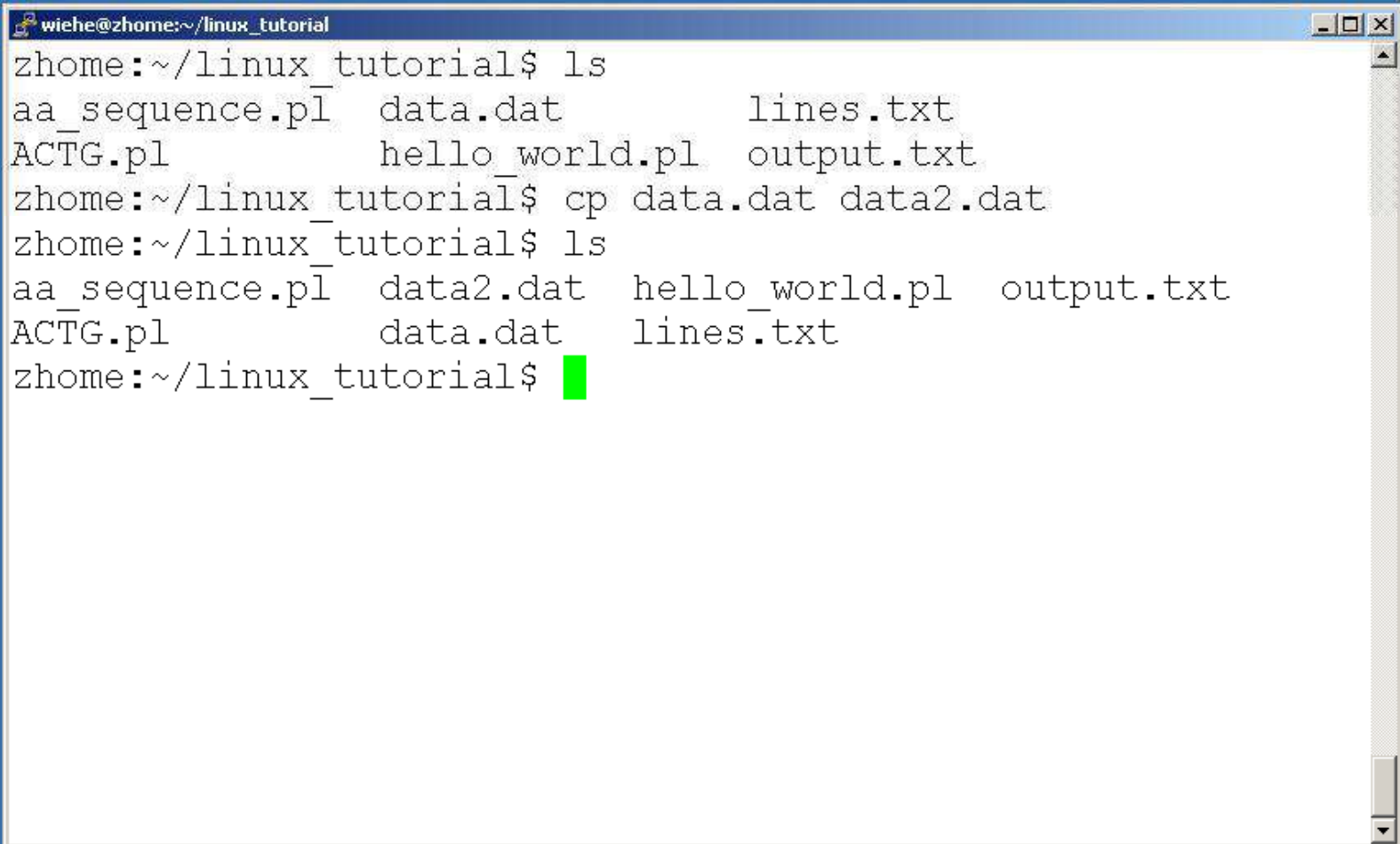
```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ tail lines.txt
p
q
r
s
t
u
v
x
y
z
zhome:~/linux_tutorial$ █
```

# File Commands

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

# Command: cp

- To copy files use “cp”



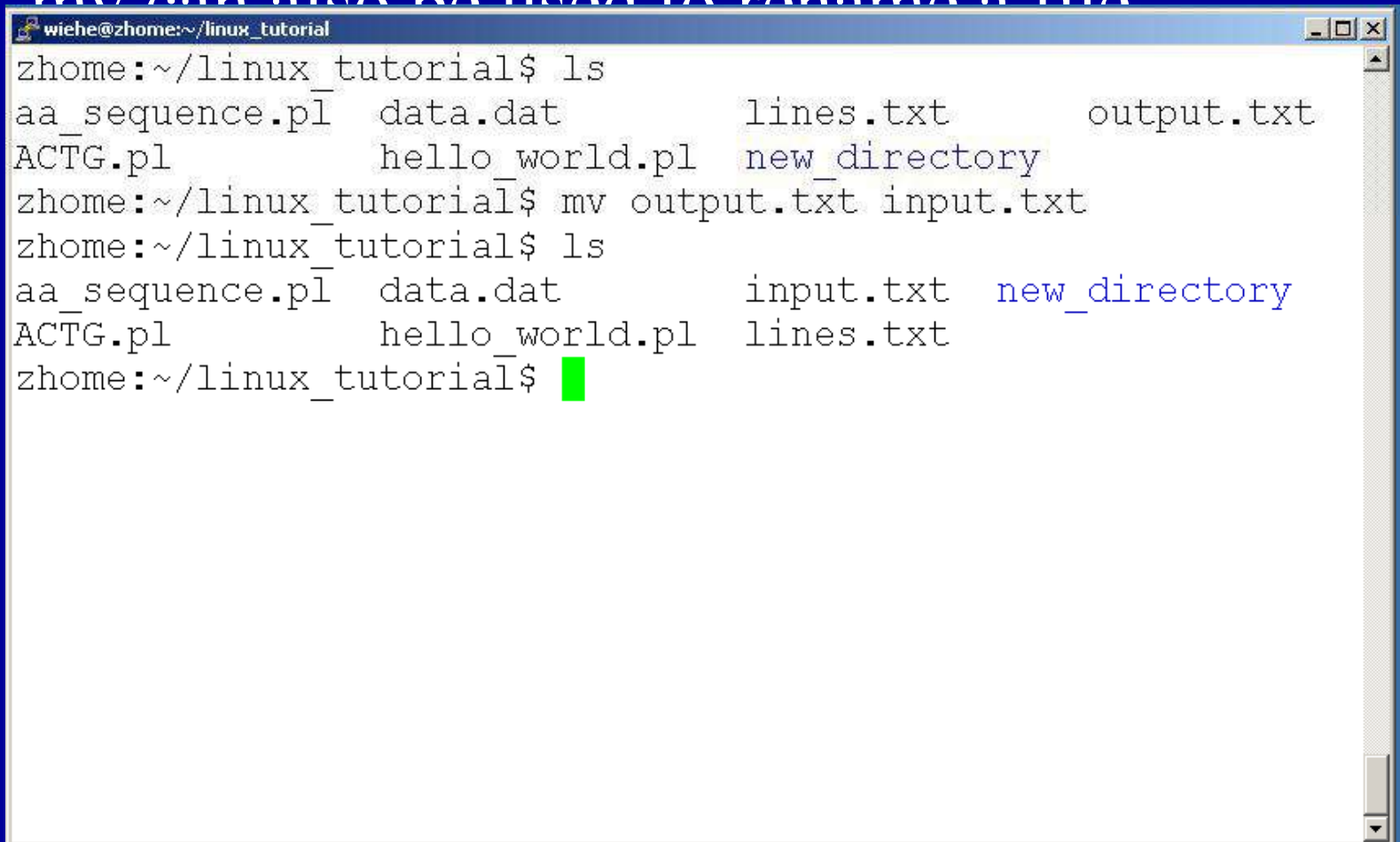
```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat      lines.txt
ACTG.pl        hello_world.pl  output.txt
zhome:~/linux_tutorial$ cp data.dat data2.dat
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data2.dat  hello_world.pl  output.txt
ACTG.pl        data.dat  lines.txt
zhome:~/linux_tutorial$ █
```

# 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



```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat      lines.txt     output.txt
ACTG.pl        hello_world.pl new_directory
zhome:~/linux_tutorial$ mv output.txt input.txt
zhome:~/linux_tutorial$ ls
aa_sequence.pl  data.dat      input.txt     new_directory
ACTG.pl        hello_world.pl lines.txt
zhome:~/linux_tutorial$ █
```

# Command: rm

```
wiehe@zhome:~/linux_tutorial/new_directory
zhome:~/linux_tutorial$ cd new_directory/
zhome:~/linux_tutorial/new_directory$ ls
data2.dat
zhome:~/linux_tutorial/new_directory$ rm data2.dat
zhome:~/linux_tutorial/new_directory$ ls
zhome:~/linux_tutorial/new_directory$ █
```

# 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

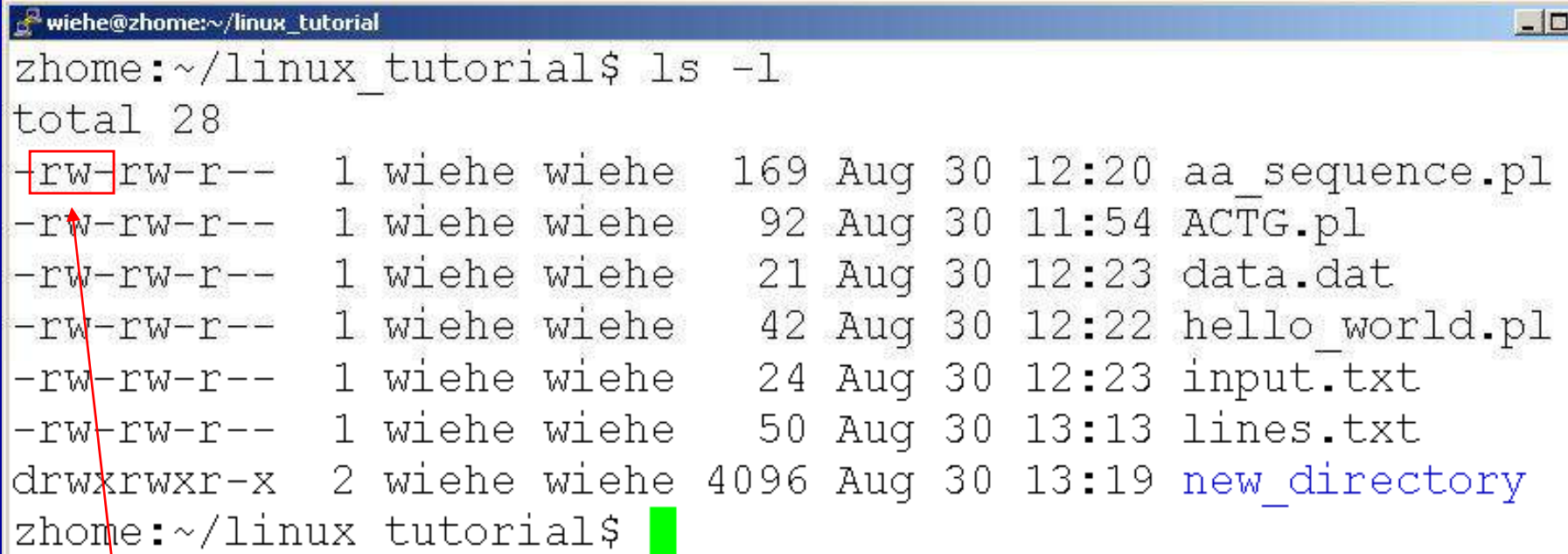
- Each file in Unix/Linux has an associated permission level
- This allows the user to prevent others from reading/writing/executing their files or directories
- Use “`ls -l filename`” to find the permission level of that file

# 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$
```

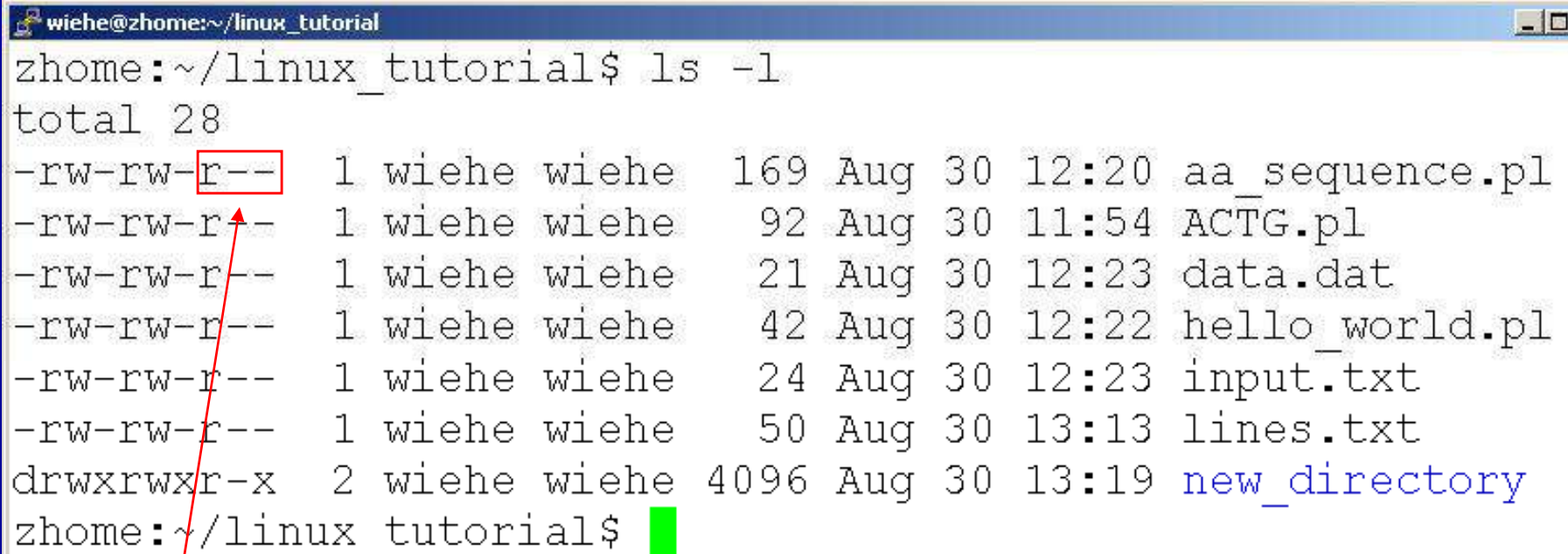


# File Permissions

```
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

```
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$
```



# Command: chmod

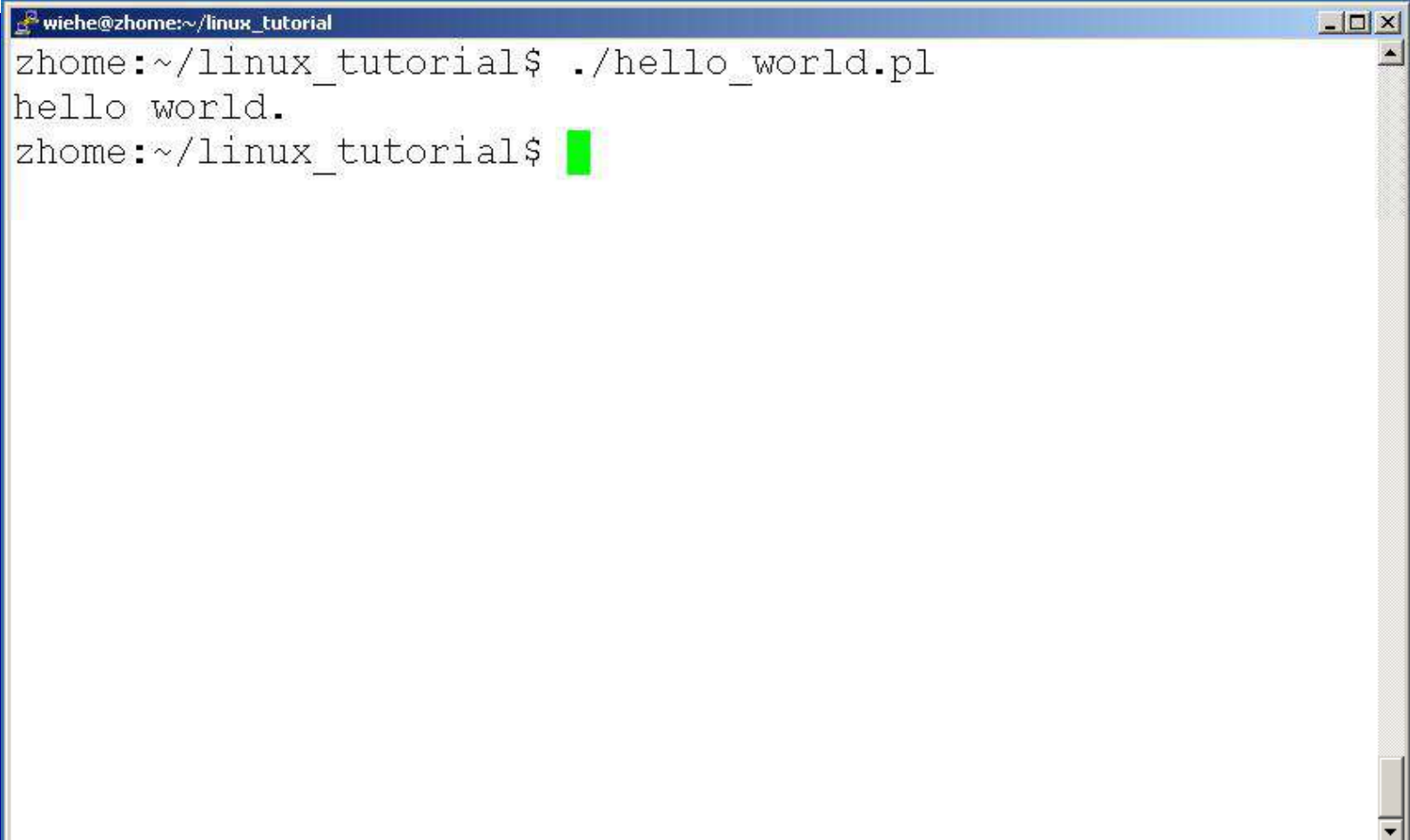
- If you own the file, you can change it's permissions with “chmod”
  - Syntax: chmod [**u**sser/**g**roup/**o**thers/**a**ll]+[permission] [file(s)]
  - Below we grant execute permission to all:

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls -l hello_world.pl
-rw-rw-r-- 1 wiehe wiehe 42 Aug 30 12:22 hello_world.pl
zhome:~/linux_tutorial$ chmod a+x hello_world.pl
zhome:~/linux_tutorial$ ls -l hello_world.pl
-rwxrwxr-x 1 wiehe wiehe 42 Aug 30 12:22 hello_world.pl
zhome:~/linux_tutorial$ █
```

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

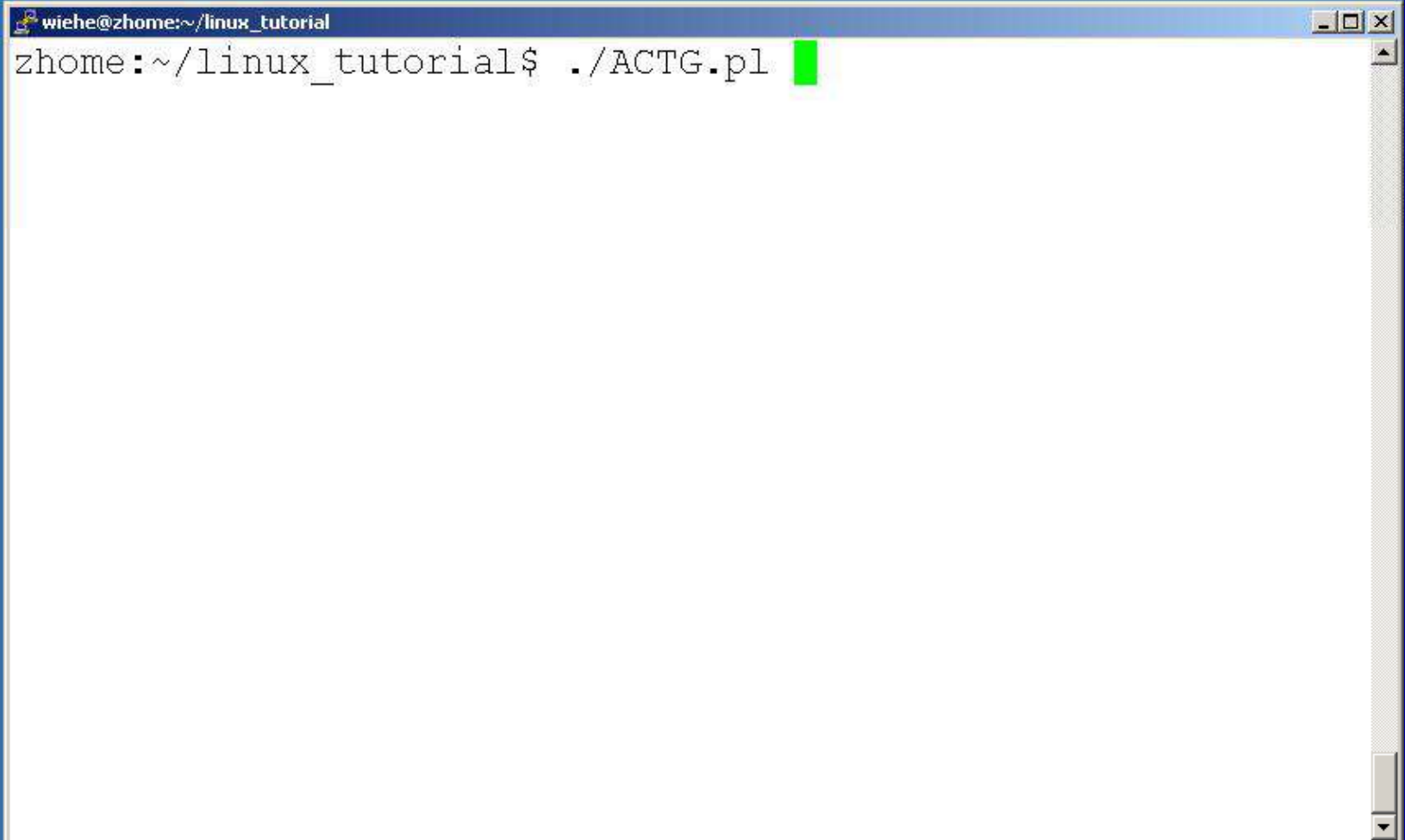
- Make sure the program has executable permissions
- Use “./” to run the program

# Running a program: an example

A terminal window with a title bar that reads "wiehe@zhome:~/linux\_tutorial". The window contains three lines of text: the first line is the prompt "zhome:~/linux\_tutorial\$" followed by the command "./hello\_world.pl"; the second line is the output "hello world."; and the third line is the prompt "zhome:~/linux\_tutorial\$" followed by a green cursor. The window has standard Linux window controls (minimize, maximize, close) in the top right corner and a scrollbar on the right side.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ./hello_world.pl
hello world.
zhome:~/linux_tutorial$ █
```

# Ending a program

- A terminal window with a blue title bar. The title bar text is "wiehe@zhome:~/linux\_tutorial". The terminal content shows the prompt "zhome:~/linux\_tutorial\$" followed by the command "./ACTG.pl" and a green cursor block. The window has standard Linux window controls (minimize, maximize, close) in the top right and a scrollbar on the right side.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ./ACTG.pl
```

# 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

```
wiehe@zhome:~/linux_tutorial
top - 13:46:33 up 50 days,  4:26,  2 users,  load average
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

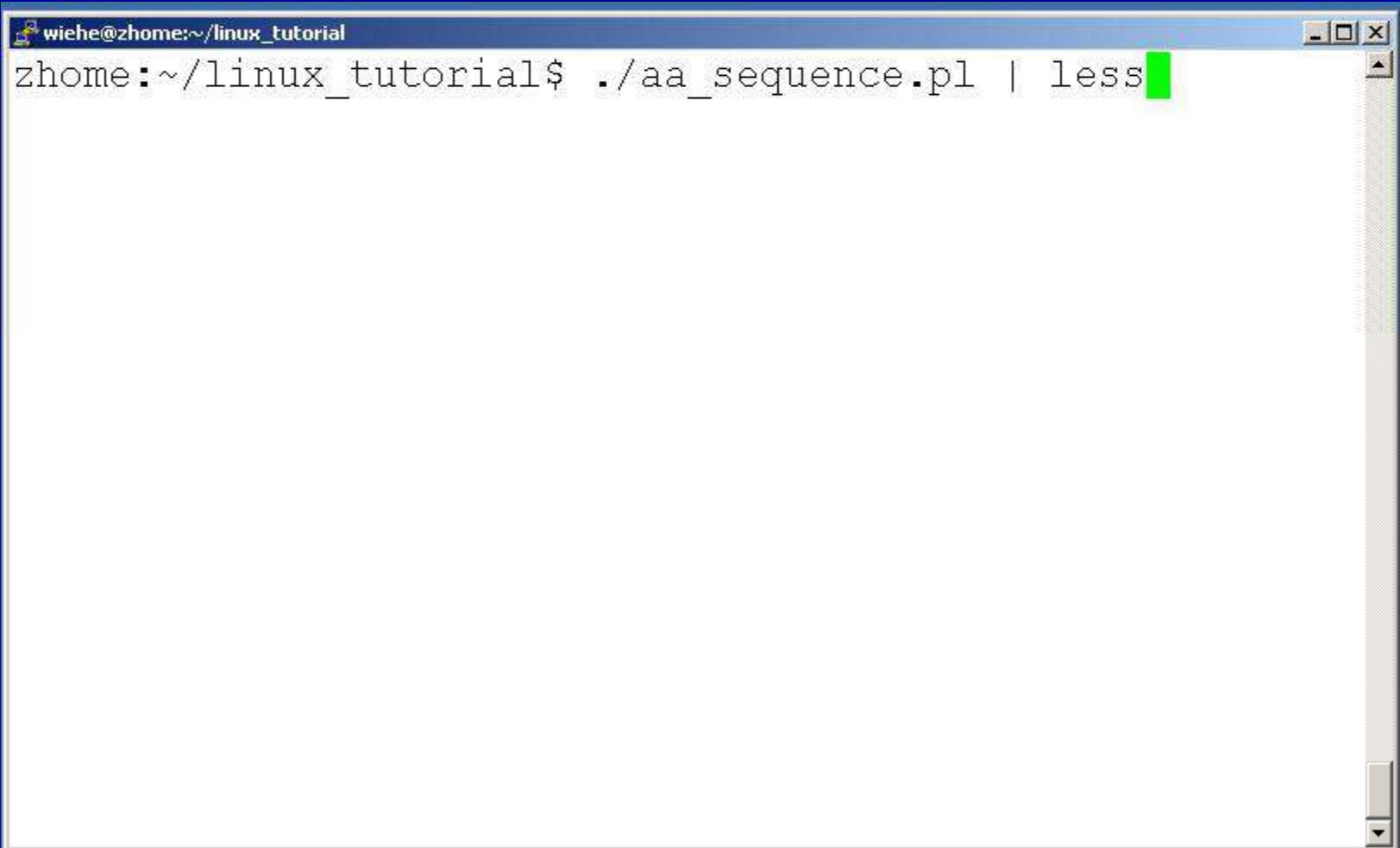
# Command: kill

```
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
 1287 pts/2        00:00:00 ps
zhome:~/linux_tutorial$ kill -9 1255
[1]+  Killed                  ./ACTG.pl
zhome:~/linux_tutorial$ ps -u wiehe
  PID TTY          TIME CMD
 1194 ?            00:00:00 sshd
 1196 pts/2        00:00:00 bash
 1289 pts/2        00:00:00 ps
zhome:~/linux_tutorial$ █
```

# 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 terminal window with a blue title bar. The title bar text is "wiehe@zhome:~/linux\_tutorial". The terminal content shows the command ". /aa\_sequence.pl | less" followed by a green cursor. The terminal has standard window controls (minimize, maximize, close) in the top right and a scrollbar on the right side.

```
wiehe@zhome:~/linux_tutorial  
zhome:~/linux_tutorial$ ./aa_sequence.pl | less
```

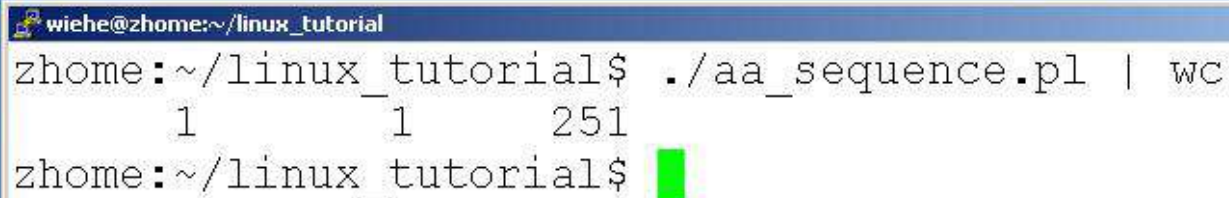
# A few examples of piping

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls
aa_sequence.pl  hello_world.pl  new_directory
ACTG.pl        input.txt
data.dat       lines.txt
zhome:~/linux_tutorial$ ./aa_sequence.pl > sequence.txt
zhome:~/linux_tutorial$ ls
aa_sequence.pl  hello_world.pl  new_directory
ACTG.pl        input.txt       sequence.txt
data.dat       lines.txt
zhome:~/linux_tutorial$ less sequence.txt
```

# 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

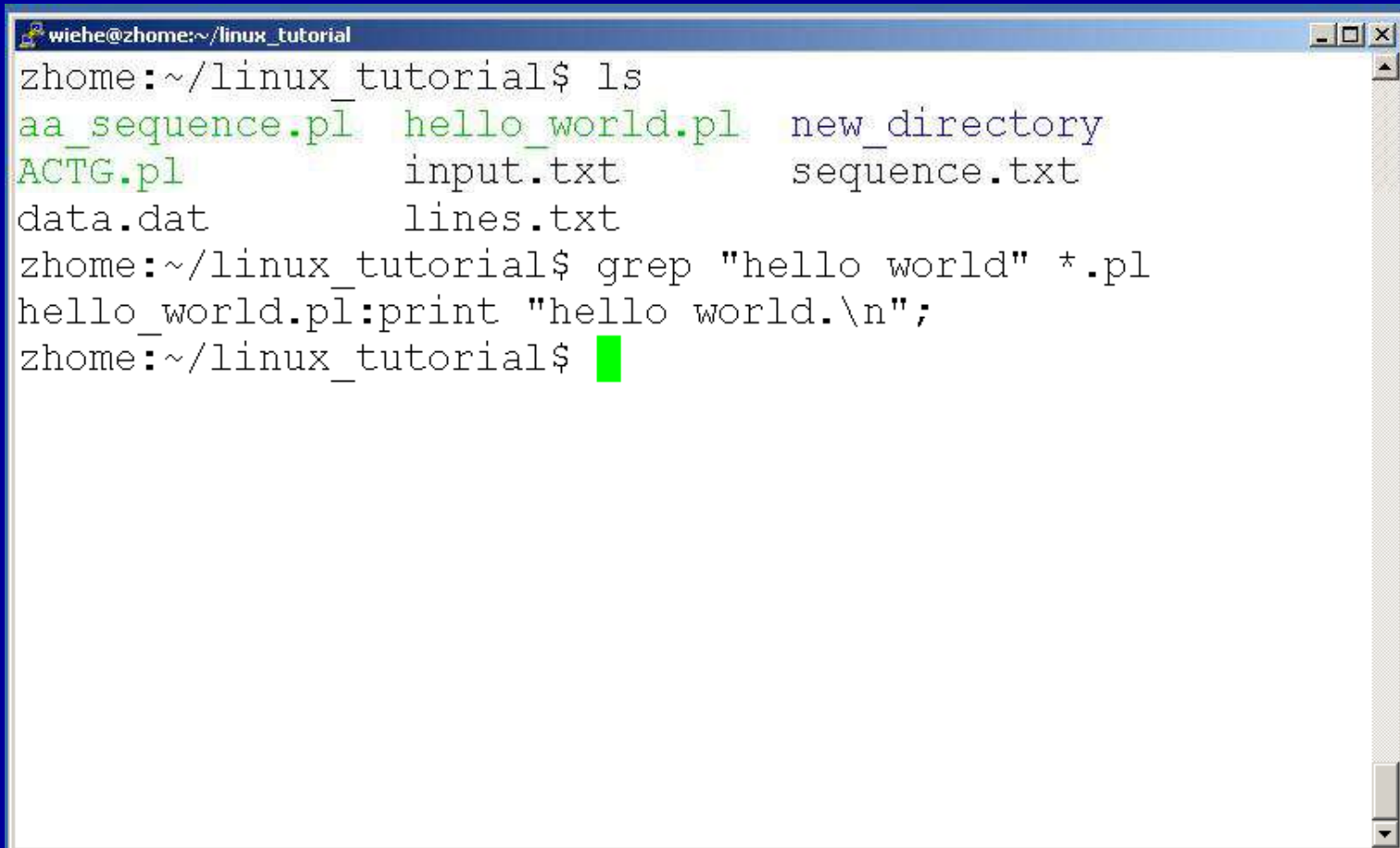
# A few examples of piping

A terminal window with a blue title bar containing the text 'wiehe@zhome:~/linux\_tutorial'. The window has standard window controls (minimize, maximize, close) in the top right corner. The terminal content shows a command being executed and its output.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ./aa_sequence.pl | wc
      1      1     251
zhome:~/linux_tutorial$ █
```

# Command: grep

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

A terminal window titled 'wiehe@zhome:~/linux\_tutorial' showing a series of commands and their outputs. The first command is 'ls', which lists files: 'aa\_sequence.pl', 'hello\_world.pl', 'new\_directory', 'ACTG.pl', 'input.txt', 'sequence.txt', 'data.dat', and 'lines.txt'. The second command is 'grep "hello world" \*.pl', which outputs 'hello\_world.pl:print "hello world.\n";'. The prompt is then ready for the next command.

```
wiehe@zhome:~/linux_tutorial
zhome:~/linux_tutorial$ ls
aa_sequence.pl  hello_world.pl  new_directory
ACTG.pl        input.txt       sequence.txt
data.dat       lines.txt
zhome:~/linux_tutorial$ grep "hello world" *.pl
hello_world.pl:print "hello world.\n";
zhome:~/linux_tutorial$
```

# 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.

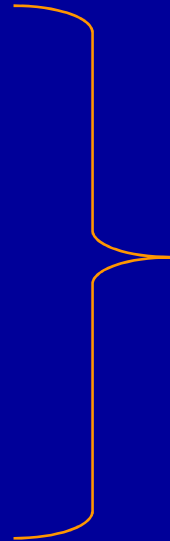
## What is Sequence analysis?

- ▶ Comparing one or more sequences to identify the similarity.

Biological background...?

# Biomolecules

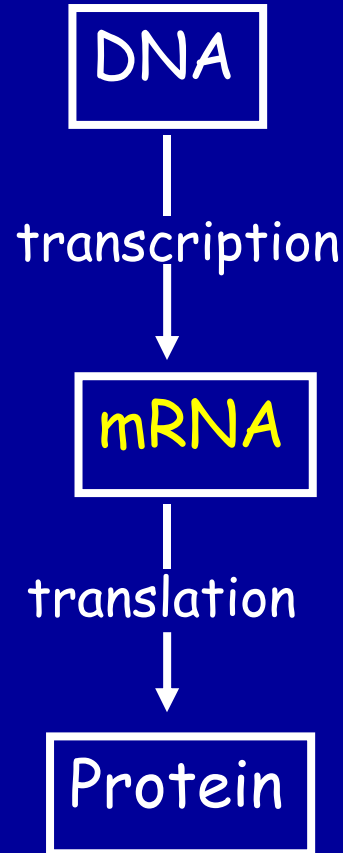
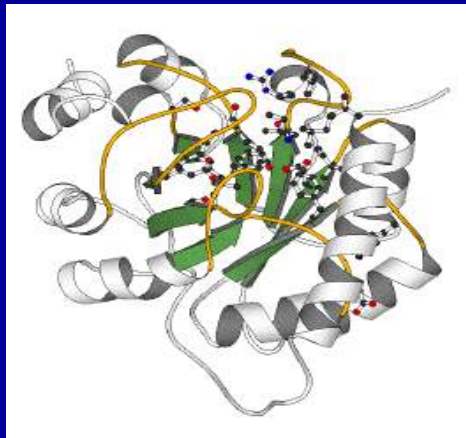
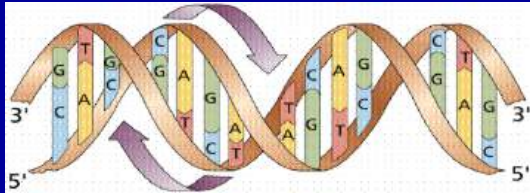
- Nucleic acids
- Proteins
- Enzymes
- Carbohydrates
- Lipids



Macromolecules

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

# The central dogma



CCTGAGCCAAC TATTGATGAA



CCUGAGCCAACUAUUGAUGAA

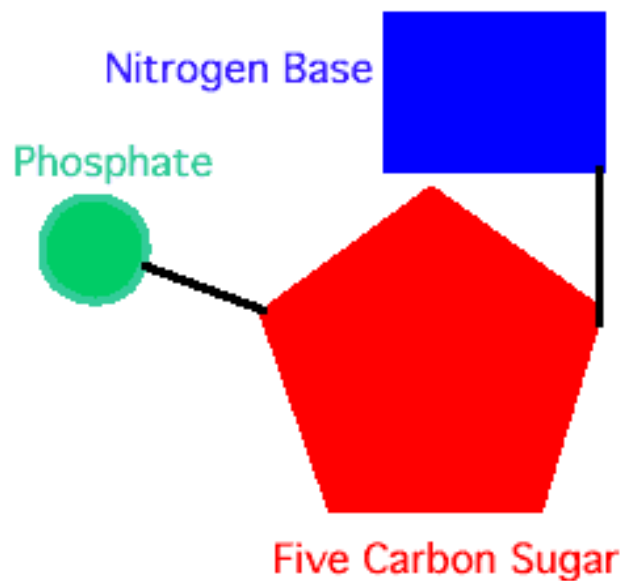


PEPTIDE

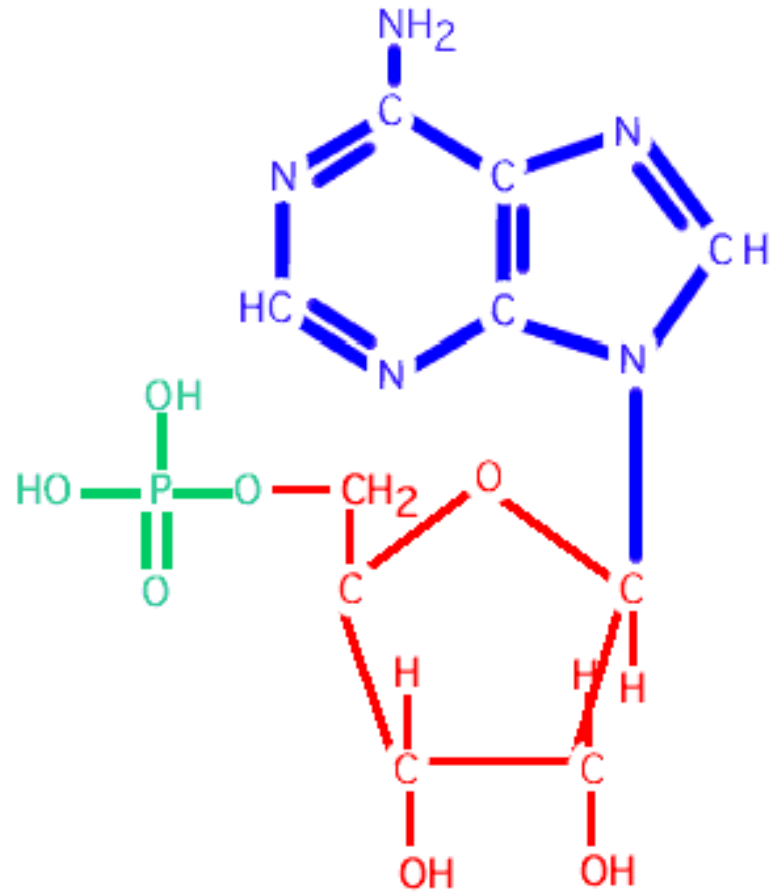
- **DNA:** units are the nucleotide residues A, C, G and T
- **RNA:** units are the nucleotide residues A, C, G and U
- **Proteins:** units are the amino acid residues A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y.

# Structure of nucleotide

Basic Nucleotide Structure



Example



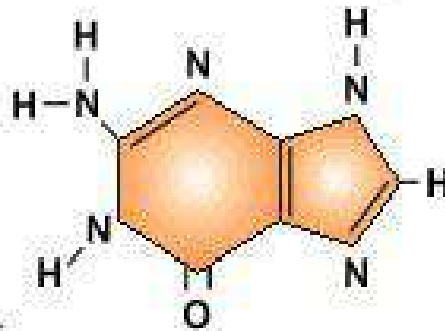
Adenosine 5' phosphoric acid

# Bases

adenine



guanine

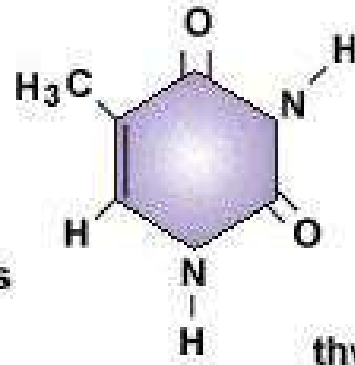


purines

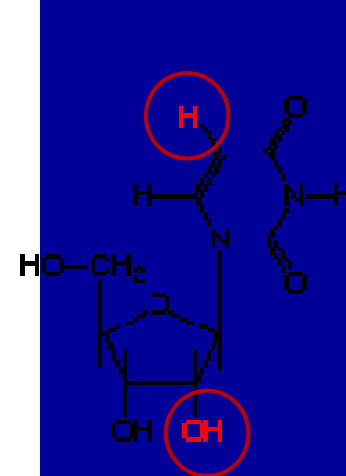
cytosine



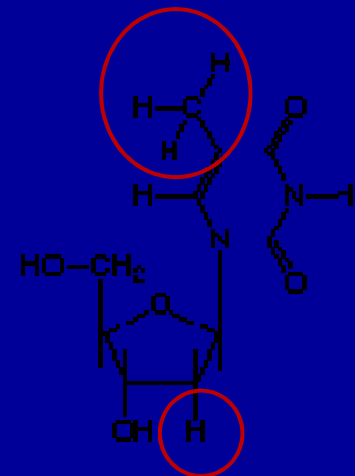
pyrimidines



thymine

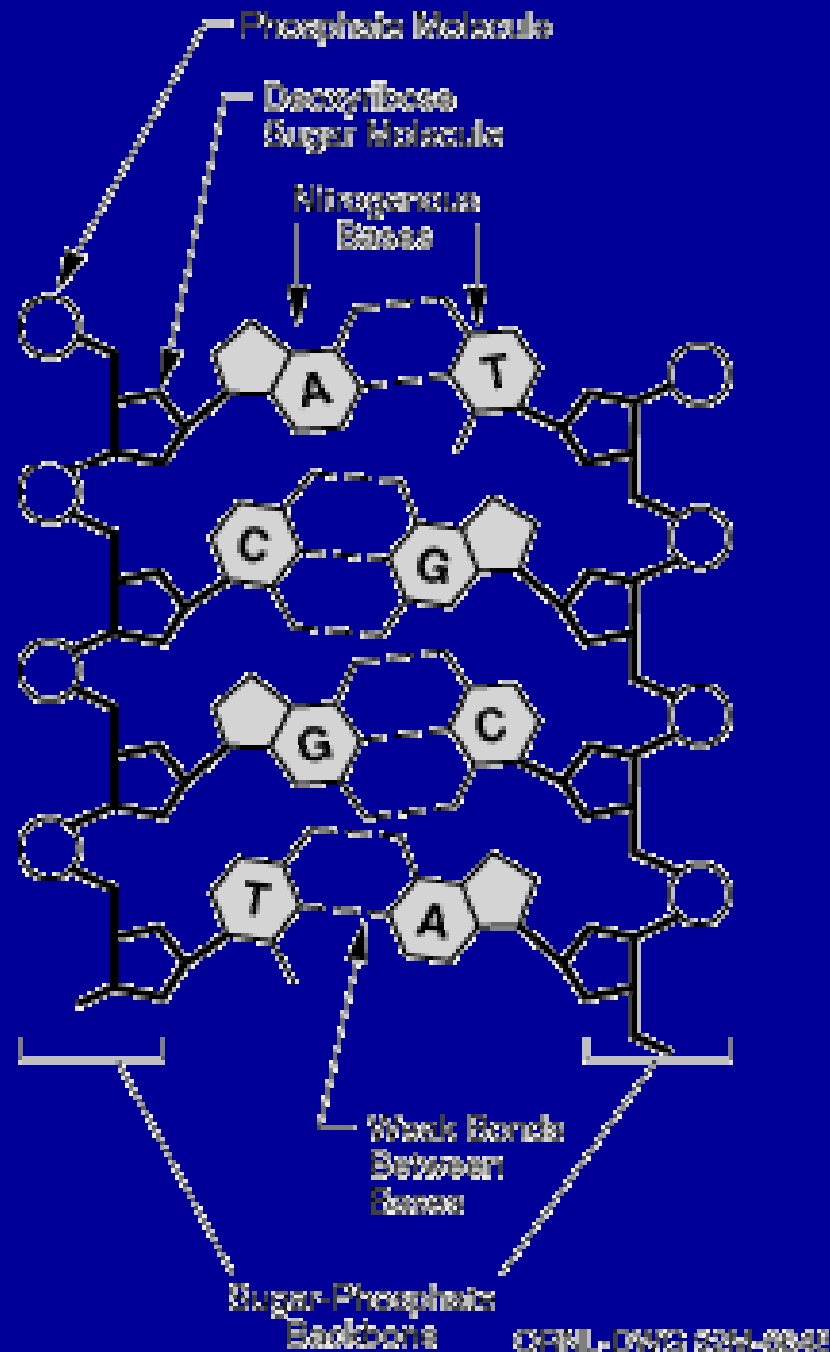


uracil (RNA)



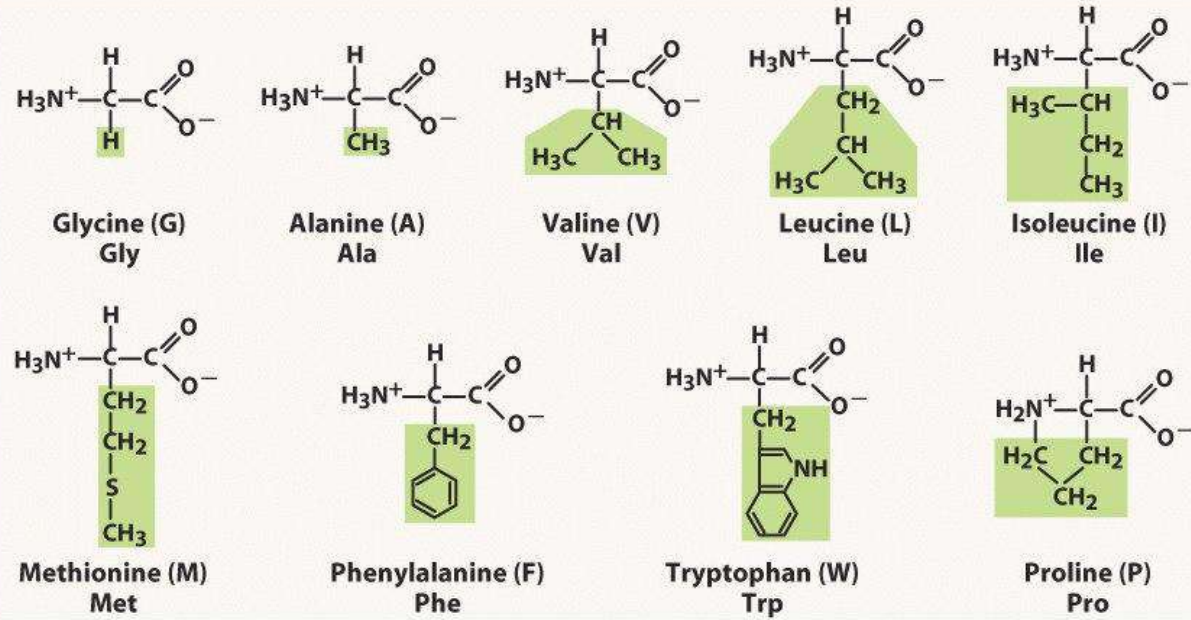
thymine (DNA)

# Complimentary base pairing

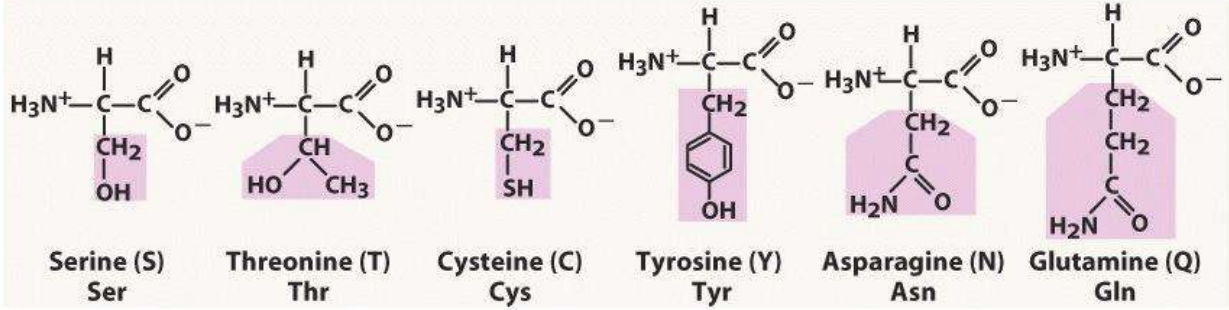


# Structure of Amino acids

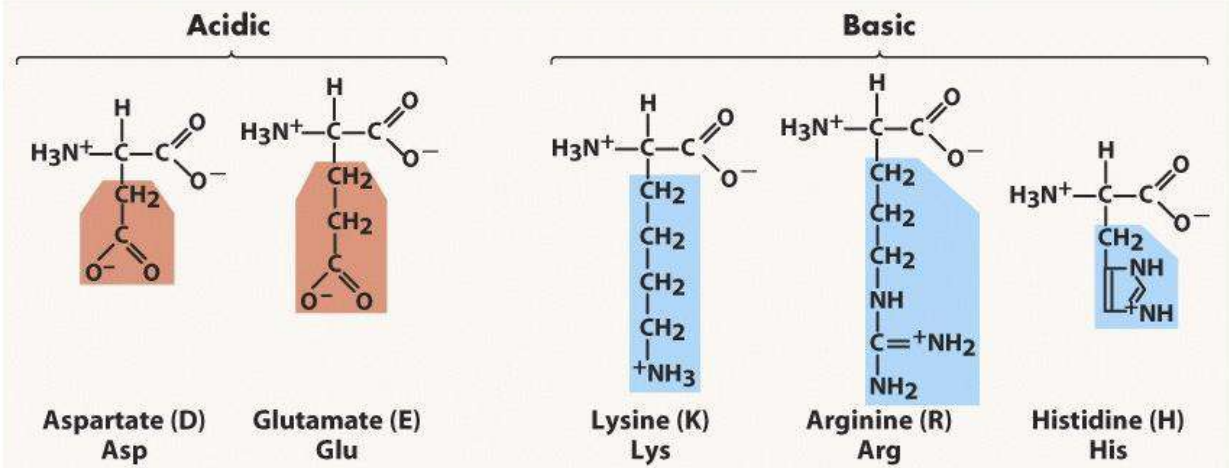
## Nonpolar side chains



## Polar side chains



## Electrically charged side chains



	U	C	A	G	
U	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	U C A G
C	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	U C A G
A	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	U C A G
G	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	U C A 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

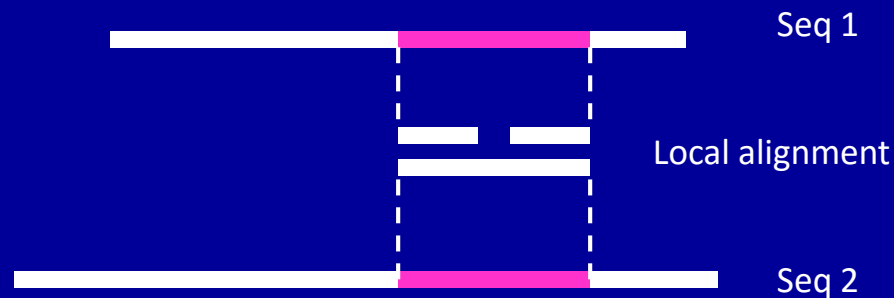
- Pairwise alignment
- Multiple alignment
  
- Local alignment
- Global alignment

# 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

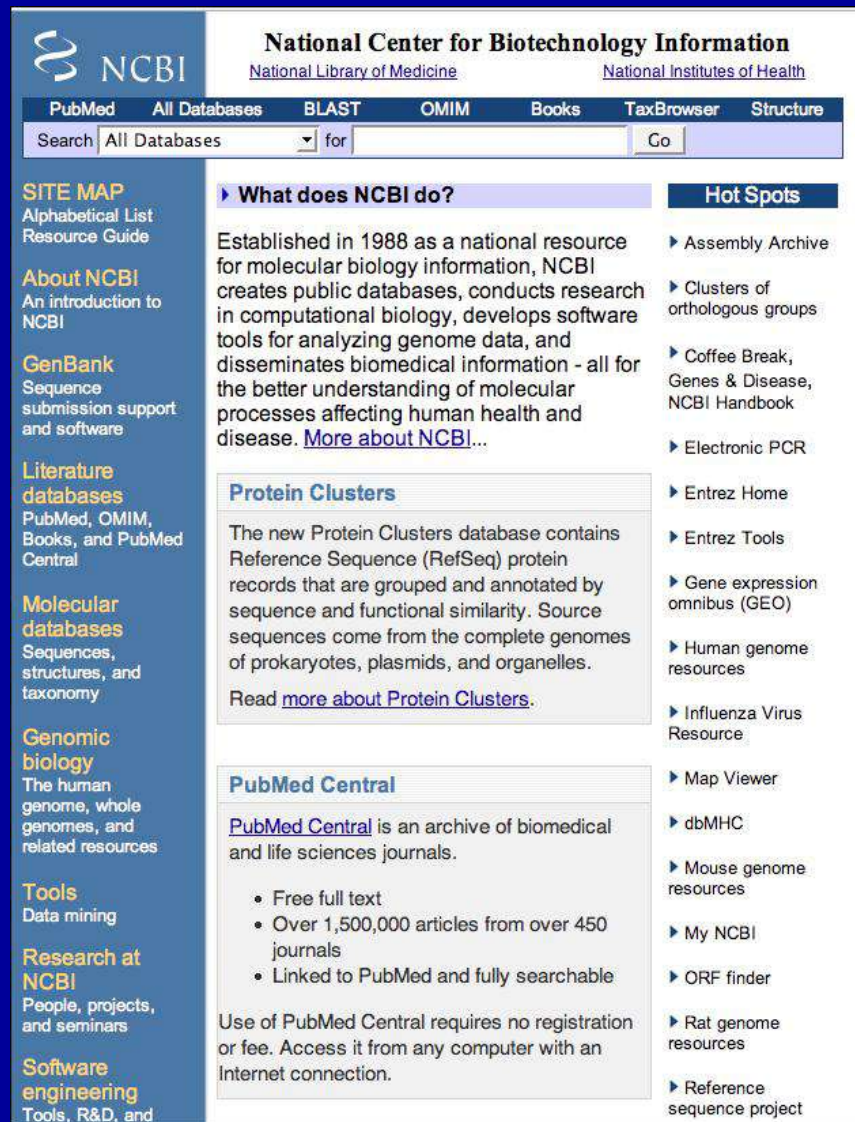
# The National Center for Biotechnology Information



***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](http://www.ncbi.nlm.nih.gov)



The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The page features a navigation bar with links to PubMed, All Databases, BLAST, OMIM, Books, TaxBrowser, and Structure. A search bar is present with a dropdown menu set to 'All Databases' and a 'Go' button. The main content area is divided into several sections: 'What does NCBI do?' (describing the center's mission), 'Protein Clusters' (describing a new database), 'PubMed Central' (describing an archive of journals), and 'Hot Spots' (a list of featured resources). A left sidebar contains a 'SITE MAP' with links to various resources like 'About NCBI', 'GenBank', 'Literature databases', 'Molecular databases', 'Genomic biology', 'Tools', 'Research at NCBI', and 'Software engineering'.

**NCBI**  
National Center for Biotechnology Information  
National Library of Medicine      National Institutes of Health

PubMed   All Databases   BLAST   OMIM   Books   TaxBrowser   Structure

Search All Databases for  Go

**SITE MAP**  
Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to NCBI

**GenBank**  
Sequence submission support and software

**Literature databases**  
PubMed, OMIM, Books, and PubMed Central

**Molecular databases**  
Sequences, structures, and taxonomy

**Genomic biology**  
The human genome, whole genomes, and related resources

**Tools**  
Data mining

**Research at NCBI**  
People, projects, and seminars

**Software engineering**  
Tools, R&D, and

**What does NCBI do?**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More about NCBI...](#)

**Protein Clusters**

The new Protein Clusters database contains Reference Sequence (RefSeq) protein records that are grouped and annotated by sequence and functional similarity. Source sequences come from the complete genomes of prokaryotes, plasmids, and organelles.

Read [more about Protein Clusters](#).

**PubMed Central**

[PubMed Central](#) is an archive of biomedical and life sciences journals.

- Free full text
- Over 1,500,000 articles from over 450 journals
- Linked to PubMed and fully searchable

Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

**Hot Spots**

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools
- ▶ Gene expression omnibus (GEO)
- ▶ Human genome resources
- ▶ Influenza Virus Resource
- ▶ Map Viewer
- ▶ dbMHC
- ▶ Mouse genome resources
- ▶ My NCBI
- ▶ ORF finder
- ▶ Rat genome resources
- ▶ Reference sequence project

# 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

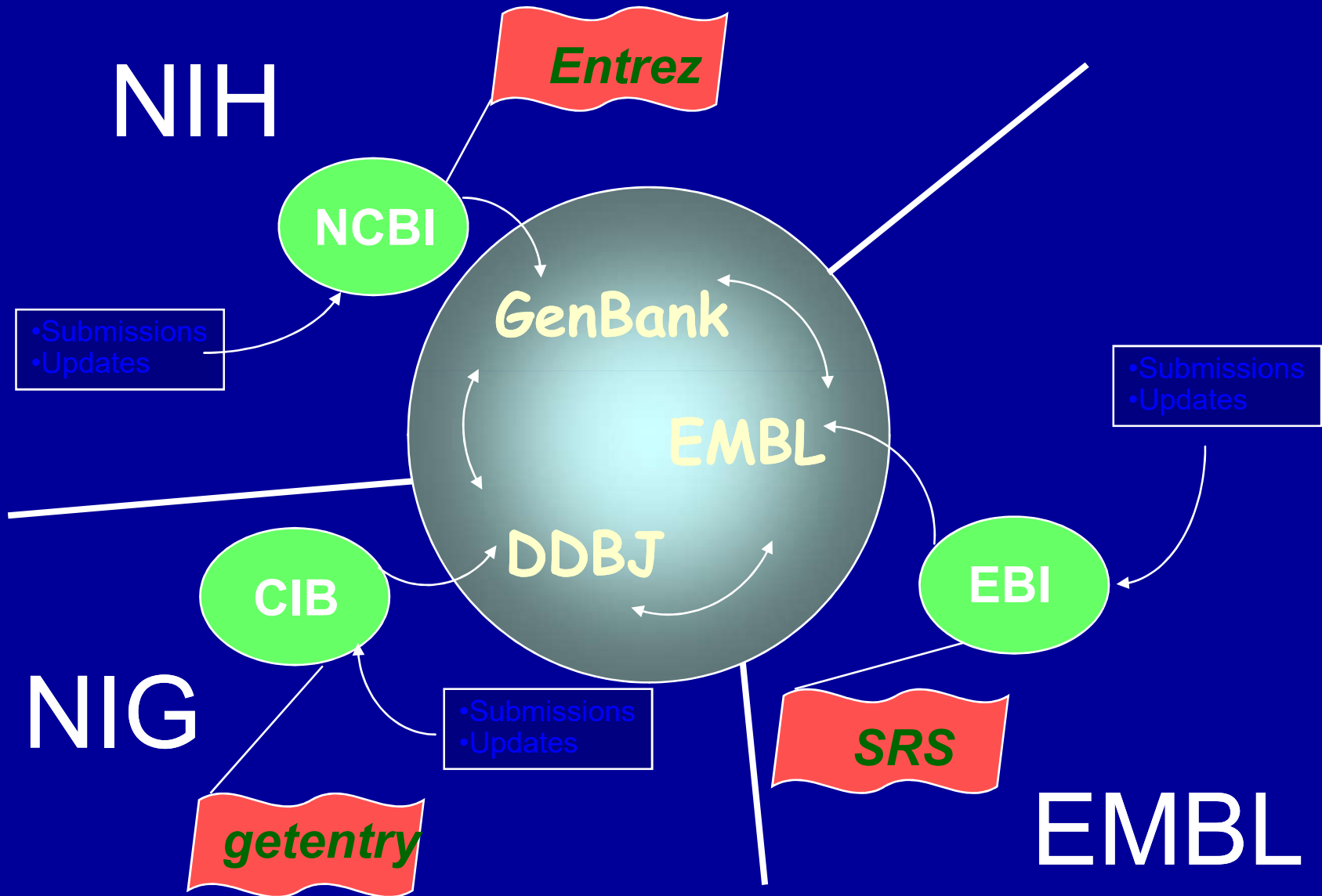
- Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the submitter
    - Examples: **GenBank, SNP, GEO**
- Derivative Databases
  - Built from primary data
  - Content controlled by third party (NCBI)
    - Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, 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="MEACNCIEPQWPADELLMKYQYISDFFIALAYFSI PLELIYFVK
            KSAVFPYRWVLVQFGAFIVLCGATHLINLWTFMSHSRTVAIVMTAKVLTAVVSCATA
            LMLVHIIPDLLSVKTRFLKNKAAELDREMGLIRTQEETGRHVRLMTEIRSTLDRH
            TILKTTLVLELGRTLALEECALWMPTRTGLELQLSYTLRQQNPVGYTVPIHLPVINQVF
            SSNRALKISPNSPVARMRPLACKHMPGEVVAVRVPLLHLSNFQINDWPELSTKRYALM
            VLMLPSDSARQVHVHELELVEVVADQVAVALS HAAILEESMRARDLLMEQNIALDLAR
            REAETAIRARNDFLAVMNHMRTPMHAI IALSSLLOETELTPEQRLMVETILKSHLL
            ATLINDVLDLSRLEDGSLQLEIATFNLHSVFREVHNLKPVASVKKLSVSLNLAADLP
            VQAVGDEKRLMQIVLNVVGNVAKFSKEGSI SI TAFVAKSESLRDFRAPEFFPAQSDNH
            FYLRVQVKDSSGGINPQDIPKLF TKFAQTQSLATRNSGGSGLGLAICRFRVNLMEGHI
            WIESEGPGRGCTAIFIVKLGFAERSNESKLPFLTKVQANHVQTNFPGKLVLMDDNGS
            VTKGLLVHLGCDVTTVSS IDEFLHVISQEHKVVFMVDCMPGIDGYELAVRIHEKFTKR
            HERPVLVALTGNIDKMTKENCMRVGMDGVILKPVSVDKMRSVLSELLEHRVLFEM"
ORIGIN
     1  gcacgagggc  tcaccgagcg  agctagctot  tcaggagtca  aggcttctgg  gtgaggggaa
     61  gaagaagaag  cttctttgat  gtgttggggg  gccaatctaa  agaggaagaa  gaaggcctct
    121  aatgtattga  ggtcggctgt  ctgggctgcc  gatctgtggt  gaatggatag  tttggtagag
    181  atgcttcaac  gacatagggg  ggctgaaaag  ggtttgaaga  aagtgaagga  ggaaccaaac
           ...
    2401 tatactgaaa  cctgtctcag  ttgataaaat  gaggagtgtt  ttatcagaac  tgttgagaca
    2461 tcgagtttta  tttgaggcta  tgtaagatat  aggaaaattg  ttctagtgaa  ggaaagattt
    2521 aaatgaaaaa  aaaaaaaaaa
//
```

} Header

## The Flatfile Format

} Feature Table

} Sequence



# 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	<b>DNA</b>
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	<b>RNA</b>
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	<b>protein</b>
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**  
National Center for Biotechnology Information  
National Library of Medicine  
National Institutes of Health

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search Entrez for rbp4 Go

**What does NCBI do?**  
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

**Hot Spots**

- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools
- ▶ Gene expression omnibus (GEO)

**The Genetic Landscape of Diabetes**  
Over 17 million Americans have diabetes. Explore the genes discovered thus far with "The Genetic

















**SITE MAP**  
Guide to NCBI resources

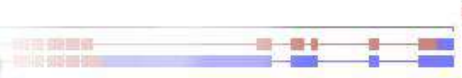
**About NCBI**  
An introduction for researchers, educators and the public

**GenBank**  
Sequence submission support and software

**Literature databases**

Search across databases    Help

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



Entrez **Gene**     current records only

Limits

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  Show:  Send to

Items 1-8 of 8 One page.

- [1: RBP4](#) Links  
 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](#) Links  
 retinol-binding protein [*Sus scrofa*]  
**GeneID:** 397124
- [3: RBP4](#) Links  
 retinol binding protein 4, plasma [*Bos taurus*]  
**GeneID:** 281444
- [4: rbp4](#) Links  
 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](#) Links  
 RNA-binding protein 4 [*Drosophila melanogaster*]  
**Other Aliases:** CG9654, RRM4, TSR  
**Other Designations:** CG9654-PA  
**Chromosome:** 3R; **Location:** 87F7-87F7  
**GeneID:** 41668
- [6: Rbp4](#) Links  
 retinol binding protein 4 [*Rattus norvegicus*]  
**Other Aliases:** RGD:3546, RBPA  
**Chromosome:** 1; **Location:** 1q53  
**GeneID:** 25703
- [7: Rbp4](#) Links  
 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



Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

Search  for     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](#)
- |                                                          |                                                            |                                                                  |                                                       |
|----------------------------------------------------------|------------------------------------------------------------|------------------------------------------------------------------|-------------------------------------------------------|
| <input type="checkbox"/> <a href="#">Bos taurus</a>      | <input type="checkbox"/> <a href="#">Canis familiaris</a>  | <input checked="" type="checkbox"/> <a href="#">Homo sapiens</a> | <input type="checkbox"/> <a href="#">Mus musculus</a> |
| <input type="checkbox"/> <a href="#">Pan troglodytes</a> | <input type="checkbox"/> <a href="#">Rattus norvegicus</a> | <input type="checkbox"/> <a href="#">Sus scrofa</a>              |                                                       |
- [Non-mammalian vertebrates](#)
- |                                                      |                                                        |                                                         |                                                             |
|------------------------------------------------------|--------------------------------------------------------|---------------------------------------------------------|-------------------------------------------------------------|
| <input type="checkbox"/> <a href="#">Danio rerio</a> | <input type="checkbox"/> <a href="#">Gallus gallus</a> | <input type="checkbox"/> <a href="#">Xenopus laevis</a> | <input type="checkbox"/> <a href="#">Xenopus tropicalis</a> |
|------------------------------------------------------|--------------------------------------------------------|---------------------------------------------------------|-------------------------------------------------------------|
- [Invertebrates](#)
- |                                                            |                                                                 |                                                                  |                                                                |
|------------------------------------------------------------|-----------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|
| <input type="checkbox"/> <a href="#">Anopheles gambiae</a> | <input type="checkbox"/> <a href="#">Caenorhabditis elegans</a> | <input type="checkbox"/> <a href="#">Drosophila melanogaster</a> | <input type="checkbox"/> <a href="#">Plasmodium falciparum</a> |
|------------------------------------------------------------|-----------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|
- [Plants](#)
- |                                                               |                                                       |  |  |
|---------------------------------------------------------------|-------------------------------------------------------|--|--|
| <input type="checkbox"/> <a href="#">Arabidopsis thaliana</a> | <input type="checkbox"/> <a href="#">Oryza sativa</a> |  |  |
|---------------------------------------------------------------|-------------------------------------------------------|--|--|
- [Fungi](#)
- |                                                             |                                                            |                                                                   |                                                                    |
|-------------------------------------------------------------|------------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------------------|
| <input type="checkbox"/> <a href="#">Magnaporthe grisea</a> | <input type="checkbox"/> <a href="#">Neurospora crassa</a> | <input type="checkbox"/> <a href="#">Saccharomyces cerevisiae</a> | <input type="checkbox"/> <a href="#">Schizosaccharomyces pombe</a> |
|-------------------------------------------------------------|------------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------------------|
- [Bacteria](#)
- |                                                           |                                                                     |                                                          |                                                         |
|-----------------------------------------------------------|---------------------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------|
| <input type="checkbox"/> <a href="#">Actinobacteria</a>   | <input type="checkbox"/> <a href="#">Bacillus</a>                   | <input type="checkbox"/> <a href="#">Chlamydia</a>       | <input type="checkbox"/> <a href="#">Cyanobacteria</a>  |
| <input type="checkbox"/> <a href="#">Escherichia coli</a> | <input type="checkbox"/> <a href="#">Mycobacterium tuberculosis</a> | <input type="checkbox"/> <a href="#">Mycoplasmatales</a> | <input type="checkbox"/> <a href="#">Proteobacteria</a> |
| <input type="checkbox"/> <a href="#">Pseudomonas</a>      | <input type="checkbox"/> <a href="#">Salmonella</a>                 | <input type="checkbox"/> <a href="#">Spirochaetes</a>    | <input type="checkbox"/> <a href="#">Streptococcus</a>  |
| <input type="checkbox"/> <a href="#">Yersinia</a>         |                                                                     |                                                          |                                                         |

Entrez  
[SITE MAP](#)  
[Entrez Help](#)

[Gene](#)  
[Search](#)  
[Gene Help](#)

[FAQ](#)

[FTP site](#)

[Related sites](#)  
[Entrez Genome](#)  
[Genomic Biology](#)  
[Homolo Gene](#)  
[LocusLink](#)  
[Map Viewer](#)  
[OMIM](#)  
[RefSeq](#)  
[UniGene](#)

[Feedback](#)

[Help Desk](#)  
[Connections](#)  
[About GeneRIFs](#)

[Subscriptions](#)

[RefSeq](#)  
[Gene](#)  
[Map Viewer](#)

# By applying limits, there are now just two entries

The screenshot shows the NCBI Entrez Gene search interface. The search query is 'Gene' for 'rbp4'. The results are limited to 'Homo sapiens' and show 2 items. The first entry is RBP4 (GeneID: 5950) and the second is POLR2D (GeneID: 5433).

**NCBI Entrez Gene**

Search  for     current records only

Limits [Preview/Index](#) [History](#) [Clipboard](#) [Details](#)

Limits: **Homo sapiens**

Display  Show:  Send to  One page.

Items 1-2 of 2

1: [RBP4](#) Links

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: [POLR2D](#) Links

polymerase (RNA) II (DNA directed) polypeptide D [*Homo sapiens*]  
**Other Aliases:** HGNC:9191, HSRBP4, HSRPB4, RBP4  
**Other Designations:** DNA directed RNA polymerase II polypeptide D; RNA polymerase II subunit hsRBP4  
**Chromosome:** 2; **Location:** 2q21  
**GeneID:** 5433

**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

# Entrez Gene (top of page)

NCBI Entrez Gene

My NCBI  
[Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

Search Gene for [ ] Go Clear [x] current records only

Limits Preview/Index History Clipboard Details

Display Graphics Show 5 Send to

All: 1 Genes Genomes: 1 SNP GeneView: 1

1: **RBP4 retinol binding protein 4, plasma** [*Homo sapiens*]  
GeneID: 5950 Locus tag: [HGNC:9922](#); [MIM:180250](#) updated 02-Aug-2005

Entrez Gene Home

Table Of Contents

- Summary
- Transcripts and products
- Genomic context
- Bibliography
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links

Links

- Conserved Domains
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (GeneRIF)
- SNP
- SNP: Genotype
- SNP: GeneView
- Taxonomy
- UniSTS
- AceView
- Ensembl
- Evidence Viewer
- GDB
- HGNC
- KEGG
- MGC
- ModelMaker
- Retina International RB...
- UCSC
- UniGene
- LinkOut

Entrez Gene Info

Feedback

Subscriptions

**Summary**

Official Symbol: RBP4 and Name: retinol binding protein 4, plasma provided by [HUGO Gene Nomenclature Committee](#)

Gene type: protein coding

Gene name: RBP4

Gene description: retinol binding protein 4, plasma

RefSeq status: Reviewed

Organism: [Homo sapiens](#)

Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Homimidae; Homo*

Summary: This protein belongs to the lipocalin family and is the specific carrier for retinol (vitamin A alcohol) in the blood. It delivers retinol from the liver stores to the peripheral tissues. In plasma, the RBP-retinol complex interacts with transthyretin which prevents its loss by filtration through the kidney glomeruli. A deficiency of vitamin A blocks secretion of the binding protein postranslationally and results in defective delivery and supply to the epidermal cells.

**Transcripts and products**

(minus strand) [RefSeq below](#)

NC\_000010

95350987 5' 95341591 3'

NM\_006744 NP\_006735 precursor

■ - coding region ■ - untranslated region

**Genomic context** [See RBP4 in MapViewer](#)

chromosome: 10; Location: 10q23-q24

95246399 95452319

C10orf3 GPR120 PDE6C RBP4 C10orf4

Note that links to many other RBP4 database entries are available

# Entrez Gene (middle of page)

The screenshot shows the Entrez Gene web page for the gene RBP4. The browser window title is "Gene - Microsoft Internet Explorer". The address bar shows the URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve>. The page content includes:

- Genomic context**: Shows the gene's location on chromosome 10, specifically at 10q23-q24. A genomic map displays the RBP4 gene structure with exons represented by boxes and introns by lines with arrows. Neighboring genes shown are C10orf43, GFR120, PDE6C, and C11orf4. A link "See RBP4 in Map Viewer" is provided.
- Bibliography**: Includes a link to "Gene References into Function (GeneRIF): Submit".
- PubMed links**: Lists one GeneRIF entry: "Genetic variation in RBP4 is associated with amyloid polyneuropathy" with a PubMed link.
- General gene information**: A section with a help icon.
- GeneOntology**:
  - Provided by [GOA](#)
  - Function**:
    - [binding](#) IEA
    - [retinal binding](#) IEA
    - [retinol binding](#) IEA
    - [transporter activity](#) IEA
  - Process**:
    - [sensory perception](#) IEA
    - [transport](#) IEA
    - [visual perception](#) IEA
  - Component**:
    - [extracellular space](#) TAS [PubMed](#)
- Homology**:
  - Mouse, Rat
  - [Map Viewer](#)
- Phenotypes**:
  - Retinol binding protein, deficiency of [MIM: 180250](#)
- Markers (Sequence Tagged Sites/STS)**:
  - [SHGC-170](#) (e-PCR)
  - Alternate name RH8396
  - Alternate name gdb:214050
  - Alternate name gdb:223450
  - [D10S2145](#) (e-PCR)
  - Alternate name GDB:675125
  - Alternate name RP8621

# Entrez Gene (bottom of page)

Gene - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve> Go Google Links

Alternate name [gdb:214050](#)  
Alternate name [gdb:223450](#)  
[D10S2145 \(e-PCR\)](#)  
Alternate name [GDB:675125](#)  
Alternate name [RH8621](#)  
Alternate name [SHGC-10696](#)  
Alternate name [gdb:675661](#)  
[SHGC-132033 \(e-PCR\)](#)  
[MARC\\_24240-24241:1032810986.5 \(e-PCR\)](#)

**General protein information** ? ↑

Names: [retinol-binding protein 4, plasma](#)  
[retinol-binding protein 4, plasma](#); [retinol-binding protein 4, interstitial](#)

**NCBI Reference Sequences (RefSeq)** ? ↑

mRNA Sequence [NM\\_006744](#)  
Source Sequence [BC020633](#), [BX495987](#), [X00129](#)  
Product [NP\\_006735](#) [retinol-binding protein 4, plasma precursor](#)  
Conserved Domains (1) [summary](#)  
[pfam00061: Lipocalin; Lipocalin / cytosolic fatty-acid binding protein family](#)  
Location: 37 - 192 Blast Score: 174

**Related Sequences** ? ↑

Nucleotide	Protein
Genomic <a href="#">AF025334</a>	<a href="#">AAC02945</a>
Genomic <a href="#">AF025335</a>	<a href="#">AAC02946</a>
Genomic <a href="#">AL356214</a>	<a href="#">CAH72328</a>
Genomic <a href="#">X02775</a>	<a href="#">CAA26553</a>
Genomic <a href="#">X02824</a>	<a href="#">CAB46489</a>
mRNA <a href="#">BC020633</a>	<a href="#">AAH20633</a>
mRNA <a href="#">X00129</a>	<a href="#">CAA24959</a>
None	<a href="#">P02753</a>

**Additional Links** ? ↑

[UniGene Hs.50223](#)  
[MIM 180250](#)  
Retina International RBP4 Mutation Database [Retina International RBP4 Mutation Database](#)

Display [Graphics](#) Show [5](#) Send to

[Restrictions on Use](#) | [Write to the Help Desk](#)  
[NCBI](#) | [NLM](#) | [NIH](#)

Internet

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

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

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 mewwwalvll aalgsgrger dcrvssfrvk enfdkarfag twyamakkdp egflfqdniv
61 aefsvdengh msatakgrvr llnmwvdcad mvgtftdted pakfkmywg vasflqrgnd
121 dhwiidtdyd tfavqyscrl lnfdgtcads ysfvfrdph glppdvqkly rqrqealcsls
181 rqyrlivhng ycddksvrnl l
```

//

Document: Done

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](#)

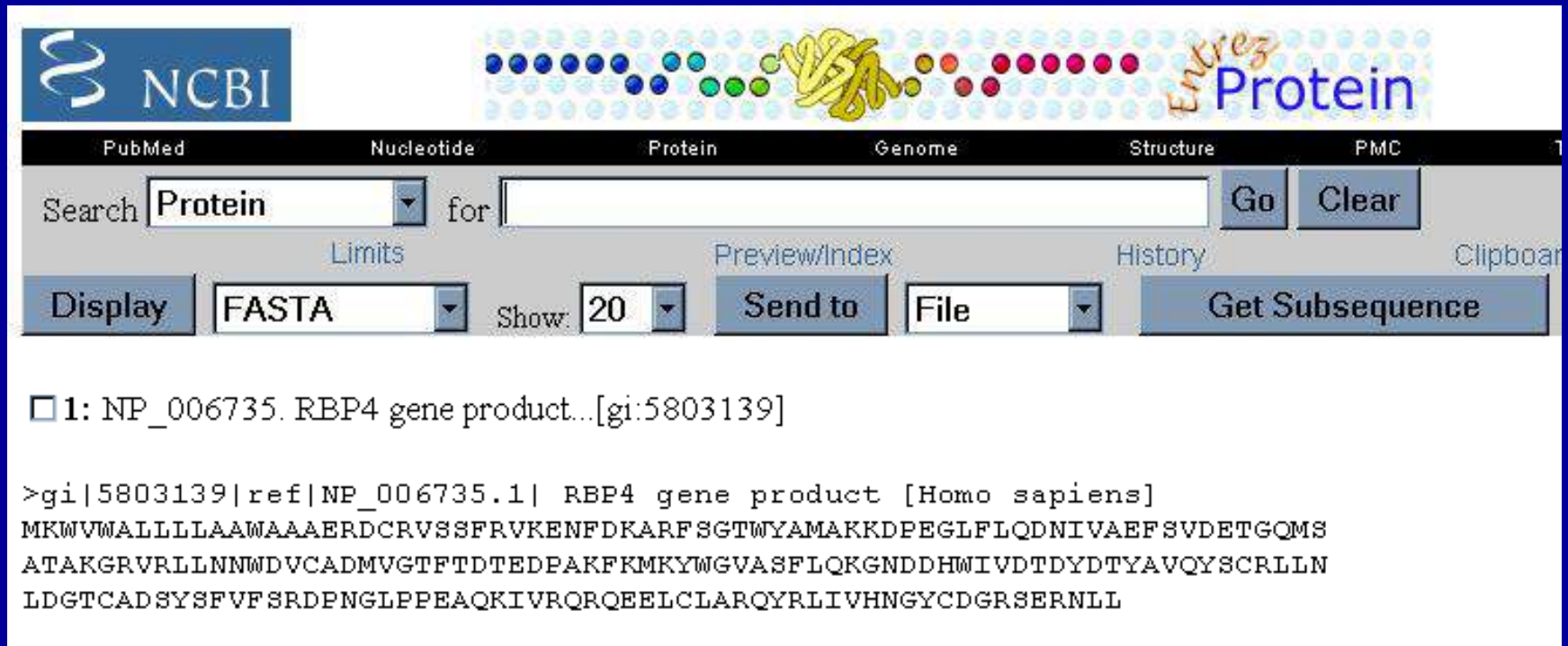
REMARK SEQUENCE FROM N.A.

REFERENCE 2 (residues 1 to 199)

Document: Done



# FASTA format



The screenshot shows the NCBI Entrez Protein search interface. At the top, there is the NCBI logo and the Entrez Protein logo. Below the logos, there are navigation tabs for PubMed, Nucleotide, Protein, Genome, Structure, and PMC. The search bar is set to "Protein" and is empty. To the right of the search bar are "Go" and "Clear" buttons. Below the search bar, there are options for "Limits", "Preview/Index", "History", and "Clipboard". The "Display" dropdown is set to "FASTA", "Show" is set to "20", "Send to" is set to "File", and "Get Subsequence" is a button. Below the search bar, there is a checkbox and the text "1: NP\_006735. RBP4 gene product...[gi:5803139]". Below this, the FASTA format output is shown:

```
>gi|5803139|ref|NP_006735.1| RBP4 gene product [Homo sapiens]
MKWVWALLLLLAAMAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEF SVDETGQMS
ATAKGRVRLNNDVDCADMVGTFTDTEDEPAKFKMKYWGVASFLQKGNDDHMIVDTDYDTYAVQYSCRLLN
LDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERLL
```

# 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

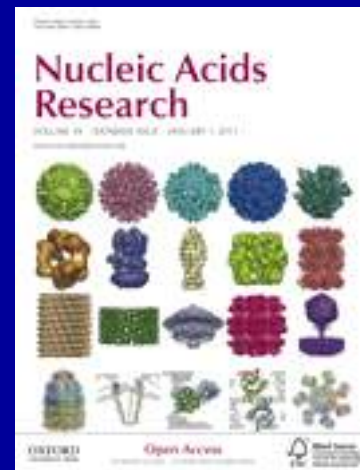
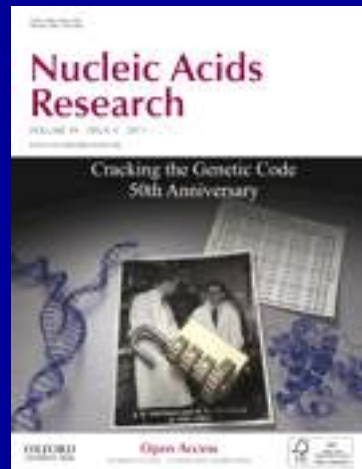
- repeated submission
  - identical or overlapping sequences by the same or different authors
  - revision of annotations
  - dumping of expressed sequence tags (EST) data
  - poor database management

# 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](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/>)

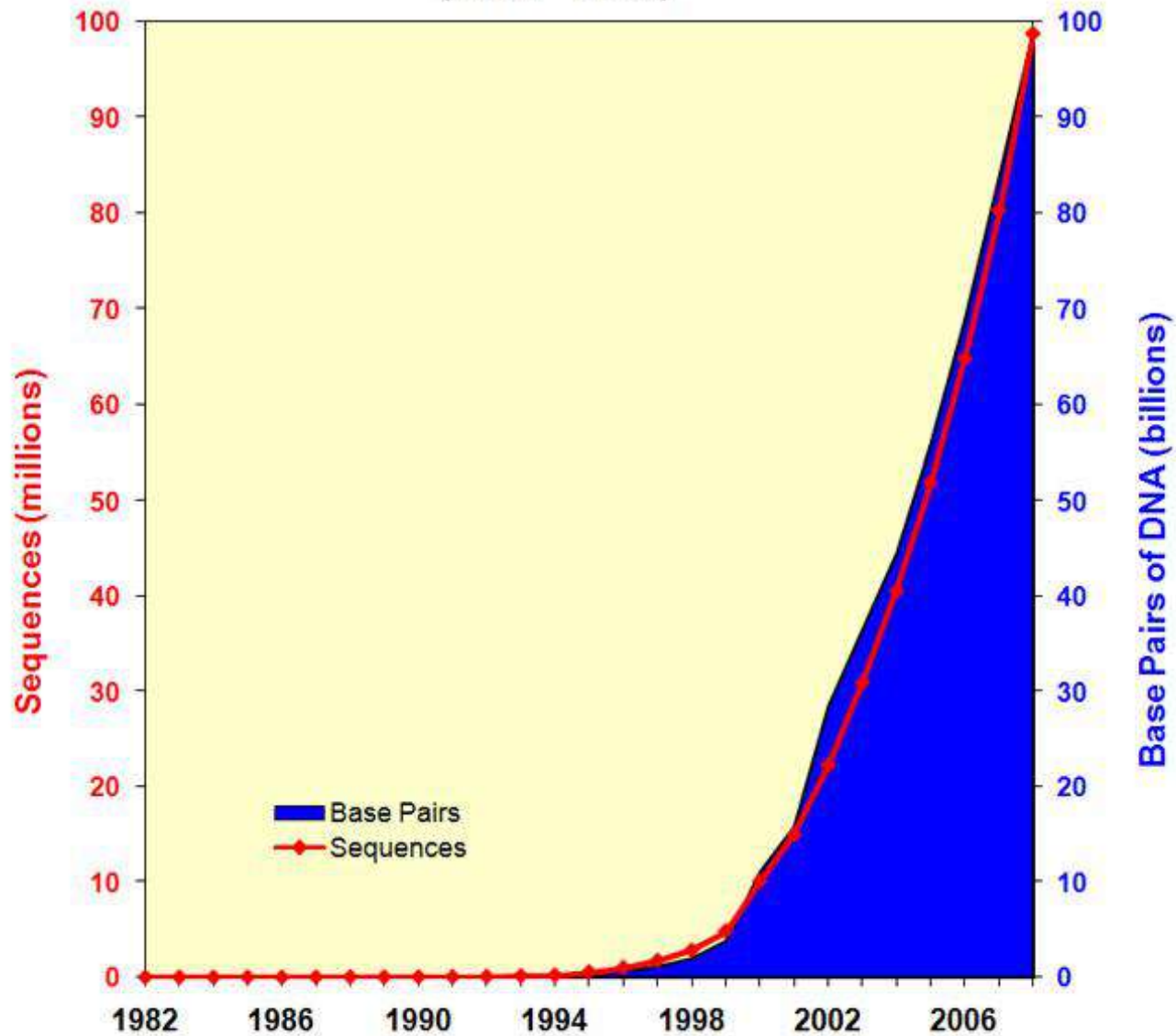
- DDBJ (Japan)

- (<http://www.ddbj.nig.ac.jp/>)

- Primary databases

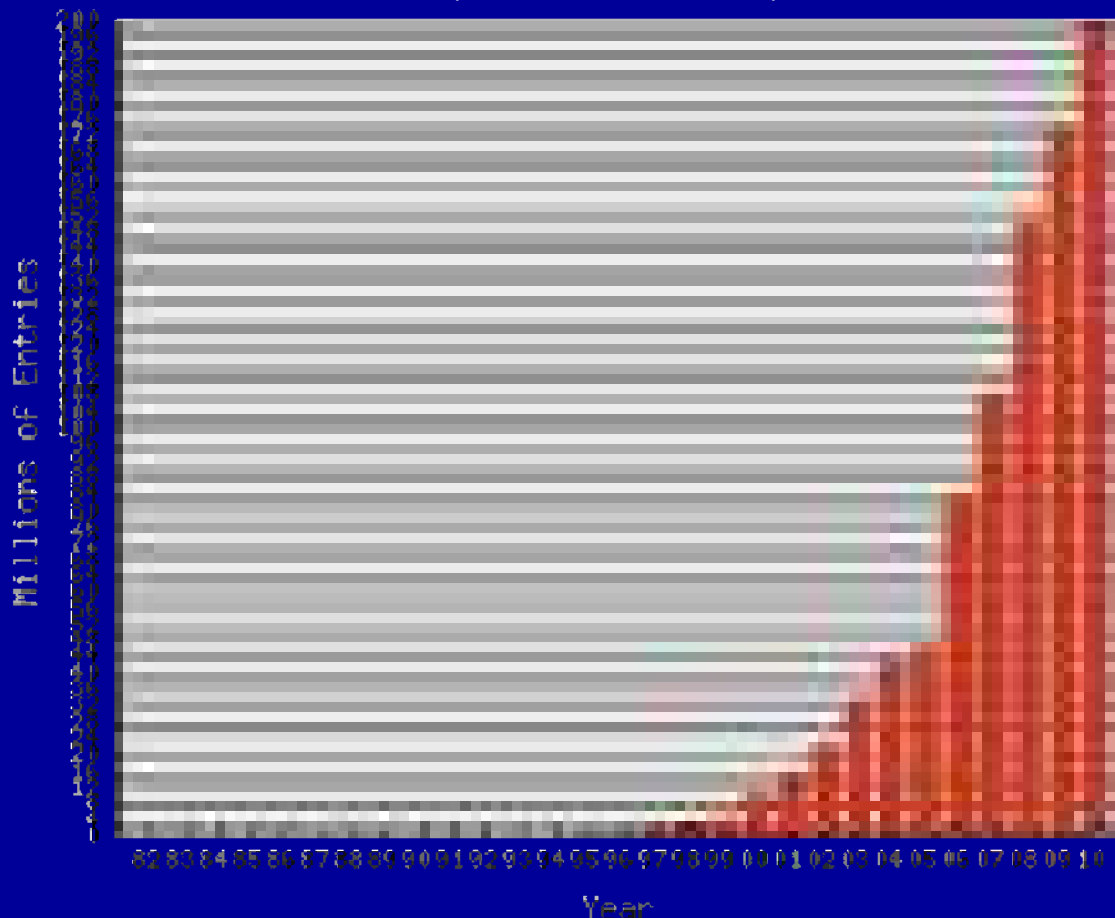
- DNA sequences are identical

# Growth of GenBank (1982 - 2008)



# EMBL Database

Number of entries  
(current 199,575,971)



Graphs created on 22 November 2010

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



National Center for  
Biotechnology Information

Search All Databases

Search Clear

#### NCBI Home

#### Site Map (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

## Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

### Human Microbiome Project

NIH Roadmap Initiative designed to characterize the community of microorganisms living on and in the human body.



1 2 3 4 5

#### Popular Resources

- [BLAST](#)
- [Bookshelf](#)
- [Gene](#)
- [Genome](#)
- [Nucleotide](#)
- [OMIM](#)
- [Protein](#)
- [PubChem](#)
- [PubMed](#)
- [PubMed Central](#)
- [SNP](#)

#### NCBI News

[Retirement of Peptidome, SRA & Trace Archive](#)

16 Feb 2011

[Due to budget constraints, NCBI will be discontinuing the](#)

[The Bookshelf has a new design & Browsing Tool](#)

09 Feb 2011

[Featuring a new homepage, search results display, limits and](#)

[More...](#)

- 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, 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

<b>PubMed:</b> biomedical literature citations and abstracts	<b>Books:</b> online books
<b>PubMed Central:</b> free, full text journal articles	<b>Images:</b> images from full text resources at NCBI
<b>Site Search:</b> NCBI web and FTP sites	<b>OMIM:</b> online Mendelian Inheritance in Man
<b>Nucleotide:</b> Core subset of nucleotide sequence records	<b>dbGaP:</b> genotype and phenotype
<b>EST:</b> Expressed Sequence Tag records	<b>UniGene:</b> gene-oriented clusters of transcript sequences
<b>GSS:</b> Genome Survey Sequence records	<b>CDD:</b> conserved protein domain database
<b>Protein:</b> sequence database	<b>UniSTS:</b> markers and mapping data
<b>Genome:</b> whole genome sequences	<b>PopSet:</b> population study data sets
<b>Structure:</b> three-dimensional macromolecular structures	<b>GEO Profiles:</b> expression and molecular abundance profiles
<b>Taxonomy:</b> organisms in GenBank	<b>GEO DataSets:</b> experimental sets of GEO data
<b>SNP:</b> single nucleotide polymorphism	<b>Epigenomics:</b> Epigenetic maps and data sets
<b>dbVar:</b> Genomic structural variation	<b>Cancer Chromosomes:</b> cytogenetic databases
<b>Gene:</b> gene-centered information	<b>PubChem BioAssay:</b> bioactivity screens of chemical substances



U.S. National Library of Medicine  
National Institutes of Health

Search: PubMed

[Limits](#) [Advanced search](#) [Help](#)

# PubMed is...

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

All Databases

PubMed

Nucleotide

Protein

Genome

Search

PubMed



for

Go

Clear

Limits

Preview/Index

History

Clipboard

Details

## 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

QuickTime™ and a  
TIFF (Uncompressed) decompressor  
are needed to see this picture.

# 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="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPGRGFI
FT FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGMK
FT DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYNNLINELLANGIQPFVTLFHWDLPO
FT VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWWFNSNGYALGTNAPGR
FT CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKKGIGITLVSNWLMPLD
FT DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIKVNRLPKFSESSLVNGSFDF
FT IGINYSSSYISNAPSHGNAPSYSTNPMTNISFEKHGIPLGPRASIIWYVYPYMFIQ
FT EDFEIFCYILKINITILQFSITENGMNEFNATLPVEEALLNTYRIDYRYRHLYYIRSA
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 tgggtgctgga tcttcagcat accaatttga 180
aggtgacagta aacgaaggcg gtagaggacc aagtatttg gataccttca cccataata 240
tcagaaaaa ataaggatg gaagcaatgc agacatcac gttgacaaat atcacgcta 300
caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc 360
ttggccaaga atactcccaa agggaaagtt gagcggagggc ataatcacg 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 Chlamydomophila caviae.  
OC Bacteria; Chlamydiae; Chlamydiales; Chlamydiaceae; Chlamydomophila.  
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/nr  
RA Read T.D., Myers G.S.A., Brunham R.C., Nelson W.C., Paulsen I.T.,  
Heidelberg J.F., Holtzapple E.K., Khouri H.M., Federova N.B.,  
Carty H.A., Umayam L.A., Haft D.H., Peterson J.D., Beanan M.J.,  
White O., Salzberg S.L., Hsia R.-C., McClarty G., Rank R.G.,  
Bavoil P.M., Fraser C.M.;  
RT "Genome sequence of Chlamydomophila caviae (Chlamydia psittaci GPIC):  
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 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/>

or send an email to [license@isb-sib.ch](mailto:license@isb-sib.ch)).  
-----  
DR EMBL; AEO16995; AAP05115.1; -. [EMBL / GenBank / DDBJ] [CodingSequence]  
DR HSSP; PO3004; 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; PFO0308; Bac DnaA; 1.  
DR Pfam; Graphical view of domain structure.  
DR PRINTS; PRO0051; DNAA.  
DR SMART; SMO0382; 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; CF440A7B300210D8 CRC64;  
MLTCSDCSTW EQFVNVYVKT RSKTAFENWI SPIQIIEETQ EKIRLEVPNI FVQNYLLDNY  
KQDLCSFVPL DAQGEPALEF VVAEIKKAPA QPIAPREPQE SPAETFEESK DFELKLNAAAY  
RFDNFIEGPS NQFVKSAAVG IAGRPGRSYN PLFIHGGVGL GKTHLLHAVG HYVREHHKNL  
RVHCITTEAF INDLVQHLRL KSIDKMKNFY RSLDLLLVDD IQFLQNRQNF EEEFCNTFET  
LINLNKQIVI TSDKPPGQLK LSERITARME WGLVAHVIGIP DLETRVAILQ HKAQKGLHI  
PNEIAFYIAD HIYGNVRQLE GAINKLTAYC RLFKGTLTES IVRDTLRELF RSPSKQKQVSV  
ESILKSVATV FQVKLQDLKG TSRSELVLA RQVAMYLAKT LITDSLVVIG SAFGKTHSTV  
LYACKTIEQK IERDETLTRQ 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](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.expasy.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](http://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://mbelserver.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](http://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](http://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



# 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/>

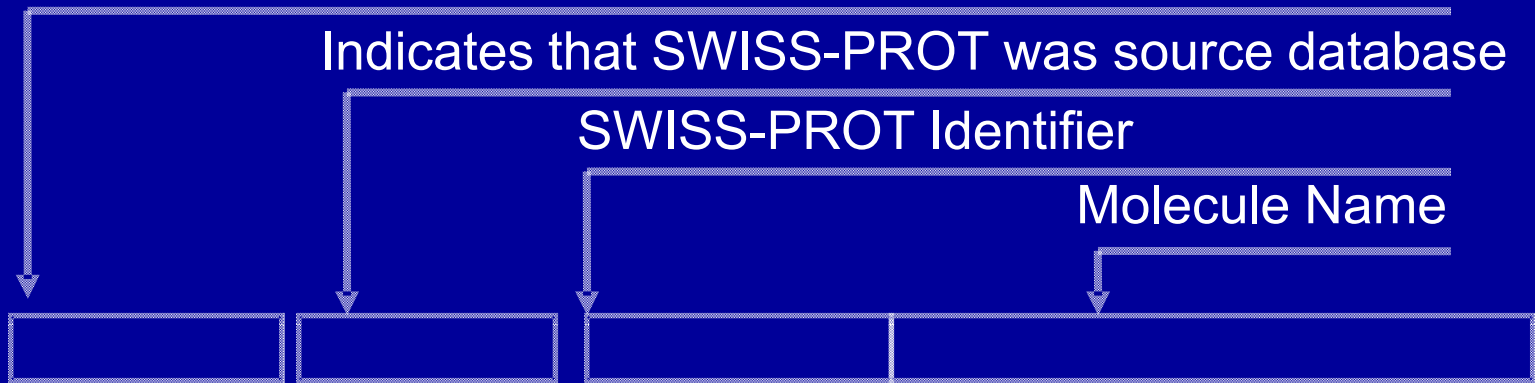
- SCOP
  - Structural Classification of Proteins
  - Hierarchically ordered and manually curated
  - 38221 PDB Entries
  - 110800 Domains
- CATH
  - Classification of protein domain structures
  - 124 folds
  - 226 Superfamily
  - 1148 Sequence family
  - 14473 Domain

# 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
- mcaaqrsoaaalaaaaprtvyafsarplaggepfnlsslrgkvllienvak
- slcgttvrdytqmndlqrrlgprglvvlgfpcnqfghqenakneeilncl
- yvrpgggfepnfmftekcevngakahplfaflrevlptpsddatalmtdp
- kfitwspvcrndvswnfekflvgpdgvpvrrysrrfltidiiepdielll
- qgasa

# Analyzing Results Using PERL Scripts

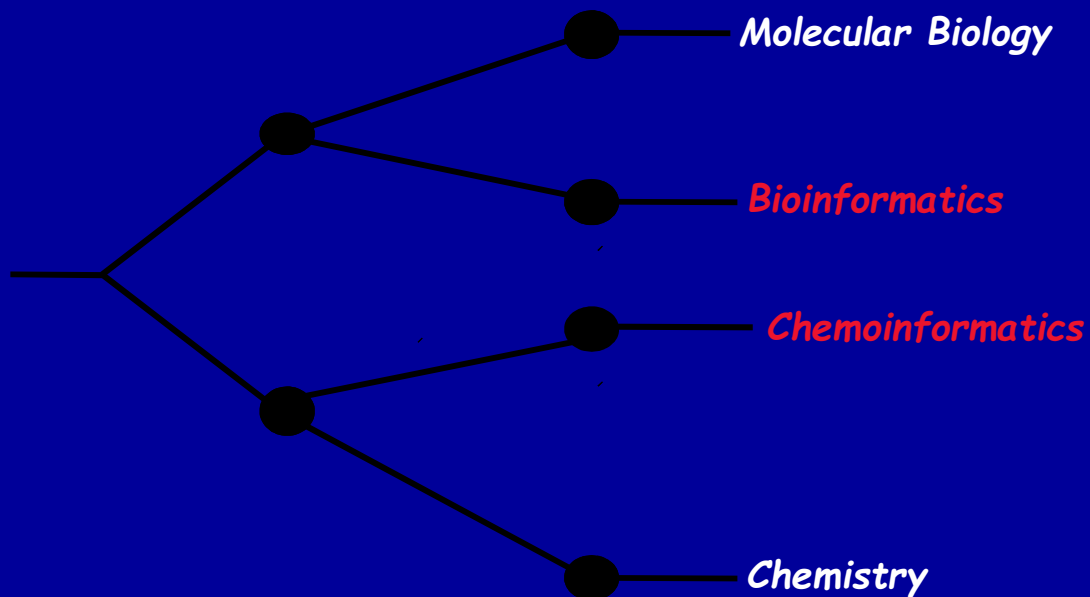
- Database servers now do:
  - Increasingly specific analysis of your results
- But you will eventually need to do analysis
- Ideal programming language is PERL
  - Designed to manipulate text and files
  - Can use it to play around with (manipulate) strings
- Will be using it in the coursework
  - PERL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

Sequences --- (Structures) --- Ligands/Lead

Genes  
Genomes

Proteins

Drugs



# Bioinformatics and Chemoinformatics Describing the DATA

Bioinformatics

Chemoinformatics

Sequences --- (Structures) --- Ligands

Chain

Chain  
Atom

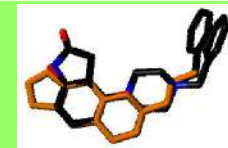
Each  
Atom

AGCTGTCGAGGGATAGGACA  
TATACATAAAATTAATATAAT

Strings



3D-Structure



SMILES, Sybyl,  
Matrices...

Structure Descriptors

# Bioinformatics and Chemoinformatics

## Storing the DATA

Bioinformatics

Chemoinformatics

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

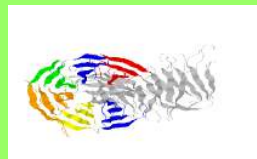
Chain

Chain  
Atom

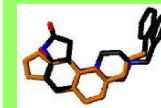
Each  
Atom

AGCTGTCGAGGGATAGGACA  
TATACATAAAATTAATATAAT

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

The screenshot shows the NCBI homepage in a Microsoft Internet Explorer browser window. The address bar displays "http://www.ncbi.nlm.nih.gov/". The page features the NCBI logo and navigation links for PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure. A search bar is visible with "Entrez" selected. The main content area includes a section titled "What does NCBI do?" which states: "Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)". A "Hot Spots" section lists: "Clusters of orthologous groups", "Coffee Break, Genes & Disease, NCBI Handbook", "Electronic PCR", and "Entrez Home".

## Chemoinformatics CAS/Beilstein/MDL

The screenshot shows the CAS website, a division of the American Chemical Society. The header includes the CAS logo and the tagline "SCIENTISTS SERVING SCIENCE". Navigation links for Products, Customer Care, About CAS, What's New, and CAS Learning Center are present. The main content area features a "Product Links" section with links to SciFinder, SciFinder Scholar, STN on the Web, STN, STN Express with Discover!, STN Easy, CA Salacts on the Web, Science Spotlight, ChemPort, and Science IP. A central text block states: "CAS... we are scientists, creating and delivering the most complete and effective digital information environment for scientific research and discovery. We provide pathways to published research in the world's journal and patent literature—virtually everything relevant to chemistry plus a wealth of information in the life sciences and a wide range of other scientific disciplines back to the beginning of the 20th century." Below this, a list of services is provided: "Whether you are: beginning a research project, looking for prior art in assessing patentability, uncovering information on other industry participants, providing research support to key decision makers." A sidebar on the right lists "Substance Database Information" including CAS Registry, Chemical Reactions, Commercially Available Chemicals, Compounds Claimed in Patents, and Regulated Chemicals Information. At the bottom, there is a banner for the "227th ACS National Meeting" in Anaheim, California, from March 28 to April 1, with CAS at Booth #552.

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

# Bioinformatics and Chemoinformatics Comparing the DATA

Bioinformatics

Chemoinformatics

Sequences --- (Structures) --- Ligands

Similarity

Fold

Descriptor

Evolutionary model

Dynamic Programming

BLAST

3D-Model

Structure/Structure  
Comparison

LSQman  
DALI  
SAP

Descriptor Similarity

Backtracking Algorithm

CACTUS  
(Cactus.nci.nih.gov)

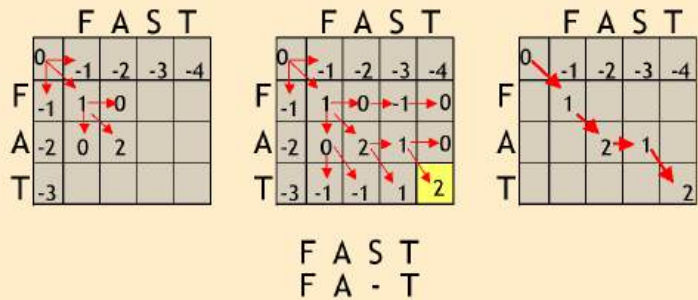
# Related Techniques

## Searching Databases

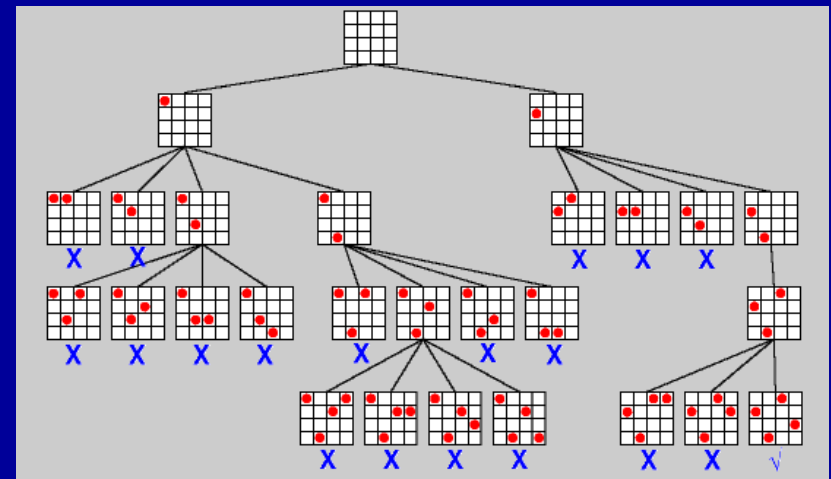
### Bioinformatics: Dynamic Programming

#### Dynamic Programming (Needlman and Wunsch)

Match=1 MisMatch=-1 Gap=-1



### Chemoinformatics Backtracking



# Bioinformatics and Chemoinformatics Building Models

Bioinformatics

Chemoinformatics

Sequences --- (Structures) --- Ligands

MSA

Fold

Descriptor

Multiple  
Sequence  
Alignments

ClustalW, TCoffee

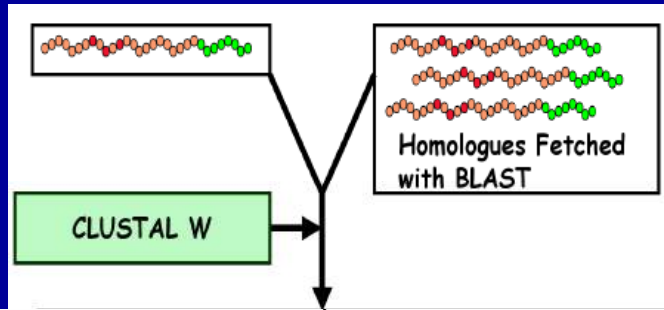
3D-Model  
Classifications

DALI  
SCOP  
CAT

Descriptor  
Computation

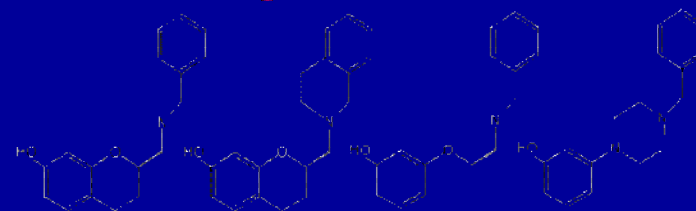
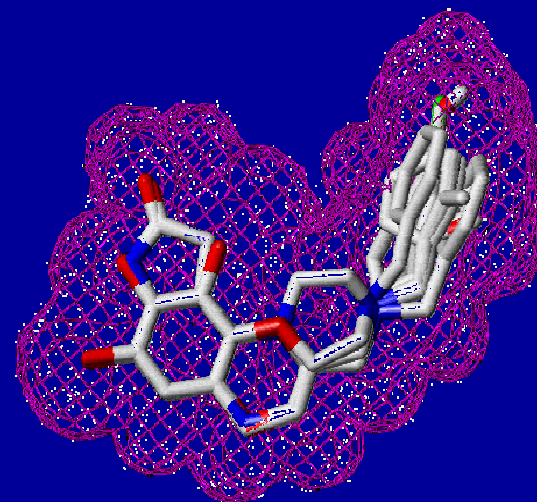
# Related Techniques Profiles

## Bioinformatics:



```
chite ---ADKPKRPLSAYMLWINSARES IKRENPDFK-VTEVAKKGGELWRGLKD
wheat --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKS LSE
trybr KKDSNAPKRAMTSEMFSSDFRS----KHS DLS-IVEMSKAAGA AWKELGP
mouse ----KPKRPRSAVNIYVSESFQ----EAKDDS-AQGK LKLVNEAWKNLSP
      ***. ::: :. . . . : . . * . * : *
```

## Chemoinformatics



# Bioinformatics and Chemoinformatics Predicting Properties

Bioinformatics

Chemoinformatics

Sequences --- (Structures) --- Ligands

Function

Binding

Descriptor

Multiple Comparisons

Profiles Pfam  
Function  
Structure  
Phylogeny

Docking/Mecanisms

Molecular Dynamic  
Docking Predictions

QSAR

Relate descriptors to  
Activity/Toxicity

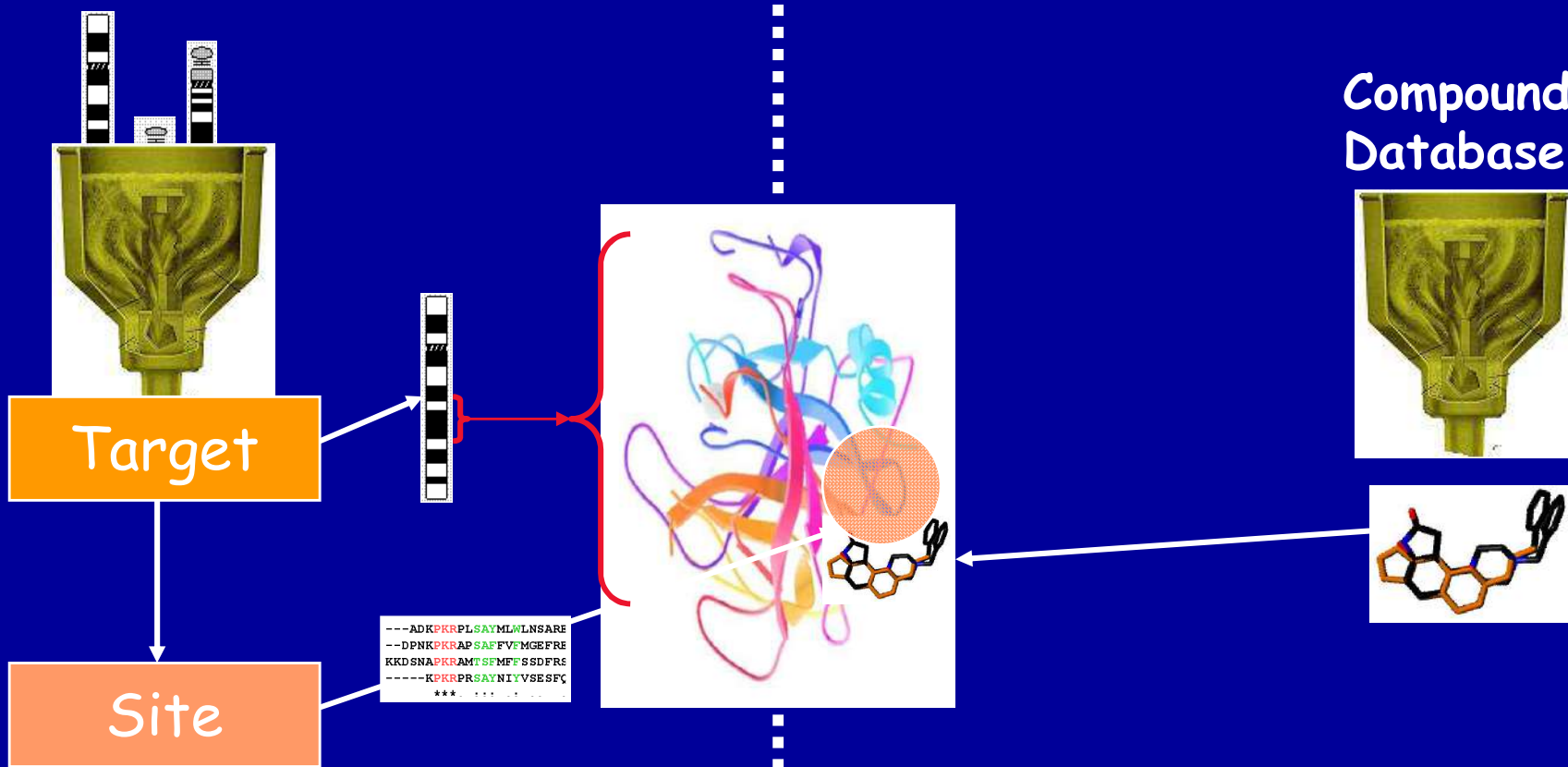


# Bioinformatics and Chemoinformatics The Questions

Bioinformatics:  
Target

Chemoinformatics  
Lead

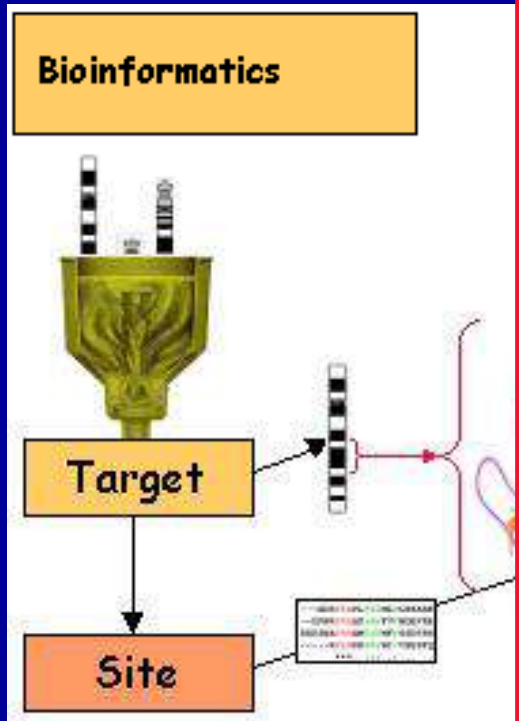
Compound  
Database



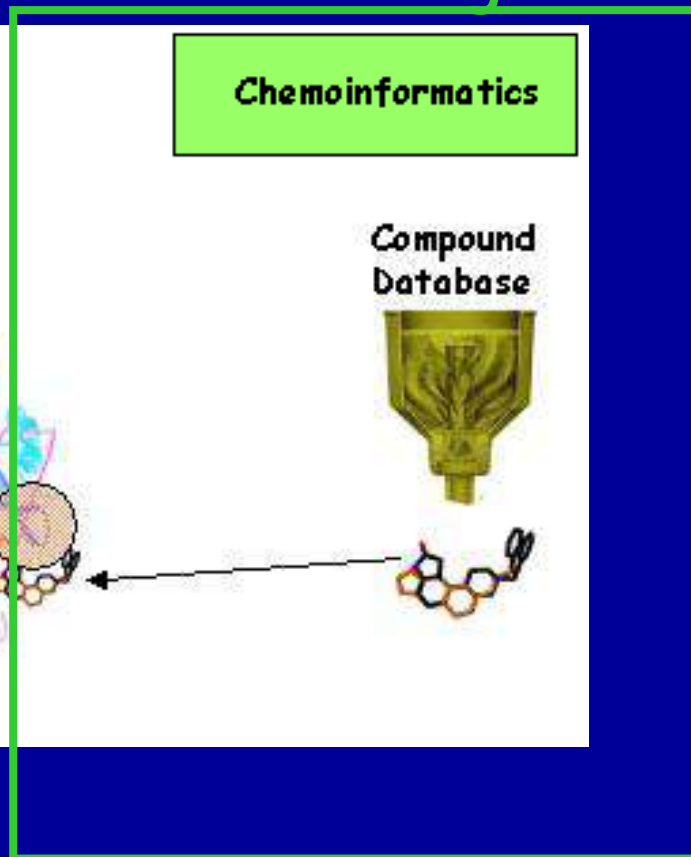
# Bioinformatics and Chemoinformatics

## The Questions

### Evolutionary Trace



### Chemical Modelling



# Bioinformatics and Chemoinformatics

## The Questions

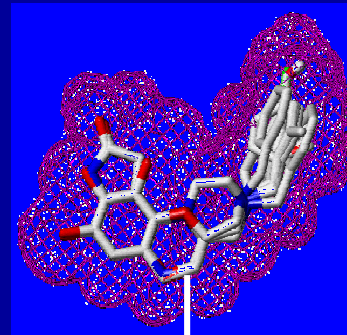
**Bioinformatics:  
Function**



**Comparative Genomics**

**Function**

**Chemoinformatics  
QSAR**



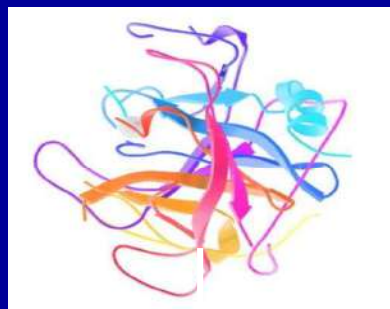
**Quantitative  
Structure-Activity  
Relationship**

**Activity/Toxicity**

# Bioinformatics and Chemoinformatics

## The Questions

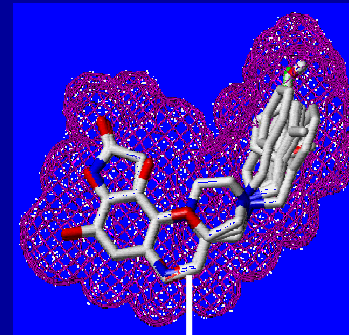
**Bioinformatics:  
Target**



**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

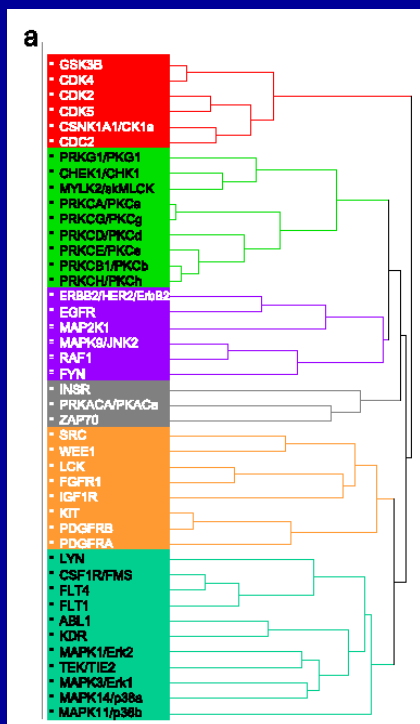
# Homology based SAR predictions

- SAR Matrices contain affinity Data

<b>Targets / Compounds</b>	<b>T1</b>	<b>T2</b>	<b>T3</b>	<b>T4</b>	<b>T5</b>	<b>T6</b>	<b>T7</b>
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

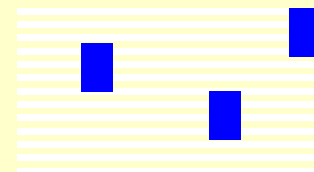
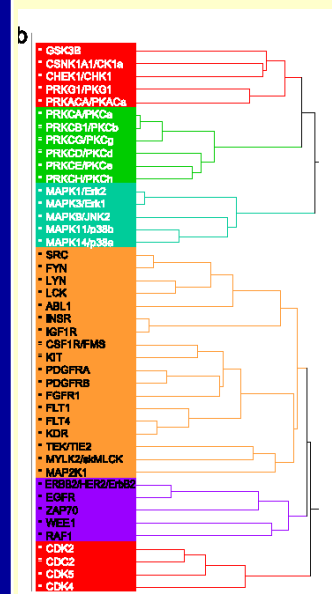
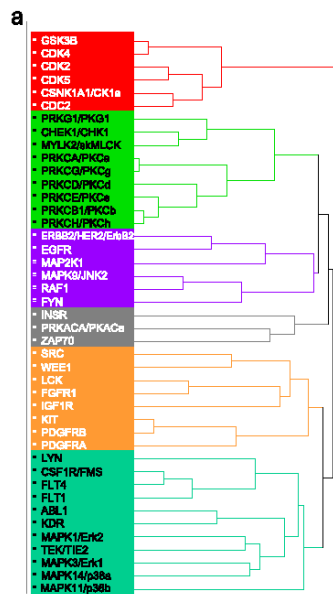


Targets / Compounds	T1	T2	T3	T4	T5	T6	T7
C1	8	4	>20	2			
C2	5	7		>20		>20	
C3	1.1	3		>20		>20	
C4	>20	>20		>20			
C5			2	0.9			
C6	>20	>20	0.1	0.09			
C7	0.1	0.2	0.1	0.1	0.3		
C8	0.77	0.2	>20	>20	>20		
C9	0.57	0.27	>20	>20			
...	0.2	>20	>20	>20	>20	>20	

# Homology based SAR predictions

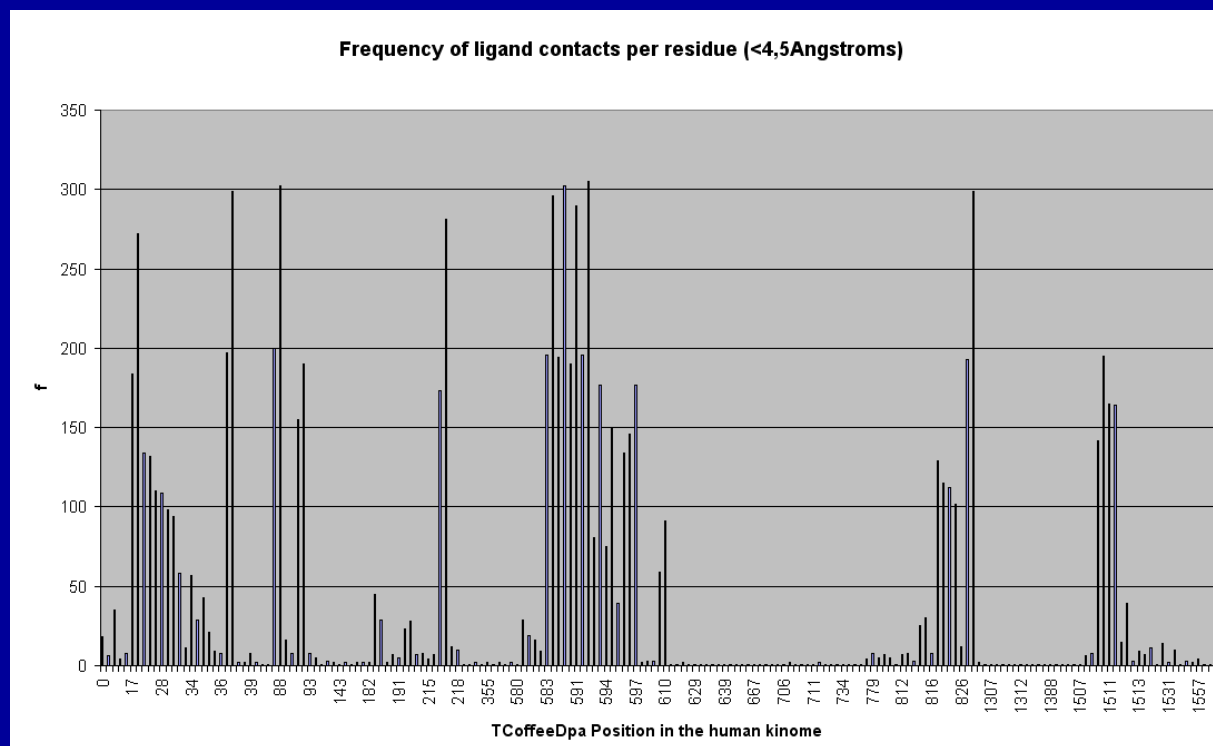
- Targets can be clustered using similarity

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	0.8			
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	>20		
—	0.2	>20	>20	>20	>20	>20	



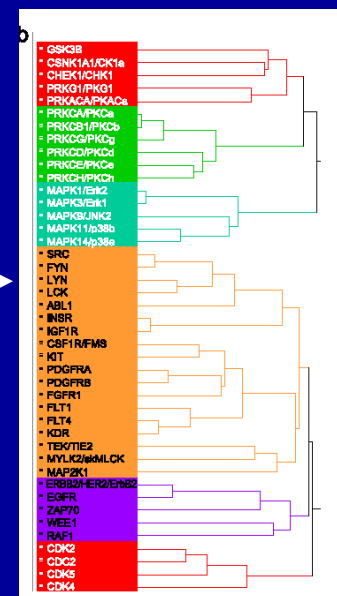
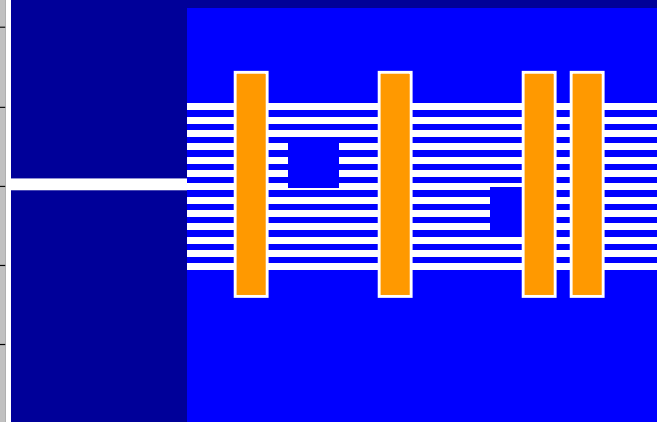
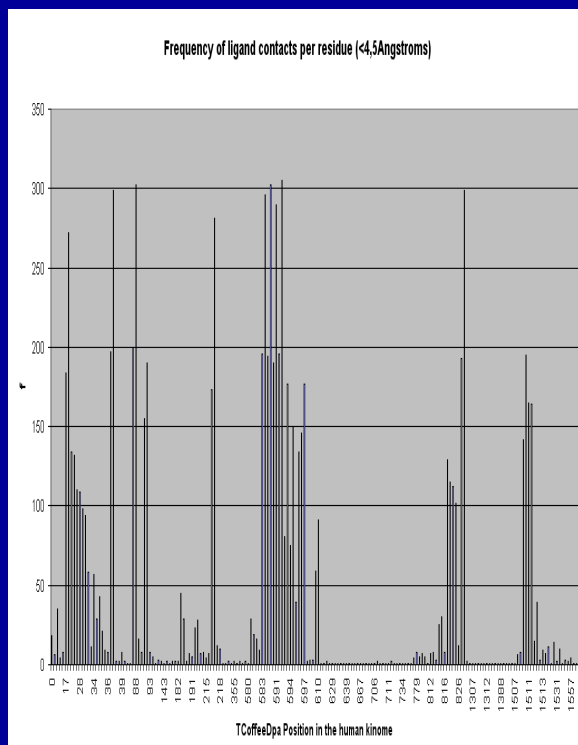
# Homology based SAR predictions

- 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

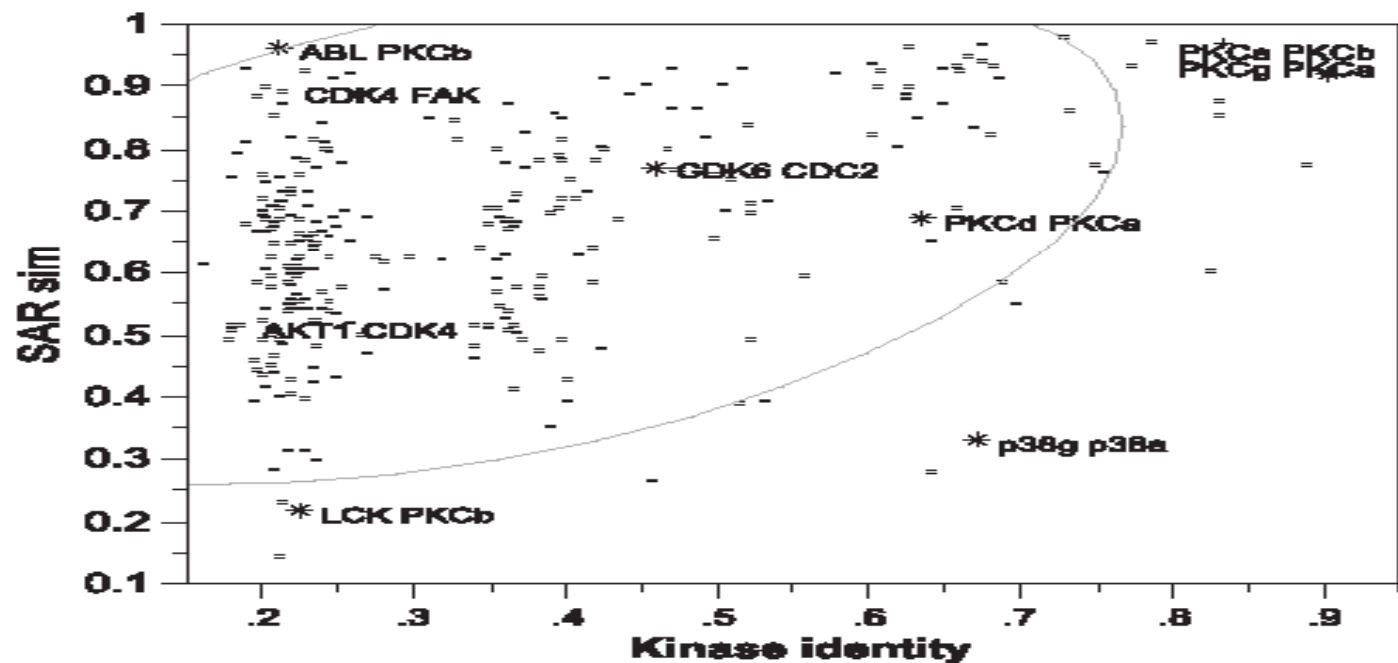


Fig. 4. Relationship between catalytic domain sequence identity and SAR similarity between pairs of kinase targets for 292 pairs with selectivity data. For this plot, only kinase pairs are shown for which data on at least two active compounds were available. Note that the available selectivity data to 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

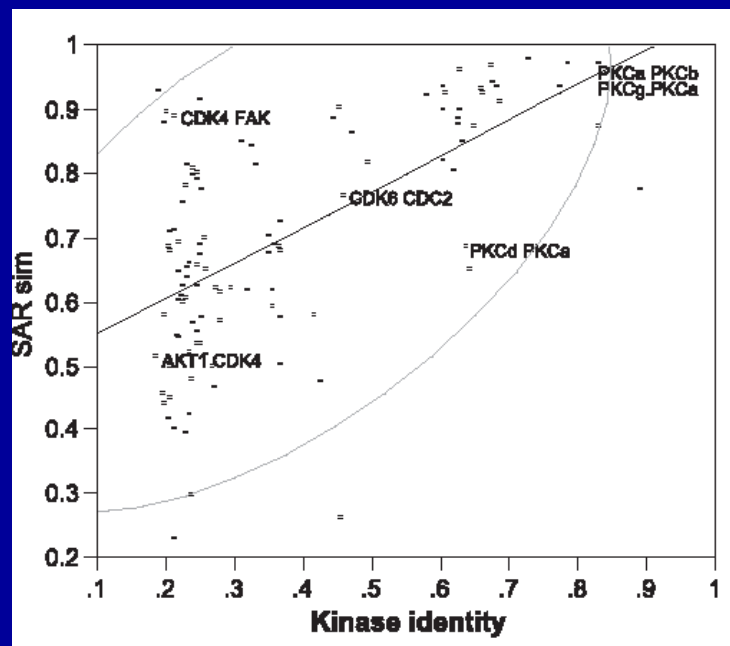
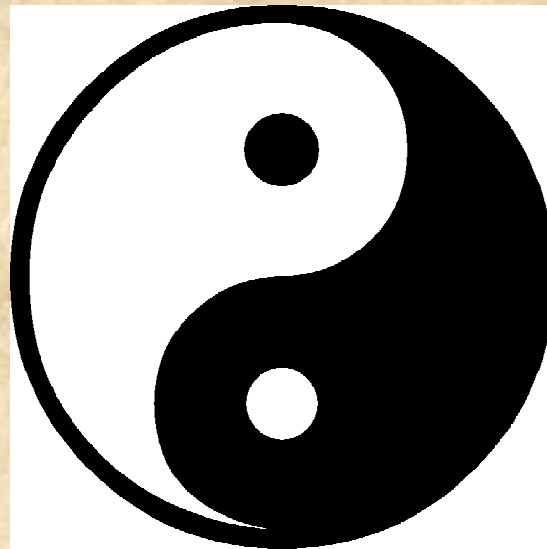


Fig. 5. Relationship between catalytic domain sequence identity and SAR similarity between 111 pairs of targets for kinases with large gatekeeper sizes. The  $R^2$  value for the linear fitting is 0.37.

# 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.”



## 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?

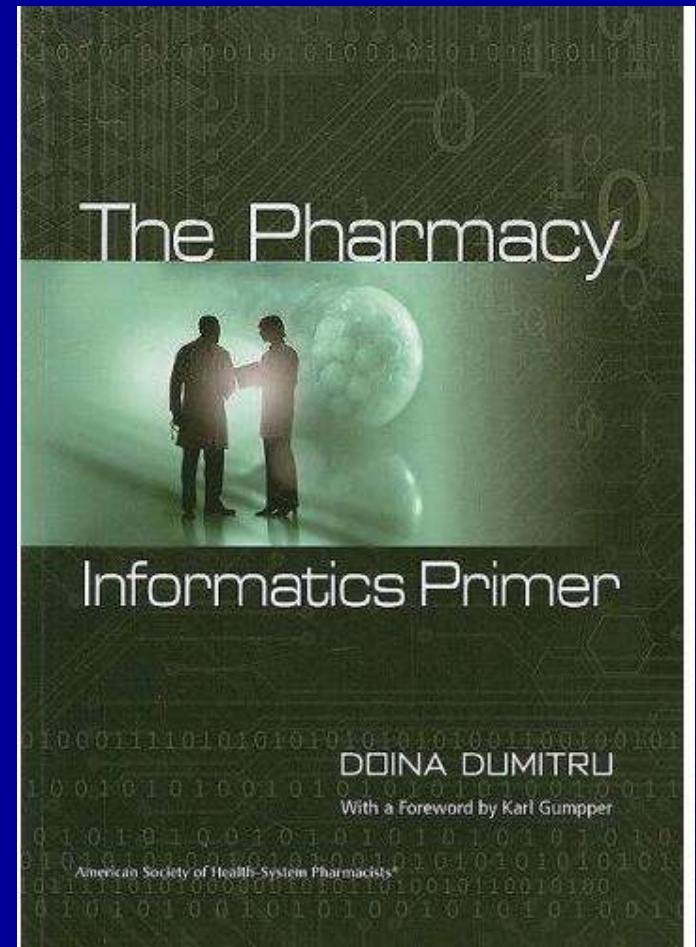
# Additional Resources

The screenshot displays the ASHP website interface. At the top left is the ASHP logo with the tagline "American Society of Health-System Pharmacists®" and "TOGETHER WE MAKE A GREAT TEAM". To the right are links for "Login", "Shopping Cart", and "Register". A search bar with a "Search" button and "Advanced Search" link is also present. Below the header is a navigation menu with buttons for "About Us", "News", "Member Center", "Advocacy", "Practice and Policy", "Meetings", "Continuing Education", "Accreditation", and "Bookstore". The main content area features a "MEMBER CENTER" sidebar with links for "Join ASHP", "Update Profile", "Renew Membership", and "Members Only". The "Sections" list includes "Ambulatory Care Practitioners", "Clinical Specialists and Scientists", "Inpatient Care Practitioners", "Pharmacy Informatics and Technology", "Pharmacy Practice Managers", and "Webinars and Podcasts". The main content area is titled "Pharmacy Informatics and Technology" and includes a graphic of a pie chart with one slice highlighted, the text "Section of PHARMACY INFORMATICS AND TECHNOLOGY<sup>SM</sup>", and a paragraph about the section's purpose. Below this are links for "Section Directory" and "ASHP Connect". Two "In the Spotlight" and "Message from the Chair" sections are also visible, each featuring a portrait and a short bio.

- 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



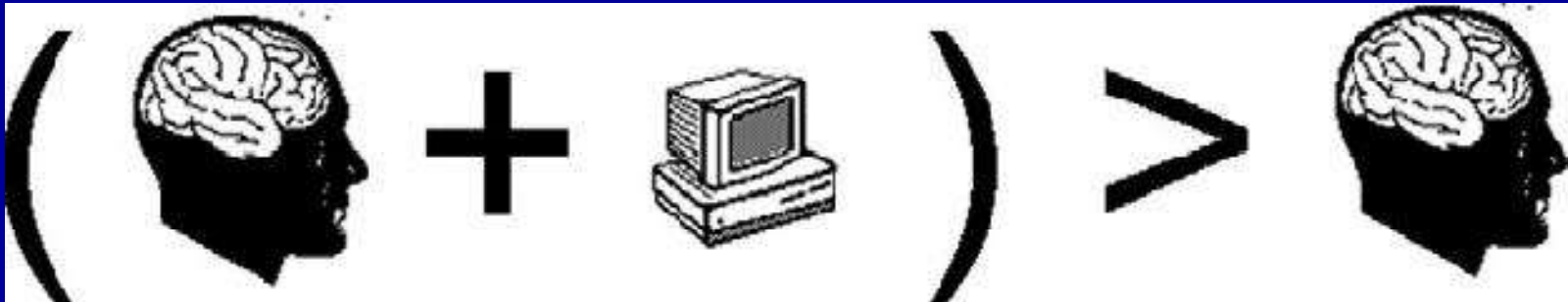
# Health Informatics – Electronic Medical Record Advantages

# 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.

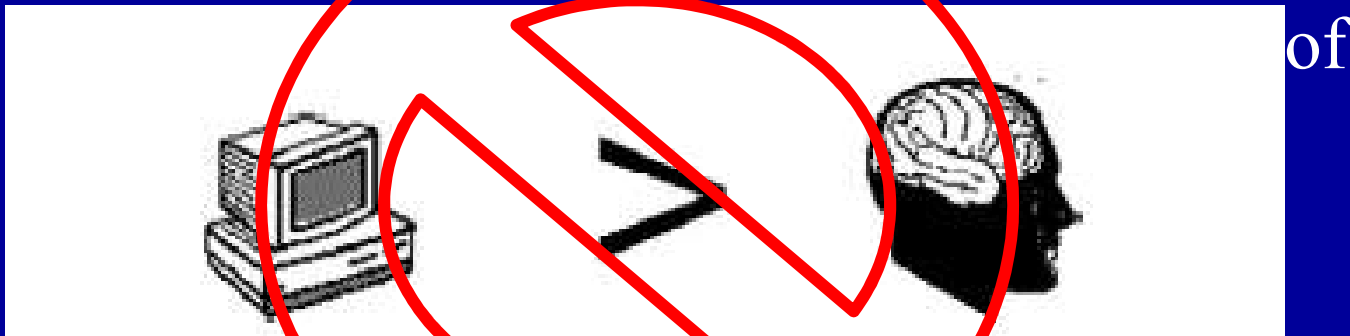
# What is Health Informatics?

- The fundamental theorem:
  - “A person working in partnership with an information resource is ‘better’ than that same person unassisted.”



# What is Health Informatics?

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



# 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

# Assessment

The Fundamental Theorem of Health Informatics states that:

- A. Information resources are more efficient than 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

VISN 20 > Reporting Services > Performance Report Document Set

Actions | 1 of 1 | Find Next | 100%



## V20 Opioid Risk - Summary Report For Authorized VA Personnel Only

Developed by Alex Linko (V20 DAG) and Gerald Kohler (NWC)  
 Default Parameters: Assigned PCP, Active Opioid Rx for 90+ Days, and 20+ MED, OEP: Yes/No, Gender: F/M  
 Note: Report excludes Tramadol Rx only patients and Suboxone patients.  
 Data as of 1/4/2016 9:04:00 PM

[Feedback](#)

[Report Documentation & Updates](#)

[VA Form - 10-0431c Consent for Long-Term Opioid Therapy for Pain](#)

[Opioid Risk Registry Data Use Agreement](#)

Custom Query | Upcoming PC Appt | Individual Patient Lookup | **Stats** | Resident Panel | MH Prescriber Panel | Report Usage | User Access

### COHORT IDENTIFIERS

Parent Site	PCP Patients	Opioid Patients	%	MED >= 120	Risk Score Top 10%	Benzo	%	State PMP Entry	V20 OSCR	Naloxone Kit	PTSD, MDD, & Tobacco	Apnea	ETOH or SUD Dx	Drug Seek Flag Active	SI Risk	SI Flag	Early Rx Refil
<input checked="" type="checkbox"/> BOISE																	
Total																	

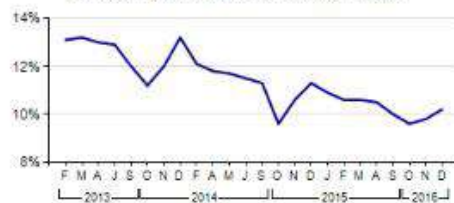
### METRICS

Parent Site	Opiate iMed Consent			Urinary Drug Screen			PC Visit Target >= 97%			State PMP Entry			Opioid Safety Case Review (OSCR)			OSCR Appropriate for Taper		
	No	Yes	%	No	Yes	%	No	Yes	%	No	Yes	%	No	Yes	%	No	Yes	%
<input checked="" type="checkbox"/> BOISE																		
Total																		

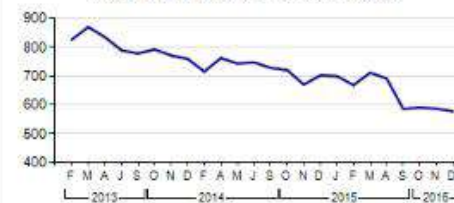
V20 % of Patients w/an Opioid Rx (3-yr trend)



V20 % of Patients w/MED >=20 (3-yr trend)



V20 Patients w/MED >=200 (3-yr trend)



# High Risk Opioid Registry

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

Morph Equiv	Risk Ranks			Recent Screens			SI			Active Flags		Admits/ED Visits	
	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	Yes	2			1	2	

Other Risk Indicators														
Pain Dxs	Cancer Dx	Terminal	MDD	PTSD	OTP	PT	Benzo Rx Active	Tobacco Use	Apnea	Homeless	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	

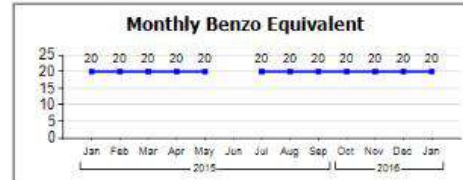
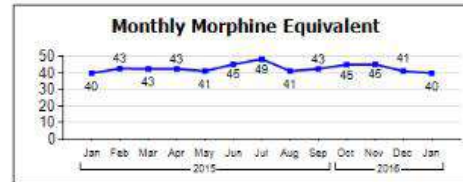
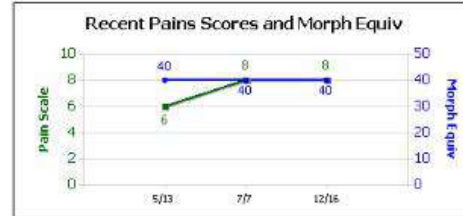
Resources	Agreement Notes	iMed Consents	UDS Tests	PMP Entries	OSRB Entries	PC Visits	MH Visits	SUD Dxs
			2	2		15	16	1

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	6.2%	-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 325MG TAB	11/13/15	DISCONTINUED	112	28	40
	CLONAZEPAM 2MG TAB (GEN)	12/31/15	ACTIVE	15	30	20
	HYDROCODONE 10/ACETAMINOPHEN 325MG 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
1/18/14		NEG	NEG		NEG			NEG

Upcoming Apointments	Date



# 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

# Other VA Initiatives



**VA Academic Detailing Service**



**Overdose Education and Naloxone Distribution (OEND): Naloxone Kit Distribution Report**

[Definitions](#)

[Help Desk](#)



**Insomnia Pharmacotherapy Dashboard**



[Definitions](#)



[Feedback](#)



[Print](#)



[Subscribe](#)



**PDSI Summary Report**

**Psychotropic Drug Safety Initiative**

Note: Please encrypt email when using PHI/PII.

[About](#)

[Feedback](#)

[Definitions](#)

[SSN Access](#)

[Trend Data](#)

[Home Page](#)

# 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

