Informatica – concetti di base

Sistemi operativi

Reti di computer

Uso della riga di comando

Testo e bioinformatica

Come funziona un PC

I personal computer si compongono di varie parti che è bene conoscere per capire come la macchina funziona. E' utile per prima cosa vedere cosa succede all'accensione:

Accensione

BIOS

Controllo elementi (scheda madre, processore, HD, ram ecc.)

Controllo memoria e tastiera

Ricerca dell'MBR sul disco master

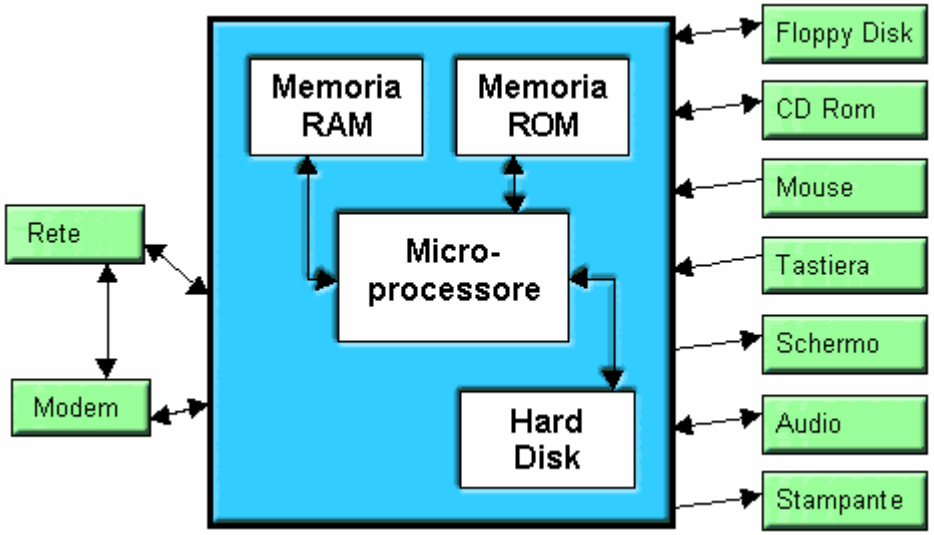
Avvio sistema operativo (kernel)

Caricamento driver periferiche

Caricamento driver grafici

Avvio dei servizi (demoni)

Macchina pronta.



Formattazione e file system

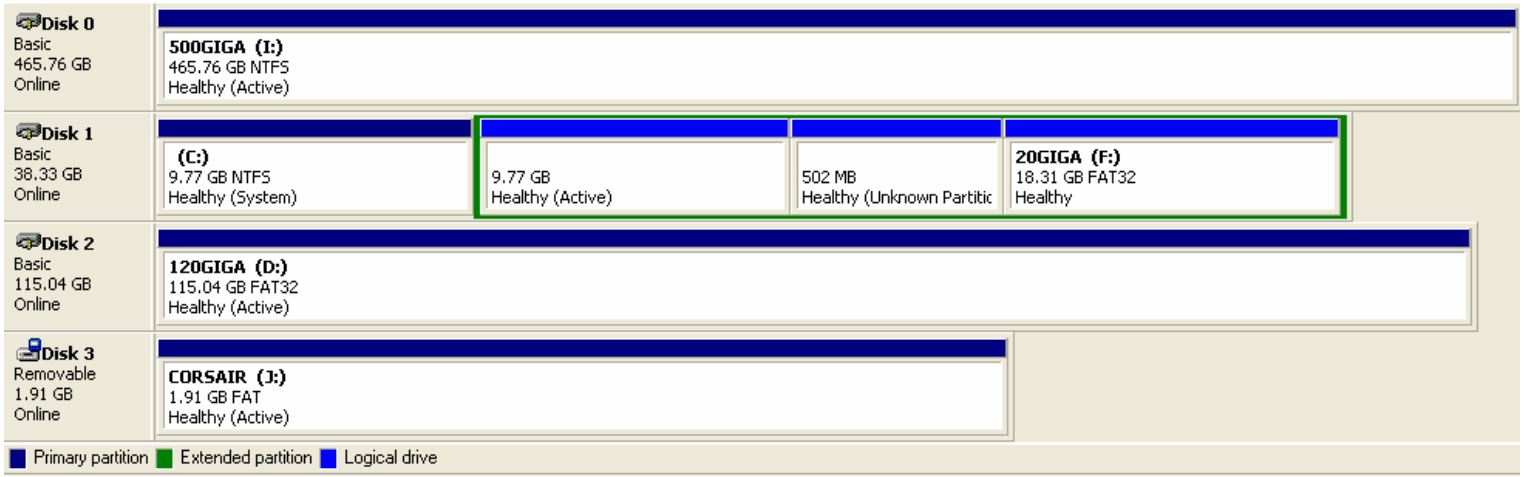
Il disco fisso non è altro che un supporto su cui memorizzare bit.

La formattazione è la procedura che permette di **organizzare lo spazio** all'interno di un disco, il quale dalla fabbrica esce come un supporto magnetico "casuale". La formattazione imposta le tracce e i blocchi e fornisce un "percorso" alle testine di lettura/scrittura.

La formattazione in genere è seguita o accompagnata dalla creazione del file system.

Il **file system** comprende **i metodi e le strutture di dati usate da un sistema operativo per tenere traccia dei file** su un hard disk o su una sua partizione. Di fatto è come una lingua che deve essere appresa dal disco affinché esso risponda un modo coerente quando gli si chiede dove trovare un file al suo interno.

Esistono vari tipi di filesystem, in base al tipo di sistema operativo che si installa (FAT, FAT32, NTFS per windows, SWAP, EXT2, RAISERFS ecc. per linux).

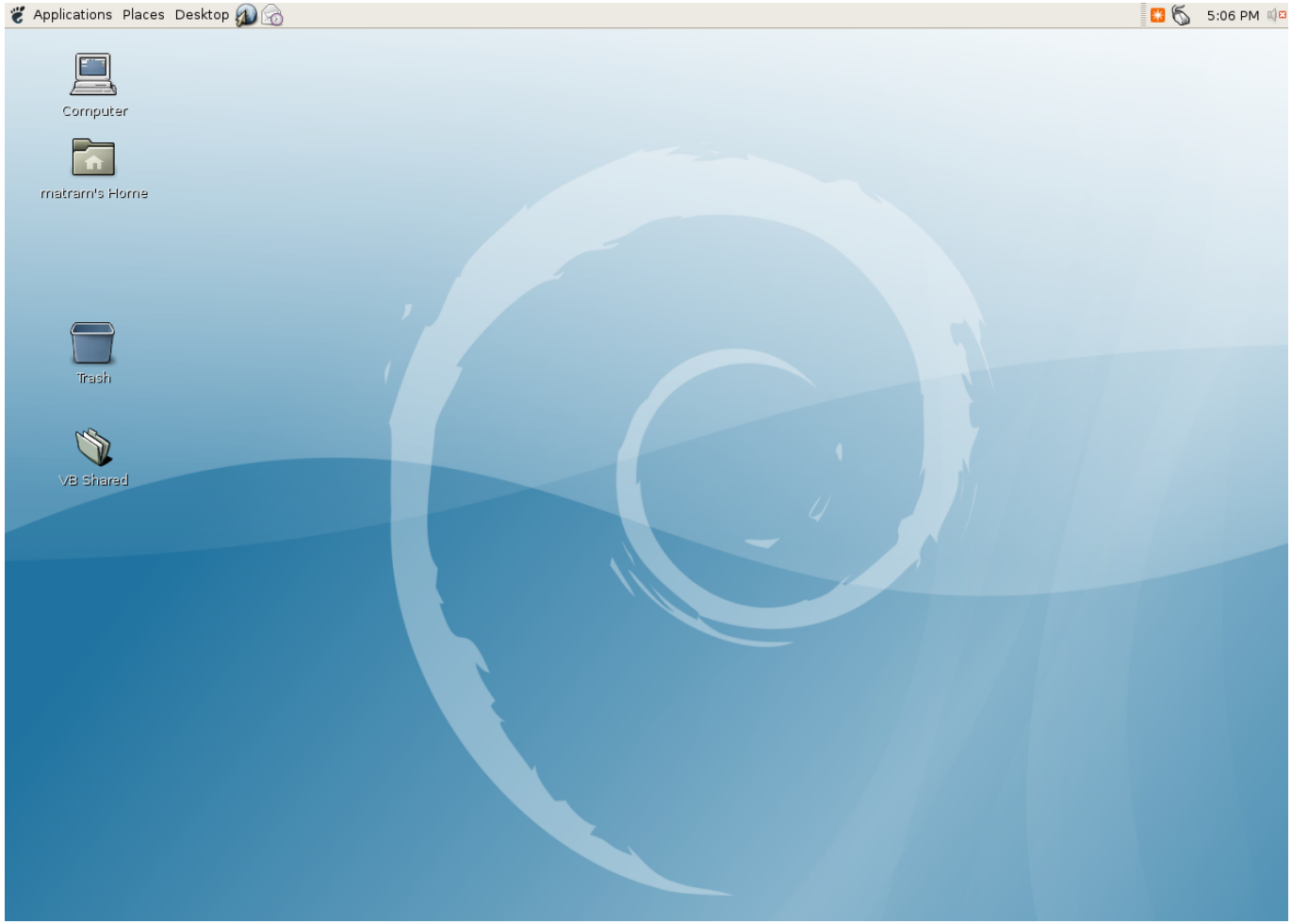


Windows

Il sistema operativo Microsoft, “ben” consolidato e uniforme, nelle varianti XP, Vista e 7. Nonostante sia lo standard di riferimento in Italia e in gran parte del mondo, non offre la possibilità di eseguire alcune applicazioni fondamentali in bioinformatica: gli sviluppatori molto spesso lavorano in altri ambienti, e non sono tenuti a rilasciare versioni per windows delle loro applicazioni…

Linux

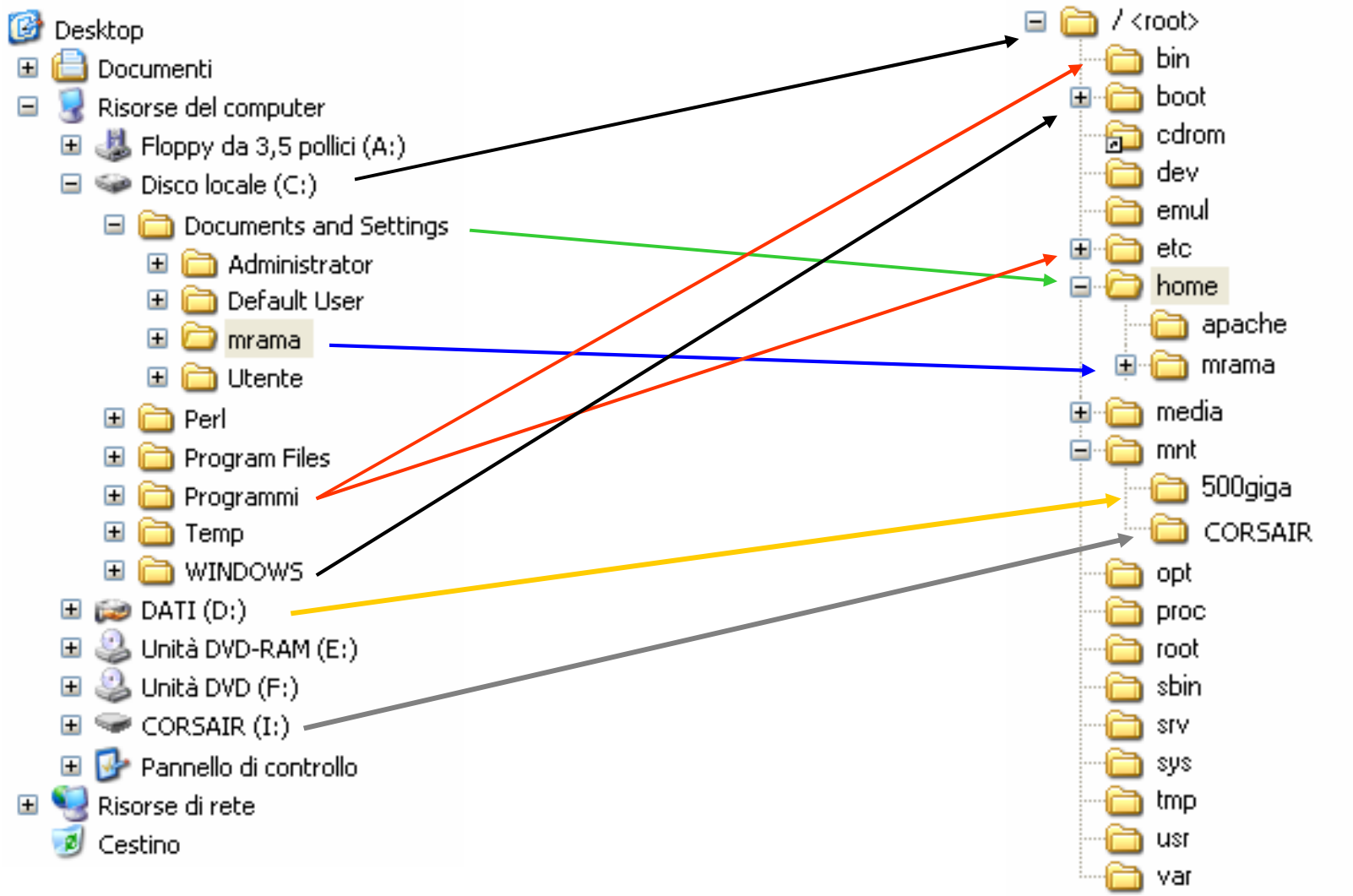
Linux è un sistema operativo aperto e gratuito, anche se non facilmente fruibile dall'utente medio, prevalentemente per motivi di abitudine. Le versioni grafiche si offrono con molti aspetti diversi (distribuzioni, es. Ubuntu, Debian, Fedora, Red Hat, Suse), ma il sistema di fondo e i principi di funzionamento sono sempre gli stessi.



Sistemi operativi a confronto

windows

linux



Tipi di file

Il sistema Windows (ad anche i sistemi linux grafici, es. Ubuntu) conta sulla cosiddetta “estensione” dei file per associare ad essi dei programmi di apertura: classici esempi sono i file .doc (MSWord), .zip (7Zip), .txt (Notepad) ecc.

Le icone assegnate alle estensioni vengono “ereditate” dal programma a cui sono associate.


file.exe - eseguibile


file.doc - MSWord


file.txt - notepad


file.zip - WinZip

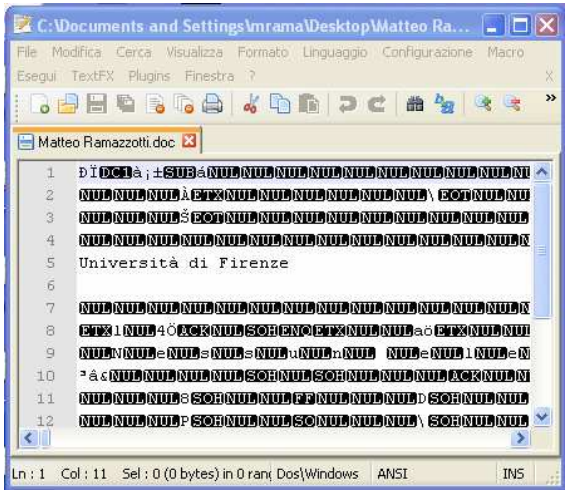

file nessuna

E’ però possibile configurare il sistema affinché il sistema presenti delle alternative: è il cosiddetto “Apri con”, che permette di aprire es. un file .txt con MSExcel. Non tutti i tipi di file però sono leggibili da tutti i programmi.

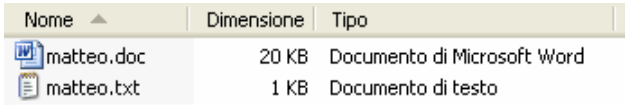
Ecco come appare un documento Word all’editor di testo NotePad++

Nel file.doc
c’è scritto

Matteo Ramazzotti
Università di Firenze



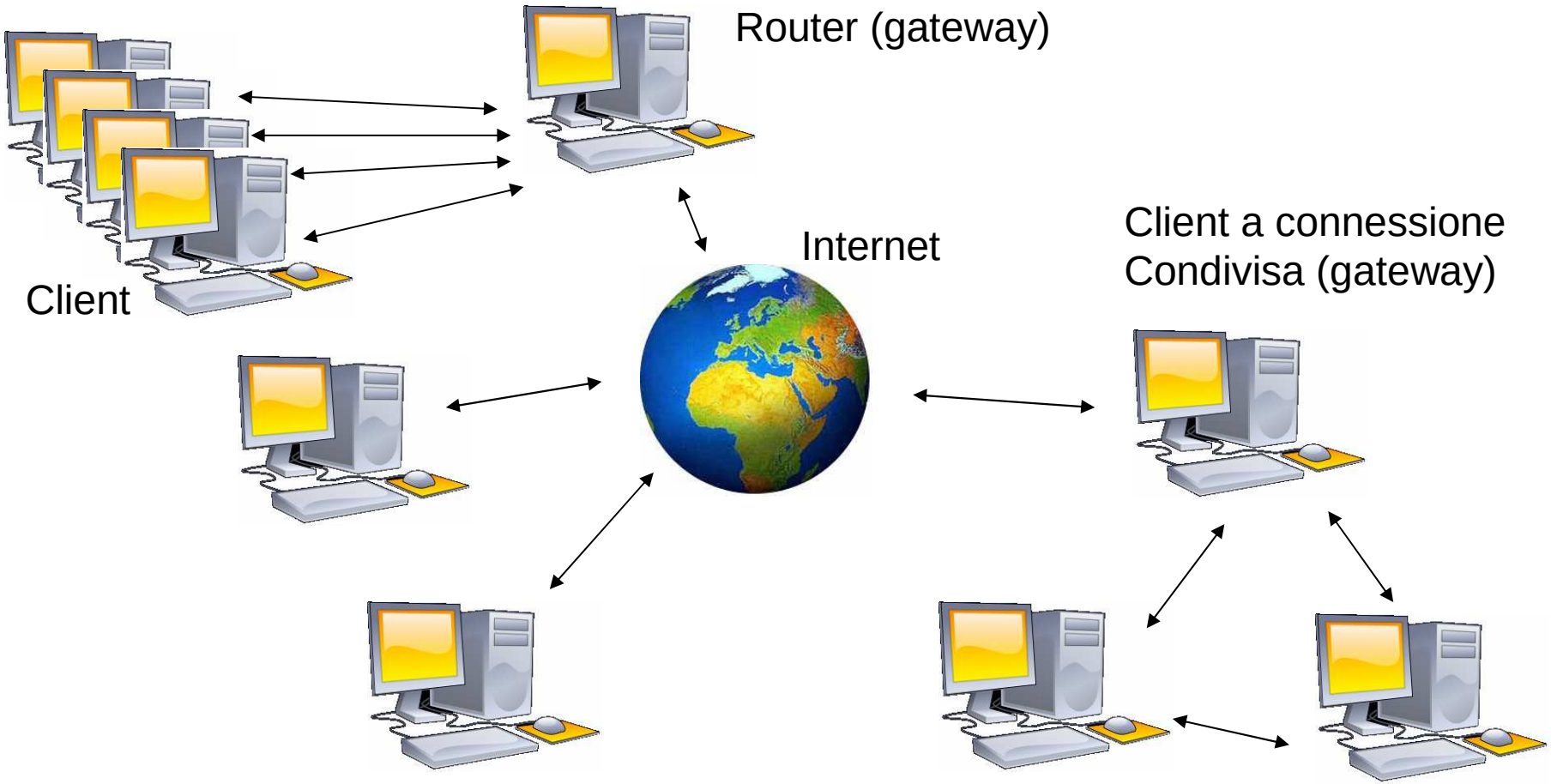
Il motivo di questo comportamento è semplice: Word interpreta una serie di informazioni scritte nel file (formattazione ecc) oltre che il testo, infatti…



Reti di computer

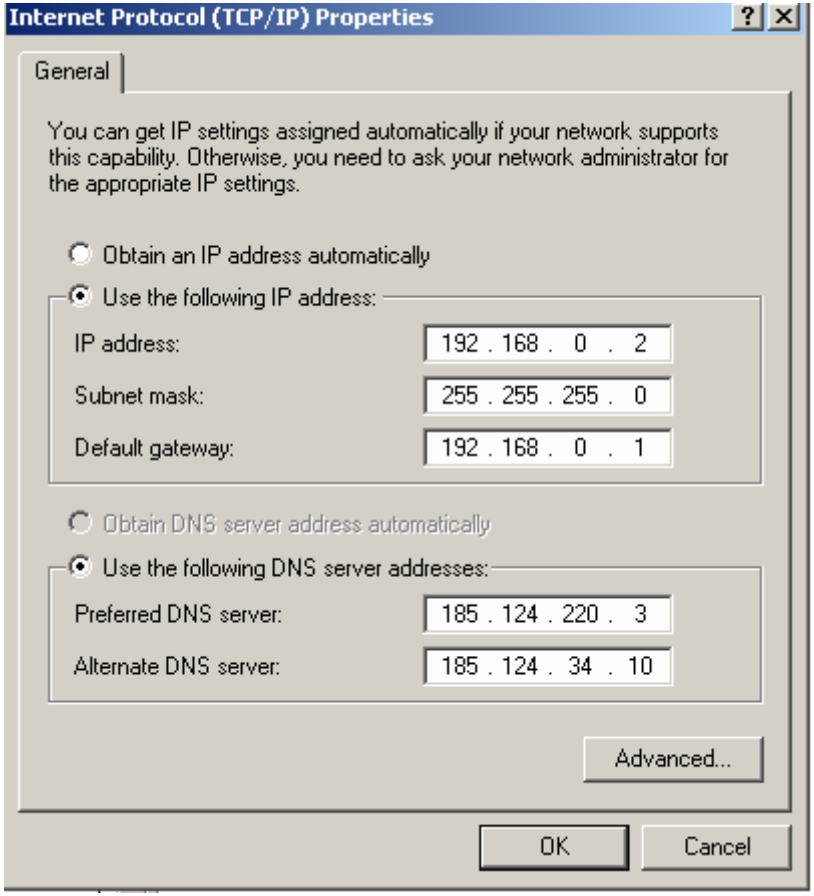
Tramite le schede di rete vari computer possono essere connessi tra loro a formare le cosiddette reti informatiche. Esistono **reti locali** (intrantet) e la **rete globale** (Internet).

Le macchine fisicamente connesse a internet (es. un PC, ma anche solo il modem), sono dette **gateway**, e possono o meno are da “ponte” per altre mecchine neslla stessa rete locale.



Configurare una rete

Le reti si basano su degli identificativi standard di tipo numerico basato su 4 numeri di 8 bit ciascuno (=> 0..255) separati da punti “.”. A ciascuna macchina connessa ad internet viene assignato uno di questo codici unico a livello mondiale.



IP (Internet protocol): può essere privato, cioè visibile solo nella rete locale o pubblico, cioè visto da tutto il mondo.

Subnet mask: determina il tipo di rete locale in cui la macchina è inserita

Gateway: IP del router connsesso ad internet nella propria rete (es. Unaltro computer, il modem ecc.)

DNS (Domain Name System): IP del computer che si occupa di tradurre gli indirizzi IP in testo (es. www.google.it)

Server e client

L'"atteggiamento" di un computer connesso ad altri può essere di

1. fruizione di un servizio: client

2. erogazione di un servizio: server

Un **server** NON è un computer, bensì un programma che sta su un computer e che risponde alle chiamate dei client erogando il servizio per cui è configurato (es distruibuzione pagine HTML).

Ogni volta hce il nostro PC visualizza una pagina internet, in realtà si è comportato da client ed ha chiesto al server (es. www.google.it, che senza il server DNS si chiamerebbe es. 74.125.43.104...) di dargli le informazioni per visualizzare la pagina.

Esempi di server sono:

APACHE: http://www.apache.org/, il più famoso server HTTP

SQL e mySQL: http://www.mysql.it, il più famoso server di database

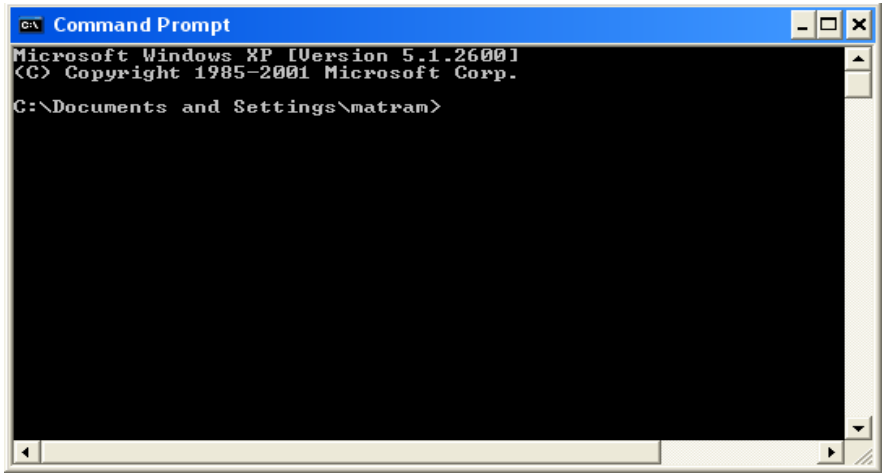
FTP e SFPT: server di condivisione file tra 2 macchine

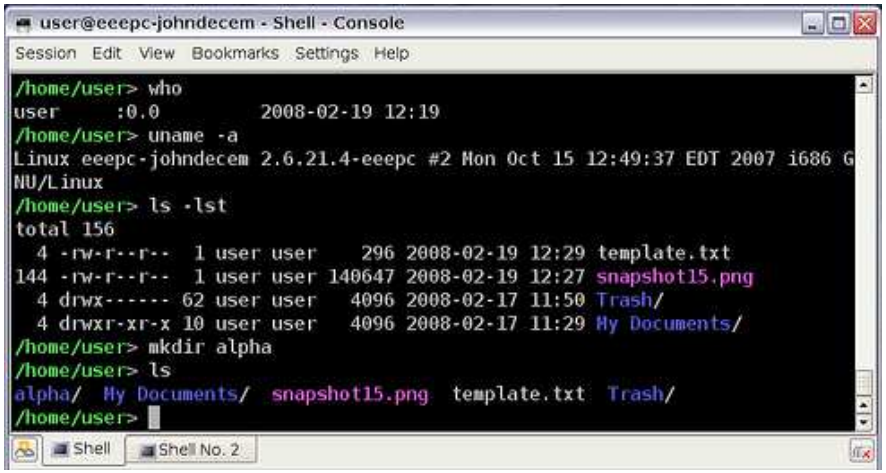
SSH: server di criptazione per connessioni tra due macchine.

Le shell di comando

E’ lo strumento basilare della bioinformatica, bisogna imparare a conviverci e ad apprezzare le potenzialità di eseguire comandi e programmi “scrivendoli” anziché cliccandoci sopra.

Windows e Linux hanno due shell molto diverse: il cosiddetto “**prompt**” di windows è la versione brutta e stupida della ben più potente “**bash**” di Linux con la quale è possibile fare davvero tutto.





Alcuni esempi di comandi shell

Comando	Windows	Linux
Aiuto	help	man
Copiare	copy	cp
Spostare	move	mv
Cancellare	del	Rm
Visualizzare	type / more	cat / tail / head
Cercare	find	grep
Lista file	dir	ls
Creare cartelle	mkdir	mkdir
Eliminare cartelle	rmdir	rmdir

La home folder

Le home sono le posizioni (cartelle) sul disco dove stanno i file e le configurazioni degli utenti di un computer.

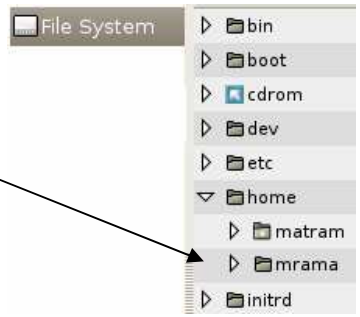
Ogni utente ha la sua home. Quello e **solo quello** che sta in questa cartella può essere gestito dall'utente (aggiungere files, rinominarli, cancellarli ecc.).

Anche l'amministratore del computer ha la sua home (anch'esso è un utente) ma può modificate i file anche fuori da essa (si dice che ha i privilegi per farlo), quindi può lavorare sul sistema (es. modificare i file di sistema contenuti in C:\windows oppure in /etc/)

Supponiamo che **mrama** sia un utente di un computer: a seconda del sistema operativo utilizzato, la sua home sarà

In windows: C:\Documents and Settings\mrama

In linux: /home/mrama



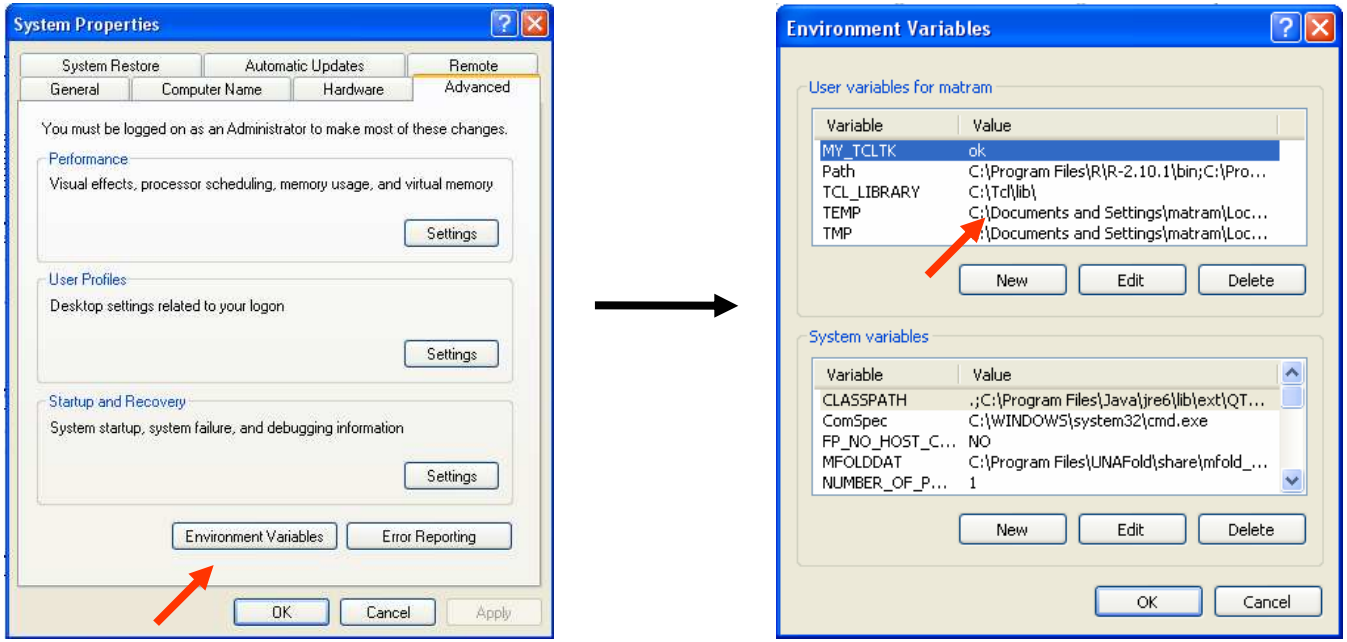
Le variabili d’ambiente

Una variabile (es v) è un oggetto che può assumere diversi valori (es. v = s/t) di diverso tipo (es. v = 10 : un numero intero; v = atgcagcatgc : una stringa).

Quando una shell viene avviata, legge le variabili d’ambiente ed è tutto quello che “sa” su chi e come la sta utilizzando.

Es. imposto VAR = Matteo, poi lancio la shell e scrivo es. “set VAR”. Il sistema scrive Matteo, cioè il valore della variabile VAR

In Windows le variabili d’ambiente si impostano da “Pannello di Controllo – Sistema – Avanzate - Variabili d’ambiente” oppure con il comando “set VAR = …” dal prompt.



Il concetto di PATH

Il “path” o percorso di esecuzione è una delle più importanti “variabili d’ambiente”: dice dove stanno i programmi e i comandi ed evita di doverne digitare il percorso quando si voglia lanciarli.

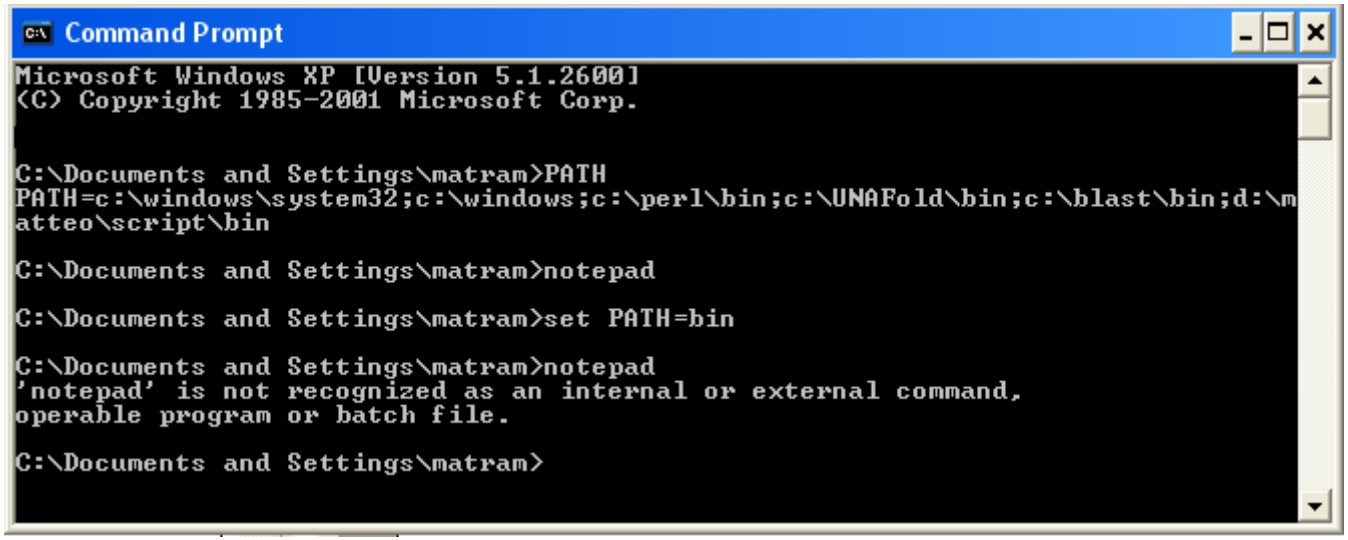
Es. PATH= c:\windows\system32;c:\windows;c:\perl\bin;c:\blast\bin;d:\matteo\script\bin

Osservare la sintassi: i vari percorsi sono separati dal “;”

Nell’esempio sotto, il PATH originale (quello scritto nelle variabili d’ambiente viste prima) permette di lanciare il blocco note semplicemente con “notepad” perché l’eseguibile notepad è nella cartella c:windows, che è specificata nel PATH

Se non fosse stato così avremmo dovuto scrivere “c:\windows\notepad”.

Se si modifica il path specificando “bin”, il nuovo path è in una cartella chiamata “bin” e il comando notepad non funzionerà più.

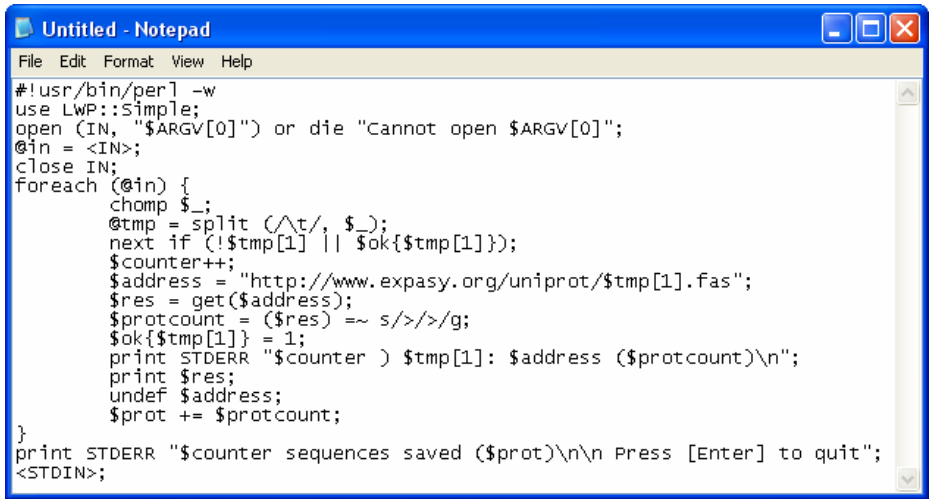


Il formato testo semplice

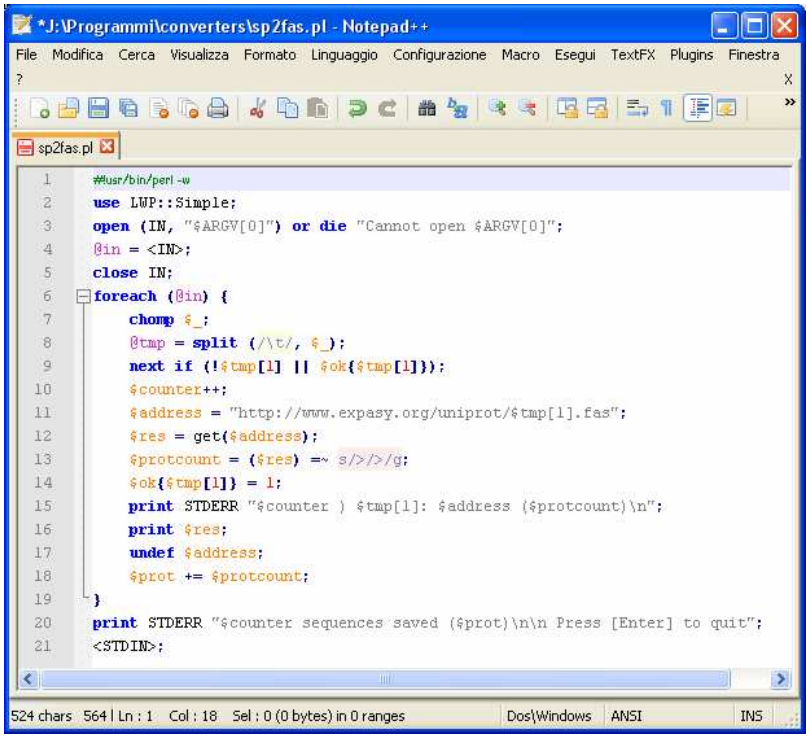
Quasi tutti i programmi che utilizzeremo hanno input ed output con formati molto diversi fra loro, ma tutti sono accomunati dal fatto che sono scritti nello stesso tipo di file, il

TESTO SEMPLICE

In bioinformatica non si utilizza (quasi) mai un editor di testo complesso come Word, si ricorre sempre a strumenti semplicissimi come Notepad o ai più complessi Notepad++ o Crimson Editor: questi sono editor che hanno delle modalità di visualizzazione ricche (highlight), ma salvano sempre i file in testo semplice.



Notepad



Notepad++

Linguaggi di programmazione

Sono degli insiemi di regole che determinano una sintassi simile al linguaggio corrente ma che servono ad indicare ad un computer una serie di operazioni per eseguire una certa procedura. Le regole possono riguardare vari livelli di astrazione che man mano si allontatano sempre di più dal linguaggio proprio del computer chiamato **Assembler**.

Tra i linguaggi di alto livello, si possono individuare due grosse categorie:

Compilati (es. C, C++): il codice sorgente viene letto da un **programma compilatore** che lo trasforma in **codice macchina** e genera un file **eseguibile**. Il file risultante non necessita più del compilatore e può essere copiato su altre macchine ed eseguito.

Interpretati (es. Perl, Ruby, Python): detti anche linguaggi di **scripting**, il codice sorgente viene eseguito da **un interprete** il quale esegue direttamente le istruzioni, senza generare un eseguibile. Non generando un eseguibile, il sorgente stesso può essere copiato su altre macchine a fatto funzionare **purchè nella macchina stessa sia presente l'interprete.**

Durante il corso proveremo ad utilizzare il linguaggio interpretato

Perl

(*Practical Extraction and Report Language*)

molto utilizzato dai biologi bioinfromatici per la sua facilità di utilizzo e la sua sintassi relativamente facile.

Tipi di caratteri

In bioinformatica è molto importante la posizione dei caratteri, sia in senso assoluto (dove stanno nel testo) sia in modo relativo (come si relazionano tra loro).

Esistono caratteri a **spaziatura ineguale** (un punto "." occupa meno spazio di una "M") e caratteri a **spaziatura uguale** ("." ed "M" occupano lo stesso spazio): solo i secondi sono accettabili !!!

Il motivo è chiaro se si osservano le scritte sotto: sono due frammenti proteici allineati, e solo nel caso dei caratteri a spaziatura uguale si può apprezzare l'allineamento…

Spaziatura ineguale

Arial
VILAMHLKIMNH
VKLAMRLKILNH

Times
VILAMHLKIMNH
VKLAMRLKILNH

Comic
VILAMHLKIMNH
VKLAMRLKILNH

Tahoma
VILAMHLKIMNH
VKLAMRLKILNH

Spaziatura uguale

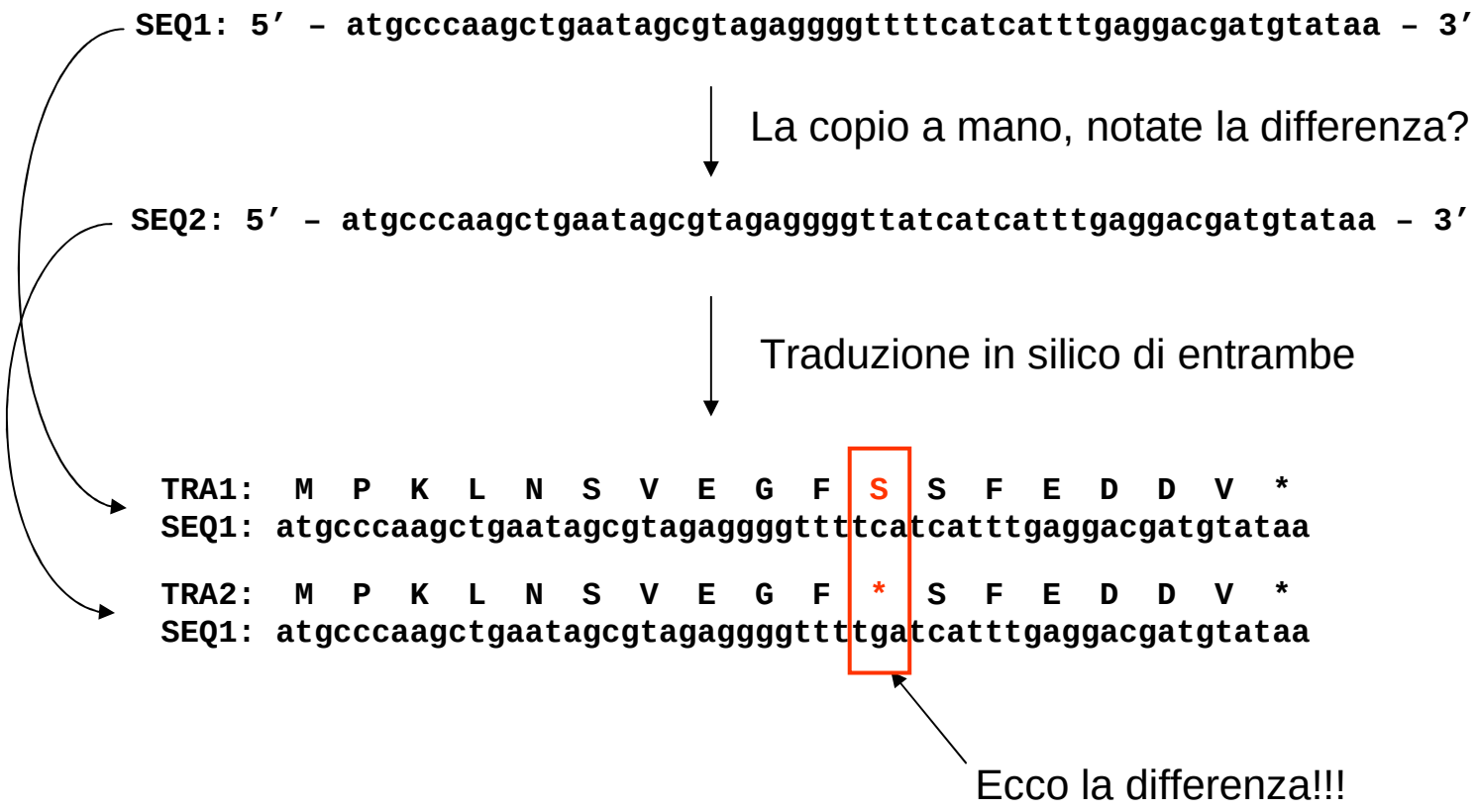
```
Courier New
VILAMHLKIMNH
VKLAMRLKILNH
```

```
Courier New
VILAMHLKIMNH
| ||| ||| ||
VKLAMRLKILNH
```

Copia e incolla

In bioinformatica non si devono **MAI** scrivere a mano le sequenze di proteine o aminoacidi: perdere o sbagliare una lettera significa combinare grossi guai per le analisi successive.

L'esempio più classico che si può fare è legato alla traduzione in silico (vedremo meglio poi…)



La tripletta tca codificava una serina (S) e si è trasformata in una tripletta di stop tga !!!

Forme e dimensioni

Quanto detto prima sul non copiare o trascrivere a mano le sequenze vale anche se le si devono cambiare di “formas”: supponiamo di avere una sequenza di 100 basi raggruppata in gruppi da 10. Ma la voglio raggruppata in gruppi di 5.

```
  1 taacatactt attgtttta actactcgtt ttccattcga ctcatcacgc
 51 tccccccccc cccccccccc cttatccgtt ccgttcgacg tatttcgttg
101 tctaatttct gacgtaactt gttcctgtt aagtaccgtt tatggcctat
151 actccggtat ttaaaacgac gacgattcca ccgtaaagcc gtcaaccaga
```

E’ chiaro che è possibile farlo a mano, ci vorranno 2 minuti, ma non ha senso e si rischia di sbagliare

```
  1 taaca tactt attgt tttta actac tcgtt ttcca ttcga ctcat cacgc
 51 tcccc ccccc ccccc ccccc cttat ccgtt ccgtt cgacg tattt cgttg
101 tctaa tttct gacgt aactt gttcc ctgtt aagta ccgtt tatgg cctat
151 actcc ggtat ttaaa acgac gacga ttcca ccgta aagcc gtcaa ccaga
```

Esistono appositi formattatori on-line che permettono di effettuare questo tipo di operazioni in modo veloce e sicuro.

Programmi ed eseguibili

Molti programmi di analisi bioinformatica sono disponibili via web con interfacce semplici e chiare. Spesso però i server sono intasati e le analisi richieste sono lente ad arrivare. In molti casi è meglio lavorare “in locale”, cioè scaricare i programmi e lavorare sulla propria macchina.

Ecco una lista dei principali programmi che verranno utilizzati durante il corso: impareremo a configurarli e ad utilizzarli durante le esercitazioni…

Blast2:	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.2.22+-win32.exe
ClustalW:	ftp://ftp.ebi.ac.uk/pub/software/clustalw2/2.0.10/clustalw-2.0.10-win.msi
ClustalX:	ftp://ftp.ebi.ac.uk/pub/software/clustalw2/2.0.10/clustalx-2.0.10-win.msi o
BioEdit:	http://www.mbio.ncsu.edu/BioEdit/BioEdit.zip
RasMol:	http://www.rasmol.org/software/RasMol_Latest_Windows_Installer.exe
Swiss PDB Viewer:	http://spdbv.vital-it.ch/download/binaries/SPDBV_4.01_PC.zip
Perl:	http://downloads.activestate.com/ActivePerl/releases/5.10.1.1007/ActivePerl-5.10.1.1007-MSWin32-x86-291969.msi
Crimson Editor:	http://www.crimsoneditor.com/download/cedt370r.exe