

Bioinformatics - a definition ?

The design, construction and use of software tools to generate, store, annotate, access and analyse data and information relating to Molecular Biology

OR

Biologists doing “stuff” with computers?

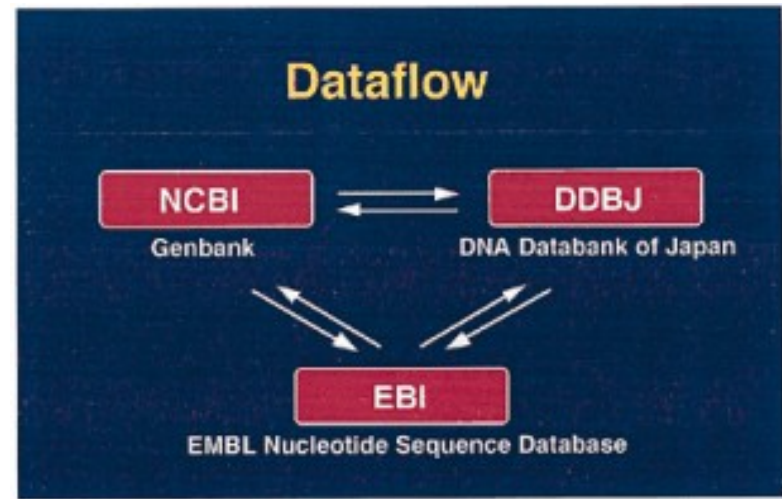
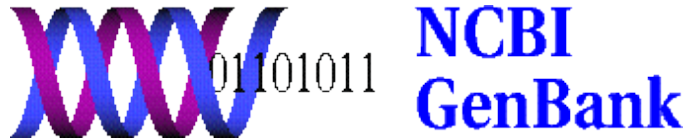
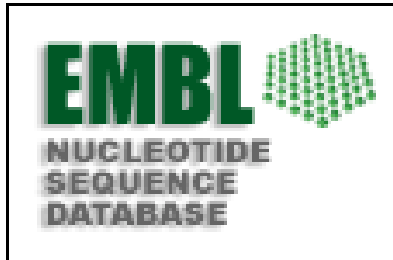
Here we consider the use of Bioinformatics tools rather than their design and construction

Here we consider the access and analysis of data and information items rather than their generation, storage or annotation

Databases - Genes to Genomes

Primary DNA Sequence Databases

Original submission by experimentalists
Content controlled by the submitter



Primary Protein Sequence Databases



Protein knowledgebase
consists of two sections:

- Swiss-Prot, manually annotated, reviewed.
- TrEMBL, automatically annotated, **not** reviewed.

Derivative Databases

Built from primary data



RefSeq

Submission by
experimentalists
Controlled by the
submitter

**akin to the
primary
research
literature**

non-redundant
richly
annotated
DNA, RNA,
protein
diverse taxa

**akin to the
review
literature**

Derivative Databases

Protein domains, motifs, families



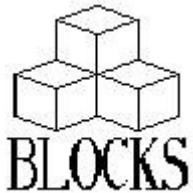
Protein domains/families represented as alignments and HMMs

Derived primarily from **UniprotKB** and **Genpept**



Aligned protein domains and consensus sequences

Derived automatically from **UniprotKB**



Conserved "blocks" of protein domain alignments

Derived from a subset of
UniprotKB



Manually curated models for several hundred protein domains

Derived from proteins from completely
sequenced genomes

Derivative Databases

Protein domains, motifs, families



Protein motifs/domains represented as **Patterns** and/or **HMMs**

Both derived from **UniprotKB/Swissprot**

Patterns are for highly conserved short regions. Example:

R-P-C-x(11)-C-V-S

HMMs are for less conserved longer regions.

Often there will be pattern(s) and an HMM for one domain.

P26367
(PAX6_HUMAN)



(422 aa) **HMM matches**

P26367
(PAX6_HUMAN)



(422 aa) **Pattern matches**

Derivative Databases

Protein domains, motifs,
families

Representations of domains by motif patterns (finger**PRINTS**)



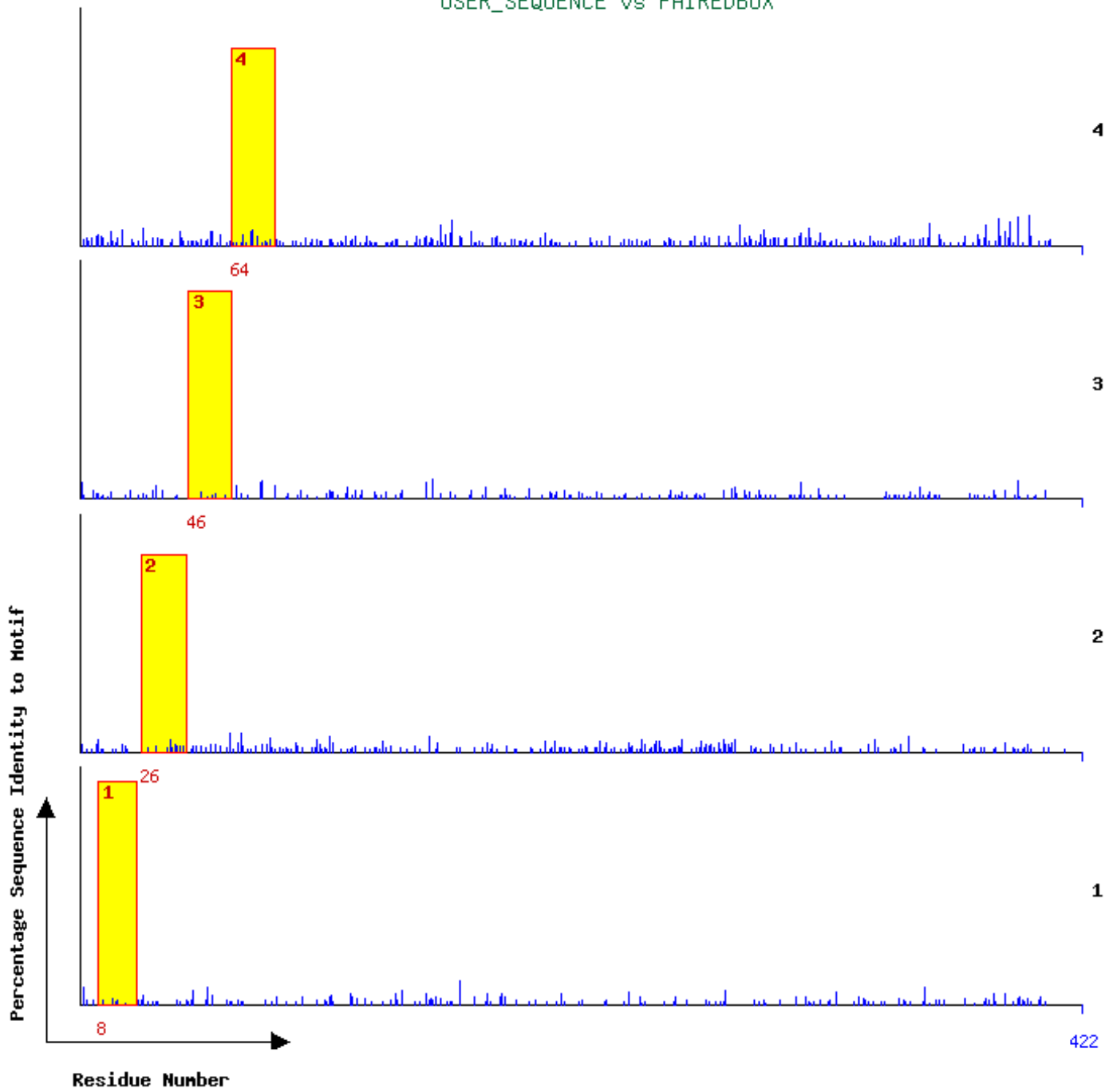
PRINTS

Derived from
UniprotKB

Each FingerPrint is composed of a series of conserved regions (motifs)

A match with a FingerPrint is thus an order set of motif matches

USER_SEQUENCE vs PAIREDBOX



For example:

PAX6_HUMAN
matching the
Paired Box,
4 motif,
Fingerprint

Database Access



Interpro is a consortium of member databases

Interpro defines protein families, domains, regions, repeats and sites according to matches against member databases

Interpro enables any subset of member databases to be searched together

RESULTS YOUR EMAIL

interactive ▾

APPLICATIONS TO RUN Clear all Check all

- | | | | | | |
|---|---|--|---|---|--|
| <input checked="" type="checkbox"/> BlastProDom | <input checked="" type="checkbox"/> FPrintScan | <input checked="" type="checkbox"/> HMMPiR | <input checked="" type="checkbox"/> HMMPfam | <input checked="" type="checkbox"/> HMMSmart | |
| <input checked="" type="checkbox"/> HMMTigr | <input checked="" type="checkbox"/> ProfileScan | <input checked="" type="checkbox"/> HAMAP | <input checked="" type="checkbox"/> patternScan | <input checked="" type="checkbox"/> SuperFamily | <input checked="" type="checkbox"/> SignalPHMM |
| <input checked="" type="checkbox"/> TMHMM | <input checked="" type="checkbox"/> HMMPanther | <input checked="" type="checkbox"/> Gene3D | | | |

Enter or Paste a PROTEIN Sequence in any format:

Help

```
>sp|P26367|PAX6_HUMAN Paired box protein Pax-6;  
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY  
YETGSIRPRAIGGSKPRVATPEVVS KIAQYKRECPSIFAW EIRDRLLESGVCTNDNIPSV  
SSINRVLRLNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ  
EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQE QIEALEKEFERTHYPDV FAR  
ERLAAKIDLPEARIQVWF SNRRAKWRREEKLRNQRRQASNT PSHIPISSSFSTSVYQPI P  
QPTTPVSSFTSGSMLGRDTDALNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCLPT  
SPSVNGRSYDITYTPPHMQTHMNSQPMGTS GTTSTGLISPGVSVVPVQVPGSEPDM SQYWPR  
LQ
```

Upload a file:

Browse...




Submit Job


Reset

SEQUENCE: [PAX6](#) [HUMAN](#) [CRC64](#): C33CDD2C1B13C397 LENGTH: 422 aa  

InterPro
[IPR001356](#)
 Domain






Homeobox


PF00046		Homeobox
SM00389		HOX
PS50071		HOMEBOX_2

InterPro 

InterPro
[IPR001523](#)
 Domain

Paired box protein, N-terminal

PR00027		PAIREDBOX
PF00292		PAX
SM00351		PAX
PS00034		PAIRED_1
PS51057		PAIRED_2

InterPro 

InterPro
[IPR009057](#)
 Domain


Homeodomain-like


SSF46689		Homeodomain-like
--------------------------	--	------------------

InterPro 

InterPro
[IPR011991](#)
 Domain

Winged helix repressor DNA-binding

G3DSA:1.10.10.10		no description
----------------------------------	---	----------------

InterPro 

InterPro
[IPR012287](#)
 Domain


Homeodomain-related


G3DSA:1.10.10.60		no description
----------------------------------	--	----------------

InterPro 

InterPro
[IPR017970](#)
 Conserved_site



Homeobox, conserved site

PS00027		HOMEBOX_1
-------------------------	---	-----------

InterPro 

noIPR
 unintegrated

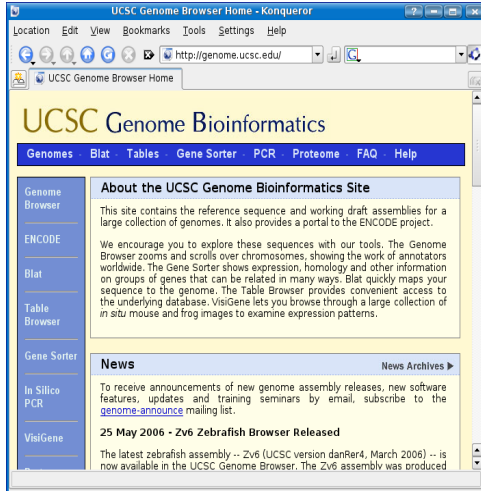
unintegrated

PTHR19418		HOMEBOX PROTEIN
PTHR19418:SF73		PAIRED BOX PROTEIN PAX-6

Genome databases



NCBI Map Viewer



e!Ensembl

EBI / Sanger Institute

Ensembl Species



Alpaca
Vicugna pacos



Anole Lizard
Anolis carolinensis



Armadillo
Dasyus novemcinctus



Bushbaby
Otolemur gamettii



Caenorhabditis elegans



Clona Intestinalis



Clona savignyi



Cat
Felis catus



Chicken
Gallus gallus



Chimpanzee
Pan troglodytes



Cow
Bos taurus



Dog
Canis familiaris



Dolphin
Tursiops truncatus



Elephant
Loxodonta africana



Fruitfly
Drosophila melanogaster



Fugu
Takifugu rubripes



Gorilla
Gorilla gorilla



Guinea Pig
Cavia porcellus



Hedgehog
Erinaceus europaeus



Horse
Equus caballus



Human
Homo sapiens



Hyrax
Procavia capensis



Kangaroo rat
Dipodomys ordii



Lamprey ([preview - assembly only](#))
Petromyzon marinus



Lesser hedgehog tenrec
Echinops telfairi



Macaque
Macaca mulatta



Marmoset
Callithrix jacchus



Medaka
Oryzias latipes



Megabat
Pteropus vampyrus



Microbat
Myotis lucifugus



Mouse
Mus musculus



Mouse Lemur
Microcebus murinus



Opossum
Monodelphis domestica



Orangutan
Pongo pygmaeus



Pig
Sus scrofa



Plka
Ochotona princeps



Platypus
Ornithorhynchus anatinus



Rabbit
Oryctolagus cuniculus



Rat
Rattus norvegicus



Saccharomyces cerevisiae



Shrew
Sorex araneus



Sloth
Choloepus hoffmanni



Squirrel
Spermophilus tridecemlineatus



Stickleback
Gasterosteus aculeatus



Tarsler
Tarsius syrichta



Tetraodon
Tetraodon nigrovindis



Tree Shrew
Tupaia belangeri



Wallaby
Macropus eugenii



Xenopus tropicalis



Zebra Finch
Taeniopygia guttata



Zebrafish
Danio rerio

EnsemblPlants Species



Arabidopsis lyrata

[Gramene](#) | *Arabidopsis lyrata*



Arabidopsis thaliana

[Gramene](#) | *Arabidopsis thaliana*
Columbia



Brachypodium distachyon

[Gramene](#) | *Brachypodium*
distachyon (L.) Beauv



Oryza sativa

[Gramene](#) | *Oryza sativa*
Nipponbare (*Japonica* rice)



Oryza sativa Indica group

[Gramene](#) | *Oryza indica* 93-11
(*Indica* rice)



Populus trichocarpa

[Gramene](#) | *Populus trichocarpa*



Sorghum bicolor

[Gramene](#) | *Sorghum bicolor*
BTX623



Vitis vinifera

[Gramene](#) | *Vitis vinifera* PN40024

NASC e!



The End