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

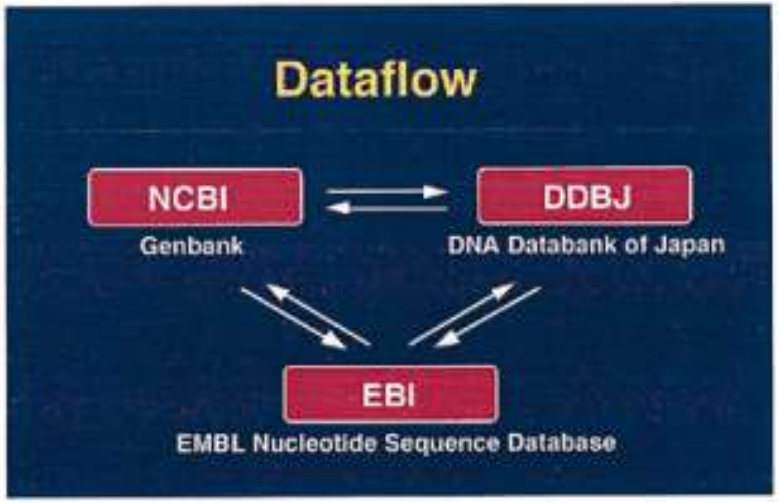
Original submission by experimentalists
Content controlled by the submitter

EMBL NUCLEOTIDE SEQUENCE DATABASE

NCBI GenBank

DDBJ DNA Data Bank of Japan

INSDC International Nucleotide Sequence Database Collaboration



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

NCBI GenBank

RefSeq

Submission by experimentalists
Controlled by the submitter

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

akin to the primary research literature

akin to the review literature

Derivative Databases

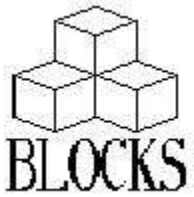
Protein domains, motifs, families

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.

Derivative Databases

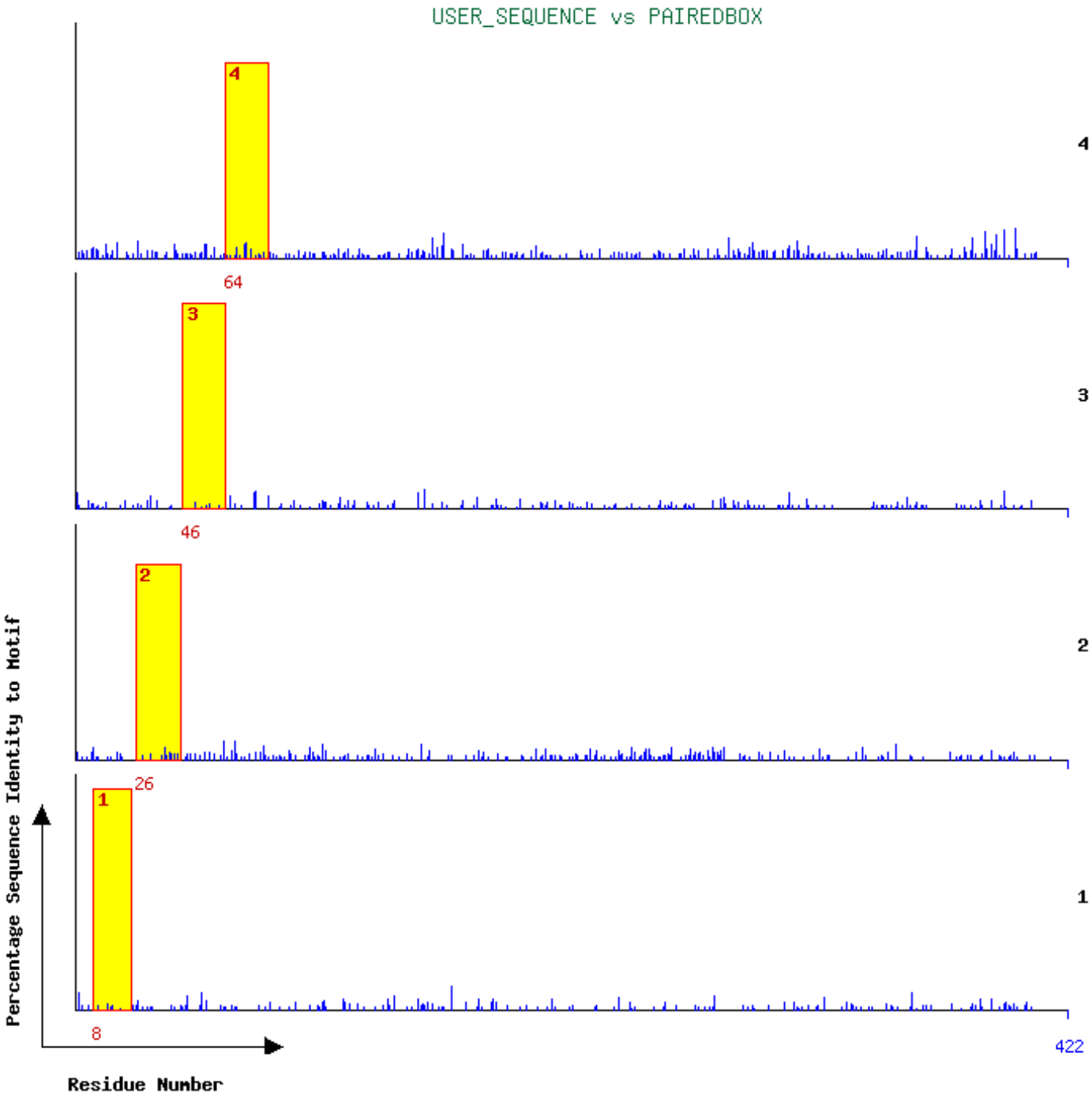
Protein domains, motifs, families

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

Derived from **UniprotKB**

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

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



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
BlastProDom
FPrintScan
HMMPIR
HMMPfam
HMMSmart
HMMTigr
ProfileScan
HAMAP
patternScan
SuperFamily
SignalPHMM
TMHMM
HMMPanther
Gene3D
Enter or Paste a PROTEIN Sequence in any format:
Help
>sp|P26367|PAX6_HUMAN Paired box protein Pax-6;
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY
YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV
SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ
EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR
ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP
QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT
SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
Upload a file:
Browse...
Submit Job
Reset

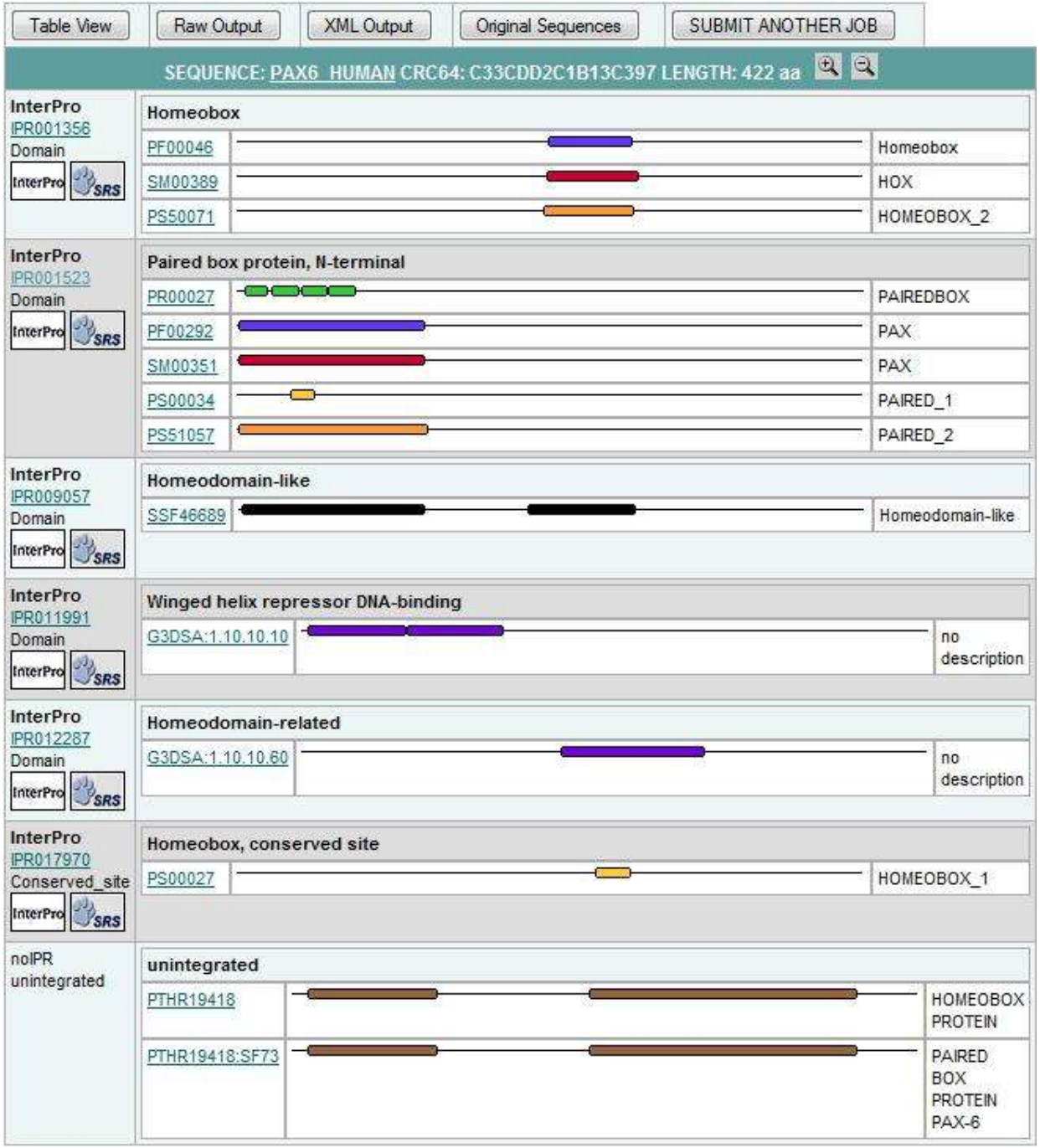
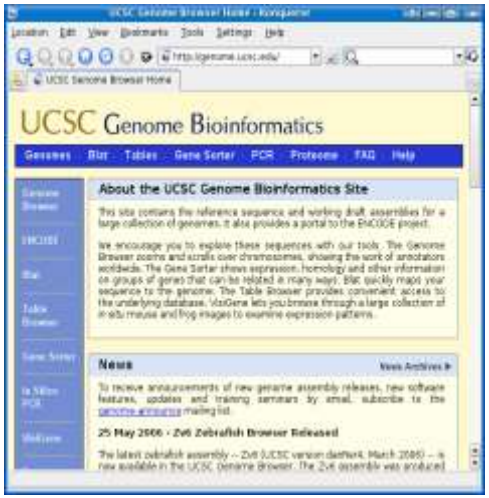


Table View
Raw Output
XML Output
Original Sequences
SUBMIT ANOTHER JOB
SEQUENCE: PAX6_HUMAN CRC64: C33CDD2C1B13C397 LENGTH: 422 aa
InterPro
IPR001356
Domain
Homeobox
PF00046
Homeobox
SM00389
HOX
PS50071
HOMEOBOX_2
InterPro
IPR001523
Domain
Paired box protein, N-terminal
PR00027
PAIREDBOX
PF00292
PAX
SM00351
PAX
PS00034
PAIRED_1
PS51057
PAIRED_2
InterPro
IPR009057
Domain
Homeodomain-like
SSF46689
Homeodomain-like
InterPro
IPR011991
Domain
Winged helix repressor DNA-binding
G3DSA:1.10.10.10
no description
InterPro
IPR012287
Domain
Homeodomain-related
G3DSA:1.10.10.60
no description
InterPro
IPR017970
Conserved_site
Homeobox, conserved site
PS00027
HOMEOBOX_1
noIPR
unintegrated
unintegrated
PTHR19418
HOMEOBOX 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
Dasypus novemcinctus


Bushbaby
Otolemur garnettii


Caenorhabditis elegans


Ciona intestinalis


Ciona 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


Pika
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

Tarsier
Tarsius syrichta

Tetraodon
Tetraodon nigroviridis

Tree Shrew
Tupaia belangeri

Wallaby
Macropus eugenii

Xenopus tropicalis

Zebra Finch
Taeniopygia guttata

Zebrafish
Danio rerio

EnsemblPlants Species

Arabidopsis lyrata
Gramene | *Arabidopsis lyrata*

Oryza sativa
Gramene | *Oryza sativa Nipponbare (Japonica rice)*

Sorghum bicolor
Gramene | *Sorghum bicolor BTX623*

Arabidopsis thaliana
Gramene | *Arabidopsis thaliana Columbia*

Oryza sativa indica group
Gramene | *Oryza indica 93-11 (Indica rice)*

Vitis vinifera
Gramene | *Vitis vinifera PN40024*

Brachypodium distachyon
Gramene | *Brachypodium distachyon (L.) Beauv*

Populus trichocarpa
Gramene | *Populus trichocarpa*

NASC e!

Arabidopsis.info

THE END