Bioinformatics

Genome information resources

Bioinformatics - lectures

- Introduction
- Information networks
- Protein information resources
- Genome information resources
- DNA sequence analysis
- Pairwise sequence alignment
- Multiple sequence alignment
- Secondary database searching
- Analysis packages
- Protein structure modelling

Genome information resources

- primary DNA sequence databases
- specialised DNA sequence databases

Primary DNA sequence databases

- EMBL
- DDBJ
- GenBank
- dbEST
- GSDB

Store DNA sequences and annotations.

Primary protein sequence databases

- EMBL - European Molecular Biology Laboratory
  - European Bioinformatics Institute (EBI)
  - collaboration with DDBJ and GenBank - exchange of new entries on daily basis
  - source of sequences: direct author submissions, genome projects, scientific literature, patents
  - rate of growth is exponential with doubling time ~9-12 months
  - most entries from model organisms
  - retrieval through SRS

Primary protein sequence databases

- DDBJ - DNA Data Bank of Japan
  - National Institute of Genetics
  - collaboration with EMBL and GenBank
  - retrieval through DBGet

- GenBank
  - National Center for Biotechnology Information (NCBI)
  - collaboration with DDBJ and EMBL
  - data split into 17 divisions
  - retrieval through Entrez

Codes for 17 divisions of GenBank

Division	*Sequence subset*
PRI	Primate
ROD	Rodent
MAM	Other mammalian
VRT	Other vertebrate
INV	Invertebrate
PLN	Plant, fungal, algal
BCT	Bacterial
RNA	Structural RNA
VRL	Viral
PHG	Bacteriophage
SYN	Synthetic
UNA	Unannotated
EST	EST (Expressed Sequence Tags)
PAT	Patent
STS	STS (Sequence Tagged Sites)
GSS	GSS (Genome Survey Sequences)
HTG	HTG (High Throughput Genomic Sequences)

```
LOCUS       DRODPPC      4001 bp    mRNA            INV       15-MAR-1990
DEFINITION  D.melanogaster decapentaplegic gene complex (DPP-C), complete cds.
ACCESSION   M30116
NID         g157291
KEYWORDS    .
SOURCE      D.melanogaster, cDNA to mRNA.
  ORGANISM  Drosophila melanogaster
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Arthropoda;
            Tracheata; Insecta; Pterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila.
REFERENCE   1  (bases 1 to 4001)
  AUTHORS   Padgett,R.W., St Johnston,R.D. and Gelbart,W.M.
  TITLE     A transcript from a Drosophila pattern gene predicts a protein
            homologous to the transforming growth factor-beta family
  JOURNAL   Nature 325, 81-84 (1987)
  MEDLINE   87090408
COMMENT     The initiation codon could be at either 1188-1190 or 1587-1589.
FEATURES             Location/Qualifiers
     source          1..4001
                     /organism="Drosophila melanogaster"
                     /db_xref="taxon:7227"
     mRNA            <1..3918
                     /gene="dpp"
                     /note="decapentaplegic protein mRNA"
                     /db_xref="FlyBase:FBgn0000490"
     gene            1..4001
                     /note="decapentaplegic"
                     /gene="dpp"
                     /allele=""
                     /db_xref="FlyBase:FBgn0000490"
     CDS             1188..2954
                     /gene="dpp"
                     /note="decapentaplegic protein (1188 could be 1587)"
                     /codon_start=1
                     /db_xref="FlyBase:FBgn0000490"
                     /db_xref="PID:g157292"
                     /translation="MRAWLLLLAVLATFQTIVRVASTEDISQRFIAAIAPVAAHIPLA
                     SASGSGSGRSGSRSVGASTSTALAKAFNPFSEPASFSDSDKSHRSKTNKKPSKSDANR
                     ..........
                     LGYDAYYCHGKCPFPLADHFNSTNHAVVQTLVNNMNPGKVPKACCVPTQLDSVAMLYL
                     NDQSTVVLKNYQEMTVVGCGCR"
BASE COUNT     1170 a   1078 c    956 g    797 t
ORIGIN
        1 gtcgttcaac agcgctgatc gagtttaaat ctataccgaa atgagcggcg gaaagtgagc
       61 cacttggcgt gaacccaaag ctttcgagga aaattctcgg acccccatat acaaatatcg
      121 gaaaaagtat cgaacagttt cgcgacgcga agcgttaaga tcgccaaaag atctccgtgc
      181 ggaaacaaag aaattgaggc actattaaga gattgttgtt gtgcgcgagt gtgtgtcttc
      241 agctgggtgt gtggaatgtc aactgacggg ttgtaaaggg aaaccctgaa atccgaacgg
      301 ccagccaaag caaataaagc tgtgaatacg aattaagtac aacaaacagt tactgaaaca
      361 gatacagatt cggattcgaa tagagaaaca gatactggag atgcccccag aaacaattca
      421 attgcaaata tagtgcgttg cgcgagtgcc agtggaaaaa tatgtggatt acctgcgaac
      481 cgtccgccca aggagccgcc gggtgacagg tgtatccccc aggataccaa cccgagccca
      541 gaccgagatc cacatccaga tcccgaccgc agggtgccag tgtgtcatgt gccgcggcat
      601 accgaccgca gccacatcta ccgaccaggt gcgcctcgaa tgcggcaaca caattttcaa
          ..........
     3841 aactgtataa acaaaacgta tgccctataa atatatgaat aactatctac atcgttatgc
     3901 gttctaagct aagctcgaat aaatccgtac acgttaatta atctagaatc gtaagaccta
     3961 acgcgtaagc tcagcatgtt ggataaatta atagaaacga g
//
```

Primary protein sequence databases

- dbEST
  - National Center for Biotechnology Information (NCBI)
  - maintains only Expressed Sequence Tag (EST) data
- GSDB - Genome Sequence DataBase
  - National Center for Genome Resourses
  - complete collections of DNA sequence for genome-sequencing laboratories
  - on-line submission of large-scale data
  - quality checks
  - format consistent with GenBank + GSDBID

Specialised DNA sequence databases

- SGD
- UniGene
- TDB
- ACeDB

Store species-specific and technique-specific DNA sequences.

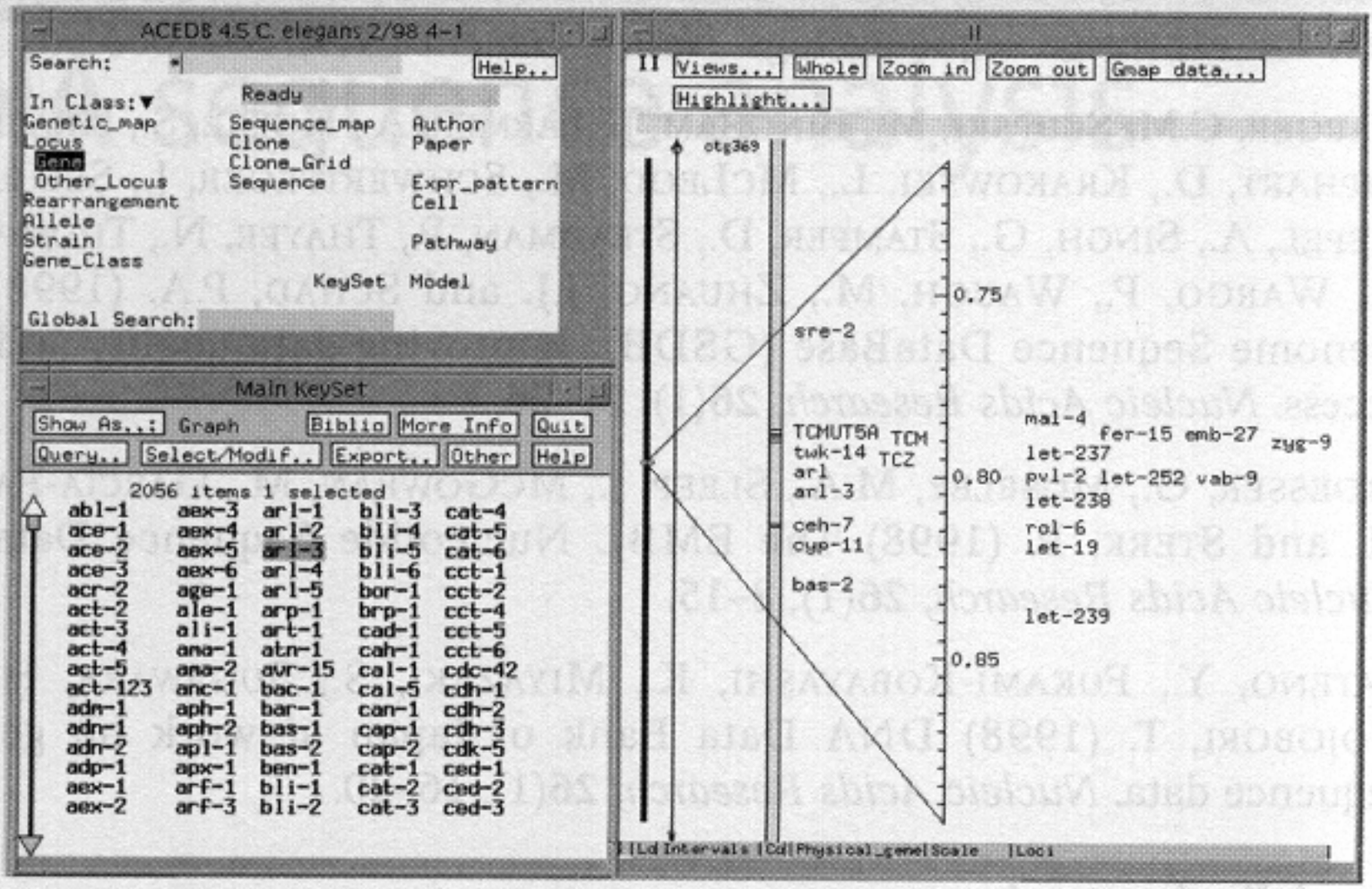
Specialised DNA sequence databases

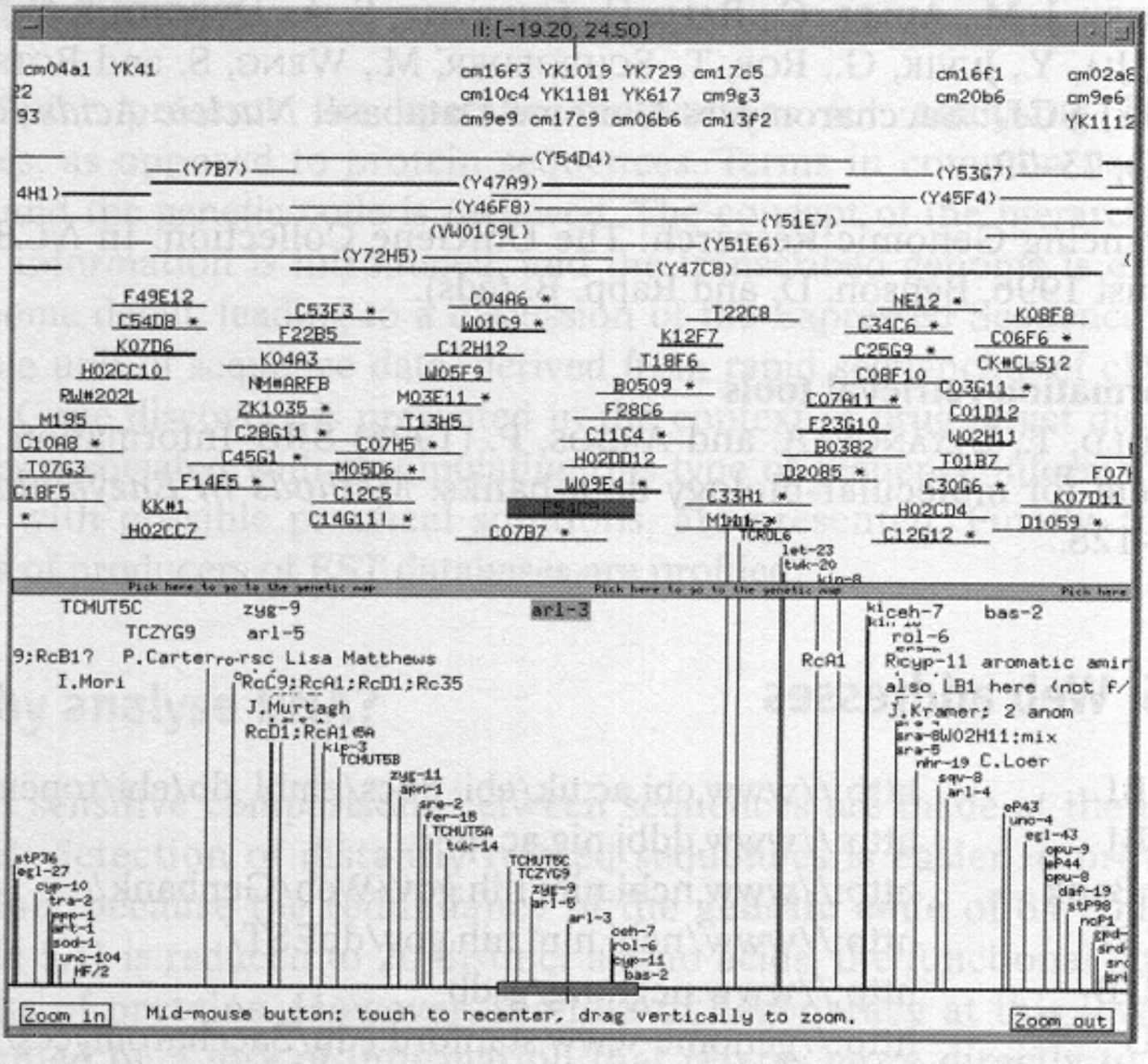
- SGD - *Saccharomyces* Genome Database
  - molecular biology and genetics of *S. cerevisiae*
  - complete genome, genes, proteins, phenotypes
  - first eukaryotic genome sequenced (1998)
  - sequence analysis, register of genes, 3D structural data, primer sequences for cloning
- UniGene
  - collection of genes encoding proteins (transcript map)
  - non-redundant; derived from GenBank
  - data organised in clusters (1 cluster = 1 unique gene)
  - gene-mapping projects and gene expression analysis

Specialised DNA sequence databases

- TDB - TIGR Database
  - suite of databases: DNA and protein sequences, gene expression, protein families, taxonomic data
  - links: TIGR microbial genome sequencing projects, parasite databases, gene index projects, *A. thaliana* database, human genomic dataset

- ACeDB - A *Cernorhabditis elegans* DataBase
  - *C. elegans* genome project
  - restriction maps, gene structural information, cosmid maps, sequence data, bibliographic information
  - software to organise data ACEDB: CGI script and perl


ACEDB 4.5 C. elegans 2/98 4-1
Search:
Help..
Ready
In Class:▼
Genetic_map
Locus
Gene
Other_Locus
Rearrangement
Allele
Strain
Gene_Class
Sequence_map
Clone
Clone_Grid
Sequence
Author
Paper
Expr_pattern
Cell
Pathway
KeySet
Model
Global Search:
Main KeySet
Show As..:
Graph
Biblio
More Info
Quit
Query..
Select/Modif..
Export..
Other
Help
2056 items 1 selected
abl-1 aex-3 arl-1 bli-3 cat-4
ace-1 aex-4 arl-2 bli-4 cat-5
ace-2 aex-5 arl-3 bli-5 cat-6
ace-3 aex-6 arl-4 bli-6 cct-1
acr-2 age-1 arl-5 bor-1 cct-2
act-2 ale-1 arp-1 brp-1 cct-4
act-3 ali-1 art-1 cad-1 cct-5
act-4 ana-1 atn-1 cah-1 cct-6
act-5 ana-2 avr-15 cal-1 cdc-42
act-123 anc-1 bac-1 cal-5 cdh-1
adn-1 aph-1 bar-1 can-1 cdh-2
adn-1 aph-2 bas-1 cap-1 cdh-3
adn-2 apl-1 bas-2 cap-2 cdk-5
adp-1 apx-1 ben-1 cat-1 ced-1
aex-1 arf-1 bli-1 cat-2 ced-2
aex-2 arf-3 bli-2 cat-3 ced-3
II
II
Views...
Whole
Zoom in
Zoom out
Gmap data...
Highlight...
otg369
0.75
0.80
0.85
sre-2
TCMUT5A
twk-14
arl
arl-3
ceh-7
cyp-11
bas-2
TCM
TCZ
mal-4
fer-15 emb-27
zyg-9
let-237
pvl-2 let-252 vab-9
let-238
rol-6
let-19
let-239
Ld Intervals
Cd Physical_gene Scale
Loci


II: [-19.20, 24.50]
cm04a1 YK41
cm16f3 YK1019 YK729 cm17c5
cm10c4 YK1181 YK617 cm9g3
cm9e9 cm17c9 cm06b6 cm13f2
cm16f1
cm20b6
cm9e6
YK1112
(Y54D4)
(Y7B7)
(Y47A9)
(Y53G7)
(Y45F4)
(Y46F8)
(Y51E7)
(VW01C9L)
(Y51E6)
(Y72H5)
(Y47C8)
F49E12
C54D8 *
K07D6
H02CC10
RW#202L
M195
C10A8 *
T07G3
C18F5 *
C53F3 *
F22B5
K04A3
NM#AREB
ZK1035 *
C28G11 *
C45G1 *
F14E5 *
KK#1
H02CC7
C04A6 *
W01C9 *
C12H12
W05F9
M03E11 *
T13H5
C07H5 *
M05D6 *
C12C5
C14G11 *
C07B7 *
T22C8
K12F7
T18F6
B0509 *
F28C6
C11C4 *
H02DD12
W09E4 *
C33H1 *
NE12 *
C34C6 *
C25G9 *
C08F10
C07A11 *
F23G10
B0382
D2085 *
C12G12 *
K08F8
C06F6 *
CK#CLS12
C03G11 *
C01D12
W02H11
T01B7
C30G6 *
H02CD4
D1059 *
K07D11
Pick here to go to the genetic map
TCMUT5C
TCZYG9
zyg-9
arl-5
arl-3
P.Carter
I.Mori
J.Murtagh
RcD1;RcA1
kip-3
TCMUT5B
zyg-11
apn-1
sre-2
fer-15
TCMUT5A
twk-14
TCMUT5C
TCZYG9
zyg-9
arl-5
arl-3
ceh-7
rol-6
cyp-11
bas-2
RcA1
ceh-7
bas-2
rol-6
also LB1 here (not f/
J.Kramer; 2 anom
sra-8W02H11:mix
nhr-19 C.Loer
sqv-8
arl-4
unc-4
egl-43
daf-19
stP36
egl-27
cyp-10
tra-2
ppp-1
art-1
sod-1
unc-104
MF/2
Zoom in
Mid-mouse button: touch to recenter, drag vertically to zoom.
Zoom out