Module 4: Function and Expression (NCBI)

i - How Do I Determine Gene Function?

Aims

- Introduce NCBI resources
- Suggest tools to mine expression data
- Provide examples of expression data in UniGene, BLink, CDD
- Show results of comparisons between UniGene and ZFIN
- Show result of pre-computed protein comparisons: BLink

Introduction

Expression datasets can be viewed and queried directly via GEO (Gene Expression Omnibus) database (http://www.ncbi.nlm.nih.gov/geo/). Additionally, links in Entrez Gene and UniGene provide access to the experimental expression data in ZFIN's Gene Expression database.

You can view a survey of tissue and developmental stages expression levels using UniGene's EST Expression Profile Viewer by following the 'Expression Profile' link from the UniGene cluster page (for example: http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Dr&CID=45278).

Additionally, gene function can be inferred based on the function of related proteins or protein domains. Related CDD domains are identified and reported in Entrez Gene records and as part of the graphical display of pre-computed protein comparisons in the BLink database (see Module 3_ii for additional information on BLink). You can also submit a query against the CDD to identify the domains present in your query sequence or to identify other sequences encoding the same domain.

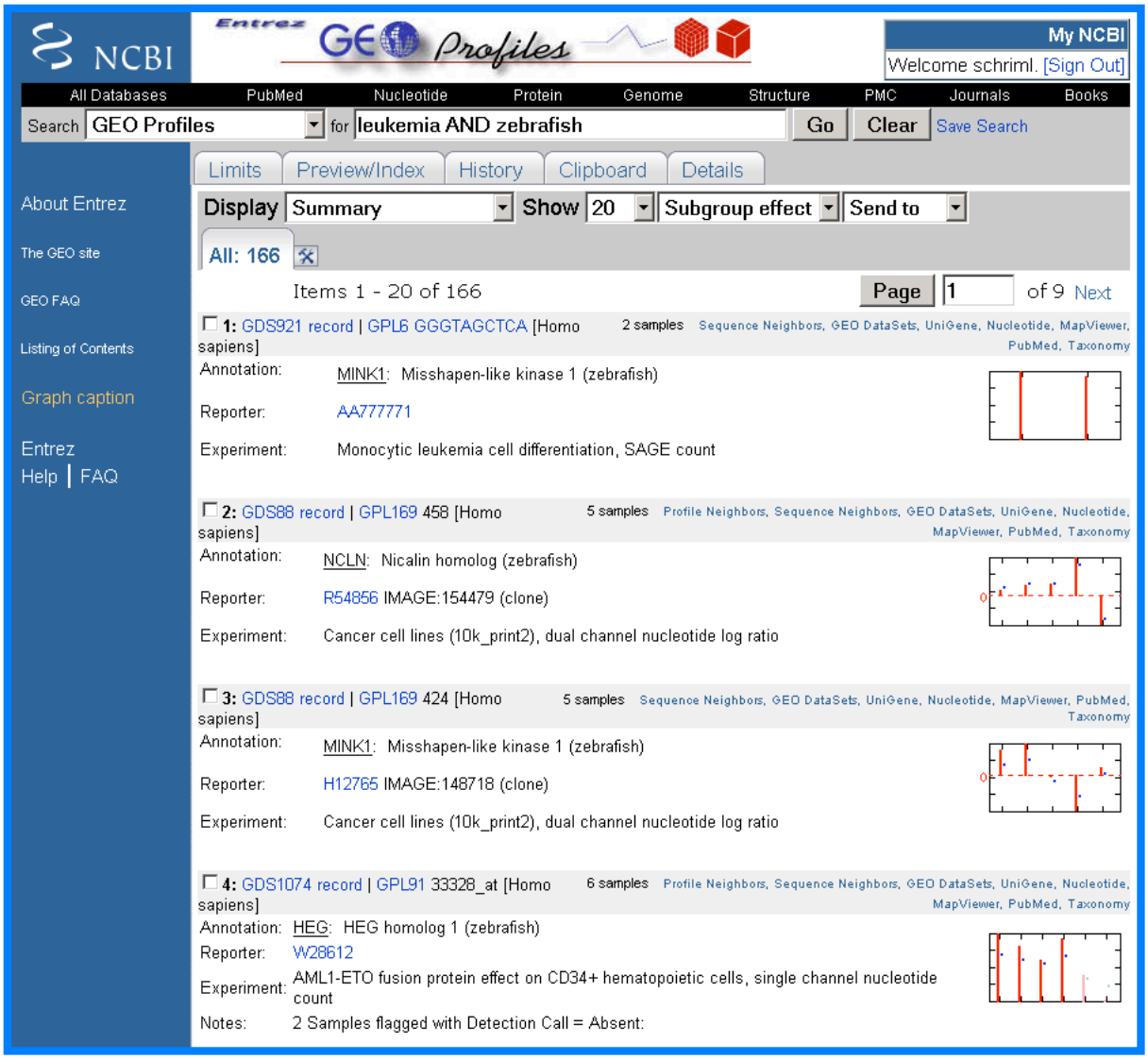
Exercises

1. Query the GEO database for datasets related to leukemia and including the keyword: zebrafish.
2. Identify links in Entrez Gene to ZFIN's Gene Expression database.
3. Identify links in Entrez Gene to the Conserved Domain Database (CDD).
4. Identify links in UniGene to ZFIN's Gene Expression database.
5. Compare expression patterns using UniGene's Expression Profile Viewer
6. Querying CDD to identify domains in your query protein

1. Query the GEO database for datasets related to leukemia and including the keyword: zebrafish.

   Result: view Geo Profiles of expression studies of zebrafish homologs. At this time there is no zebrafish expression data deposited in GEO.

   Explore the GEO datasets and GEO profiles (http://www.ncbi.nlm.nih.gov/projects/geo/)



2. Identify links in Entrez Gene to ZFIN's Gene Expression database (GXD).
3. Identify links in Entrez Gene to the Conserved Domain Database (CDD).

NCBI Entrez Gene

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

Search Gene for Go Clear current records only

Limits Preview/Index History Clipboard Details

Limits: **Danio rerio**

Display Graphics Show 5 Send to

All: 1 Genes Genomes: 0 SNP GeneView: 0

Entrez
SITE MAP
Entrez Help

Gene
About
FAQ
FTP site
Help
NCBI Handbook
Search
Statistics

My NCBI
My NCBI help

Related sites
BLAST
Entrez Genome
Genome Project
Genomic Biology
GEO
HomoloGene
Map Viewer
OMIM
Probe
RefSeq
UniGene
UniSTS

Feedback
Help Desk
Corrections
About GeneRIFs

Subscriptions
RefSeq
Gene
Map Viewer

1: cha charon [*Danio rerio*] Links

GeneID: 406203 Locus tag: ZDB-GENE-040421-2 updated 23-May-2006

Official Symbol: cha **and Name:** charon **provided by** Zebrafish Nomenclature Committee
Gene type: protein coding
Gene name: cha
Gene description: charon
RefSeq status: Provisional
Organism: *Danio rerio*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio*

Bibliography: Gene References into Function (GeneRIF): **Submit** help ?

PubMed links
GeneRIFs:
1. Antagonistic interactions between Charon and Nodal (Southpaw) play an important role in L/R patterning in zebrafish. PubMed

General gene information ?

GeneOntology
Provided by GO

Function	**Evidence**
protein binding	IPI
Process	
determination of left/right symmetry	IGI
determination of left/right symmetry	IMP
embryonic heart tube development	IMP
endoderm formation	IMP
mesoderm formation	IMP
negative regulation of signal transduction	IMP
specification of organ axis polarity	IMP

General protein information ?

Name: charon

NCBI Reference Sequences (RefSeq)

mRNA Sequence NM_212969
Source Sequence AB110416
Product NP_998134 charon
Conserved Domains (1) summary
smart00041: CT; C-terminal cystine knot-like domain (CTCK)
Location: 148 - 208 Blast Score: 92

Related Sequences

Nucleotide	**Protein**
mRNA AB110416	BAD16586

FIN ZDB-GENE-040421-2

CDD link

Scroll down to Exercise 6 to see the CT CDD page

GXD links

4. Identify links in UniGene to ZFIN's Gene Expression database (GXD)
5. Compare expression patterns using UniGene's Expression Profile Viewer

Scroll down to the link to ZFIN's Gene Expression Database and UniGene's Expression Profile Viewer

NCBI UniGene
ORGANIZED VIEW OF THE TRANSCRIP

PubMed Nucleotide Protein Genome Structure Popset
Search UniGene Go
Limits Preview/Index History Clipboard Details

NCBI

UniGene
Homepage
Query Tips
FAQs
DDD
Library Browser
Download UniGene

Related Resources
HomoloGene
dbEST
Trace Archive
BLAST
CGAP
MGC cDNA clones

UniGene Dr.24802 *Danio rerio* hspa8

Heat shock protein 8 (hspa8)

SELECTED PROTEIN SIMILARITIES

Comparison of sequences in UniGene with proteins supported by a complete genome. The alignments can suggest function of a gene.

A. thaliana	ref:NP_187864.1 - heat shock protein 70 [Arabidopsis thaliana]	75.34 % / 647 aa (see ProtEST)
C. elegans	ref:NP_503068.1 - HSP-1 heat shock 70kd protein A [Caenorhabditis elegans]	84.00 % / 649 aa (see ProtEST)
D. melanogaster	ref:NP_524356.1 - Heat shock protein cognate 4; scattered [Drosophila melanogaster]	83.90 % / 649 aa (see ProtEST)
E. coli	sp:P04475 - DNAK_ECOLI Chaperone protein dnaK	50.61 % / 567 aa (see ProtEST)
H. sapiens	sp:P11142 - HS7C_HUMAN Heat shock cognate 71 kDa protein	94.46 % / 649 aa (see ProtEST)
M. musculus	sp:P08109 - HS7C_MOUSE Heat shock cognate 71 kDa protein	94.46 % / 649 aa (see ProtEST)
R. norvegicus	pir:S07197 - S07197 dnaK-type molecular chaperone hsc73 - rat	94.46 % / 649 aa (see ProtEST)
S. cerevisiae	sp:P10591 - HS71_YEAST HEAT SHOCK PROTEIN SSA1	74.42 % / 649 aa (see ProtEST)

Click here to go to UniGene's Expression Profile Viewer

Click here to go to ZFIN to view the Gene Expression data

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

cDNA sources: brain, heart, kidney, olfactory rosettes, regenerated fin, segmentation, embryo 72 hours, adult

Restricted Expression: embryo 72 hours [Show more like this]

Expression Profile: View expression levels using UniGene's EST ProfileViewer

ZFIN: Gene Expression provided by the Zebrafish Information Network

MAPPING POSITION

Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.

Chromosome: LG 10

UniSTS entry: fb01g06.x1
UniSTS entry: fk94d02.x1
UniSTS entry: fc74e11.y1
UniSTS entry: fc02g03.x1
UniSTS entry: hsp70
UniSTS entry: fi48b06.x1
UniSTS entry: fc02g03.x1
UniSTS entry: MARC_6733-6734:992007355:3
UniSTS entry: fb61f01.x1
UniSTS entry: AI883146

SEQUENCES

Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.

mRNA sequences (7)

L77146.1	Danio rerio heat shock cognate (hsc70) mRNA, complete cds	P
Y11413.1	D.rerio hsc70 mRNA	P
BC045841.1	Danio rerio heat shock protein 8, mRNA (cDNA clone MGC:55272 IMAGE:3819770), complete cds	PA
AY422994.1	Danio rerio heat shock 70kDa protein 8 (HSPA8) mRNA, complete cds	P
BC063228.1	Danio rerio heat shock protein 8, mRNA (cDNA clone MGC:77588 IMAGE:6997248), complete cds	A

NCBI » UniGene » EST Profile Viewer

Pubmed Nucleotide Protein Genome Structure Popset Taxonomy

Search UniGene Go Clear

Limits Index History Details

Expression profile suggested by analysis of EST counts.

Dr.24802- hspa8: Heat shock protein 8

See Legend

Breakdown by Tissue

	Dr.24802	
brain	1733	1/577
heart	6163	20/3245
kidney	2557	2/782
olfactory r...	1851	1/540
regenerated...	1889	6/3175

Breakdown by Developmental Stage

	Dr.24802	
gastrula	0	0/265
segmentation	3121	5/1602
embryo 24 h...	0	0/180
pharyngula	0	0/72
embryo 72 h...	7782	20/2570
adult	1056	12/11360

Restricted Expression(contributing more than half of the EST frequency)

Dr.24802: Expression restricted to embryo 72 hours [Show more like this]

Following the Link to UniGene's Expression Profile Viewer

Compare the levels of expression by Tissue or Developmental Stage

Click on "Show more like this" link to view similarly expressed UniGene clusters

LEGEND

Restricted pools are repesented by orange border

Liver	98	13 /131488
Lung	0	0 /282332

Pool name

Transcripts per million(TPM)

Spot intensity based on TPM

Gene EST / Total EST in pool

Here is the top of the list of 12 similarly expressed UniGene clusters

NCBI UniGene
ORGANIZED VIEW OF THE TRANSCRIPTOME
My NCBI
Welcome schriml. [Sign Out]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search UniGene for embryo 72 hours[restricted] AND txid7955[org Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

All: 12 DB-Gene: 6 DB-GEO: 0 DB-HomoloGene: 12

Items 1 - 12 of 12 One page.

NCBI
UniGene
Homepage
Query Tips
FAQ
DDD
Download UniGene
Related Resources
LocusLink
HomoloGene
dbEST
Trace Archive
BLAST
CGAP
MGC cDNA clones
Full Length cDNA Projects

1: Dr.45285 HomoloGene, Nucleotide, Taxonomy, Homologous UniGene, UniSTS
CDNA clone IMAGE:4786889
Danio rerio, 674 sequence(s)

2: Dr.33885 MGC cDNA clone, Gene, HomoloGene, Nucleotide, Taxonomy, Co-expressed UniGene, Homologous UniGene, UniSTS
zgc:92872: Zgc:92872
Danio rerio, 150 sequence(s)

3: Dr.31854 MGC cDNA clone, Gene, HomoloGene, Nucleotide, SNP, Taxonomy, Co-expressed UniGene, Homologous UniGene, UniSTS
rps24: Ribosomal protein S24
Danio rerio, 415 sequence(s)

4: Dr.28230 MGC cDNA clone, HomoloGene, Nucleotide, PubMed, Taxonomy, Co-expressed UniGene, Homologous UniGene
rps7: Ribosomal protein S7
Danio rerio, 197 sequence(s)

5: Dr.24802 MGC cDNA clone, Gene, HomoloGene, Nucleotide, Protein, PubMed, Taxonomy, Co-expressed UniGene, Homologous UniGene, UniSTS
hspa8: Heat shock protein 8
Danio rerio, 1404 sequence(s)

6: Dr.24685 HomoloGene, Nucleotide, Taxonomy, Homologous UniGene, UniSTS
Zgc:111961
Danio rerio, 539 sequence(s)

6. Querying CDD to identify domains in your query protein

View alignment of other proteins containing the CT domain
Click on "CDD" to query the CDD domain

Following the link from Entrez Gene for the CT domain in CDD to view Domain details

NCBI

Conserved Domains

HOME | SEARCH | SITE MAP | Entrez | CDD | Structure | Protein | Help

smart00041.10

CT

Links:

Source: Smart
Taxonomy: Bilateria
PubMed: 8 links
Protein: smart00041 related
architectures
representatives
Related CD: 2 links

C-terminal cystine knot-like domain (CTCK); The structures of transforming growth factor-beta (TGFbeta), nerve growth factor (NGF), platelet-derived growth factor (PDGF) and gonadotropin all form 2 highly twisted antiparallel pairs of beta-strands and contain three disulphide bonds. The domain is non-globular and little is conserved among these presumed homologues except for their cysteine residues. CT domains are predicted to form homodimers.

This domain model appears to be related to other CDs:

pfam00007 ------- smart00041 ------- pfam03045

[mouse over icons to display CD accession/name and number of common hits]

Statistics:

PSSM-Id: 34
Aligned: 24 rows
PSSM: 84 columns
Status: Alignment from source
Created: 12-Dec-2003
Updated: 12-Dec-2003

Show Alignment | Format: Compact Hypertext | Row Display: up to 10 | Color Bits: 2.0 bits | Type Selection: the most diverse members

```
gi 1346296   2945 TSATEHCPDVSACDPA.[1].IVNTTC.[1].QICN.[64].CECCQAAKYSGVSVRLTCEDGTV     RPHR     V 3064
gi 117603     261 KISKPIKFELSGCTSM.[2].YRAKFC      GVCT.[ 1].GRCCTPHRTTTLPVEFKCPDGEV     MKKN     M 317
gi 7019349     96 TQPLKQTIHEEGCNSR.[1].IINRFC.[1].GQCN.[17].CSFCKPKKFTTMMVTLNCPELQP     PTKK.[2].V 170
gi 2459995    164 TVPFNQTIAHEDCQKV.[1].VQNNLC.[1].GKCS.[14].CSHCSPTKFTTVHLRLNCTSPTP     VVKM     V 233
gi 7305361   2827 KVTIRMTIRKNDCRSN.[2].VNLVSC.[1].GRCP.[15].CKCCREVGLQRRSVQLFCATNAT.[1].VPYT     V 2899
gi 585526     311 QGEYDYQNEKTNCSAN     IIMAKC.[1].GQCQ.[15].CRCCKADRVEPRKAHLVCDNGKK     KIYK     Y 380
gi 1346605    473 SSSVNVTVNYNGCKKK     VEMARC.[1].GECK.[15].CLCCQEENYEYREIDLDCPDGGT     IPYR     Y 542
gi 548342      43 HYVDSISHPLYKCSSK.[1].VLLARC.[1].GHCS.[22].CHCCRPQTSKLKALRLRCSGGMR     LTAT     Y 120
gi 1488368    110 VRINTTILWHQGCETE     VNITFC.[1].GSCP.[15].CTCCQERRVHEETVPLHCPNGSA     ILHT     Y 179
gi 33302576 1435 KEQVREYYTENDCRSR.[2].LKYAKC.[1].GGCG      NQCCAAKIVRRRKVRMVCSNNRK     YIKN     L 1491
```

Submit a protein query, choosing among the databases in the pull-down menu

NCBI

Conserved Domains

HOME | SEARCH | SITE MAP | PubMed | Entrez | CDD | Structure | Protein | Taxonomy | BLAST | Help?

Search across Entrez databases GO CLEAR Help

CDD help
NCBI Handbook
CD-Search
CDART
Pfam
SMART
COG

Find CDs
in Entrez:

MMDB
Cn3D
VAST
Research

Last Revised 05/09/05

A Conserved Domain Database and Search Service, v2.03

Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. NCBI's Conserved Domain Database is a collection of multiple sequence alignments for ancient domains and full-length proteins. The CD-Search service may be used to identify the conserved domains present in a protein query sequence:

Submit Query | Reset | Search Database CDD v2.03 - 10991 PSSMs

Enter a **Protein** query as Accession, GI, or Sequence in FASTA format

NP_998134

SMART v4.0 - 663 PSSMs
Pfam v11.0 - 7255 PSSMs
COG v1.00 - 4873 PSSMs
KOG v1.00 - 4825 PSSMs
CDD v2.03 - 10991 PSSMs

Read about the FASTA format description. Click here for advanced options.

Computational biologists define conserved domains based on recurring sequence patterns or motifs. The un-curated section of CDD contains domains imported from SMART, Pfam and COGs. The source databases also provide descriptions and links to citations. Because conserved domains correspond to compact structural units, CDs are linked to 3D structure when possible. The NCBI-curated section of CDD attempts to group ancient domains related by common descent into family hierarchies.

To identify conserved domains in a protein sequence, the CD-Search service uses the reverse position-specific BLAST algorithm. The query sequence is compared to a position-specific score matrix prepared from the underlying conserved domain alignment. Hits may be displayed as a pairwise alignments of the query sequence with representative domain sequences, or as multiple alignments. CD-Search now is run by default in parallel with protein BLAST searches. Although the user waits for the BLAST queue to further process the request, the domain architecture of the query may already be studied.

Run CDART, the Conserved Domain Architecture Retrieval Tool, to search for proteins with similar domain architectures. CDART uses pre-computed CD-Search results to quickly identify proteins with a set of domains similar to that of the query.

Read more about CDD:

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki C, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res. 2005;33 Database Issue:D192-6. [Abstract] [Full Text]

Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 2004;32(Web Server issue):W327-31. [Abstract] [Full Text]

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. 2003;31:383-7. [Abstract] [Full Text][Terms]

Databases