Perl Bioinformatics2 – Algorithms, Database Integration, Pfam, XML

BIOC 8142 Mar 4, 2013
Bill Pearson wrp@virginia.edu 4-2818 Jordan 6-057

Goals of today's lecture:

- Quick overview of alignment algorithms, scoring matrices
- Quick overview of sequence formats
- Integrating information from databases
- Mapping Accessions – RefSeq to Uniprot
- Capturing Pfam Domain Boundaries
  - parsing XML

1

Final Project (home work to be turned in) Due March 17, 2013

Using GSTT1_DROME, PAXI_HUMAN, MYC_HUMAN, or a protein of your own:

- Do a blast search against human RefSeq (`/data/slib/genomes/hum_refseq`)
- For each of the high scoring sequences (`E()<2.0`), report:
  - the description, E-value, start and stop of the alignment
  - the Pfam domains on the target/sbjct protein included in the alignment
  - the Pfam domains on the target protein NOT included in the alignment
  - based on E()-value, reverse BLAST search, and domain composition, identify the highest scoring unrelated sequence

Dynamic programming for sequence alignment

- Sequence alignments can be *global* – end-to-end, or *local*
- The *Dynamic Programming Algorithm* allows one to examine very large numbers of paths in $O(n^2)$ time
- Local alignments can also be used to find duplicated domains in proteins

Algorithms for Biological Sequence Comparison

Algorithm	Value Calculated	Scoring Matrix	Gap penalty	Time required	
Needleman-Wunsch	Global similarity	Any	Penalty / Gap	$O(n^2)$	Needleman and Wunsch, 1970
Sellers	(Global) distance	Unity	Penalty / Residue	$O(n^2)$	Sellers, 1974
Smith-Waterman	Local similarity	$S_{ij} < 0.0$	Affine (q+rk)	$O(n^2)$	Smith and Waterman, 1981 Gotoh, 1982
SRCHN	Approx. local similarity	diagonal	Penalty / Gap	$O(n) – O(n^2)$	Wilbur and Lipman, 1983
FASTP/FASTA	Approx. local similarity	$S_{ij} < 0.0$	Limited size (q+rk)	$O(n^2)/K$	Lipman and Pearson, 1985 Pearson and Lipman, 1988
BLAST	Maximum Segment Score	$S_{ij} < 0.0$	Multiple segments	$O(n^2)/K$	Altschul et al 1990
BLAST2.0	Approx. local similarity	$S_{ij} < 0.0$	(q+rk)	$O(n^2)/K$	Altschul et al 1997

Algorithms for Pairwise Comparison

		Global	Local	Distance
HBHU vs HBHU	Hemoglobin beta-chain - human	725	725	0
HAHU	Hemoglobin alpha-chain - human	314	322	152
MYHU	Myoglobin - Human	121	166	212
GYPL	Leghemoglobin - Yellow lupin	8	48	239
LZCH	Lysozyme - Chicken	-107	32	220
NRBO	Pancreatic ribonuclease	-124	31	280
CCHU	Cytochrome C - human	-160	26	321
MCHU vs MCHU	Calmondulin	671	671	0
TPHUCS	Troponin C	395	438	161
PVPK2	Parvalbumin	-57	115	313
CIHUH	Calpain heavy chain	-2085	100	2463
AQJFNV	Aequorin precursor	-65	76	391
KLSWM	Calcium binding protein	-89	52	323
LDLR vs EGF	EGF precursor	-591	655	2549

Algorithms for Global and Local Similarity Scores

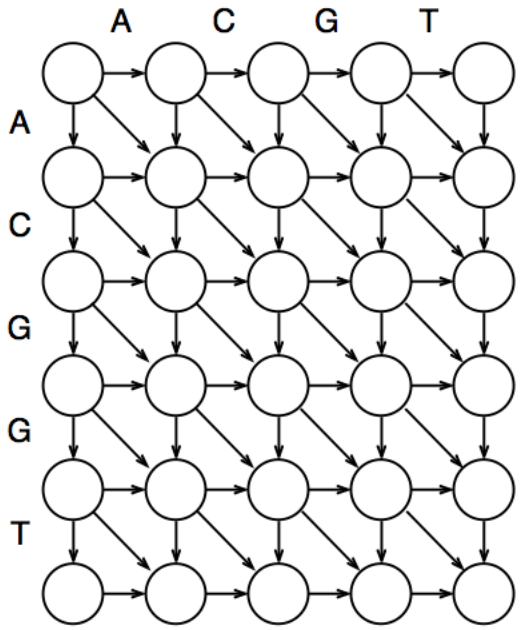
Global:

$S(0,0) \leftarrow 0$
for $j \leftarrow 1$ **to** N **do**
  $S(0,j) \leftarrow S(0,j-1) + \sigma(\ _{b_j})$
for $i \leftarrow 1$ **to** M **do**
[$S(i,0) \leftarrow S(i-1,0) + \sigma(\ ^{a_i}_{\ })$
  for $j \leftarrow 1$ **to** N **do**
    $S(i,j) \leftarrow max[S(i-1,j-1) + \sigma(\ ^{a_i}_{b_j}), S(i-1,j) + \sigma(\ ^{a_i}_{\ }), S(i,j-1) + \sigma(\ _{b_j})]$
]
write "Global similarity score is" $S(M,N)$

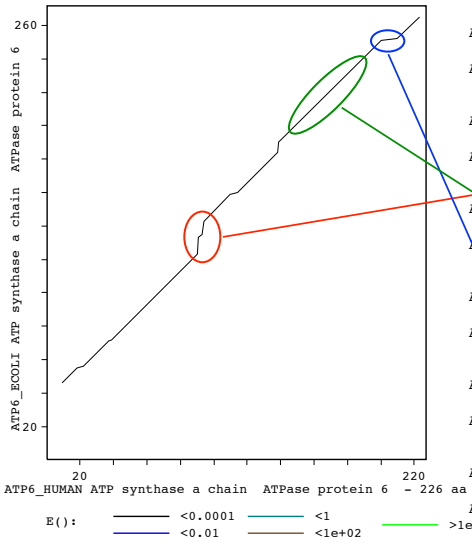
Local:

$best \leftarrow 0$
for $j \leftarrow 1$ **to** N **do**
  $S'(0,j) \leftarrow 0$
for $i \leftarrow 1$ **to** M **do**
[$S'(i,0) \leftarrow 0$
  for $j \leftarrow 1$ **to** N **do**
  [$S'(i,j) \leftarrow max[0, S'(i-1,j-1) + \sigma(\ ^{a_i}_{b_j}), S'(i-1,j) + \sigma(\ ^{a_i}_{\ }), S'(i,j-1) + \sigma(\ _{b_j})]$
    $best \leftarrow max(S'(i,j), best)$
]
write "Local similarity score is" $best$

+1 : match
-1 : mismatch
-2 : gap



alignment paths highlight indels



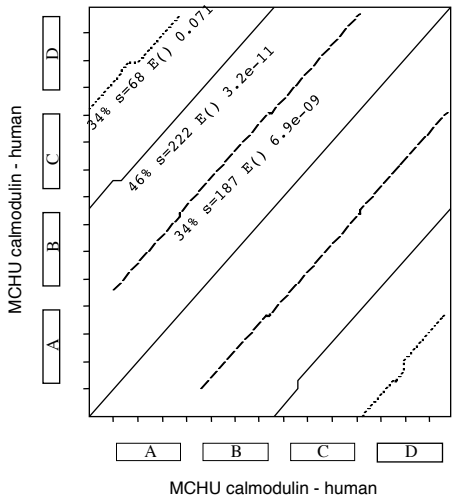

```
>>sp|P0AB98|ATP6_ECOLI ATP synthase (271 aa)
Smith-Waterman score: 178; E(): 2.1e-06
 23.3% identity in 236 aa overlap (8-222:45-264)
                                          10
ATP6_H                          MNENLFASFIAPTILGL
                                         :..  ..::: 
ATP6_E HLNNLQLDLRTFSLVDPQNPPATFWTINIDSMFFSVVLGL
           20        30        40        50
                 20        30        40        50
ATP6_H PAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQM
       ....::  .   ...  . ... :. . .: ... ..
ATP6_E ---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSV
              60        70        80        90
        60        70        80
ATP6_H MTMHNTKGRTWSLMLVSLIIFIATTNLLGLLP--------
        :.. :..  . . .......   ::. :::
ATP6_E KDMYHGKSKLIAPLALTIFVWVFLMNLMDLLPIDLLPYIA
             100       110       120       130
       90        100       110       120
ATP6_H -HSF-------TPTTQLSMNLAMAIPLWAGTVIMGFRSKI
       : .       .:.......:.::. ..   ... : :
ATP6_E EHVLGLPALRVVPSADVNVTLSMALGVF---ILILFYSIK
            140       150       160
          130               140       150
ATP6_H KNALAHFLPQGTPTPL-----IPMLVIIETISLLIQPMAL
       .... :  . :  :       ::. .:.: .::: .:..:
ATP6_E MKGIGGFTKELTLQPFNHWAFIPVNLILEGVSLLSKPVSL
       170       180       190       200
       160       170       180       190
ATP6_H AVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTIL
       . :: .:. ::.:.. ::..     :   :     :: ::
ATP6_E GLRLFGNMYAGELIFILIAGLLPWWSQWILNVPWAIFHIL
       210       220       230       240
        200       210       220
ATP6_H ILLTILEIAVALIQAYVFTLLVSLYLHDNT
       :.         .:::.: .:. .::
ATP6_E IIT---------LQAFIFMVLTIVYLSMASEEH
       250               260       270
```

8

Local alignments - calmodulin

```
>>gi|49037474|sp|P62158.2|CALM_HUMAN Calmodulin;  CaM                          (149 aa)
 Waterman-Eggert score: 220;  49.3 bits; E(1) <  3.2e-11
46.1% identity (73.7% similar) in 76 aa overlap (1-76:77-149)
               10        20        30        40        50        60        70
calm_h MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARK
       : :  .::.:   .::: .:::::.: :.. ::  :: .::.. :. :...:: :.: ::.: ... ::. ::. :
calm_h MKDTDSEEEI---REAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK
         80           90       100       110       120       130       140

 Waterman-Eggert score: 181;  41.5 bits; E(1) <  6.9e-09
34.3% identity (64.8% similar) in 105 aa overlap (11-111:47-147)
               20        30        40        50        60          70        80
calm_h AEFKEAFSLFDKDGDGTITTKELGTVM-RSLGQNPTEAELQDMINEVDADGNGTIDFPEF---LTMMARKMKDTDSEEEI
       ::... ..  : ::.:::   :. :.: :.. .. .: :... .   : :::: :.  :.   .: ...:. : . .: :
calm_h AELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMI
         50        60        70        80        90       100       110       120
         90       100       110
calm_h REAFRVFDKDGNGYISAAELRHVMT
       :::    : ::.: ..  :. ..::
calm_h REA----DIDGDGQVNYEEFVQMMT
            130       140

 Waterman-Eggert score: 64;  18.2 bits; E(1) <  0.071
34.2% identity (71.1% similar) in 38 aa overlap (1-37:113-146)
              10        20        30
calm_h MADQLTEEQIAEF-KEAFSLFDKDGDGTITTKELGTVM
       ....::.:.. :. .::    : :::: .. .:.  .:
calm_h LGEKLTDEEVDEMIREA----DIDGDGQVNYEEFVQMM
            120       130       140
```


Repeated domains with local alignments



Sequence formats (and databases)

- Protein/DNA sequence formats
  - FASTA (FASTQ)
  - Genbank/EMBL
  - formats for database searches
- Multiple sequence alignment formats
  - ClustalW, Stockholm, PHYLIP
- Sequence Databases
  - NCBI/Entrez Primary (Genbank)
  - NCBI/Entrez Derivative (RefSeq, Gene, …)
  - Uniprot (SwissProt)
  - Domain databases (InterPro, Pfam)

FASTA minimal format

description line ➔ `>GT8.7 transl. of pa875.con, 19 to 675`

sequence ➔
```
MPMILGYWNVRGLTHPIRMLLEYTDSSYDEKR
YTMGDAPDFDRSQWLNEKFKLGLDFPNLPYLI
dgshkitqsnailryLARKHHLDGETEEERIR
ADIVENQVMDTRMQLIMLCYNPDFEKQKPEFL
KTIPEKMKLYSEFLGKRPWFAGDKVTYVDFLA
YDILDQYRMFEPKCLDAFPNLRDFLARFEGLK
KISAYMKSSRYIATPIFSKMAHWSNK
```

next sequence➔ `>seq2 description`

`...`

File must be plain text, no special formatting characters. In MS Word, save as text file.

FASTA NCBI format

gi|number db|accession.version description

```
>gi|85725204|ref|NP_001034042.1| glutathione S transferase D1, ...
MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFA
LWESRAIQVYLVEKYGKTDSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPAD
PEAFKKIEAAFEFLNTFLEGQDYAAGDSLTVADIALVATVSTFEVAKFEISKYANVNRWYEN
AKKVTPGWEENWAGCLEFKKYFE
```

gi|number db|acc.ver name description

```
>gi|121694|sp|P20432.1|GSTT1_DROME RecName: Full=Glutathione S-
   transferase 1-1; AltName: Full=GST class-theta
MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFA
LWESRAIQVYLVEKYGKTDSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPAD
PEAFKKIEAAFEFLNTFLEGQDYAAGDSLTVADIALVATVSTFEVAKFEISKYANVNRWYEN
AKKVTPGWEENWAGCLEFKKYFE
```

FASTQ format

sequence start➔ `@SEQ_ID`
sequence data➔ `GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT`
quality start ➔ `+`
quality data ➔ `!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65`

All data are on one line, to avoid confusion of a start of sequence ‘@’ and a quality ‘@’.

Other “flat file” formats – GenBank

```
LOCUS       MUSGST                  1018 bp    mRNA    linear   ROD 12-JUN-1993
DEFINITION  Mouse, glutathione transferase GT8.7 mRNA, complete cds.
ACCESSION   J03952
VERSION     J03952.1  GI:193687
KEYWORDS    glutathione transferase.
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 1018)
  AUTHORS   Pearson,W.R., Reinhart,J., Sisk,S.C., Anderson,K.S. and Adler,P.N.
  TITLE     Tissue-specific induction of murine glutathione transferase mRNAs
            by butylated hydroxyanisole
  JOURNAL   J. Biol. Chem. 263 (26), 13324-13332 (1988)
   PUBMED   3417659
COMMENT     Original source text: Mouse (Female CD-1 (Charles River, outbred),
            7-8 weeks old) liver, cDNA to mRNA, clone pGT875.
            Draft entry and computer-readable sequence for [1] kindly provided
            by W.Pearson, 16-JUN-1988.
```

Other “flat file” formats – GenBank

```
FEATURES             Location/Qualifiers
     source          1..1018
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10090"
     CDS             13..669
                     /note="glutathione transferase (EC 2.5.1.18)"
                     /codon_start=1
                     /protein_id="AAA37747.1"
                     /db_xref="GI:309278"
                     /translation="MPMILGYWNVRGLTHPIRMLLEYTDSSYDEKRYTMGDAPDFDRS
                     QWLNEKFKLGLDFPNLPYLIDGSHKITQSNAILRYLARKHHLDGETEEERIRADIVEN
                     QVMDTRMQLIMLCYNPDFEKQKPEFLKTIPEKMKLYSEFLGKRPWFAGDKVTYVDFLA

    YDILDQYRMFEPKCLDAFPNLRDFLARFEGLKKISAYMKSSRYIATPIFSKMAHWSNK"
ORIGIN      Unreported.
        1 cagcacagca ccatgcctat gatactggga tactggaacg tccgcggact gacacacccg
       61 atccgcatgc tcctggaata cacagactca agctatgatg agaagagata caccatgggt
      121 gacgctcccg actttgacag aagccagtgg ctgaatgaga agttcaagct gggcctggac
...
      901 gacccaatct cacagcccgt tttctgcgaa gtgaggtctg tcctgaacta gtgcttccta
      961 gaattacccc gattggtcac atatcttagt gctagccctc ctagagttac ccgtaaag
//
```

Other “flat file” formats – EMBL

```
ID   J03952; SV 1; linear; mRNA; STD; ROD; 1018 BP.
XX
AC   J03952;
XX
DT   12-JUN-1993
XX
DE   Mouse, glutathione transferase GT8.7 mRNA, complete cds.
XX
KW   glutathione transferase
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Muridae; Murinae; Mus.
XX
RN   [1]
RP   1-1018
RX   PUBMED; 3417659.
RA   Pearson,W.R., Reinhart,J., Sisk,S.C., Anderson,K.S. and Adler,P.N.;
RT   Tissue-specific induction of murine glutathione transferase mRNAs by
RT   butylated hydroxyanisole;
RL   J. Biol. Chem. 263 (26), 13324-13332 (1988)
XX
CC   Original source text: Mouse (Female CD-1 (Charles River, outbred), 7-8
CC   weeks old) liver, cDNA to mRNA, clone pGT875. Draft entry and
CC   computer-readable sequence for [1] kindly provided by W.Pearson,
CC   16-JUN-1988.
XX
```

Other “flat file” formats – EMBL

```
FH   Key             Location/Qualifiers
FH
FT   source          1..1018
FT                   /mol_type="mRNA"
FT                   /db_xref="taxon:10090"
FT                   /organism="Mus musculus"
FT   CDS             13..669
FT                   /db_xref="GI:309278"
FT                   /codon_start=1
FT                   /protein_id="AAA37747.1"
FT                   /translation="MPMILGYWNVRGLTHPIRMLLEYTDSSYDEKRYTMGDAPDFDRSQ
FT                   WLNEKFKLGLDFPNLPYLIDGSHKITQSNAILRYLARKHHLDGETEEERIRADIVENQV
FT                   MDTRMQLIMLCYNPDFEKQKPEFLKTIPEKMKLYSEFLGKRPWFAGDKVTYVDFLAYDI
FT                   LDQYRMFEPKCLDAFPNLRDFLARFEGLKKISAYMKSSRYIATPIFSKMAHWSNK"
FT                   /note="glutathione transferase (EC 2.5.1.18)"
XX
SQ   Sequence 1018 BP; 237 A; 306 C; 241 G; 234 T; 0 other;
     cagcacagca ccatgcctat gatactggga tactggaacg tccgcggact gacacacccg        60
     atccgcatgc tcctggaata cacagactca agctatgatg agaagagata caccatgggt       120
     gacgctcccg actttgacag aagccagtgg ctgaatgaga agttcaagct gggcctggac       180
...
     tccacacagc cttcattgtc cccagtttct ttcacatggc ccctttgca ttggtccctc       900
     gacccaatct cacagcccgt tttctgcgaa gtgaggtctg tcctgaacta gtgcttccta       960
     gaattacccc gattggtcac atatcttagt gctagccctc ctagagttac ccgtaaag        1018
//
```

Other “flat file” formats – EMBL

```
DR   EMBL; X14233; CAA32449.1.
DR   EMBL; S51044; AAA04220.1.
DR   EMBL; M97702; -.
DR   EMBL; AE014297; AAF54786.1.
DR   EMBL; AE014297; ABC66168.1.
DR   EMBL; AY121705; AAM52032.1.
DR   PIR; A34798; XUFF11.
DR   RefSeq; NP_001034042.1; -.
DR   RefSeq; NP_524326.1; -.
DR   UniGene; Dm.2439; -.
DR   PDB; 3EIN; X-ray.
DR   PDBsum; 3EIN; -.
DR   DIP; DIP-17237N; -.
DR   IntAct; P20432; 9.
DR   STRING; P20432; -.
DR   PRIDE; P20432; -.
DR   Ensembl; FBtr0082607; FBpp0082077.
DR   Ensembl; FBtr0100410; FBpp0099824.
DR   GeneID; 41503; -.
DR   KEGG; dme:Dmel_CG10045; -.
DR   CTD; 41503; -.
DR   FlyBase; FBgn0001149; GstD1.
DR   eggNOG; inNOG09765; -.
DR   InParanoid; P20432; -.
DR   OMA; WAGCLEF; -.
DR   OrthoDB; EOG9DJK2V; -.
DR   PhylomeDB; P20432; -.
DR   BioCyc; DMEL-XXX-02:DMEL-XXX-02-011246-
MONOMER; -.
DR   BRENDA; 2.5.1.18; 48.
DR   NextBio; 824191; -.
DR   ArrayExpress; P20432; -.
DR   GermOnline; CG10045; Drosophila
melanogaster.
DR   GO; GO:0004364; F:glutathione
transferase activity.
DR   GO; GO:0005515; F:protein binding.
DR   InterPro; IPR010987; Glutathione-S-
Trfase_C-like.
DR   InterPro; IPR004045; Glutathione_S-
Trfase_N.
DR   InterPro; IPR017933;
Glutathione_S_Trfase/Cl_chnl_C.
DR   InterPro; IPR004046; GST_C.
DR   InterPro; IPR012336; Thioredoxin-
like_fold.
DR   Gene3D; G3DSA:1.20.1050.10; GST_C_like.
DR   Pfam; PF00043; GST_C.
DR   Pfam; PF02798; GST_N.
DR   SUPFAM; SSF47616; GST_C_like.
DR   SUPFAM; SSF52833; Thiordxn-like_fd.
DR   PROSITE; PS50405; GST_CTER.
DR   PROSITE; PS50404; GST_NTER.
XX
CC   -!- FUNCTION: Conjugation of reduced
glutathione to a wide number of
CC   exogenous and endogenous hydrophobic
electrophiles.
```

Multiple Sequence Alignment formats

- CLUSTAL (ClustalW, Muscle, T-Coffee)
- Stockholm (HMMER, Pfam)
- Phylip (PHYLIP)

Use BioPerl to convert between formats

MSA – Clustal W format

```
CLUSTAL W (1.81) multiple sequence alignment


GTM1_HUMAN      CCCATGATACTGGGGTACTGGGACATCCGCGGGCTGGCCCACGCCATCCGCCTGCTCCTG
GTM4_HUMAN      TCCATGACACTGGGGTACTGGGACATCCGCGGGCTGGCCCACGCCATCCGCCTGCTCCTG
GTM5_HUMAN      CCCATGACTCTGGGGTACTGGGACATCCGTGGGCTGGCCCACGCCATCCGCTTGCTCCTG
GTM2_HUMAN      CCCATGACACTGGGGTACTGGAACATCCGCGGGCTGGCCCATTCCATCCGCCTGCTCCTG
GTB3_RAT        CCCATGACACTGGGTTACTGGGACATCCGTGGGCTAGCGCATGCCATCCGCCTGCTCCTG
GTB1_MOUSE      CCTATGATACTGGGATACTGGAACGTCCGCGGACTGACACACCCGATCCGCATGCTCCTG
GTB1_RAT        CCTATGATACTGGGATACTGGAACGTCCGCGGGCTGACACACCCGATCCGCCTGCTCCTG
GTB3_MOUSE      CCTATGACACTGGGCTATTGGAACACCCGCGGACTGACTCACTCCATCCGCTTGCTCCTG
GTB1_CRILO      CCTATGATACTGGGATACTGGAATGTCCGCGGTCTGACAAACCCCATCCGCCTGCTCCTG
GTB2_MOUSE      CCTATGACACTAGGTTACTGGGACATCCGTGGGCTGGCTCACGCCATCCGCCTGCTCCTG
GTB2_RAT        CCTATGACACTGGGTTACTGGGACATCCGTGGGCTGGCTCACGCCATTCGCCTGTTCCTG
GTMU_MESAU      CCTGTGACACTGGGTTACTGGGACATCCGTGGGCTGGCTCATGCCATCCGCCTGCTCTTG
GTM3_HUMAN      TCTATGGTTCTCGGGTACTGGGATATTCGTGGGCTGGCGCACGCCATCCGCCTGCTCCTG
GT2_CHICK       GTGGTCACGTTGGGTTATTGGGACATCCGCGGGTTGGCCCACGCCATCCGCCTGCTGCTG
                    *     * ** ** *** *    ** **  *  *  *  * ** *** ** *  **

GTM1_HUMAN      GAATACACAGACTCAAGCTATGAGGAAAAGAAGTACACGATGGGGGACGCTCCTGATTAT
GTM4_HUMAN      GAATACACAGACTCAAGCTACGAGGAAAAGAAGTATACGATGGGGGACGCTCCTGACTAT
GTM5_HUMAN      GAATACACAGACTCAAGCTATGTGGAAAAGAAGTACACGATGGGGGACGCTCCTGACTAT
GTM2_HUMAN      GAATACACAGACTCAAGCTACGAGGAAAAGAAGTACACGATGGGGGACGCTCCTGATTAT
```

Interleaved format used by MUSCLE, T-COFFEE, PSI-BLAST

MSA – Stockholm format

```
# Comment
  # STOCKHOLM 1.0
  #=GF ID    UPSK
  #=GF SE    Predicted; Infernal
  #=GF SS    Published; PMID:9223489
  #=GF RN    [1]
  #=GF RM    9223489
  #=GF RT    The role of the pseudoknot at the 3' end of turnip yellow mosaic
  #=GF RT    virus RNA in minus-strand synthesis by the viral RNA-dependent RNA
  #=GF RT    polymerase.
  #=GF RA    Deiman BA, Kortlever RM, Pleij CW;
  #=GF RL    J Virol 1997;71:5990-5996.
seq-id/start-stop
  AF035635.1/619-641            UGAGUUCUCGAUCUCUAAAAUCG
  M24804.1/82-104               UGAGUUCUCUAUCUCUAAAAUCG
  J04373.1/6212-6234            UAAGUUCUCGAUCUUUAAAAUCG
  M24803.1/1-23                 UAAGUUCUCGAUCUCUAAAAUCG
  #=GC SS_cons                  .AAA....<<<<aaa....>>>>
  //
```

Sequence databases

- NCBI/Entrez Primary (Genbank)
- NCBI/Entrez Derivative (RefSeq, Gene, …)
- Uniprot (SwissProt)
- Domain databases (InterPro, Pfam)

The Problem – data retrieval and integration

- Given a set of BLAST hits, how do I:
  - find the domains on each protein?
  - find active sites/enzyme activity for the protein (or domain)?
  - find the genes for the protein?

Data retrieval and integration

1. Run a BLAST search
   - what do I need to save?
2. Getting data from NCBI/EBI/Ensembl-BioMart
   - using Web based interfaces, LWP, NCBI efetch
   - using other API's, Web services, BioMart
3. Documenting a search/analysis

Final Project (home work to be turned in)
Due March 17, 2013

Using GSTT1_DROME, PAXI_HUMAN, MYC_HUMAN, or a protein of your own:

- Do a blast search against human RefSeq (`/data/slib/genomes/hum_refseq`)
- For each of the high scoring sequences (`E()<2.0`), report:
  - the description, E-value, start and stop of the alignment
  - the Pfam domains on the target/sbjct protein included in the alignment
  - the Pfam domains on the target protein NOT included in the alignment
  - based on E()-value, reverse BLAST search, and domain composition, identify the highest scoring unrelated sequence

From alignments to domains, what needs to be done?

- How to get from RefSeq to Pfam?
- How to check for alignment/domain overlap
  - sort domains by location
  - sort alignments by location

From alignments to domains, How to get from RefSeq to Pfam?

- Pfam uses Uniprot Id's and Uniprot accession numbers:

```
   |acc  |id
>sp|O43708|MAAI_HUMAN Maleylacetoacetate isomerase GN=GSTZ1 PE=1
SV=3MQAGKPILYSYFRSSCSWRVRIALALKGIDYKTVPINLIKDRGQQFSKDFQALNPMKQVPTLKIDGITIHQSLAI
IEYLEEMRPTPRLLPQDPKKRASVRMISDLIAGGIQPLQNLSVLKQVGEEMQLTWAQNAITCGFNALEQILQSTAGIYC
VGDEVTMADLCLVPQVANAERFKVDLTPYPTISSINKRLLVLEAFQVSHPCRQPDTPTELRA
```

- UniProt provides a utility for mapping from other accession numbers to UniProt accessions/ids

http://www.uniprot.org/faq/28#id_mapping_examples

Mapping to/from UniProt accessions

http://www.uniprot.org/faq/28#id_mapping_examples

Name	Abbreviation		Direction
UniProt			
UniProtKB AC/ID		ACC+ID	from
UniProtKB AC		ACC	to
UniProtKB ID		ID	to
UniParc	UPARC	both	
UniRef50	NF50	both	
UniRef90	NF90	both	
UniRef100	NF100	both	

Name	Abbreviation	Direction
Other sequence databases		
DNA	EMBL_ID	both
DNA CDS	EMBL	both
PIR	PIR	both
UniGene	UNIGENE_ID	both
Entrez Gene	P_ENTREZGENEID	both
GI number*	P_GI	both
IPI	P_IPI	both
RefSeq	P_REFSEQ_AC	both
3D structure databases		
PDB	PDB_ID	both
DisProt	DISPROT_ID	both
HSSP	HSSP_ID	both

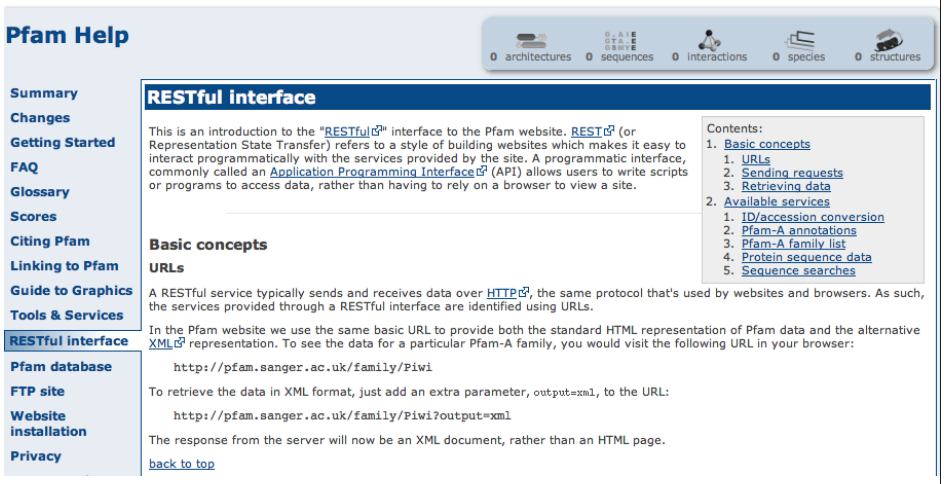
Mapping to/ from UniProt

```
use LWP::UserAgent;

my $base = 'http://www.uniprot.org';
my $tool = 'mapping';

my @refseq_list = ();
while (my $refseq_id = <>) {
  chomp $refseq_id;
  push @refseq_list, $refseq_id;
}

my $params = {
   to => 'ACC', from => 'P_REFSEQ_AC',
   format => 'tab',
   query => join(" ", @refseq_list),
};

my $agent = LWP::UserAgent->new;
push @{$agent->requests_redirectable}, 'POST';

my $response = $agent->post("$base/$tool/", $params);

while (my $wait = $response->header('Retry-After')) {
  print STDERR "Waiting ($wait)...\n";  sleep $wait;
  $response = $agent->get($response->base);
}

if ($response->is_success) {
  print $response->content;
}
else {
  die 'Failed, got ' . $response->status_line .
    ' for ' . $response->request->uri . "\n";
}
```

Mapping to/from UniProt

```
24% uniprot_map.pl ref_seq.list
From    To
NP_001504.2     O43708
Np_001504.2     A6NNB8
NP_714543.1     Q7RTV2
NP_665877.1     O43708
NP_001503.1     O15217
NP_001503.1     Q6P4G1
NP_001395.1     P26641
NP_001395.1     Q53YD7
NP_671488.1     Q8NE79
NP_665683.1     P08263
NP_665683.1     Q5SZC1
NP_004271.1     O43324
NP_000844.2     P30711
NP_000845.1     P30712
NP_000838.3     Q16772
NP_006294.2     Q13155
```

What about Pfam?



What about Pfam?

```
#!/usr/bin/perl -w

use strict;
use LWP::Simple;

for my $acc ( @ARGV ) {

    my $loc="http://pfam.sanger.ac.uk/";
    my $url = "protein/id/$acc";

    my $entry = get $loc . $url;

    print $entry;
}
```

Pfam protein annotation:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
  <head>
   <title>Pfam:
Protein: GSTT1_HUMAN (P30711)
</title>
    <meta name="verify-v1" content="GjV+z51f7mSCShhAOJZh1UW8J+iiCgWmbxIFg2GkG0Q=" />
<meta name="verify-v1" content="FA9AR+bh3BmS05vcSp0mbiAB80DgELEAkFvu4q9ViC8=" />

<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="Description" content="Pfam is a large collection of protein families, represented by
     multiple sequence alignments and hidden Markov models (HMMs)" />


<!-- ========================================================================== -->
<!-- make the site RSS feed discoverable -->

<link href="http://xfam.wordpress.com/tag/pfam/feed/"
      rel="alternate"
      type="application/rss+xml"
      title="Pfam News" />

<!-- ========================================================================== -->
<!-- third-party javascript libraries

     we are now loading third-party libraries from remote sites. We get
     prototype and scriptaculous from googleapis and the YUI components
     for tree viewing directly from yahoo
-->
```

Pfam protein annotation:

Summary

GSTT1_HUMAN

This is the summary of UniProt entry GSTT1_HUMAN (P30711).

Description:	Glutathione S-transferase theta-1 EC=2.5.1.18
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	240 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

Source	Domain	Start	End
Pfam A	GST_N	6	76
Pfam A	GST_C	111	201

What about Pfam?

```
#!/usr/bin/perl -w

use strict;
use LWP::Simple;

for my $acc ( @ARGV ) {

    my $loc="http://pfam.sanger.ac.uk/";
    my $url = "protein/$acc?output=xml";

    my $entry = get($loc . $url);

    print $entry;
}
```

Pfam protein annotation (xml):

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- information on UniProt entry P30711 (GSTT1_HUMAN), generated: 17:38:49 31-Mar-2010 -->
<pfam xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://pfam.sanger.ac.uk/"
      xsi:schemaLocation="http://pfam.sanger.ac.uk/
                          http://pfam.sanger.ac.uk/static/documents/schemas/protein.xsd"
      release="24.0"
      release_date="2009-10-07">
  <entry entry_type="sequence" db="uniprot" db_release="57.6" accession="P30711" id="GSTT1_HUMAN">
    <description>
<![CDATA[
Glutathione S-transferase theta-1 EC=2.5.1.18
]]>
    </description>
    <taxonomy tax_id="9606" species_name="Homo sapiens (Human)">Eukaryota; Metazoa; Chordata;
     Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates;
     Haplorrhini; Catarrhini; Hominidae; Homo.</taxonomy>
    <sequence length="240" md5="a9cdedfd8f1dce1b7d6c106be78cbc73" crc64="BD19F2BFDEF9F619"
     version="4">MGLELYLDLLSQPCRAVYIFAKKNDIPFELRIVDLIKGQHLSDAFAQVNPLKKVPALKDGDFTLTESVAILLYLTRKYKVPDY
     WYPQDLQARARVDEYLAWQHTTLRRSCLRALWHKVMFPVFLGEPVSPQTLAATLAELDVTLQLLEDKFLQNKAFLTGPHISLADLVAITELMHPV
     GAGCQVFEGRPKLATWRQRVEAAVGEDLFQEAHEVILKAKDFPPADPTIKQKLMPWVLAMIR</sequence>
    <matches>
      <match accession="PF02798" id="GST_N" type="Pfam-A">
        <location start="6" end="76" ali_start="17" ali_end="75" hmm_start="15" hmm_end="74"
     evalue="4.2e-08" bitscore="42.20" />
      </match>
      <match accession="PF00043" id="GST_C" type="Pfam-A">
        <location start="111" end="201" ali_start="119" ali_end="200" hmm_start="9" hmm_end="93"
     evalue="0.00019" bitscore="30.30" />
      </match>
    </matches>
  </entry>
</pfam>
```

Dealing with XML – XML::Simple

```
#!/usr/bin/perl -w

use strict;
use LWP::Simple;
use XML::Simple;
use Data::Dumper;

for my $acc ( @ARGV ) {

    my $loc="http://pfam.sanger.ac.uk/";
    my $url = "protein?id=$acc&output=xml";

    my $entry = get $loc . $url;

    my $xml = XML::Simple->new();

    my $xml_data = $xml->XMLin($entry);

    print Dumper($xml_data);
}
```

Dealing with XML – XML::Simple

```
pfam_getprot2.pl gstt1_human
$VAR1 = {
          'xmlns' => 'http://pfam.sanger.ac.uk/',
          'entry' => {
                       'matches' => {
                                      'match' => {
                                                   'GST_C' => {
                                                                'location' => {
                                                                               'ali_end' => '200',
                                                                               'bitscore' => '30.30',
                                                                               'end' => '201',
                                                                               'evalue' => '0.00019',
                                                                               'hmm_end' => '93',
                                                                               'ali_start' => '119',
                                                                               'start' => '111',
                                                                               'hmm_start' => '9'
                                                                             },
                                                                'type' => 'Pfam-A',
                                                                'accession' => 'PF00043'
                                                              },
                                                   'GST_N' => {
                                                                'location' => {
                                                                               'ali_end' => '75',
                                                                               'bitscore' => '42.20',
                                                                               'end' => '76',
                                                                               'evalue' => '4.2e-08',
                                                                               'hmm_end' => '74',
                                                                               'ali_start' => '17',
                                                                               'start' => '6',
                                                                               'hmm_start' => '15'
                                                                             },
                                                                'type' => 'Pfam-A',
                                                                'accession' => 'PF02798'
                                                              }
                                                 }
                                    },
```

Pfam protein annotation:

Summary

GSTT1_HUMAN

This is the summary of UniProt entry GSTT1_HUMAN (P30711).

Description:	Glutathione S-transferase theta-1 EC=2.5.1.18
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	240 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

Source	Domain	Start	End
Pfam A	GST_N	6	76
Pfam A	GST_C	111	201

Dealing with XML – XML::Simple

```
pfam_getprot2.pl
warning: <match> element has non-unique value in 'id' key attribute: HS1_rep at pfam_getprot2.pl
    line 31
...
$VAR1 = {
          'xmlns' => 'http://pfam.sanger.ac.uk/',
          'entry' => {
                       'matches' => {
                                      'match' => {
                                                   'SH3_1' => {
                                                                'location' => {
                                                                               'ali_end' => '543',
                                                                               'bitscore' => '58.30',
                                                                               'end' => '543',
                                                                               'evalue' => '3.7e-13',
                                                                               'hmm_end' => '48',
                                                                               'ali_start' => '498',
                                                                               'start' => '498',
                                                                               'hmm_start' => '1'
                                                                             },
                                                                'type' => 'Pfam-A',
                                                                'accession' => 'PF00018'
                                                              },
                                                   'HS1_rep' => {
                                                                  'location' => {
                                                                                 'ali_end' => '325',
                                                                                 'bitscore' => '31.30',
                                                                                 'end' => '333',
                                                                                 'evalue' => '0.00014',
                                                                                 'hmm_end' => '21',
                                                                                 'ali_start' => '305',
                                                                                 'start' => '305',
                                                                                 'hmm_start' => '1'
                                                                               },
                                                                  'type' => 'Pfam-A',
                                                                  'accession' => 'PF02218'
                                                                },
```

SRC8_HUMAN

This is the summary of UniProt entry SRC8_HUMAN (Q14247).

Description:	Src substrate cortactin
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	550 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

Source	Domain	Start	End
Pfam A	HS1_rep	83	119
Pfam A	HS1_rep	120	156
Pfam A	HS1_rep	157	193
Pfam A	HS1_rep	194	230
Pfam A	HS1_rep	231	267
Pfam A	HS1_rep	268	304
Pfam A	HS1_rep	305	333
coiled_coil	n/a	352	401
low_complexity	n/a	358	398
	n/a	403	484
Pfam A	SH3_1	498	543

Dealing with XML – XML::Twig

- all we want to do is find:

```
<matches>
<match accession="PF02798" id="GST_N" type="Pfam-A">
   <location start="6" end="76" ali_start="17" ali_end="75"
   hmm_start="15" hmm_end="74" evalue="4.2e-08" bitscore="42.20" />
   </match>
   <match accession="PF00043" id="GST_C" type="Pfam-A">
   <location start="111" end="201" ali_start="119" ali_end="200"
   hmm_start="9" hmm_end="93" evalue="0.00019" bitscore="30.30" />
   </match>
</matches>
```

```
 <location start="6" end="76" ali_start="17" ali_end="75"
hmm_start="15" hmm_end="74" evalue="4.2e-08" bitscore="42.20" />
<location /> -> tag (with end);  attribute1="this" attr2="that"
```

Dealing with XML – XML::Twig

```
#!/usr/bin/perl -w
# pfam_getprot3.pl GSTT1_DROME

use strict;
use LWP::Simple;
use XML::Twig;

my $loc="http://pfam.sanger.ac.uk/";
my $url;
my @match_list;

for my $acc ( @ARGV ) {
  @match_list = ();
  if ($acc =~ m/_/) {$url = "protein?id=$acc&output=xml"; }
  else {$url = "protein/$acc?output=xml"; }
  my $res = get($loc . $url);
  my $twig = XML::Twig->new(twig_roots => {matches => 1},
                        twig_handlers => {match => \&push_match},
                        pretty_print => 'indented');
  my $xml = $twig->parse($res);
  @match_list = sort { $a->{start} <=> $b->{start} } @match_list;
  for my $match ( @match_list ) {
    print join("\t",($match->{id},$match->{start},$match->{end})),"\n";
  }
}

sub push_match {
    my ($t, $elt) = @_;
    my $attr_ref =  $elt->{att};
    my $loc_ref = $elt->first_child('location')->{att};
    push @match_list, { %$attr_ref, %$loc_ref };
}
```

Dealing with XML – XML::Twig

```
 pfam_getprot1.pl GSTT1_HUMAN
$VAR1 = [
          {
            'ali_end' => '82',
            'bitscore' => '74.10',
            'end' => '82',
            'accession' => 'PF02798',
            'hmm_end' => '76',
            'evalue' => '7.8e-18',
            'type' => 'Pfam-A',
            'id' => 'GST_N',
            'ali_start' => '4',
            'hmm_start' => '2',
            'start' => 3
          },
          {
            'ali_end' => '189',
            'bitscore' => '58.80',
            'end' => '192',
            'accession' => 'PF00043',
            'hmm_end' => '92',
            'evalue' => '4.1e-13',
            'type' => 'Pfam-A',
            'id' => 'GST_C',
            'ali_start' => '105',
            'hmm_start' => '2',
            'start' => 104
          }
        ];
```

Dealing with XML – XML::Twig

```
# ~wrp/new_data/pfam_getprot3.pl uses XML::Twig

pfam_getprot3.pl src8_human
HS1_rep  83      119
HS1_rep  120     156
HS1_rep  157     193
HS1_rep  194     230
HS1_rep  231     267
HS1_rep  268     304
HS1_rep  305     333
Pfam-B_30333     403     484
SH3_1    498     543
```

A simpler strategy for Pfam domain coordinates

```
#!/usr/bin/perl -w

# pfam_getprot0.pl ACC
# only works for accessions P09488

use strict;
use LWP::Simple;

for my $acc ( @ARGV ) {

  my $url = "http://pfam.sanger.ac.uk/protein/id/$acc";

  print get($url);
}
```

```
% perl pfam_getprot0.pl P30711
P  P30711        GSTT1_HUMAN      Glutathione S-transferase theta-1 ...
A  GST_N   PF02798   6    76
A  GST_C   PF00043   111  201
```

Uniprot Database CrossRefs

```
DR   EMBL; X14233; CAA32449.1.
DR   EMBL; S51044; AAA04220.1.
DR   EMBL; M97702; -.
DR   EMBL; AE014297; AAF54786.1.
DR   EMBL; AE014297; ABC66168.1.
DR   EMBL; AY121705; AAM52032.1.
DR   PIR; A34798; XUFF11.
DR   RefSeq; NP_001034042.1; -.
DR   RefSeq; NP_524326.1; -.
DR   UniGene; Dm.2439; -.
DR   PDB; 3EIN; X-ray.
DR   PDBsum; 3EIN; -.
DR   DIP; DIP-17237N; -.
DR   IntAct; P20432; 9.
DR   STRING; P20432; -.
DR   PRIDE; P20432; -.
DR   Ensembl; FBtr0082607; FBpp0082077.
DR   Ensembl; FBtr0100410; FBpp0099824.
DR   GeneID; 41503; -.
DR   KEGG; dme:Dmel_CG10045; -.
DR   CTD; 41503; -.
DR   FlyBase; FBgn0001149; GstD1.
DR   eggNOG; inNOG09765; -.
DR   InParanoid; P20432; -.
DR   OMA; WAGCLEF; -.
DR   OrthoDB; EOG9DJK2V; -.
DR   PhylomeDB; P20432; -.
DR   BioCyc; DMEL-XXX-02:DMEL-XXX-02-011246-
MONOMER; -.
DR   BRENDA; 2.5.1.18; 48.
DR   NextBio; 824191; -.
DR   ArrayExpress; P20432; -.
DR   GermOnline; CG10045; Drosophila
melanogaster.
DR   GO; GO:0004364; F:glutathione
transferase activity.
DR   GO; GO:0005515; F:protein binding.
DR   InterPro; IPR010987; Glutathione-S-
Trfase_C-like.
DR   InterPro; IPR004045; Glutathione_S-
Trfase_N.
DR   InterPro; IPR017933;
Glutathione_S_Trfase/Cl_chnl_C.
DR   InterPro; IPR004046; GST_C.
DR   InterPro; IPR012336; Thioredoxin-
like_fold.
DR   Gene3D; G3DSA:1.20.1050.10; GST_C_like.
DR   Pfam; PF00043; GST_C.
DR   Pfam; PF02798; GST_N.
DR   SUPFAM; SSF47616; GST_C_like.
DR   SUPFAM; SSF52833; Thiordxn-like_fd.
DR   PROSITE; PS50405; GST_CTER.
DR   PROSITE; PS50404; GST_NTER.
XX
CC   -!- FUNCTION: Conjugation of reduced
glutathione to a wide number of
CC   exogenous and endogenous hydrophobic
electrophiles.
```

Uniprot Database Cross References

www.uniprot.org/uniprot/P09488.xml

```
<entry dataset="Swiss-Prot" created="1989-07-01" modified="2012-11-28" version="148">
<accession>P09488</accession>
<name>GSTM1_HUMAN</name>
<protein>...</protein>
<gene>...</gene>
<organism>...</organism>
<reference key="2">
  <citation type="journal article" date="1988" name="Proc. Natl. Acad. Sci. U.S.A."
volume="85" first="7293" last="7297">
    <title>Hereditary differences in the expression ... </title>
    <authorList>
      <person name="Seidegaard J."/>
      <person name="Vorachek W.R."/>
      <person name="Pero R.W."/>
      <person name="Pearson W.R."/>
    </authorList>
    <dbReference type="MEDLINE" id="89017184"/>
    <dbReference type="PubMed" id="3174634"/>
    <dbReference type="DOI" id="10.1073/pnas.85.19.7293"/>
  </citation>
  <scope>NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1)</scope>
</reference>
<comment type="function">
  <text evidence="1">
    Conjugation of reduced glutathione ...
  </text>
</comment>
```

Uniprot Database Cross References

www.uniprot.org/uniprot/P09488.xml

```
<comment type="catalytic activity">...</comment>
<comment type="subunit">...</comment>
<dbReference type="EC" id="2.5.1.18"/>
<dbReference type="RefSeq" id="NP_000552.2">
  <property type="nucleotide sequence ID" value="NM_000561.3"/>
</dbReference>
<dbReference type="PDB" id="1XW6">
  <property type="method" value="X-ray"/>
  <property type="resolution" value="1.90"/>
  <property type="chains" value="A/B/C/D=2-217"/>
</dbReference>
<dbReference type="GO" id="GO:0005829">
  <property type="term" value="C:cytosol"/>
  <property type="evidence" value="TAS:Reactome"/>
</dbReference>
<evidence key="1" type="ECO:0000006">
  <source> <dbReference type="PubMed" id="16548513"/> </source>
</evidence>
<sequence length="218" mass="25712" checksum="98FB03E87B83A31B" modified="2007-01-23"
version="3">
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNL
PYLIDGAHKITQSNAILCYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEF
EKLKPKYLEELPEKLKLYSEFLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCLDAFPN
LKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK
</sequence>
</entry>
```

Uniprot Database Cross References

www.uniprot.org/uniprot/P09488.xml

```
my $acc = $ARGV[0];
my $loc="http://www.uniprot.org/uniprot/$acc.xml";
my @link_list = ();
my $result = get($loc);

my $twig = XML::Twig->new(
                                twig_roots => { "entry/dbReference" => 1},
                                twig_handlers => {dbReference => \&push_link},
                                pretty_print => 'indented'
                               );
my $xml = $twig->parse($result);

@link_list = grep {$_->{type} =~ m/RefSeq|PDB|PRIDE|InterPro|Pfam|GO/} @link_list;

my @fields = qw(type id);
print "#",join("\t",@fields),"\n";
for my $link ( @link_list ) {
  print join("\t",(@{$link}{@fields})),"\n";
}

sub push_link {
  my ($t, $elt) = @_;
  push @link_list, { %{$elt->{att}} };
}
```

Uniprot Database Cross References

www.uniprot.org/uniprot/P09488.xml

```
perl get_uniprot.pl P09488
#type    id
RefSeq   NP_000552.2
RefSeq   NP_666533.1
PDB      1GTU
PDB      1XW6
PDB      1XWK
PDB      1YJ6
PDB      2F3M
PDBsum   1GTU
PDBsum   1XW6
PDBsum   1XWK
PDBsum   1YJ6
PDBsum   2F3M
PRIDE    P09488
GO       GO:0005829
GO       GO:0004364
GO       GO:0044281
GO       GO:0006805
InterPro IPR010987
InterPro IPR004045
InterPro IPR017933
InterPro IPR004046
InterPro IPR003081
InterPro IPR012336
Pfam     PF00043
Pfam     PF02798
```

More information from the NCBI – eutils/elinks

www.ncbi.nlm.nih.gov/books/NBK1058/

- ESearch: responds to a text query with the list of UIDs matching the query in a given database, along with the term translations of the query.
- ESummary: responds to a list of UIDs with the corresponding document summaries.
- EFetch: responds to a list of UIDs with the corresponding data records. [reference documentation]
- ELink: responds to a list of UIDs in a given database with either a list of related IDs in the same database or a list of linked IDs in another Entrez database.
- EGQuery: responds to a text query with the number of records matching the query in each Entrez database.

Can't BioPerl
LWP – your script as a browser

```
use LWP::Simple;

use strict;
use LWP::Simple;

for my $n ( @ARGV ) {
    my $gid = $n;

    my $loc="http://eutils.ncbi.nih.gov/entrez/eutils/";
    my $url = "efetch.fcgi?db=Protein&rettype=fasta&id=$gid";

    my $entry = get $loc . $url;

    print $entry;
}
```

```
wrpmbp2 44% efetch1.pl 121694
>gi|121694|sp|P20432.1|GSTT1_DROME Glutathione S-transferase 1-1; GST class-theta
MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFALWESRAIQ
VYLVEKYGKTDSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPADPEAFKKIEAAFEFLNT
FLEGQDYAAGDSLTVADIALVATVSTFEVAKFEISKYANVNRWYENAKKVTPGWEENWAGCLEFKKYFE
```

NCBI Eutils – efetch.fcgi

```
#!/usr/bin/perl -w

use strict;
use LWP::Simple;

for my $n ( @ARGV ) {
    my $gid = $n;

    my $loc="http://eutils.ncbi.nih.gov/entrez/eutils/";
    my $url = "efetch.fcgi?db=Protein&id=$gid&rettype=fasta";
    my $entry = get $loc . $url;

    print $entry;
}
```

When constructing URLs for the eUtils, please use lowercase characters for all parameters except &WebEnv. There is no required order for the URL parameters in an eUtils URL, and null values or inappropriate parameters are ignored. Avoid placing spaces in the URLs, particularly in queries. If a space is required, use a plus sign (+) instead of a space:

- Incorrect: &id=352, 25125, 234, ...
- Correct: &id=352,25125,234,...
- Incorrect: &term=biomol mrna[properties] AND mouse[organism]
- Correct: &term=biomol+mrna[properties]+AND+mouse[organism]

Other special characters, such as the # symbol used in referring to a query key on the History server, should be represented by their URL encodings (%23 for #).

NCBI Eutils – einfo.fcgi/elink.fcgi – XML

```
curl 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=protein'
```

```
<?xml version="1.0"?>
<!DOCTYPE eInfoResult PUBLIC "-//NLM//DTD eInfoResult, 11 May 2002//EN" "http://
    www.ncbi.nlm.nih.gov/entrez/query/DTD/eInfo_020511.dtd">
<eInfoResult>
    <DbInfo>
    <DbName>protein</DbName>
...
    <FieldList>
        <Field>
          <Name>ALL</Name>
          <FullName>All Fields</FullName>
...
        </Field>
...
    </FieldList>
    <LinkList>
        <Link>
          <Name>protein_bioproject</Name>
          <Menu>BioProject Links</Menu>
          <Description>Proteins related to BioProjects</Description>
          <DbTo>bioproject</DbTo>
        </Link>
```

NCBI Eutils – einfo.fcgi/elink.fcgi – XML

curl 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=protein'

```
        <Link>
          <Name>protein_biosystems</Name>
          <Menu>BioSystem Links</Menu>
          <Description>Pathways and other biosystems... .</Description>
          <DbTo>biosystems</DbTo>
        </Link>
        <Link>
          <Name>protein_ccds</Name>
          <Menu>ccds</Menu>
          <Description>Link to Consensus CDS</Description>
          <DbTo>ccds</DbTo>
        </Link>
        <Link>
          <Name>protein_cdd</Name>
          <Menu>Conserved Domain Links</Menu>
          <Description>Full list ... </Description>
          <DbTo>cdd</DbTo>
        </Link>
        <Link>
          <Name>protein_omim</Name>
          <Menu>OMIM Links</Menu>
          <Description>Online Mendelian Inheritance in Man (OMIM) records </Description>
          <DbTo>omim</DbTo>
        </Link>
    </LinkList>
    </DbInfo>
</eInfoResult>
```

NCBI Eutils – einfo.fcgi/elink.fcgi – XML

curl 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=protein'

```
#!/usr/bin/perl -w
use strict;
use LWP::Simple qw(get);
use XML::Twig;

my $loc="http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=protein";

my @link_list = ();

my $result = get($loc);

my $twig = XML::Twig->new(
                      twig_roots => {LinkList => 1},
                      twig_handlers => {Link => \&push_link}
                     );
my $xml = $twig->parse($result);

my @fields = qw(DbTo Name Menu);
print "#",join("\t",@fields),"\n";
for my $link ( @link_list ) {
  print join("\t",(@{$link}{@fields})),"\n";
}

sub push_link {
  my ($t, $elt) = @_;
  my $link_info = {};
  for my $tag ( qw(Name Description DbTo Menu) ) {
    $link_info->{$tag} = $elt->first_child($tag)->text;
  }
  push @link_list, $link_info;
}
```

NCBI Eutils – einfo.fcgi/elink.fcgi – XML

curl 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=protein'

```
#DbTo     Name      Menu
bioproject          protein_bioproject  BioProject Links
biosystems          protein_biosystems  BioSystem Links
ccds      protein_ccds        ccds
cdd protein_cdd     Conserved Domain Links
cdd protein_cdd_concise_2     Concise Conserved Domain Links
gene      protein_gene        Gene Links
homologene          protein_homologene  HomoloGene Links
nuccore   protein_nuccore     Nucleotide Links
omim      protein_omim        OMIM Links
pmc protein_pmc     PMC Links
popset    protein_popset      PopSet Links
protein   protein_protein     Related Sequences
protein   protein_protein_cdart_summary
protein   protein_protein_identical     Identical Proteins
protein   protein_protein_refseq2uniprot          Protein (UniProtKB)
protein   protein_protein_uniprot2refseq          Protein (RefSeq)
proteinclusters     protein_proteinclusters       Protein Cluster Links
pubmed    protein_pubmed      PubMed Links
pubmed    protein_pubmed_refseq         PubMed (RefSeq) Links
pubmed    protein_pubmed_weighted       PubMed (Weighted) Links
snp protein_snp     SNP Links
snp protein_snp_genegenotype  Gene Genotype Links
snp protein_snp_geneview
structure protein_structure   Structure Links
structure protein_structure_related
structure protein_structure_related_list          Related Structures (List)
taxonomy  protein_taxonomy    Taxonomy Links
unigene   protein_unigene     UniGene Links
```

NCBI Eutils – elink.fcgi – XML

www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.Elink

```
curl 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?
    dbfrom=protein&id=121735&&db=protein&linkname=protein_protein_uniprot2refseq'

<?xml version="1.0"?>
<!DOCTYPE eLinkResult PUBLIC "-//NLM//DTD eLinkResult, 23 November 2010//EN" "http://
    www.ncbi.nlm.nih.gov/entrez/query/DTD/eLink_101123.dtd">
<eLinkResult>

    <LinkSet>
          <DbFrom>protein</DbFrom>
          <IdList>
                <Id>121735</Id>
          </IdList>
          <LinkSetDb>
                <DbTo>protein</DbTo>
                <LinkName>protein_protein_uniprot2refseq</LinkName>
                <Link>
                      <Id>23065547</Id>
                </Link>
                <Link>
                      <Id>23065544</Id>
                </Link>
          </LinkSetDb>
    </LinkSet>
</eLinkResult>
```

Documenting a search/analysis – (what goes in your "notebook" ?)

- What do you need to know to reproduce your results?
  - date and time
  - file/program name (version)
  - query / file name, library name
  - search parameters? (matrix, gap penalities?)
- How do you store this information?
  - scripts (directory per project, make copies for new runs and projects, don't edit scripts)
  - results files (what happens if you generate 100 files, the last 10 of which are correct?)
  - how to you document/log your mistakes?
- Can you reproduce the process (with new data or updated databases) 6 months later?

Documenting a search/analysis –

- FASTA results:

```
# fasta35_t -c -1 -p -q -w 80 -m 9i -m 6 -H -S -f -10 -g -2 TMP.q H 2
FASTA searches a protein or DNA sequence data bank
 version 35.04 Feb. 20, 2010
Please cite: W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448
Query: TMP.q
  1>>>gi|21264427|sp|P30711.4|GSTT1_HUMAN Glutathione S-transferase theta-1; GST
    class-theta-1; Glutathione transf - 240 aa
Library: Human/Refseq proteins 17401176 residues in 38000 sequences
  17401176 residues in 38000 sequences
Statistics:  Expectation_n fit: rho(ln(x))= 4.2367+/-0.000222; mu= 13.0410+/- 0.013
    mean_var=62.7681+/-11.550, 0's: 7 Z-trim: 48  B-trim: 2 in 1/59 Lambda=
    0.161884
Algorithm: FASTA (3.5 Sept 2006) [optimized]Parameters: BL50 matrix (15:-5)xS ktup:
    2 join: 42, opt: -1, open/ext: -10/-2, width:  16 Scan time:  1.640
```

```
perl -e 'print `date`."\n";'
Thu Apr  1 16:44:11 EDT 2010
```

Homework 1

Write a program that takes a list of sequence accessions/names from refseq and:

1. Download the sequences from the appropriate sequence database
2. Identify the Uniprot accession numbers for the refseq sequences
3. Identify the locations of the Pfam domains on the Uniprot sequences

63

Final Project (home work to be turned in) Due March 17, 2013

Using GSTT1_DROME, PAXI_HUMAN, MYC_HUMAN, or a protein of your own:

- Do a blast search against human RefSeq (`/data/slib/genomes/hum_refseq`)
- For each of the high scoring sequences (`E()<2.0`), report:
  - the description, E-value, start and stop of the alignment
  - the Pfam domains on the target/sbjct protein included in the alignment
  - the Pfam domains on the target protein NOT included in the alignment
  - based on E()-value, reverse BLAST search, and domain composition, identify the highest scoring unrelated sequence