TODAY'S TOPICS:

- Bio::Seq, Bio::SeqIO Review
- Bio::SeqFeatures
  - what are features
  - accessing, manipulating features
- Bio::SearchIO
  - digging into BLAST results

THE BIO::SEQIO OBJECT

- **Creating a Bio::SeqIO object:**

```
$InSeqIO = Bio::SeqIO->new(-file => "$infile",
                          -format => ‘Genbank’);

$OutSeqIO = Bio::SeqIO->new(-file => ">$infile.FASTA",
                           -format => ‘FASTA’);
```

- **Can do either *input* or *output*. Remember the “>“!**

- **The Bio::SeqIO->new() method simply instantiates a connection – no I/O has been done!**
- ***next_seq*, *write_seq* methods do the I/O**

BIO::SEQ OBJECT

- The Bio::Seq Object can store a description, accession, version, alphabet, species, and features.

```
LOCUS    ECORHO    1880 bp    DNA linear    BCT 26-APR-1993
DEFINITION  E.coli rho gene coding for transcription termination factor.
ACCESSION  J01673
VERSION  J01673.1 GI:147605
…
ORGANISM  Escherichia coli
    Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
 Enterobacteriaceae; Escherichia.
…
ORIGIN 15 bp upstream from HhaI site.
 1       aaccctagca ctgcgccgaa atatggcatc cgtggtatcc cgactctgct gctgttcaaa
61       aacggtgaag tggcggcaac caaagtgggt gcactgtcta aaggtcagtt gaaagagttc
…
```

BIO::SEQFEATURE

- **Feature is a specific location in a biological sequence.**
- **For example, the Bio::SeqFeature object can store a sequence range, as well as a description of feature ‘type’ and other information.**

GENBANK FILE

```
LOCUS  SCU60829  2303 bp  DNA linear  PLN  24-JUN-1997
DEFINITION Saccharomyces cerevisiae Mre11p (MRE11) gene, complete cds.
ACCESSION U60829
VERSION U60829.1  GI:2209264
KEYWORDS .
SOURCE Saccharomyces cerevisiae (baker's yeast)
ORGANISM Saccharomyces cerevisiae
…
FEATURES Location/Qualifiers
 source   1..2303
          /organism="Saccharomyces cerevisiae"
          /mol_type="genomic DNA"
          /strain="SK1"
          /db_xref="taxon:4932"
          /chromosome="XIII"
   gene   168..2246
          /gene="MRE11"
   CDS    168..2246
          /gene="MRE11"
          /note="SK1 derived coding sequence; recombination enzyme"
…
ORIGIN
 1        aaactgactt aaggtttaaa tagtatggcc aatcgaatag aacccaaaca ttatagccat
61        attaaattac tctttacgct tgtaaggaag acaatgtgga aacaacatta agagaatgca
```

BIO::SEQFEATURE

- Feature is a specific location in a biological sequence.
- For example, the Bio::SeqFeature object can store a sequence range, as well as a description of feature ‘type’ and other information.
- Feature attributes include a seq ID, location, a primary tag, and some hash tags

BIO::SEQFEATURE

```
$feat_obj = Bio::SeqFeature::Generic->new( -start => 10,
                                      -end => 100,
                                      -strand => -1,
                                 -primary => 'repeat',
                         -source_tag => 'repeatmasker',
                          -display_name => 'alu family',
                                      -score => 1000,
                  -tag => { new => 1,
                         author => 'someone',
                          sillytag => 'this is silly!' }
 );
```

BIO::SEQFEATURE METHODS

attach_seq

Usage : $feat_obj->**attach_seq**($seq)

Function: Attaches a Bio::Seq object to this feature. This Bio::Seq object is for the *entire* sequence: ie from 1 to 10000

Returns : TRUE on success

Args : a Bio::PrimarySeqI compliant object

BIO::SEQFEATURE METHODS

attach_seq – provide the full sequence whose features are to be described

seq

Usage : $tseq = $feat_obj->**seq**()

Function: returns the truncated sequence (if there) for this feature

Returns : sub seq (a Bio::PrimarySeqI compliant object) on attached sequence bounded by start & end, or undef if there is no sequence attached

Args : none

BIO::SEQFEATURE METHODS

attach_seq – provide the full sequence whose features are to be described

seq – return a Bio::Seq object of the sub-sequence with the feature

get_tag_values

Usage : @values = $feat_obj->**get_tag_values**('note');

Function: Returns a list of all the values stored under a particular tag.

Returns : A list of scalars

Args : The name of the tag

BIO::SEQFEATURE METHODS

attach_seq – provide the full sequence whose features are to be described

seq – return a Bio::Seq object of the sub-sequence with the feature

get_tag_values – returns a list of scalar values for a tag

add_tag_value

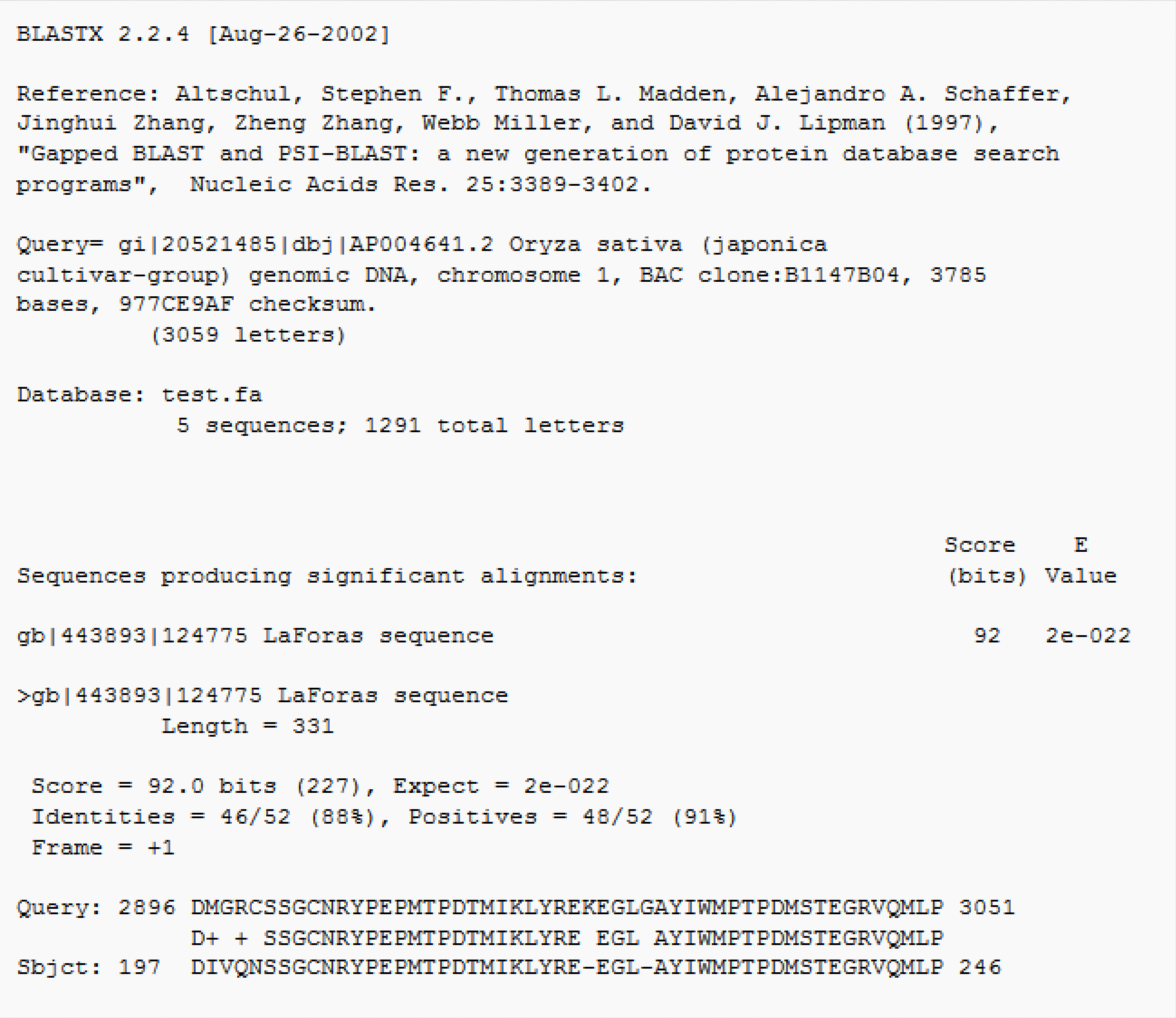
Usage : $feat_obj->**add_tag_value**('note',"this is a note");
Returns : TRUE on success

Args : tag (string) and one or more values (any scalar(s))

BIO::SEARCHIO

- **SearchIO refers to the processing of search results**
- **The Bio::SearchIO object can read BLAST output and allow you to access individual BLAST results, hits, and even HSPs.**

BLAST OUTPUT FILE



```
BLASTX 2.2.4 [Aug-26-2002]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|20521485|dbj|AP004641.2 Oryza sativa (japonica
cultivar-group) genomic DNA, chromosome 1, BAC clone:B1147B04, 3785
bases, 977CE9AF checksum.
         (3059 letters)

Database: test.fa
           5 sequences; 1291 total letters



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gb|443893|124775 LaForas sequence                                  92   2e-022

>gb|443893|124775 LaForas sequence
          Length = 331

 Score = 92.0 bits (227), Expect = 2e-022
 Identities = 46/52 (88%), Positives = 48/52 (91%)
 Frame = +1

Query: 2896 DMGRCSSGCNRYPEPMTPDTMIKLYREKEGLGAYIWMPTPDMSTEGRVQMLP 3051
            D+ + SSGCNRYPEPMTPDTMIKLYRE EGL AYIWMPTPDMSTEGRVQMLP
Sbjct: 197  DIVQNSSGCNRYPEPMTPDTMIKLYRE-EGL-AYIWMPTPDMSTEGRVQMLP 246
```

BIO::SEARCHIO

- **SearchIO refers to the processing of search results**

- **The Bio::SearchIO object can read BLAST output and allow you to access individual BLAST results, hits, and even HSPs.**

- **Reads HMMer, Exonerate and rnamotif output, too!**

-
  ```
  use Bio::SearchIO;
  my $in_searchIOobj = new Bio::SearchIO(-format => 'blast',
                                        -file => 'report.bls');
  ```

BIO::SEARCHIO METHODS

next_result

Usage : my $hit = $in_searchIOobj->**next_result**;

Function: Returns the next Result from a search

Returns : Bio::Search::Result::ResultI object

Args : none

BIO::SEARCHIO METHODS

next_result – returns a result for a single query

next_hit

Usage : while($hit = $result->**next_hit**()) { ... }

Function: Returns the next available Hit object, representing potential matches between the query and various entities from the database.

Returns : a Bio::Search::Hit::HitI object or undef if there are no more.

Args : none

BIO::SEARCHIO METHODS

next_result – returns a result for a single query

next_hit – returns a list of hits for a single result

next_hsp

Usage : while($hsp = $obj->**next_hsp**()) { ... }

Function : Returns the next available High Scoring Pair
Returns : Bio::Search::HSP::HSPI object or null if finished
Args : none

BIO::SEARCHIO METHODS

next_result – returns a result for a single query

next_hit – returns a hit for a some result

next_hsp – returns a High-Scoring Segment Pair for some hit

get_aln

Usage : my $aln = $hsp->**get_aln**;

Function: Returns a Bio::SimpleAlign representing the HSP alignment

HOMEWORK, ETC.

See today's handout for in-class exercises. Data here:

```
cp -r /gsfs1/xdisk/ssolonen/ecol553_nov6/ .
```

Homework 9, on Bio::SeqIO and Bio::SearchIO, is due Tuesday, Nov 13th