CAP 5510: Introduction to Bioinformatics

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfS07.html
Perl: Practical Extraction & Report Language

- Created by Larry Wall, early 90s
- Portable, “glue” language for interfacing C/Fortran code, WWW/CGI, graphics, numerical analysis and much more
- Easy to use and extensible
- OOP support, simple databases, simple data structures.
- From interpreted to compiled
- High-level features, and relieves you from manual memory management, segmentation faults, bus errors, most portability problems, etc, etc.
- Competitors: Python, Tcl, Java
Perl Features

- Bit Operations
- Pattern Matching
- Subroutines
- Packages & Modules
- Objects
- Interprocess Communication
- Threads, Process control
- Compiling
Managing a Large Project

- Devise a common data exchange format.
- Use modules that have already been developed.
- Write Perl scripts to convert to and from common data exchange format.
- Write Perl scripts to “glue” it all together.
What is BioPerl?

- Toolkit of Perl modules useful for bioinformatics
- Open source; Current version: BioPerl 1.5.2
- Routines for handling biosequence and alignment data.
- Why? Human Genome Project: Same project, same data. different data formats! Different input formats. Different output formats for comparable utility programs.
  - BioPerl was useful to interchange data and meaningfully exchange results.
    - “Perl Saved the Human Genome Project”
- Many routine tasks automated using BioPerl.
- String manipulations (string operations: substring, match, etc.; handling string data: names, annotations, comments, bibliographical references; regular expression operations)
- Modular: modules in any language
pTk - to enable building Perl-driven GUIs for X-Window systems.

BioJava

BioPython

The BioCORBA Project provides an object-oriented, language neutral, platform-independent method for describing and solving bioinformatics problems.
#!/usr/bin/perl -w

# Storing DNA in a variable, and printing it out

# First we store the DNA in a variable called $DNA
$DNA = 'ACGGGAGGACGGGAAAATTACTACGGCATTAGC';

# Next, we print the DNA onto the screen
print $DNA;

# Finally, we'll specifically tell the program to exit.
exit;  #perl1.pl
#!/usr/bin/perl -w
$DNA1 = 'ACGGGAGGACGCGGAAAATTACTACGGGCATTAGC';
$DNA2 = 'ATAGTGCCGTGAGAGTGATGTAGTA';
# Concatenate the DNA fragments
$DNA3 = "$DNA1$DNA2";
print "Concatenation 1):

$DNA3

";
# An alternative way using the "dot operator":
$DNA3 = $DNA1 . $DNA2;
print "Concatenation 2):

$DNA3

";
# transcribe from DNA to RNA; make rev comp; print;
$RNA = $DNA3; $RNA =~ s/T/U/g;
$rev = reverse $DNA3; $rev =~ tr/AGCTacgt/TCGAtgca/;
print "$RNA\n$rev\n";
exit;  #perl2.pl
#!/usr/bin/perl -w
# Read filename & remove newline from string
$protFile = <STDIN>; chomp $protFile;
# First we have to "open" the file
unless (open(PROTEINFILE, $protFile) {
    print "File $protFile does not exist"; exit;})
# Each line becomes an element of array @protein
@protein = <PROTEINFILE>;
print @protein;
# Print line #3 and number of lines
print $protein[2], "File contained ", scalar @protein, " lines\n";
# Close the file.
close PROTEINFILE;
exit; #perl3.pl
Perl: subroutines

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

# using command line argument
$dna1 = $ARGV[0]; $dna2 = $ARGV[1];

# Call subroutine with arguments; result in $dna
$dna = addGACAGT($dna1, $dna2);

print "Add GACAGT to $dna1 & $dna2 to get $dna\n\n";
exit;

##### addACGT: concat $dna1, $dna2, & "ACGT". ######
sub addGACAGT {
    my($dnaA, $dnaB) = @_; my($dnaC) = $dnaA.$dnaB;
    $dnaC .= 'GACAGT';
    return $dnaC;
}

#perl4.pl
```
How Widely is Bioperl Used?
What Can You Do with Bioperl?

- Accessing sequence data from local and remote databases
- Transforming formats of database/file records
- Manipulating individual sequences
- Searching for similar sequences
- Creating and manipulating sequence alignments
- Searching for genes and other structures on genomic DNA
- Developing machine readable sequence annotations

Types of Perl Objects:

- Sequence objects
- Location objects
- Interface objects
- Implementation objects
BioPerl Modules

- **Bio::PreSeq**, module for reading, accessing, manipulating, analyzing single sequences.
- **Bio::UnivAln**, module for reading, parsing, writing, slicing, and manipulating multiple biosequences (sequence multisets & alignments).
- **Bio::Struct**, module for reading, writing, accessing, and manipulating 3D structures.
- Support for invoking **BLAST** & other programs.
- Download URL [http://www.bioperl.org/Core/Latest/]
- Tutorial [http://www.bioperl.org/Core/Latest/bptutorial.html]
- Course [http://www.pasteur.fr/recherche/unites/sis/formation/bioperl/index.html]
BioPerl Sequence Object

```perl
$seqobj->display_id(); # readable id of sequence
$seqobj->seq(); # string of sequence
$seqobj->subseq(5, 10); # part of the sequence as a string
$seqobj->accession_number(); # if present, accession number
$seqobj->moltype(); # one of 'dna','rna','protein'
$seqobj->primary_id(); # unique id for sequence independent
                        # of its display_id or accession number
```
Example 1: Convert SwissProt to fasta format

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

use strict;
use Bio::SeqIO;
my $in  = Bio::SeqIO->newFh ( -file   => '<seqs.html',
                              -format => 'swiss' );
my $out = Bio::SeqIO->newFh ( -file   => '>seqs.fasta',
                              -format => 'fasta' );

print $out $_ while <$in>;

exit; #bioperl1.pl
```
Example 2 : Load sequence from remote server

```perl
#!/usr/bin/perl -w
use Bio::DB::SwissProt;

$database = new Bio::DB::SwissProt;

$seq = $database->get_Seq_by_id('MALK_ECOLI');

my $out = Bio::SeqIO->newFh(-fh => STDOUT,
                            -format => 'fasta');

print $out $seq;

exit;
```

```perl
#!/local/bin/perl -w
use Bio::DB::GenBank;

my $gb =
    new Bio::DB::GenBank(
                  -retrievaltype=>'tempfile',
                  -format=>'Fasta');

my ($seq) = $seq =
    $gb->get_Seq_by_id("5802612");

print $seq->id, "\n";
print $seq->desc(), "Sequence: \n";
print $seq->seq(), "\n";
exit;
```
#! /local/bin/perl -w
use strict;
use Bio::SeqIO;
my $in  = Bio::SeqIO->new ( -file   => 'seqs.html', -format => 'swiss' );
my $out = Bio::SeqIO->new ( -file   => 'seqs.fas', -format => 'fasta' );

while ($seq = $in->next_seq()) {
    $accNum = $seq->accession_number();
    print "Accession# = $accNum\n";
    $out->write_seq($seq);
}

exit; #bioperl2.pl
#!/usr/bin/perl –w

# define a DNA sequence object with given sequence
$seq = Bio::Seq->new('-seq'=>'actgtggcgtaacct',
    '-desc'=>'Sample Bio::Seq object',
    '-display_id' => 'somethingxxx',
    '-accession_number' => 'accnumxxx',
    '-alphabet' => 'dna' );
$gb = new Bio::DB::GenBank();

$seq = $gb->get_Seq_by_id('MUSICGHBA1'); #returns Seq object
$seq = $gb->get_Seq_by_acc('AF303112'); #returns Seq object
# this returns a SeqIO object :
$seqio = $gb->get_Stream_by_batch([ qw(J00522 AF303112)]);
exit; #bioperl3.pl
#!/usr/local/bin/perl -w

use Bio::DB::GenBank;

$gb = new Bio::DB::GenBank;

$seq1 = $gb->get_Seq_by_acc('AF303112');
$seq2 = $seq1->trunc(1,90);
$seq2 = $seq2->revcom();

print $seq2->seq(), "\n";
$seq3 = $seq2->translate;

print $seq3->seq(), "\n";
exit; #bioperl4.pl
use Bio::Structure::IO;

$in = Bio::Structure::IO->new(-file => "inputfilename", '-format' => 'pdb');

$out = Bio::Structure::IO->new(-file => "outputfilename", '-format' => 'pdb');

# note: we quote -format to keep older perl's from complaining.
while ( my $struc = $in->next_structure() ) {
    $out->write_structure($struc);
    print "Structure ", $struc->id, " number of models: ",
    scalar $struc->model, "\n";
}
use Bio::SeqIO;
$seqin  = Bio::SeqIO->new(-format =>'EMBL', -file =>'f1');
$seqout = Bio::SeqIO->new(-format =>'Fasta',-file =>'f1.fa');
while ((my $seqobj = $seqin->next_seq())) {
    print "Seq: ", $seqobj->display_id, ", Start of seq ",
    substr($seqobj->seq,1,10),"\n";
    if ( $seqobj->moltype eq 'dna') {
        $rev = $seqobj->revcom;
        $id = $seqobj->display_id();
        $id = "$id.rev";
        $rev->display_id($id);
        $seqout->write_seq($rev); } #end if
    foreach $feat ( $seqobj->top_SeqFeatures() ) {
        if( $feat->primary_tag eq 'exon' ) {
            print STDOUT "Location ", $feat->start,"."
            $feat->end," GFF[",$feat->gff_string,"]\n";
        } # end foreach
    } # end foreach
} # end while
exit; #bioperl6.pl
Example 3: Read alignment file using AlignIO class

#!/usr/bin/perl -w
use strict;
use Bio::AlignIO;
my $in = new Bio::AlignIO(-file => '<data/infile.aln',
   -format => 'clustalw');

# returns an alignI (alignment interface)
my $aln = $in->next_aln();
print "same length of all sequences: ",
   ($aln->is_flush()) ? "yes" : "no", "\n";
print "alignment length: ", $aln->length, "\n";
printf "identity: %.2f \%\n", $aln->percentage_identity();
printf "identity of conserved columns: %.2f \%\n", $aln->overall_percentage_identity();
Example 4: Standalone BLAST

```perl
#!/usr/bin/perl -w
use strict;
use Bio::SeqIO;
use Bio::Tools::Run::StandAloneBlast;
my $seq_in = Bio::SeqIO褫 new(-file => '<data/prot1.fasta',
    -format => 'fasta');
my $query = $seq_in -> next_seq();
my $factory = Bio::Tools::Run::StandAloneBlast -> new('program' => 'blastp',
    'database' => 'swissprot',
    _READMETHOD => 'Blast');
my $blast_report = $factory->blastall($query);
my $result = $blast_report->next_result();
while (my $hit = $result->next_hit()){
    print "hit name: ", $hit->name(), "Significance: ", $hit->significance(), "\n";
    while (my $hsp = $hit->next_hsp()){
        print "E: ", $hsp->evalue(), "frac_identical: ", $hsp->frac_identical(), "\n";
    }
}
exit;
```
CpG Islands

- Regions in DNA sequences with increased occurrences of substring “CG”
- Rare: typically C gets methylated and then mutated into a T.
- Often around promoter or “start” regions of genes
- Few hundred to a few thousand bases long
Problem 1:

• **Input**: Small sequence \( S \)

• **Output**: Is \( S \) from a CpG island?

  • Build Markov models: \( M^+ \) and \( M^- \)
  • Then compare
### Markov Models

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td>0.322</td>
<td>0.298</td>
<td>0.078</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>0.248</td>
<td>0.246</td>
<td>0.298</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>0.177</td>
<td>0.239</td>
<td>0.292</td>
<td>0.292</td>
</tr>
</tbody>
</table>
How to distinguish?

- **Compute**

\[
S(x) = \log \left( \frac{P(x|M_+)}{P(x|M_-)} \right) = \sum_{i=1}^{L} \log \left( \frac{p_{x(i-1),x_i}}{m_{x(i-1),x_i}} \right) = \sum_{i=1}^{L} r_{x(i-1),x_i}
\]

<table>
<thead>
<tr>
<th>r=p/m</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.740</td>
<td>0.419</td>
<td>0.580</td>
<td>-0.803</td>
</tr>
<tr>
<td>C</td>
<td>-0.913</td>
<td>0.302</td>
<td>1.812</td>
<td>-0.685</td>
</tr>
<tr>
<td>G</td>
<td>-0.624</td>
<td>0.461</td>
<td>0.331</td>
<td>-0.730</td>
</tr>
<tr>
<td>T</td>
<td>-1.169</td>
<td>0.573</td>
<td>0.393</td>
<td>-0.679</td>
</tr>
</tbody>
</table>

**Score(GCAC)**

\[= 0.461 - 0.913 + 0.419 < 0.\]

**GCAC** not from CpG island.

**Score(GCTC)**

\[= 0.461 - 0.685 + 0.573 > 0.\]

**GCTC** from CpG island.
Problem 1:
- **Input:** Small sequence $S$
- **Output:** Is $S$ from a CpG island?
  - Build Markov Models: $M^+$ & $M^-$
  - Then compare

Problem 2:
- **Input:** Long sequence $S$
- **Output:** Identify the CpG islands in $S$.
  - Markov models are inadequate.
  - Need Hidden Markov Models.
Markov Models

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td>C</td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td>G</td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td>T</td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
</tbody>
</table>

$P(A^+|A^+)$
$P(G^+|C^+)$
CpG Island + in an ocean of –
First order Hidden Markov Model

MM=16, HMM= 64 transition probabilities (adjacent bp)