Genomic Databases and Bioperl

Presentation for ECS289A Winter ’03

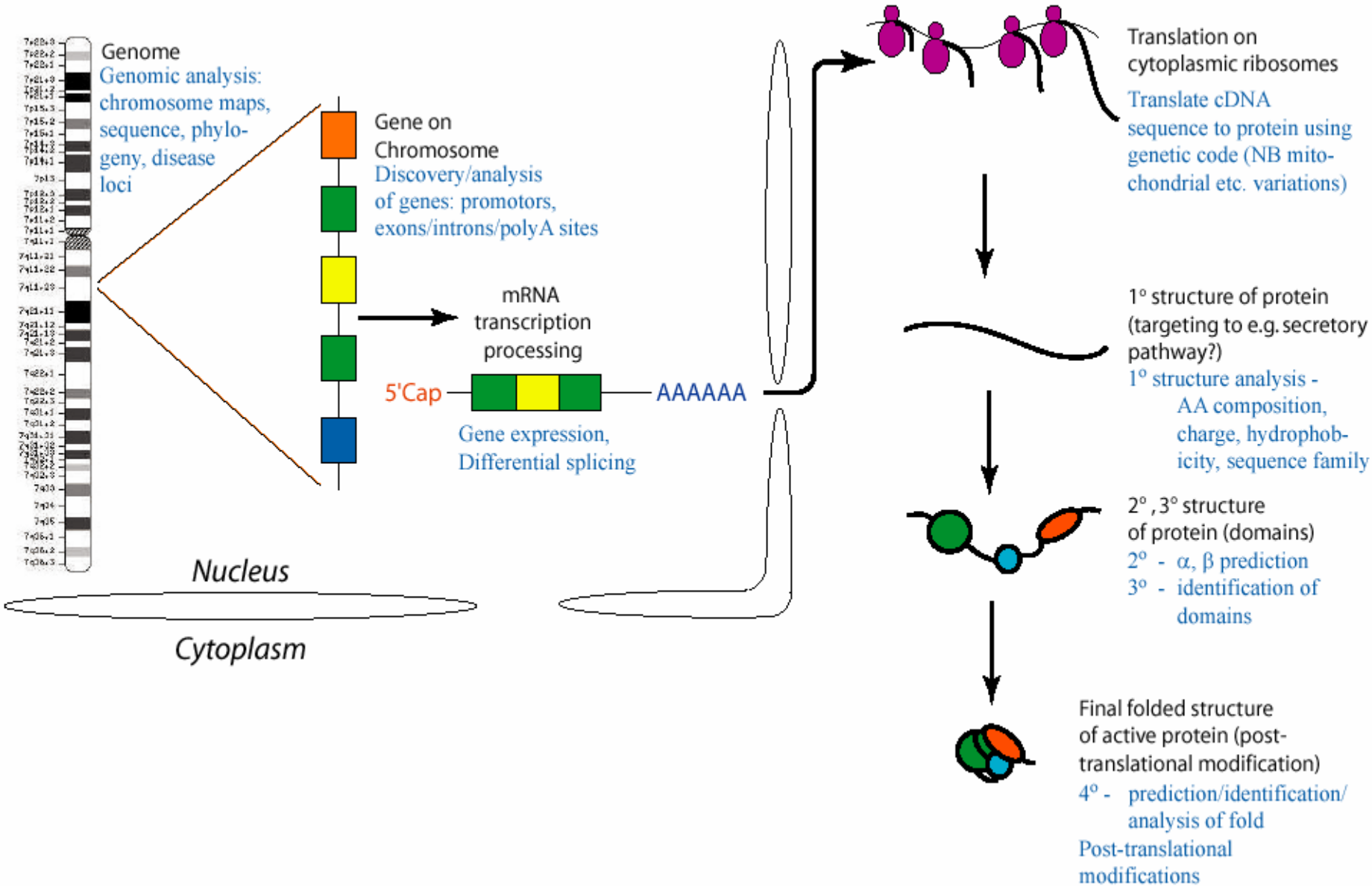
by

Nameeta Shah

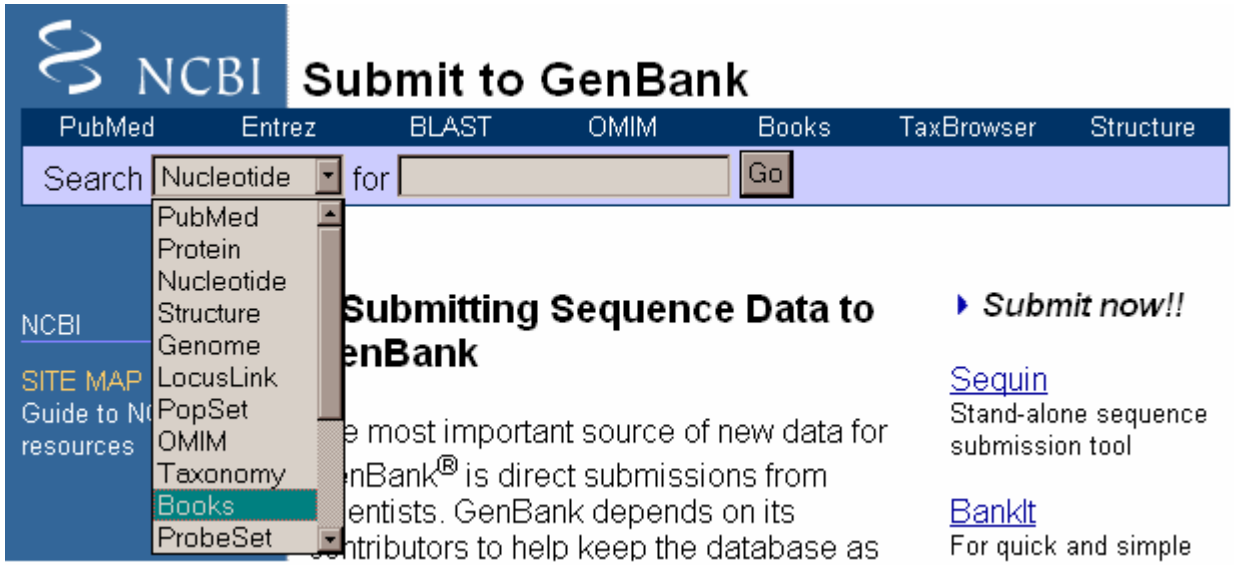
Overview

- Genomic Databases
  - GenBank
  - PIR
  - SMD
- Bioperl
  - Motivation
  - What is Bioperl?
  - Main capabilities
  - Major objects
  - An example
  - Disadvantages

Genomic Data



GenBank



- GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.
- part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI.

Typical record

Protein Information Resource (PIR)

PIR NREF - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss Messenger

Address http://pir.georgetown.edu/pirwww/search/pirnref.shtml

PIR Non-Redundant Reference Protein Database

Text Search Protein Databases: GO!

About PIR | Databases | Search & Retrieval | Download | Support

Database Description

The PIR-NREF is a Non-redundant REFerence protein database designed to provide a timely and comprehensive collection of all protein sequence data, keeping pace with the genome sequencing projects and containing source attribution and minimal redundancy. The database contains all sequences in PIR-PSD, SwissProt, TrEMBL, RefSeq, GenPept, and PDB. Identical sequences from the same source organism (species) reported in different databases are presented as a single NREF entry with protein IDs and names from each underlying database, in addition to protein sequence, taxonomy, and composite bibliography. Related sequences identified by all-against-all FASTA search are listed for each NREF entry. The web site provides direct entry retrieval (based on protein IDs), text search (protein or species names), and sequence search (BLAST, peptide match, and pattern match) for full-scale and species-based protein identification. Species-based browsing and searching are supported for about 100 organisms, including over 70 complete genomes. PIR-NREF is available for free downloading and redistribution from our FTP site in XML format (data file) and FASTA format (sequence file). The database is updated biweekly and the release 1.14, 20-Jan-2003, contains 1,130,196 entries from:

Database	Release#	Date	# of Entries
PIR	75.01	20-Jan-2003	283,269
SwissProt	40.39	10-Jan-2003	120,961
TrEMBL	22.8	10-Jan-2003	728,713
GenPept	133.0	15-Dec-2002	1,277,378
RefSeq		15-Jan-2003	463,539
PDB		13-Jan-2003	20,261

Release History

More Description

PIR-NREF current release 1.14, 20-Jan-2003 contains 1,130,196 entries. Download

Related NREF sequences identified by all-against-all FASTA search were pre-computed in collaboration with the Advanced Biomedical Computing Center at the National Cancer Institute - Frederick and the DuPont Bioinformatics Team.

Find Proteins by Name, Organism or UID

Retrieve a matching list (a summary report if only one entry found) by protein name and organism name using substring match, or by UID using exact match.

All UID Fields | Submit | Example

Protein Sequence Similarity Search

Search NREF by Species/Organism

BLAST Search

Retrieve a matching list of entries by searching your query protein sequence against the NREF database.

Paste query sequence (single-letter amino acid code) or ">" followed by unique sequence identifier from any underlying database:

E-value: 0.0001 **Filter** Yes Submit Example

Peptide Match

Retrieve a list of entries with exact matches to your query peptide sequence.

Enter a string of single-letter amino acid codes below: Submit Example

Pattern Match

Retrieve a list of entries matching your query pattern or a ProSite pattern.

Insert a user-defined pattern below: Click here for help on how to write a protein pattern

Or, alternatively, type in a valid PROSITE code for a query pattern (e.g., PS00888):

Internet

Stanford Microarray Database

Stanford Microarray Database

- SMD Home
- Public Search
- Published Data
- Software & Tools
- Microarray Links
- Stanford Genomics
- S.O.U.R.C.E.
- Staff
- SMD Code
- About SMD
- SMD Specifications
- Help Index

Help

For Citation, there are 98 matching your query

Page Navigation	List Navigation	List Display
Top / Bot, Next	Alizadeh A... --> Keller G, ... Go! or Download full list	Limit to: Organism : none selected Sorted by Header : Citation Filtered (on Sort Header) : Re-list Revert List

Citation	Organisms(s)	Web Supplement	PubMed Link	Full Text	Data in SMD
Alizadeh AA, et al. (2000) Nature 403(6769):503-11	*Homo sapiens*		PubMed	nature	SMD
Arbeitman MN, et al.(2002) Science 297(5590):2270-5	*Drosophila melanogaster*		PubMed	Science	SMD
Baldwin et al.(2002) Genome Biology 2002 4(1):R2	*Homo sapiens*			GenomeBiology	SMD
Bernstein JA, et al. (2002) Proc Natl Acad Sci U S A 99(15):9697-702	*Escherichia coli*		PubMed	PNAS	SMD
Bjorkholm B, et al. (2001) Infect Immun 69(12):7832-8	*Helicobacter pylori*		PubMed	IAI	SMD

Motivation for Bioperl

- GenBank Growth
- PDB Growth
- Ways to “mine” databases for the discovery of:
  - Genes
  - Proteins
  - Evolutionary relationships
  - Biochemical pathways
- Web interfaces
  - Ease of use but they don’t scale
  - Non-standard
  - Queries tailored for specific user needs are not possible

Motivation…

- Why a computer scripting language?
  - Easy to scale up to large numbers of sequences
  - Easy to handle data in multiple formats
  - Easier to see patterns in sequence data
- Why Perl?
  - Ease of use by novice programmers
  - Fast software prototyping
    - Flexible language
    - Compact code
- Powerful pattern matching via “regular expressions”
- Availability of many ready-to-use modules
- Portability
- Open Source – easy to extend and customize, No licensing fees

What is Bioperl?

- Bioperl is a group of open-source-software developers for bioinformatics
- Bioperl is also a collection of Perl “objects” for simplifying Perl scripts for bioinformatics tasks such as:
  - Parsing database (e.g. GenBank) files
  - Parsing results of sequence analysis programs – Blast, Genscan, Hmmer, etc
  - Sequence manipulation and analysis
  - Obtaining multiple database entries over the internet

Main capabilities of the Bioperl package

- Automated parsing of major database formats
- Automated parsing of reports from BLAST, Genscan, HMMER, etc.
- Sequence manipulation operations including:
  - Sequence translation
  - Reverse complementation
  - Restriction site identification
  - Signal sequence identification
  - Molecular weight calculations
- Batch retrieval of records from remote databases
- Sequence annotation capabilities

Main capabilities of the Bioperl package…

- Simple, uniform Perl interfaces to running BLAST, Smith-Waterman, Clustalw and Tcoffee locally.
- Manipulation of genomic-size sequences on memory-limited computers.

Bioperl Objects

- Sequence Objects

Bioperl has several different objects for handling protein, DNA, and RNA sequence data

- Seq
  - Principle sequence object in Bioperl
  - Includes sequence and annotation data
- PrimarySeq
  - Seq object stripped of its annotations
  - Useful with large sequences
- LocatableSeq
  - Sequence object with start, end and strand attributes, part of a multiple sequence alignment

Bioperl Objects

- Sequence Objects
  - LiveSeq
    - Used for sequences whose feature locations may change over time
    - Typically used for newly sequenced genomes
    - Same interface as Seq object
  - LargeSeq
    - Special type of Seq object for handling long (>100MB) sequences
  - SeqI
    - Sequence interface object

Bioperl Objects

- Sequence IO Objects

Bioperl's SeqIO objects make sequence data-format conversion simple:

  - Data formats currently supported: Fasta, EMBL, Genbank, Swissprot, PIR and GCG
  - SeqIO can read a stream of sequences – located in a single or in multiple files
  - Once the sequence data has been read in with SeqIO, it is available to Bioperl in the form of Seq objects

Sequence Manipulation

- Once a sequence is available to Bioperl in the form of a sequence object, many standard bioinformatics calculations can be performed on the sequence including:
  - Extracting subsequences
  - Performing reverse complementation
  - Translating a DNA sequence
  - Identifying restriction enzyme sites
  - Identifying signal cleavage sites
  - Obtaining molecular weights and sequence statistics

More Bioperl objects

- SeqStats and SeqWords
- RestrictionEnzyme
- Sigcleave
- SeqPattern
- SeqFeature, Location, Structure
- Alignment Objects
  - The “SimpleAlign” object
- Alignment IO
  - “AlignIO” is the object for data conversion of alignment files
- Accessing remote databases
  - Objects for Genbank, Genpept, swissprot, gdb and acedb
- Tools

An Example

```
#!/usr/local/bin/perl -w
#
# How to retrieve GenBank entries over the Web
#
# by Jason Stajich
#
use Bio::DB::GenBank;
use Bio::SeqIO;
   $gb = new Bio::DB::GenBank;

# the output stream for your seqs, this can be a file
# instead or STDOUT, see the Bio::SeqIO module for info

   $seqout = new Bio::SeqIO(-fh => \*STDOUT, -format => 'fasta');

# if you want a single seq
   $seq = $gb->get_Seq_by_id('MUSIGHBA1');
$seqout->write_seq($seq);
# or by accession
$seq = $gb->get_Seq_by_acc('AF303112');

$seqout->write_seq($seq);

# if you want to get a bunch of sequences use the batch method
   $seqio = $gb->get_Stream_by_batch([ qw(J00522 AF303112 2981014)]);

while( defined ($seq = $seqio->next_seq )) {
        $seqout->write_seq($seq);
}
```

An Example

- DNA sequence analysis problem
  - We're looking for genes related to the BRCA1

```
$gb = new Bio::DB::GenBank(-retrievaltype =>
   'infile' , -format => 'genebank');
$in = Bio::SeqIO->new('-file'=>"infile", '-
   format'=>'genebank');
$seqobj = $in->next_seq();
@ann = ($seqobj->annotation)->each_gene_name();
foreach $a (@ann){
   $cond = ($a~={BRCA1});
}
```

An Example

- DNA sequence analysis problem contd…
  - We have reason to look on chromosome V of C.elegans

```
If($cond && $seqobj->species->{common_name} =
  ~{elegans}) {
  $seq = $seqobj->primary_seq->{seq}
  $id = $seqobj->id
}
```

  - We are interested in looking at possible "promoter" sequences 100 to 200 base pairs upstream

```
@allfeatures = $seqobj->all_SeqFeatures();
$feature_start = ($allfeatures[0])->start;
$subsequence = $seqobj->subseq($feature_start-200,
  100);
```

An Example

- DNA sequence analysis problem contd…
  - We want to align the resulting sequences with Clustalw and find regions of identity and similarity

```
@params = ('ktuple'=>2);
$factory = Bio::Tools::Run::Alignment::Clustalw-
  >new(@params);
$aln = $factory->align($unaligned_seq_file);
$threshold = 60;
$string = $aln->consensus_string($threshold);
Print "Consensus string with threshold = $threshold is
  $string\n";
```

Disadvantages of Bioperl

- Efficiency
- Not 100% object-oriented
- Code becomes public
- More overhead

References

- http://www.ukc.ac.uk/bio/baines/bi821/Default.htm
- http://www.ncbi.nlm.nih.gov/Genbank/index.html
- http://pir.georgetown.edu/
- http://genome-www5.stanford.edu/MicroArray/SMD/
- http://www.bioperl.org/Core/bptutorial.html
- Peter Schattner, Perl and Bioperl: Tools for Automated Analysis of Biological Sequence Data, O'Reilly Bioinformatics Technology Conference 2002