Bioinformatics

Biological databases

Martin Saturka

http://www.bioplexity.org/lectures/

EBI version 0.4

Data sources for bioinformatics

Main types of biological databases with utilization tools.

- communication with databases. database usage.
- genomes, structure families, expression maps.

Main topics

- database technics
  - file types, sql, biodas
  - protocols, bioperl
- particular databases
  - sequences, structures
  - expression, ontology
- digest approaches
  - constraint programming
  - filters, data structures

File types

- sequences
  - Fasta - multiple sequences
    one sequence: first line - header `>...`, next lines - per 70 nt
    GenBank header: `gi|`*gi-number*`|gb|`*accession*`|` *locus*
  - GFF - sequence features

    ```
    seqname source feature start end score strand frame
    ```
  - GenBank flat files or ASN.1
    flat files: multiline description, 6x10 nt per line
    ASN.1: structured description `{...{...}}`, optionally packed
- structures
  - PDB - line/column-wise data
    line start: line decription - comment / atom

    ```
    atom rank role residuum chain res-rank coordinates
    ```
- expression
  - tables: lines - genes, columns - tissues / experiments
  - MAGE: `http://www.mged.org/Workgroups/MAGE/mage.html`

Annotation access

annotation systems

`http://www.open-bio.org/wiki/Projects`

- BioDAS - distributed annotation system
  - `http://biodas.org/`
  - data access

    `http://example.org/das/organism/features?segment=CHR_I:1,500`

    `site prefix das data command arguments`
  - system composition
    - a reference sequence server plus annotation servers
- other projects
  - MOBY - interoperability for biological data server services
  - OBDA - sequence access standardization
  - myGrid - grid and middleware for bioinformatics
    - `http://www.mygrid.org.uk/`
    - myExperiment - myGrid spin-off

SQL

- tabular data storage
  - main open-source database systems
    small: sqlite, large: postgresql, firebirdsql
  - access: sql - structured query language
  - suitable for data structured into regular tables
  - different approaches
    - linear data (sequences): flat files
    - deeply structured data: ASN.1, HDF, NetCDF

```
> sqlite3
CREATE TABLE genes (gi INTEGER, chr CHAR(3), ori CHAR(1));
INSERT INTO genes VALUES (826, '19', '+');
INSERT INTO genes VALUES (827, 'X', '-');
SELECT * FROM genes;
.quit
```

Genomes

where to get data on sequenced chromosomes

gene specific: gene id sequence specefic: accession

- main genome database sites
  - NCBI - National center for biotechnology information
    - http://www.ncbi.nlm.nih.gov/Entrez/
  - EMBL - European bioinformatics institute
    - http://www.ebi.ac.uk/embl/
  - DDBJ - DNA databank of Japan
    - http://www.ddbj.nig.ac.jp/
- NCBI ftp://ftp.ncbi.nlm.nih.gov/
  - directory /genomes/H_sapiens/
    - assembled reference sequences: Assembled_chromosomes
  - file /gene/DATA/gene2refseq.gz
    - gene IDs with positions along chromosomes

DNA variations

SNPs, CNVs

- many projects set to deal with intra-species variation
  - dbSNP
    `http://www.ncbi.nlm.nih.gov/SNP/`
  - the SNP consortium
    `http://snp.cshl.org/`
  - haplotypes
    `http://www.hapmap.org/`
  - glovar - human variations
    `http://www.glovar.org/`
  - human variome
    `http://www.humanvariomeproject.org/`
  - general variomes
    `http://variome.net/`

Gene prediction

open-source and on-line gene prediction

- Glimmer - bacteria, archea, viruses
  - `http://cbcb.umd.edu/software/glimmer/`
- GlimmerHMM - eukaryotic genes
  - `http://cbcb.umd.edu/software/GlimmerHMM/`
- GeneZilla (TIGRscan) - eukaryotic genes
  - `http://www.genezilla.org/`
- GenScan - human genes
  - `http://genes.mit.edu/GENSCAN.html`
- software lists

  `http://www.genefinding.org/`

Nucleic structures

RNAs and 3D nucleic structural databases

- 3D structures of nucleic acids
  - RNABase
    `http://www.rnabase.org/`
  - NDB nucleic acids database
    `http://ndbserver.rutgers.edu/`
- SCOR - structural classification of RNA
  `http://scor.berkeley.edu/`
  - RNA motifs, structures and interactions
- other databases
  - Small RNA database
    `http://condor.bcm.tmc.edu/smallRNA/`
  - Noncoding RNA database
    `http://biobases.ibch.poznan.pl/ncRNA/`

Proteins

protein structures

- 3D structures
  - RCSB `http://www.rcsb.org/`
- protein domains
  - Expasy `http://www.expasy.ch/`
  - UniProt `http://www.uniprot.org/`
- structures
  - SCOP, CATH, FSSP, CASP, PFAM hierarchical classification

Protein families

systematics on protein structures

- SCOP `http://scop.berkeley.edu/`
  - structural classification of proteins
  - alpha, beta, alpha/beta, alpha+beta, ... superfamilies
  - folds: cca 1000, superfamilies: cca 1500, families: cca 3000

- CATH `http://www.cathdb.info/`
  - class (C), architecture (A), topology (T), homologous superfamily (H)
  - cca 1400 families
  - C: main secondary structure composition
  - A: orientation of secondary structures
  - T: folds with sec. structure connectivity
  - H: similarity superfamilies

Expasy

ExPASy (expert protein analysis system)

- UniProt - the universal protein resource
  http://www.expasy.uniprot.org/
  - knowledgebase, reference clusters, archives
- swissprot
  http://www.expasy.ch/sprot/
  - database of protein sequences together with annotations
  - structure and function of proteins
- prosite
  http://www.expasy.ch/prosite/
  - documentation on protein domains, folds, families

Expressions

expression microarrays repositories

- not a central repository
  - every institution wants to have a main microarray database
- some of the repositories
  - GEO - gene expression omnibus
    http://www.ncbi.nlm.nih.gov/geo/
  - Stanford microarray database
    http://genome-www.stanford.edu/
  - Broad (MIT/Harvard) institute
    http://www.broad.mit.edu/tools/data.html
  - EBI ArrayExpress
    http://www.ebi.ac.uk/arrayexpress/
  - ChipDB
    http://staffa.wi.mit.edu/chipdb/public/

Expression atlases

- expression mapping projects
  - BrainAtlas (mouse oriented)
    http://www.brainatlas.org/
    http://www.brain-map.org/
  - RAD - RNA abundance database
    http://www.cbil.upenn.edu/RAD3/
  - BodyMap
    http://bodymap.ims.u-tokyo.ac.jp/
  - GNF gene expression atlas
    http://expression.gnf.org/
  - 3D developmental gene expression
    http://www.univie.ac.at/GeneEMAC/
  - TissueInfo
    http://pbtest.med.cornell.edu/services/tissueinfo/query

- relational schema
  - GUS - genomics unified schema
    http://www.gusdb.org/

Ontology

overall description of bio-systems

- Gene Ontology
  - `http://www.geneontology.org/`
  - description of gene products for various databases
  - the main bio-ontology project
- Gene Cards
  - `http://www.genecards.org/`
  - human genes information / ontology database
  - one of the first ontology projects
- KEGG
  - `http://www.geneobjects.org/`
  - Kyoto encyclopedia of genes and genomes
  - mainly known for molecular interaction pathways

Organisms

sites dedicated to particular model organisms

- the sites:
  - the generic model organism database project http://www.gmod.org/
  - Escherichia coli http://ecocyc.org/
  - Saccharomyces cerevisiae http://www.yeastgenome.org/
  - Arabidopsis thaliana http://www.arabidopsis.org
  - Drosophila melanogaster http://www.flybase.org/ http://www.fruitfly.org/
  - Caenorhabditis elegans http://www.wormbase.org/
  - Danio rerio http://zfin.org/
  - Mus musculus http://www.informatics.jax.org/
  - Rattus sp. http://rgd.mcw.edu/

EBI

European bioinformatics institute

- EBI `http://www.ebi.ac.uk/`
  - part of EMBL `http://www.embl.org/`

- EBI databses
  - EMBL nucleotide database
    `http://www.ebi.ac.uk/embl/`
  - UniProt (together with Expasy and PIR)
  - ArrayExpress
    `http://www.ebi.ac.uk/arrayexpress/`
    public repository for microarray data
  - Ensembl
    `http://www.ensembl.org/`
    genomes and annotation for metazoa

NCBI

National center for biotechnology information

`http://www.ncbi.nlm.nih.gov/`

- the main bioinformatics institute / web site
- databases and services
  - sequence databases
    GenBank, ESTs, SNPs, etc.
  - PubMed - literature database
  - Entrez
    `http://www.ncbi.nlm.nih.gov/entrez/`
    retrieval system connecting together plethora of databases including PubMed, genomes, ontologies
  - Blast - the search engine, OMIM, etc.
  - Science primer
    `http://www.ncbi.nlm.nih.gov/About/primer/`
    introductions into molecular biology and bioinformatics

Blast

Blast - basic local alignment search tool

`http://www.ncbi.nlm.nih.gov/BLAST/`

- two examples
  - **blast** - nucleotide - blastn (for short queries)
    `ATCAGTGTAGTCATCGATACCGTAGTCA`
    - short random sequence
  - results
    nothing significant, use mouse sequence gi 83999722
  - display graph
  - **genomes** - human - megablast (for related sequences)
    `GACACCTTCTCTCCTCCCAGATTCCAGTAACTCCCAATCTTCTCTCTGCAG`
    - part of an immunoglobulin sequence
  - results
    two very significant matches, use ref NT_026437.11
  - click on IGHG1 for information
    click on 'blue box' to zoom in 8x

Perl

standard language for biosequences

- Perl scripting
  - pros:
    - for fast access to various configuration and log files
    - suitable for short to middle programs on structured data
    - huge amount of packages for varius database systems, datastructures, including formats of biological data and connections to biological databases
    - regularily used for parsing datafiles and program outputs in common daily bioinformatics
  - cons:
    - usually hard to read and sustain scripts
    - object oriented approach rather rudimentary

simple perl scripting

- example.pl

```
#/usr/bin/env perl
use strict;
use warnings;

my $var01 = "GATTACA";
$var01 =~ s/T/U/g;
print substr (reverse($var01), 1, 4), "\n";      # CAUU

my @array01 = (3.14, "Pi");
my %hash01 = ("value" => 3.14, "symbol" => "Pi");
print $hash01{"symbol"}, "\n" if 3.14 == $array01[0];
```

BioPerl

modules for sequence-based work in bioinformatics

- modules
  - core, run, dbi packages
    - main parser, standard bio filetypes
    - wrapper around variety bioinformatics tools
    - connecting to biological databases
  - microarray package
    - manipulation of microarray formats (preliminary)
  - other packages
    - for linkage studies, C extensions for align algorithms, etc.

- usage
  - use Bio::Perl;

Filetype conversion

- converter
  - read a given file 'file.seq'
  - takes sequence and writes it in fasta format

```
#/usr/bin/env perl
use strict;
use warnings;
use Bio::Perl;

my $in = Bio::SeqIO->new(-file => "file.seq" ,
'-format' => 'GenBank');
my $out = Bio::SeqIO->new(-file => ">file.fa" ,
'-format' => 'Fasta');

my $seq = $in->next_seq();
$out->write_seq($seq);
```

Sequence retrieval

- example on sequence files
  - takes sequence of human protein il9r
  - makes blast request and write the results

```
#/usr/bin/env perl
use strict;
use warnings;
use Bio::Perl;

my $seq_obj = get_sequence('genbank',"il9r_human");
write_sequence(">il9r.fasta",'fasta',$seq_obj);

my $blast_result = blast_sequence($seq_obj);
write_blast(">il9r.blast", $blast_result);
```

Human gene IDs extraction

```
#/usr/bin/env perl
use strict;          #use with <gene2refseq
use warnings;
use Bio::Perl;

my @cols = (1, 2, 7, 9, 10);
my $col_last = 11;
my $done = 0;
while (<STDIN>) {
.   my @line = split /\s+/, $_;
.   if ("9606" eq $line[0]) {
.    $line[2] = lc $line[2];
.    next if $done >= $line[1];
.    foreach my $col (@cols) {
.      print STDOUT $line[$col], " ";}
.    print STDOUT $line[$col_last], "\n";
}}
```

Data structures

data storage for running programs

- one / several long sequences
  - simple string (array of chars)
  - not to put it into composite structures
    - long time to access a packed string
- sets of all oligonucleotides
  - nucleotides $\rightarrow$ numbers 0...3
  - standard array as a lookup table for the oligos
    - fast access to each cell
- table of gene ids - gaps between ids
  - hash (i.e. associative array)
  - scripting languages with suitable hash data structures

Constraint programming

how to solve problems with constraints

- kind of logic programming
  - declarative programming
    - specify what to do, not how to do it
  - a logic program with constraint specifications
    - setting relations between data
  - specific methods for constraint satisfaction
    - many general solutions but few of them obey the constraints
  - mostly combinatorial problems
    - not the technics for general optimizations

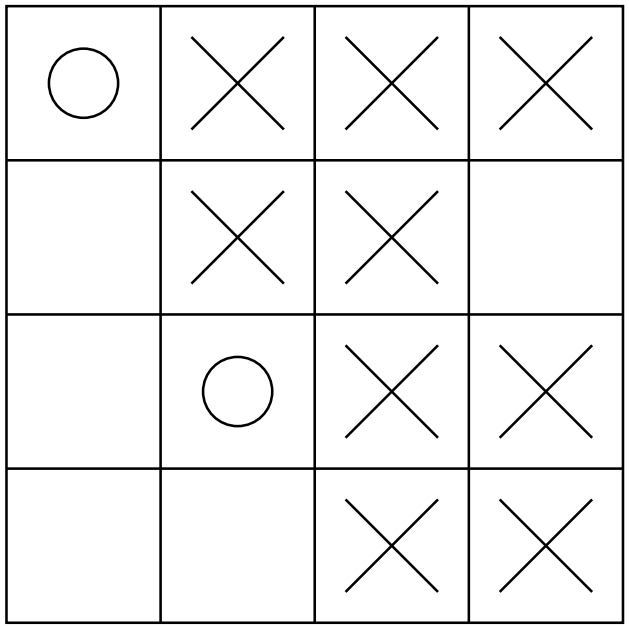
Constraint approaches

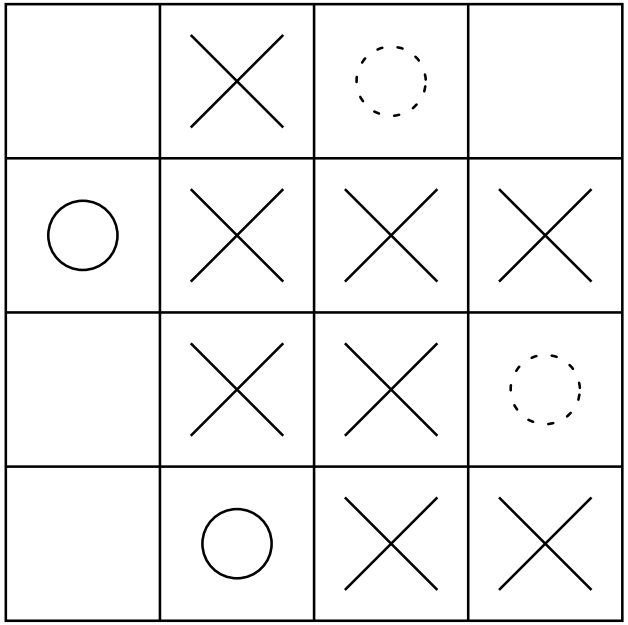
search while consistent

- depth-first search
  - better than starting paths if most of them false
  - back-tracking rather ineffective, to avoid it
  - efficiency with filtering out the wrong paths

- filter ahead
  - set as unaccessible all the recognised wrong paths
  - lesser ways for other (backtracked) search attempts
  - data reduction for a final exhaustive search

Chess board example





- the second attempt leads to the searched result
  - much faster than the search-backtrack approach

Constraints in bioinformatics

problems accessible for CP

- where (not) to use CP
  - sequence alignment - no
    - all of the alignments allowed
    - optimization case, not for the constraint programming
  - clustering, classification - no
    - many possible ways
    - optimization case, not for the constraint programming
  - sequence assembly - not exactly
    - while based on constraints, not a global filtering
  - higher structure prediction - yes
    - suitable connections of short sequences required
  - RNA gene prediction - yes
    - specific sequence characteristics required

Filters

filtering as CP examples

- non-coding RNA prediction
  - each RNA gene contain base-paired sequences
  - complementary sequences with a limited separation
  - $(7\text{nt}, 70\text{nt})$-stack used in FastR software

- structure composition
  - predicted secondary structures should be concatenatable
  - first, to thread short sequences to gain building blocks
  - combinatorial search to concatenate the sequences

Items to remember

Nota bene:

file types, data access

- Database sites
  - sequences, structures
  - expressions, ontologies
- Programming
  - bioperl, conversion
  - constraints, filters