Bioinformatics - Lecture 10

Bioinformatics

Software support

Martin Saturka

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

EBI version 0.4

Software for bioinformatics

Common computation tools and systems in bioinformatics.

- numerical, algebraic and statistical software.
- computation systems specific to bioinformatics.

Main topics

- general software
  - scripting, licenses
  - mpi, sse, gpgpu
- scientific tools
  - emboss, 3D structures
  - algebra, regression, graphs
- R system
  - syntax, statistics
  - packages, examples

Open source

IP: copyrights, trademarks, patents

- software licenses
  - public domain
  - BSD, MIT
  - LGPL, GPL
- multimedia, texts
  - FDL
  - CC - by, sa, nd, nc

- open source licensing
  - Open source initiative
    www.opensource.org/licenses/
  - Creative commons
    creativecommons.org
    sciencecommons.org

Programming

theoretical systems and actual languages

- approaches
    - imperative
        - most standard programming languages
    - declarative
        - functional, logic, constraint programming
- languages
    - compiled
        - low level work: C/C++, Fortran
    - interpreted
        - Python, Tcl/Tk, Perl, Ruby, PHP, Lisp

Floating point

- precision
  - single, double, extended precision, double double, quadruple
- hardware
  - FPU: x87, RISC, pipelines
  - SIMD: altivec, sse, gpgpu

inverted square 'magic'

```
float InvSqrt(float x) {
    float xhalf = 0.5f*x;
    int i = *(int*)&x;           // float → bits
    i = 0x5f3759df - (i>>1);     // guess on result value
    x = *(float*)&i;             // bits → float
    x = x*(1.5f-xhalf*x*x);      // result value adjusting
    return x;}                   // relative error below 0.002
```

MPI

parallel programming

- methods
  - MPI, PVM, threads
  - www.open-mpi.org

```
# include <mpi.h>
...
MPI_Init(&argc, &argv);
MPI_Comm_size(MPI_COMM_WORLD, &ntasks);
MPI_Comm_rank(MPI_COMM_WORLD, &id);
...
MPI_Send(msg, ln, MPI_INT, dest, tag, MPI_COMM_WORLD);
MPI_Recv(msg, ln, MPI_INT, MPI_ANY_SOURCE, tag, MPI_COMM_WORLD, &st);
...
MPI_Finalize();
```

Algebraic methods

non-trivial informatical / numerical methods

- hashing - data storage
  - perfect hashing
    - function from a given constant set of strings to an interval
  - cuckoo hashing
    - simple implementation, usage of two hash functions
- direct minimization
  - linear programming
    - used e.g. for robust (median) regression
  - quadratic programming
    - used e.g. for SVM - support vector machines
- eigen problems
  - eigen-vectors as linear data approximation

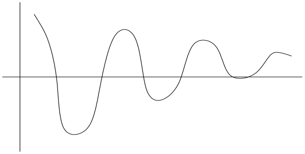
FFT NLLS

Fast Fourier transform with non-linear least squares fitting

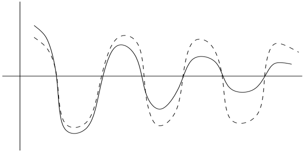
- usage
  - biological cycle / rhytm study
    - period determination, run description
- steps
  - detrending
    - arithmetic mean subtracting
  - normalization
    - variation unification
  - FFT
    - taking the greatest value
  - NLLS
    - (cosine) curve fitting to data

Data transformations

initial data preparation

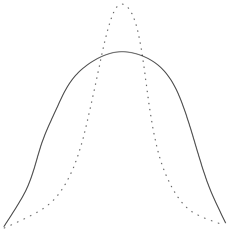


interpolated original data



detrended and standardized data

peak shapes



adjusting after the FFT and period determination are done

Non-linear fitting

regression with a class of non-linear functions

- Levenberg-Marquardt method
  - small errors generally approximated by quadratics
  - iterative method - smooth interpolation between the steepest descent and the inverse Hessian method
  - second derivatives give ’order’ information, first derivatives give the minimizing direction
  - damped iteration along the first derivatives
- software
  - implemented in many packages R system, GSL library, Octave, etc.

Algebra

linear algebra, series, groups, symbolic manipulations

- linear algebra
  - Octave (`www.octave.org`), Scilab (`www.scilab.org`)
  - arpack, lapack, scalapack, blas, atlas
    `www.netlib.org/lapack/`
    `math-atlas.sourceforge.net`
- algebra
  - Maxima, GAP, Pari-GP, Axiom, R, GSL, NumPy
    `maxima.sourceforge.net`
    `www.gnu.org/software/gsl/`

Graphics

general visualization software

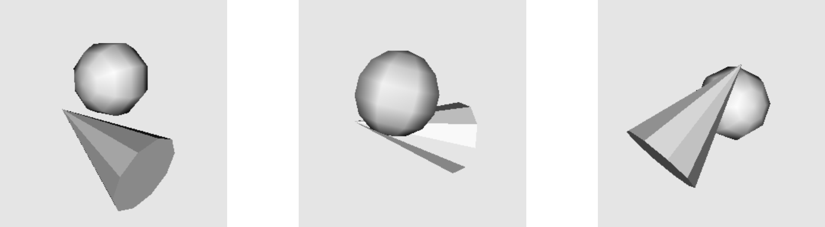
- 2D / graphs / diagrams
    - Graphviz (`www.graphviz.org`)
    - GD, Gnuplot, PLplot
    - XFig, Dia
- 3D / OpenGL graphics
    - VTK (`www.vtk.org`)
    - OpenSceneGraph (`www.openscenegraph.org`)
    - Pov-Ray (`www.povray.org`)
    - Blender, DataExplorer, Mayavi, OpenInventor, etc.

VTK / Tcl interface

```
#!/usr/bin/env wish8.4
package require vtk

vtkSphereSource sphere
sphere SetRadius 1.0
sphere SetCenter 1 1 1
sphere SetThetaResolution 8
sphere SetPhiResolution 8

vtkConeSource cone
cone SetHeight 3.0
cone SetRadius 1.0
cone SetResolution 10

vtkPolyDataMapper sphereMapper
sphereMapper SetInput [sphere GetOutput]
vtkPolyDataMapper coneMapper
coneMapper SetInput [cone GetOutput]

vtkActor sphereActor
sphereActor SetMapper sphereMapper
```

```
vtkActor coneActor
coneActor SetMapper coneMapper

vtkRenderer ren
ren AddActor sphereActor
ren AddActor coneActor
ren SetBackground 0.9 0.9 0.9

vtkRenderWindow renWin
renWin AddRenderer ren
renWin SetSize 300 300

vtkRenderWindowInteractor iren
iren SetRenderWindow renWin
vtkInteractorStyleTrackballCamera style

iren SetInteractorStyle style
iren AddObserver UserEvent \
   wm deiconify .vtkInteract
iren Initialize
wm withdraw .
```

VTK example



- VTK usage
  - suitable for complex 3D data visualization
  - interactive, screen export, many algorithms
  - C++ libs, interface to Python, Tcl/Tk, Java

MM/MD

software for molecular modelling

- classical MM/MD
  - Gromacs (`www.gromacs.org`)
  - Tinker, NAMD/VMD
- molecular visualization
  - RasMol, Raster3D, VieMol, Garlic, PyMol
    `www.openrasmol.org`
    `pymol.sourceforge.net`
- file tools
  - OpenBabel - data formats conversion
    `openbabel.sourceforge.net`

Sequences

standard sequence comparison / manipulation tools

- Blast
  - `www.ncbi.nlm.nih.gov/blast/`
  - `blast.wustl.edu`

- Emboss
  - `emboss.sourceforge.net`
    the European Molecular Biology Open Software Suite
  - usage for:
    sequence alignment, database search, motif identification, sequence patterns, presentation tools

- Clustal X/W, Phylip, Molphy, fastDNAml
  - multiple sequence alignment, phylogenies

Sequence profiles

hidden-stochastic approaches

- HMM methods
  - hmmer
    `hmmer.janelia.org`
    `emboss.sourceforge.net/embassy/hmmer/`
  - build and calibrate models
    align and extract sequences
- CM methods
  - Rfam
    `rfam.janelia.org`
    `www.sanger.ac.uk/Software/Rfam/`
  - sequence alignments, covariant models

Expression clustering

- Cluster
  - program, C library, Python/Perl interface
    `bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm`
  - clustering gene expression data
    k-means clusters, hierarchical clustering
  - original software by Eisen
    `rana.lbl.gov/EisenSoftware.htm`
  - cluster visualization by treeView
    `jtreeview.sourceforge.net`

- Lingua
  - R-system package for gene expression data-mining
    `www.bioplexity.org`
  - relation search and clustering

Open source sites

bioinformatics open-source software sites

- common / bioinformatics repositories
  - bioinformatics.org
    - lists on bioinformatics software, databases, news
  - sourceforge.net
    - general repository of open-source software

- common / bioinformatics projects
  - www.r-project.org
    - general statistics and microarray analysis software
  - www.open-bio.org
    - biological sequences oriented scripting tools

Statistics

statistics methods

- branches
  - explorative / descriptive statistics
    - data characteristics, as mean, variance, etc.
  - confirmative / inferential statistics
    - comparing achieved p-values to α significance (0.05) level
- parametric methods
  - when we assume a known class of (usually normal) distributions of random errors
  - example: Student's t-test
- robust methods
  - tests without assumption of a distribution
  - usually safe, but could be weak on distinguishing
  - example: quantile tests

R system

the standard open-source statistics software

- description
  - system for statistical computing
  - with many statistical tests, modelling, time-series, etc.
  - graphics with suitable 2D/3D plots
  - data models on matrices, arrays, data-frames
  - specific functionality by CRAN packages
- about
  - not for string processing (use Perl/Python/Ruby), not for internal processing of large databases (use respective DBMS)
  - originally S system, now R and S++ systems
  - used commonly for bioinformatics, biostatistics, econometrics

R syntax

- vectors, matrices, arrays for regular data
- data frames: matrix like-structures for database-like tables, i.e. particular columns of possibly different types

```
z1 <- c(2.3, 3.5, 12.1, 4.9, 8.2)
sum(z1)/length(z1); mean(z1); var(z1)
z2 <- 2*z1 - 1
z3 <- array (c(1:24), c(4,6))
z3[1, 3:5] <- NA
z3[is.na(z3)] <- 0

f <- function (x1, x2) {
   x3 <- (x1 * x2)^0.5
   x3
}
f(2,3)
```

R statistical models

dependency description formulae

- ˜ operator for model definition
  - Y ˜ X the Y response depends on X
  - Y ˜ X1 + X2 the Y depends on both X1 and X2
  - Y ˜ X1 - X2 the Y depends on X1, not on X2

linear regression of *y* by *x*:

```
x <- c(2.3, 3.5, 12.1, 4.9, 8.2)
y <- c(4.3, 5.6, 30.0, 12.5, 20.7)
y ˜ x
```

classification analysis of variance:

```
av <- state <- c("one", "one", "two", "one", "two")
A <- factor(av)
y ˜ A
```

classification analysis of covariance:

```
y ˜ A + x
```

R example

simple R usage on (statistics) problems

- linear regression
  - `lm(formula = y ~ x)`
- analysis of variance
  - `aov(formula = y ~ A)`
- Student’s t-test

  `t.test(c(0.1, 0.11, 0.9, 0.8), c(2.1, 2.0, 1.5))`
- graphics
  - `plot(sin, 0, 7)`

R data

methods for reading / writing data

file content:

	col1	col2	col3	...
row1	1.2	8.5	-2.0	...
row2	2.2	-6.1	3.2	...
...				

- tabular data

  ```
  read.table("file", header = TRUE, row.names = 1)
  write.table(dataframe)
  ```

- data import
  - package foreign - for e.g. Octave data
  - relational databases
    - packages RPgSQL, RdbiPgSQL, RSQLite, PL/R
  - BioConductor
    - for microarray data

R extensions

R packages system

- extending R
  - usage of the R language and compiled languages C/C++, fortran
  - extern interface

    ```
    Z <- .Fortran("fncnam", ..., PACKAGE="pkg")
    Z <- .C("functionname", ..., PACKAGE="pkg")
    ```

```
subroutine fncnam(matrix, size1, size2, result)
integer size1, size2
double precision matrix1(size1,size2), result

...
result = 3.14
end
```

R networking

parallel programming interface for R

- R snow
  - simple network of workstations
  - can be used with MPI, PVM, sockets

- functions
  - `library(snow)`

  - `cl <- makeCluster(2, type = "MPI")`
  - `clusterCall(cl, function() Sys.info())`
  - `clusterEvalQ(cl, library(boot))`
  - `clusterApply(cl, 1:2, get("+"), 3)`
  - `stopCluster(cl)`

R packages

the comprehensive R archive network

- CRAN
    - archive of packages for R
        - survival models, time-series
        - bootstrapping, sampling
        - various clustering methods
        - database interfaces
        - quadratic programming

- bioinformatics
    - microarray processing
      bioconductor.org
      bioplexity.org

Python

the current standard open-source scripting language

`www.python.org`

- characteristics
  - can be viewed as a usable Java replacement
  - dynamic, object-oriented, extensible language
  - gluing tool with many usable packages
  - high-level language, not for a low-level work
- package interfaces
  - user interface: TkInter, wxPython
  - database: DBI, SQLAlchemy, SQLObject
  - algorithms: Boost library interface
  - number cruncing: NumPy/SciPy, MPI, RPy

Python example

- sample python code - usage of VTK libraries

```
import vtk

def setImageWrite(self, ftyp, fname):
    if ("PS" == ftyp):
        wobj = vtk.vtkGL2PSExporter()
        wobj.SetFilePrefix(fname)
        wobj.SetRenderWindow(self.renWin)
        wobj.Write()
    else:
        wobj = vtk.vtkPNGWriter()
        ffname = fname + ".png"
        w2i = vtk.vtkWindowToImageFilter()
        w2i.SetInput(self.renWin)
        w2i.Update()
        wobj.SetFileName(ffname)
        wobj.SetInput(w2i.GetOutput())
        self.renWin.Render()
        wobj.Write()
    return
```

RPy

R interface to Python

`rpy.sourceforge.net`

- package description
  - simple / robust interface
  - R objects available in Python
  - all the R functions available
  - R modules available as well

- package usage

```
import rpy
rpy.r.t_test([0.1, 0.11, 0.9, 0.8], [2.1, 2.0, 1.5])
rpy.r.plot(rpy.r.sin, 0, 7)
```

Items to remember

Nota bene:

programming approaches

- Software
  - scripting, algebra systems
  - molecular, bioinformatic tools

- R system
  - statistics models, data
  - packages, python interface