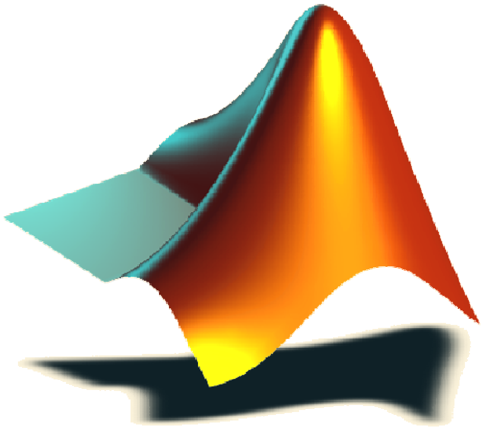
Bioinformatics with MATLAB

Noviembre 18, 2003
Pontificia Universidad Javeriana

Agenda

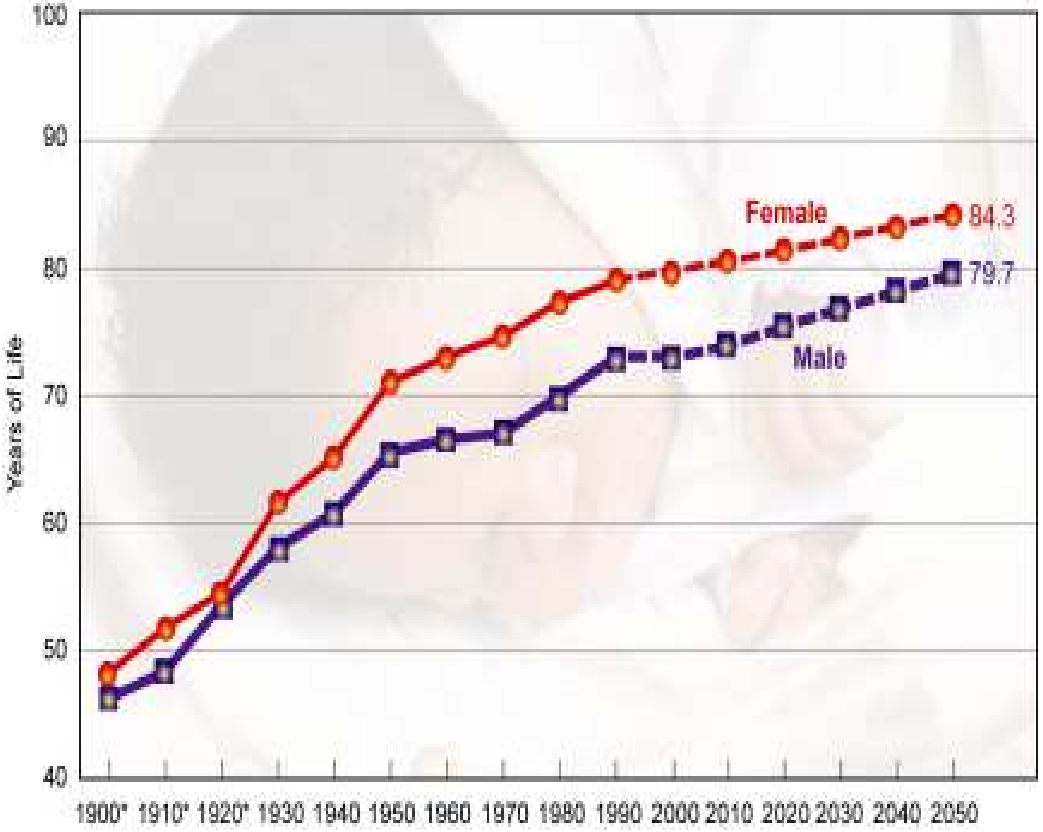
- Bioinformatics an engineering challenge
- Overview of MATLAB®
- The Bioinformatics Toolbox
- Developing and deploying applications with MATLAB.
- Product demonstrations and examples
- Questions and answer session



Improved health care technology leads to increased lifespan - with age bringing more diseases.

- 61 million Americans have some form of cardiovascular disease; 8.5 million Americans who had cancer are alive today.
- Total health care expenditures are 14% of US GDP and rising.
- Health care spending in the United States is projected to reach $3.1 trillion in 2012, up from $1.4 trillion in 2001*
- The worldwide pharma and biotech industries spent $69 billion in 2001 on R&D.
- The US government spent $22 billion on life science R&D in 2001.

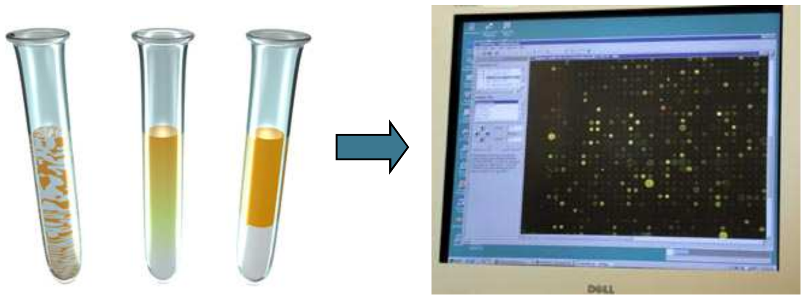
Life Expectancy at Birth USA 1900-2050



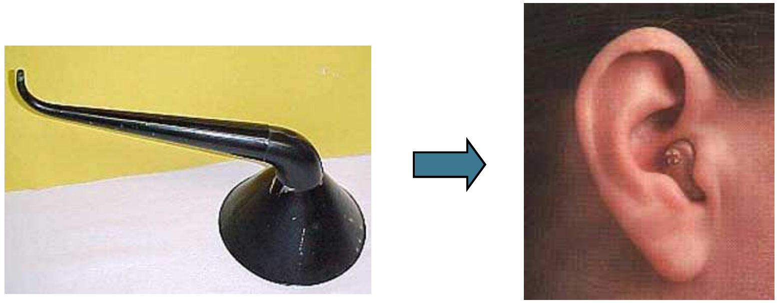
*Centers for Medicare & Medicaid Services

Life Science R&D spending is growing and R&D activities are becoming more quantitative.

- Pharmaceutical and biotech companies are starting to adopt discovery techniques using genomics and bioinformatics and are becoming more dependent on engineering methods.
- Medical instrumentation and devices companies are pushing the boundaries of mechanical, electrical and biomedical engineering and can oftentimes benefit from a variety of engineering disciplines in their work.



Drug Discovery



Medical Devices and Instrumentation

**sources:* American Association for the Advancement of Science, Ernst & Young

Bioinformatics is the application of computational methods to biology.

Combine rapidly evolving biological sciences ...

- Genomics
- Proteomics
- Metabolic pathways

... with computational methods...

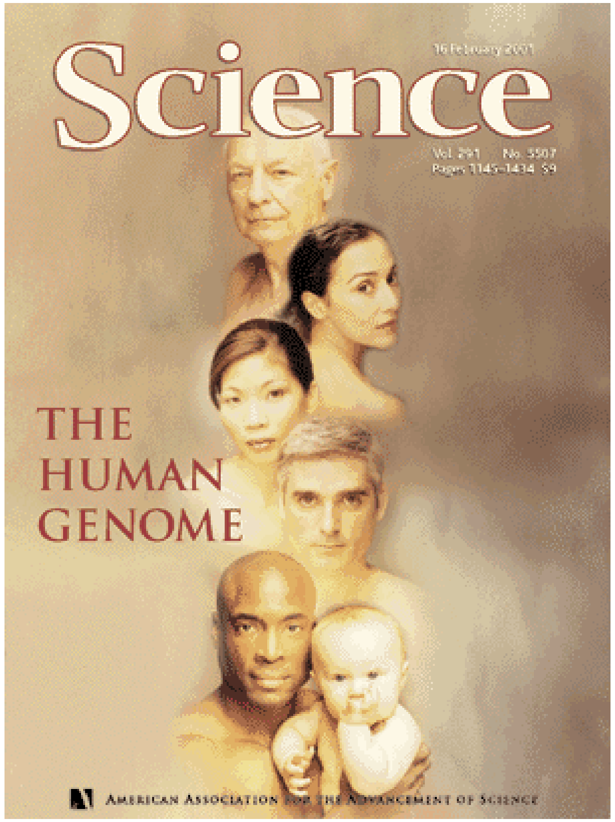
- Gene sequencing (Human Genome Project)
- Expression analysis (DNA microarrays)
- Combinatorial chemistry

... to develop engineered products.

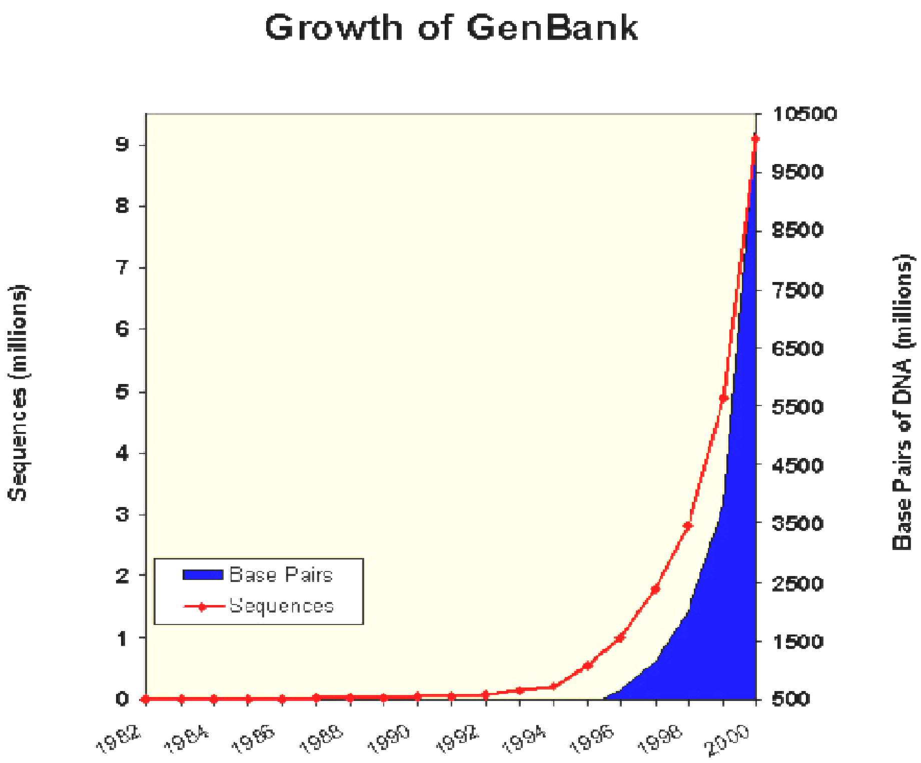
- Main application: automate drug target discovery.
- Basic research into the causes of disease.
- Genetically engineer better crops & livestock.

26-JUN-2000

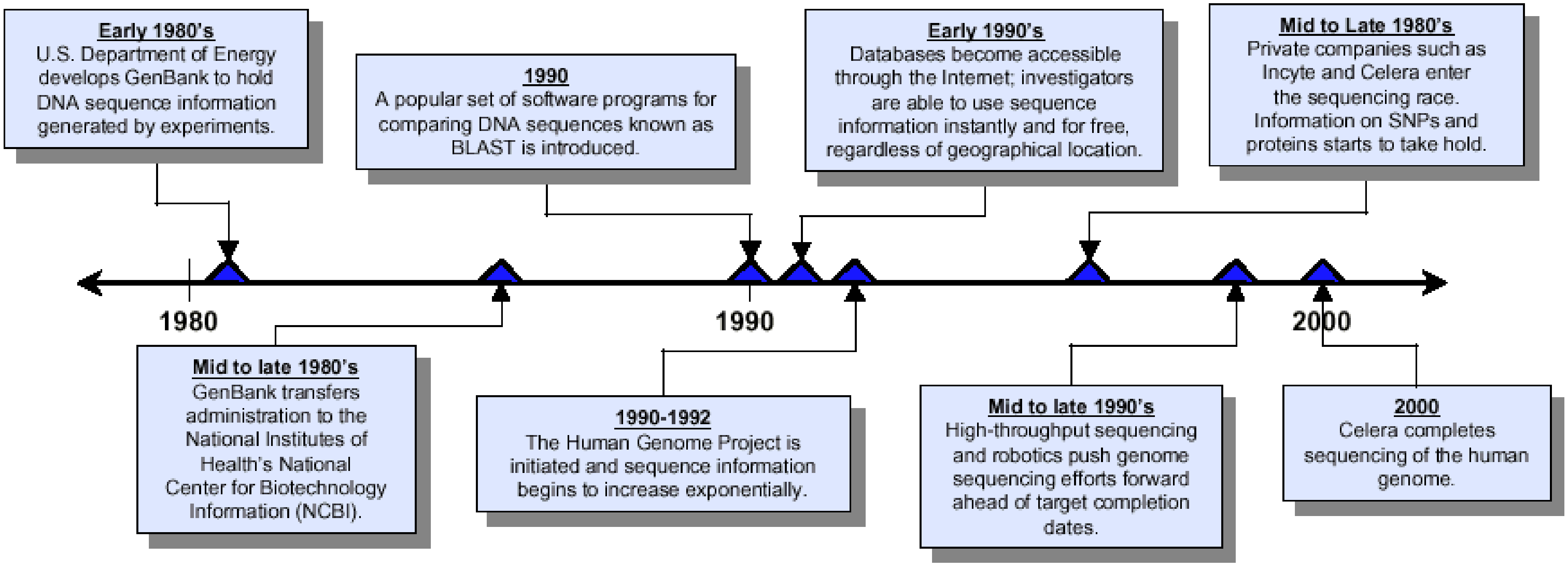
Complete draft of the human genome



Genetic Sequence Information is growing exponentially.



Important Bioinformatics Milestones



Early 1980's
U.S. Department of Energy develops GenBank to hold DNA sequence information generated by experiments.

1990
A popular set of software programs for comparing DNA sequences known as BLAST is introduced.

Early 1990's
Databases become accessible through the Internet; investigators are able to use sequence information instantly and for free, regardless of geographical location.

Mid to Late 1980's
Private companies such as Incyte and Celera enter the sequencing race. Information on SNPs and proteins starts to take hold.

Mid to late 1980's
GenBank transfers administration to the National Institutes of Health's National Center for Biotechnology Information (NCBI).

1990-1992
The Human Genome Project is initiated and sequence information begins to increase exponentially.

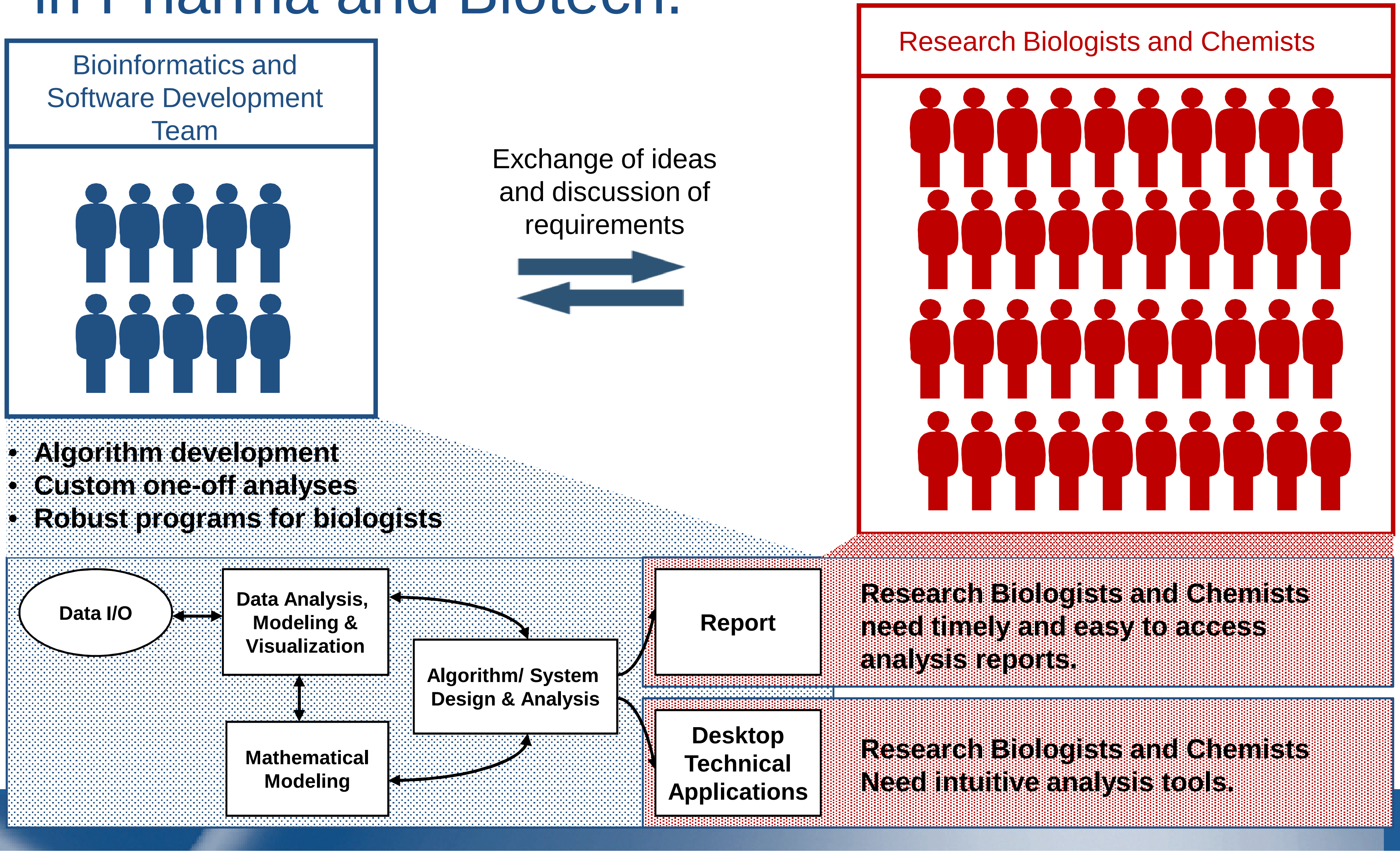
Mid to late 1990's
High-throughput sequencing and robotics push genome sequencing efforts forward ahead of target completion dates.

2000
Celera completes sequencing of the human genome.

Source: Front Line Strategic Consulting, Inc.

The practical challenge of working as a Bioinformatics Specialist

The data intensive discovery process in Pharma and Biotech.



Bioinformatics and Software Development Team

Exchange of ideas and discussion of requirements

Research Biologists and Chemists

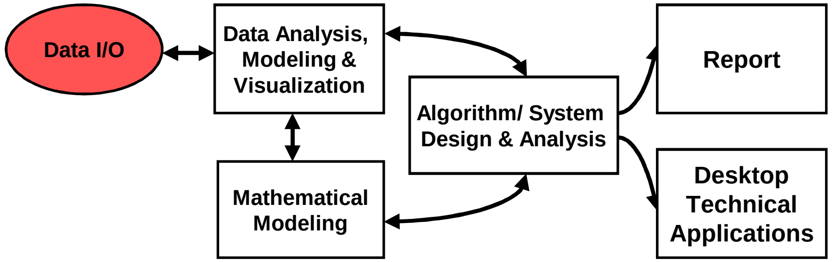
- Algorithm development
- Custom one-off analyses
- Robust programs for biologists

Data I/O
Data Analysis, Modeling & Visualization
Mathematical Modeling
Algorithm/ System Design & Analysis
Report
Desktop Technical Applications

Research Biologists and Chemists need timely and easy to access analysis reports.

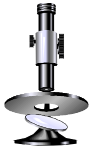
Research Biologists and Chemists Need intuitive analysis tools.

Data access and integration of data sources and applications.


Data I/O
Data Analysis, Modeling & Visualization
Mathematical Modeling
Algorithm/ System Design & Analysis
Report
Desktop Technical Applications

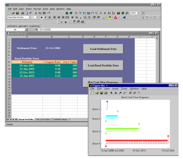
Your Databases — Execute — Read and write and run methods on your databases.

Your Files — Read / Write — Read and write files in most formats

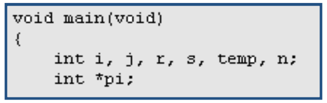


Your instruments that generate data, signals and images — Control — Control Instruments and acquire data

Internet databases you use — Execute — Read from and write to Internet sites



Your existing Excel / COM Applications — Execute — Run or interface to your applications

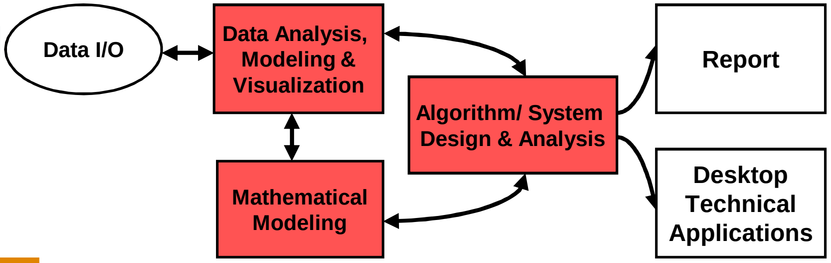


```
void main(void)
{
    int i, j, r, s, temp, n;
    int *pi;
```

Your existing C/C++, Java, Perl programs — Execute — **Run your applications**

Converting formats and integrating systems

Speed up analysis tasks with ready made functionality.



Data I/O

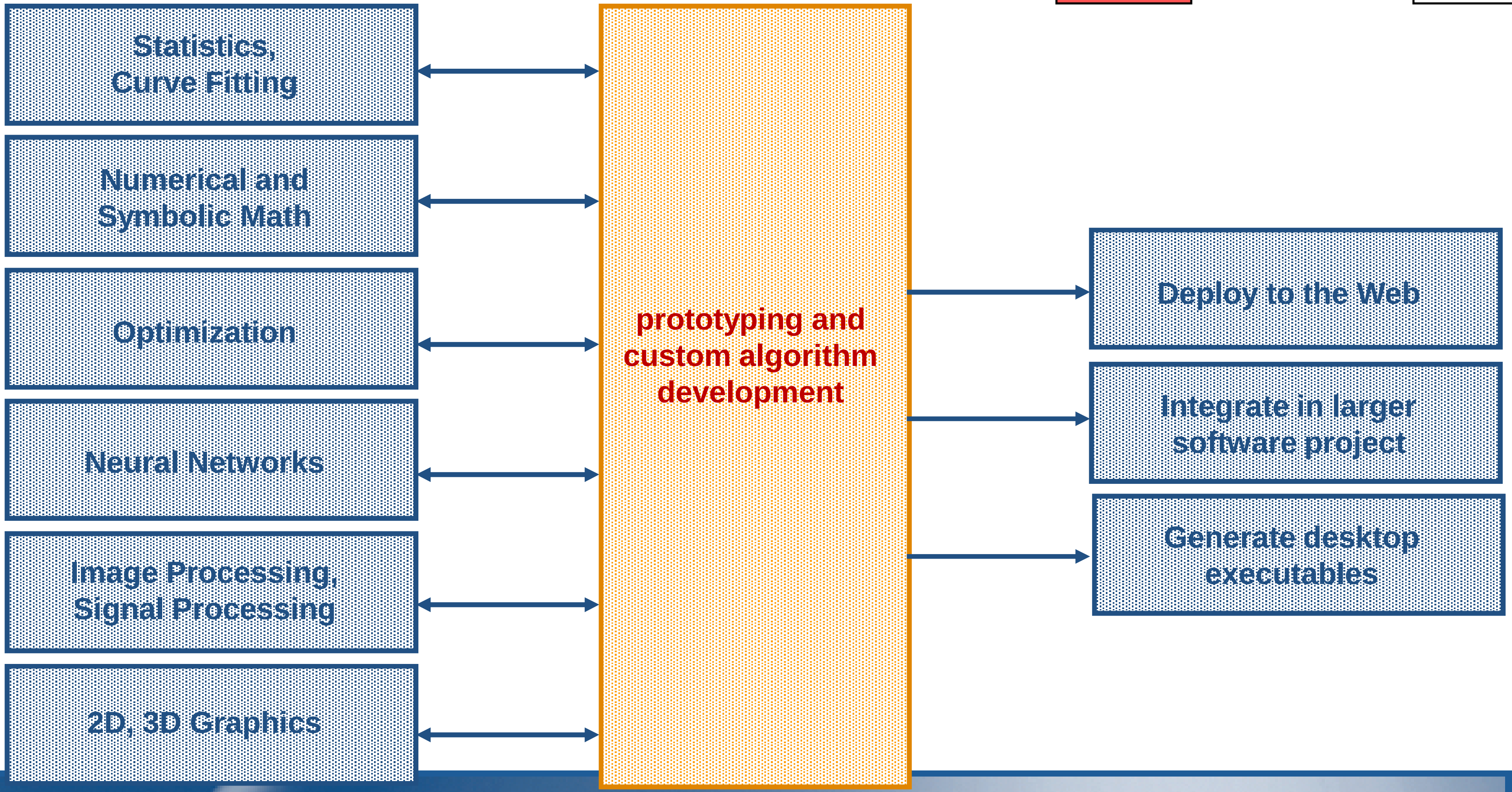
Data Analysis, Modeling & Visualization

Mathematical Modeling

Algorithm/ System Design & Analysis

Report

Desktop Technical Applications



Statistics, Curve Fitting

Numerical and Symbolic Math

Optimization

Neural Networks

Image Processing, Signal Processing

2D, 3D Graphics

prototyping and custom algorithm development

Deploy to the Web

Integrate in larger software project

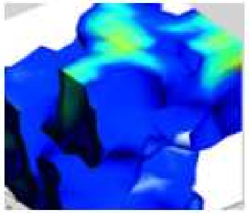
Generate desktop executables

The MathWorks at a Glance

- Founded in 1984, privately held
- Over 1000 employees, including 1/3 in product development
- Revenues exceeding $200M
- More than 500,000 users in 100 countries
- Natick, MA - World Headquarters
  - Product Development
  - Technical Support
- European Offices
  - UK, France, Germany, Italy, Switzerland, Spain, and The Netherlands
- Distributors in 21 countries

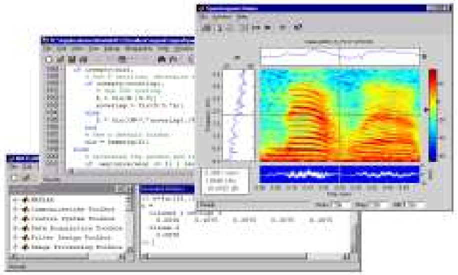
MathWorks Mission and Vision

Accelerate innovation and discovery
in engineering and science



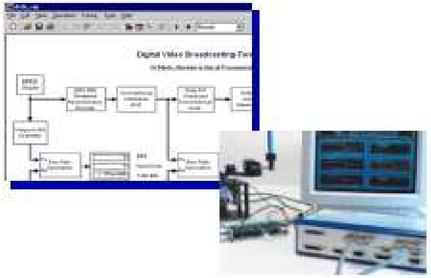
MATLAB

- a powerful, high-level language to develop algorithms, collect and analyze data, and visualize information



Simulink

- a graphical system to model and simulate complex systems, and implement real-time and embedded systems



MathWorks Products are Used in Various Industries

- Aerospace and Defense
- Automotive
- **Biotech, Pharmaceutical and Medical**
- Communications, Semiconductor
- Education
- Financial Services
- Industrial Equipment and Machinery
- Instrumentation
- **Medical Devices and Instrumentation**

Thousands of universities teach students using MathWorks products.

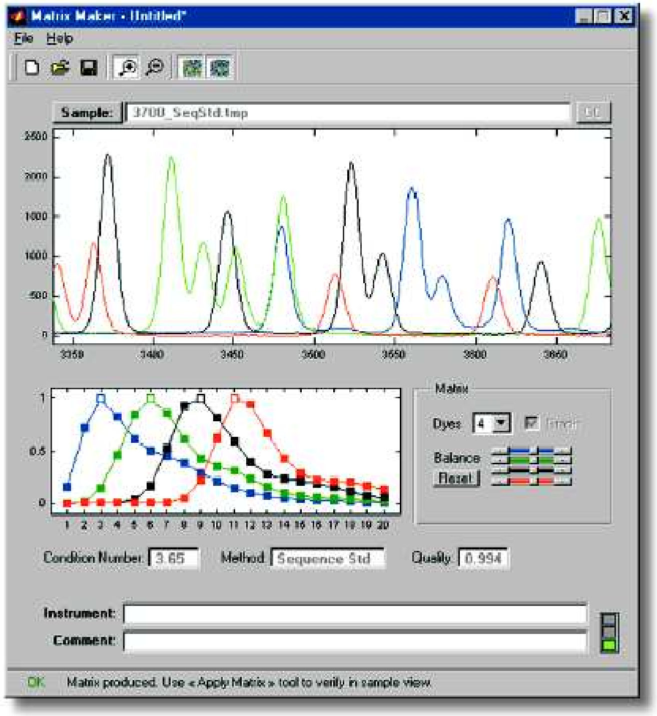
More than 450 textbooks for education and professional use, in 19 languages

- Biosciences
- Controls
- Signal Processing
- Image Processing

- Mechanical Engineering
- Mathematics
- Natural Sciences
- Environmental Sciences

Chandler Allen Phillips, M.D., P.E.
HUMAN FACTORS ENGINEERING
Applied Factor Analysis in the Natural Sciences
Richard Reyment
K. G. Jöreskog
Frank C. Hoppensteadt
Charles S. Peskin
Modeling and Simulation in Medicine and the Life Sciences
Second Edition
Springer
Nonlinear Biomedical Signal Processing
VOLUME II
Dynamic Analysis and Modeling
Edited by METIN AKAY
Biomedical Signal Processing and Signal Modeling
Eugene N. Bruce
MULTIVARIATE BAYESIAN STATISTICS
Models for Source Separation and Signal Unmixing
Daniel B. Rowe
CHAPMAN & HALL/CRC
System Theory and Practical Applications of Biomedical Signals

Technical Applications



Sample: 3700_SeqStd.tmp

Dyes: 4

Condition Number: 3.65 Method: Sequence Std Quality: 0.994

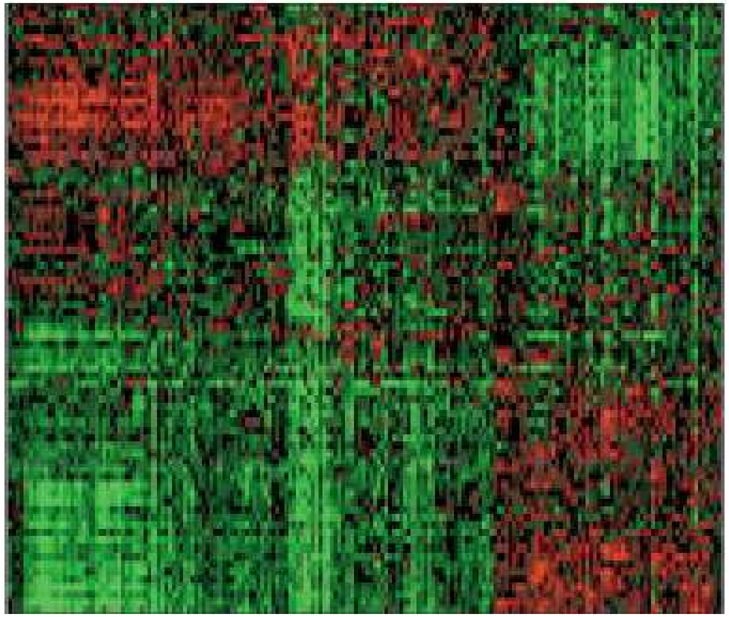
Instrument:

Comment:

OK Matrix produced. Use « Apply Matrix » tool to verify in sample view.

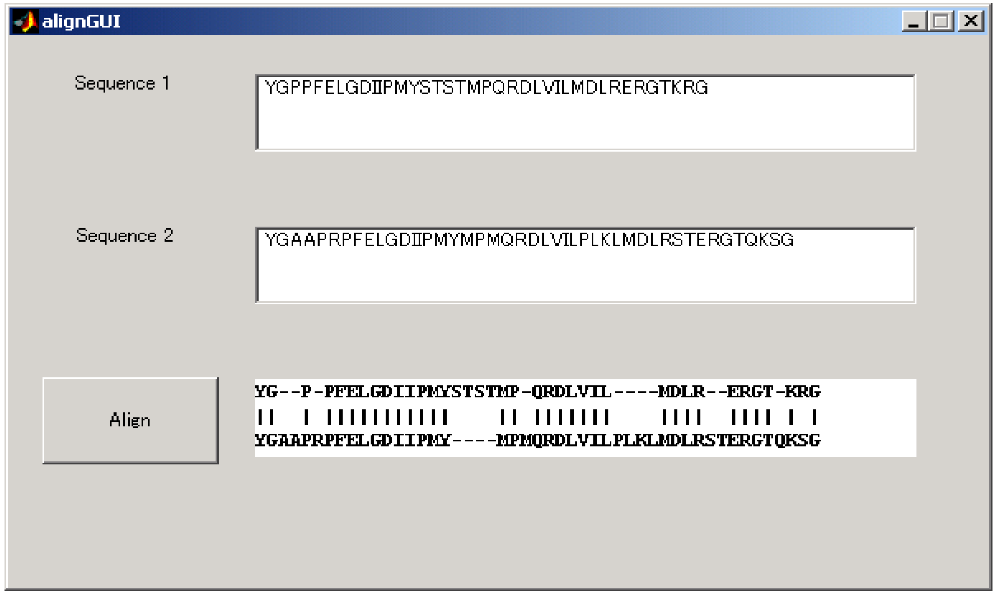
A portion of the DNA dye-label spectral profile, which allows the researcher to read the sequence of bases in a selected strand of DNA.

Rosetta Inpharmatics predicts breast cancer outcome from genetic profile



Sequence Analysis Applications

Deploying a Sequence Analysis Algorithm

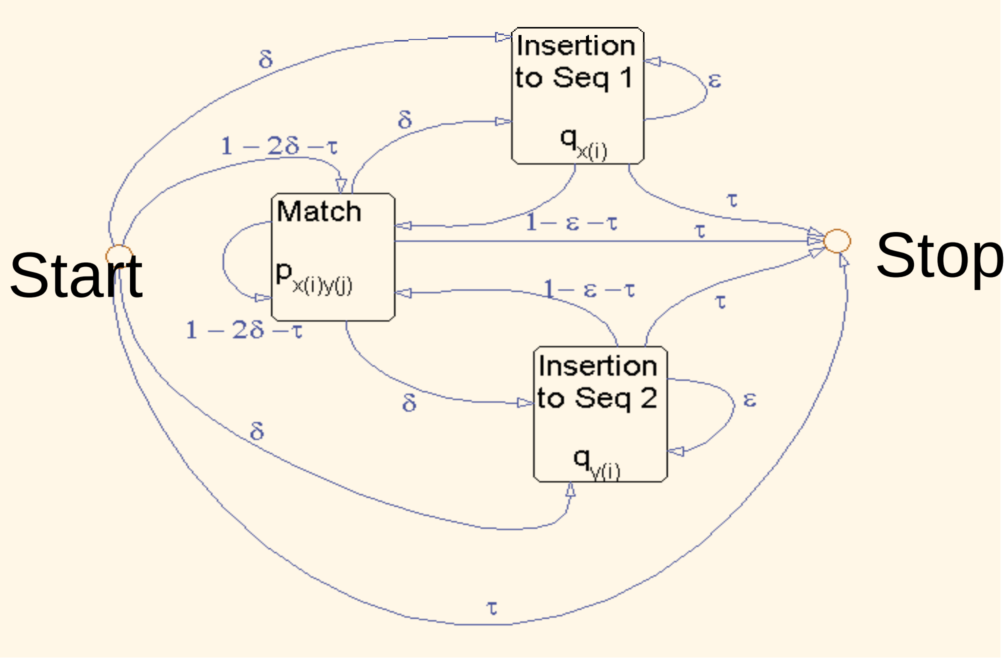


Sequence 1 YGPPFELGDIIPMYSTSTMPQRDLVILMDLRERGTKRG

Sequence 2 YGAAPRPFELGDIIPMYMPMQRDLVILPLKLMDLRSTERGTQKSG

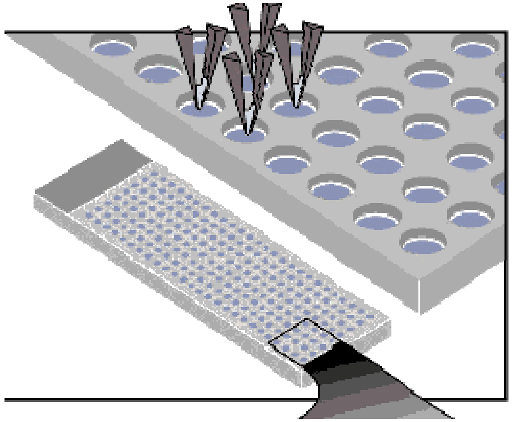
```
YG--P-PFELGDIIPMYSTSTMP-QRDLVIL----MDLR--ERGT-KRG
||  | |||||||||||    || |||||||    ||||  |||| | |
YGAAPRPFELGDIIPMY----MPMQRDLVILPLKLMDLRSTERGTQKSG
```

Hidden Markov Model for Pair-wise Alignment

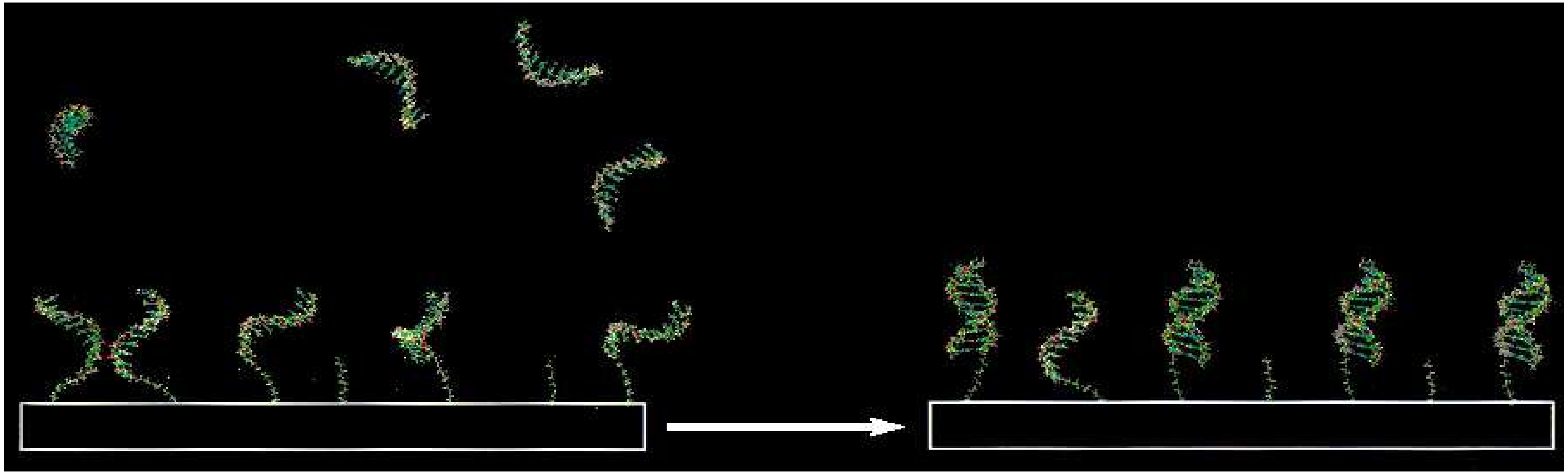


Case Study: Microarray Image Processing

How do Microarrays work?

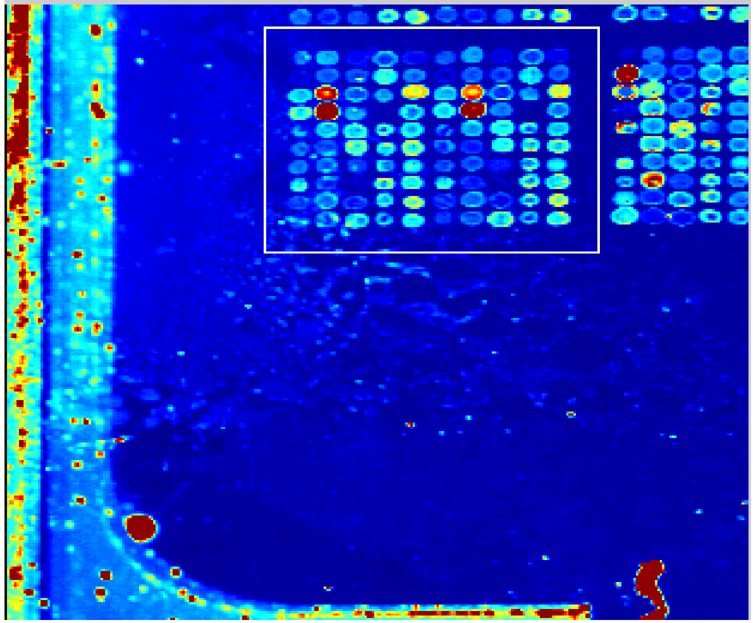


mRNA (messenger RNA) from several cell types are each tagged with a fluor emitting a different color light and then hybridized to an array of cDNA (complementary DNA).



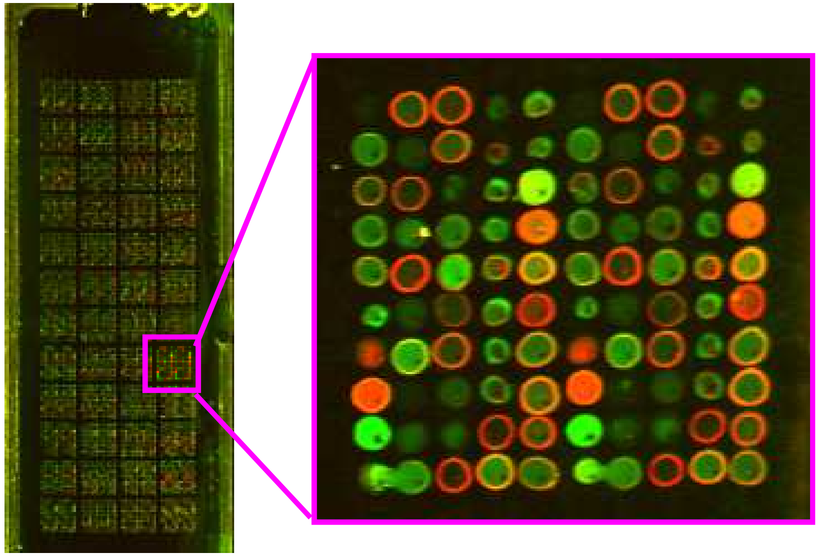
High-throughput experimental techniques require automated image analysis

- Automate image and statistical analysis
- Try out different algorithms
- Build software applications
- Gather quality control measures
- Normalize



Analyzing DNA with Microarray Imaging

Fluorescently tagged mRNA from different cells are hybridized to a microscopic array of hundreds of thousands of cDNA spots that correspond to different genes. Illuminated spots emit different color light, indicating which genes are expressed (e.g., green=control, red=sample, yellow=both).



Through image analysis, the fluorescence at the site of each immobilized cDNA can be quantified. For example, the log ratio of red-to-green intensity gives a measure of gene expression.

Application Challenges

- Clean up images with noise
- Correct for rotation, skew => regular spot spacing (rows, cols)
- Isolate sub-image array of colored spots
- Separate red and green planes
- Remove non-uniform local background
- Identify regular grid pattern of spots on slide
- Address individual spots by region of interest
- Integrate red and green intensity values
- Detect poor spot quality and flag as bad data points
- Determine gene expression from intensities
- Develop robust algorithm to automate process
- Deploy application to implement algorithm.

Solution Algorithm

1. Read image file (**`imread`**)
2. Determine horizontal spot locations (columns)
   a. Create horizontal profile using column averages (**`mean`**)
   b. Remove local background using morphology (**`imtophat`**)
   c. Segment and label spot columns (**`im2bw`**, **`bwlabel`**)
   d. Extract spot centers (**`regionprops`**, **`.Centroid`**)
   e. Calculate column boundaries between spots
3. Transpose image and repeat => spot rows
4. Display detected spot locations on top of image
5. Tabulate spot intensities.

What did this case study show?

1. **MATLAB** environment was great for developing an algorithm (environment + language + graphics)
2. **Image Processing Toolbox** provided a rich set of functions for segmentation, region properties and background removal
3. **Signal Processing Toolbox** provided autocorrelation function to determine spot periodicity.

The Bioinformatics Toolbox

New

Function Overview

- File I/O
  - Read FASTA, PDB, GenePix, Affymetrix and many more format files
- Web connectivity
  - Directly access GenBank, PDB, EMBL, PIR,…
- Sequence analysis
  - Base density, codon counts, ORF finding,…
- Sequence alignment
  - Local, global and profile HMM based alignment
- Microarray normalization & visualization
  - Normalization tools, Gene filters, expression profile cluster analysis,…
- Protein visualization
  - Hydrophobicity plots, Ramachandran plots,…

Getting data into MATLAB

“get” functions retrieve data from Internet based databases.

- `getembl` - Sequence data from EMBL.
- `getgenbank` - Sequence data from GenBank.
- `getgenpept` - Sequence data from GenPept.
- `getpdb` - Sequence data from PDB.
- `getpir` - Sequence data from PIR-PSD.
- `gethmmprof` - HMM from the PFAM database.
- `getgeodata` - Gene Expression Omnibus (GEO) data

Sequence Alignment Tutorial Example

- Get human and mouse genes from GenBank
- Look for open reading frames (ORFs)
- Convert DNA sequences to amino acid sequences
- Create a dotplot of the two sequences
- Perform global alignment
- Perform local alignment

Microarray Data Analysis Tutorial Example

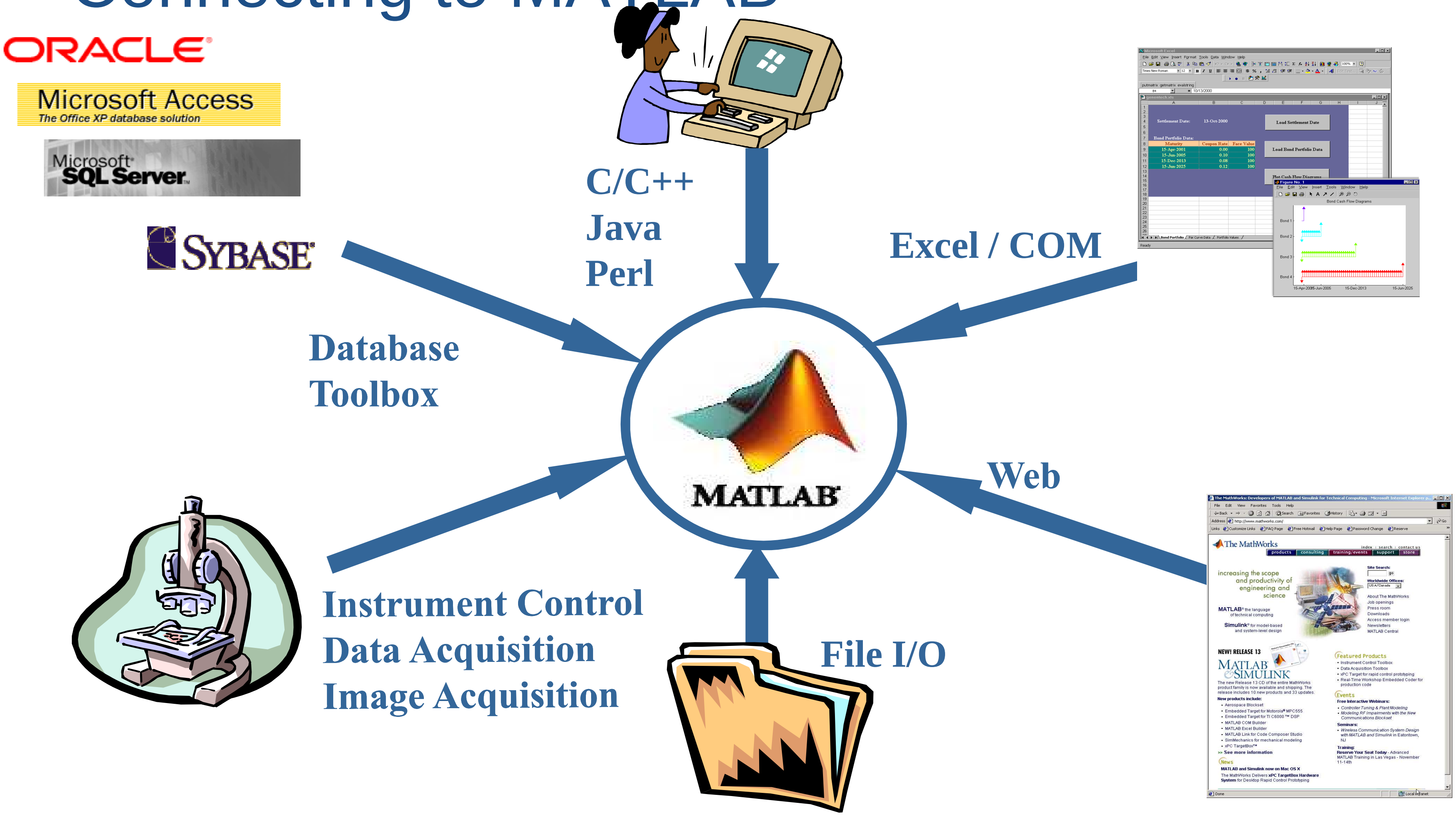
- Plot expression profiles for genes
- Filter genes based on information content of profile
- Perform hierarchical clustering
- Perform K-means clustering
- Perform Principal Component Analysis

Reference:

DeRisi, JL, Iyer, VR, Brown, PO. "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science. 1997 Oct 24;278(5338):680-6.

Integrating and Deploying Bioinformatics Tools with MATLAB

Connecting to MATLAB



C/C++
Java
Perl

Excel / COM

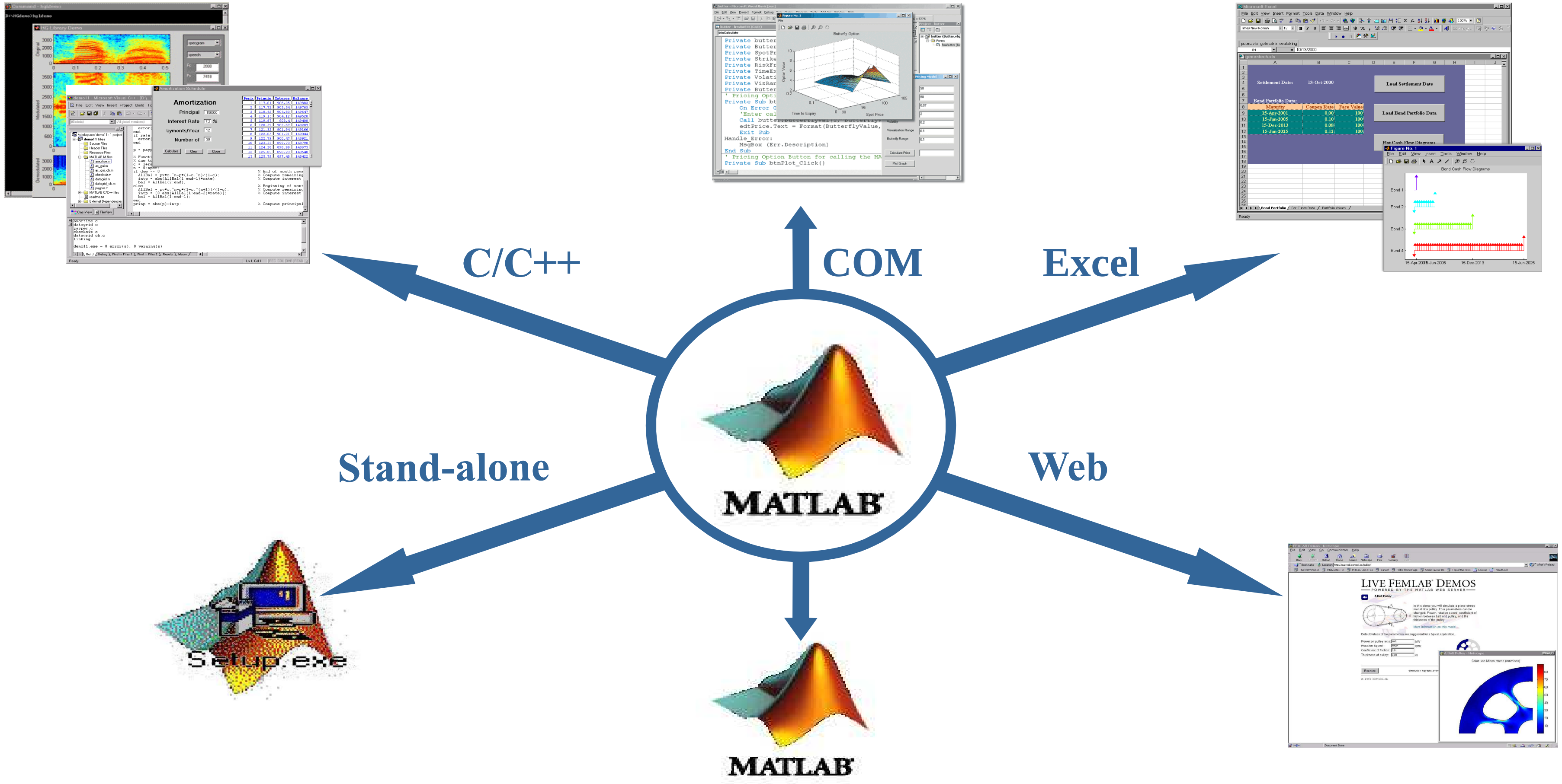
Database
Toolbox

Web

Instrument Control
Data Acquisition
Image Acquisition

File I/O

Deploying with MATLAB



C/C++

COM

Excel

Stand-alone

Web

Database Connections

ORACLE

Microsoft Access
The Office XP database solution

Microsoft SQL Server

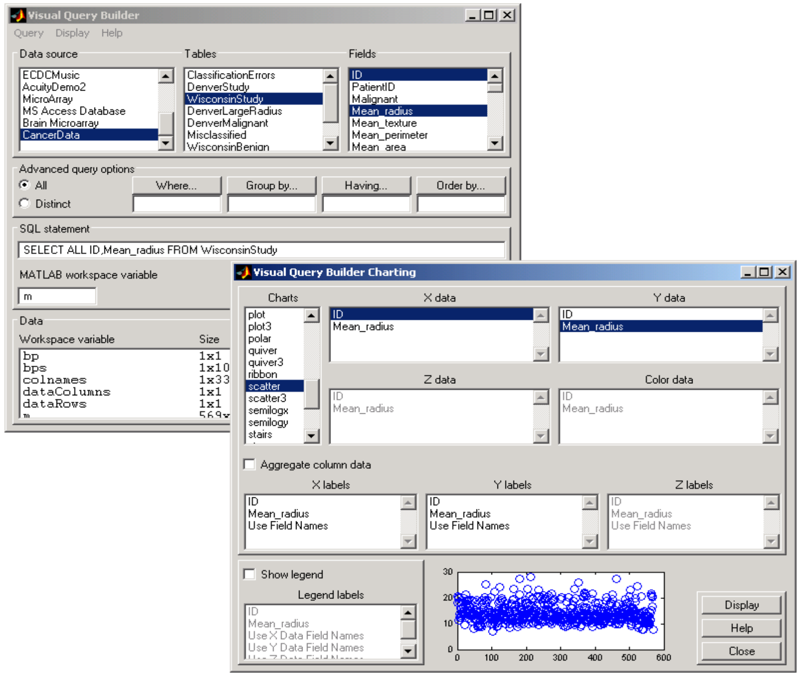
SYBASE

- ODBC or JDBC compliant database
  - ODBC and JDBC on PC
  - JDBC on UNIX
- Data types are preserved
- Retrieval of large/partial data sets
- Access multiple connections (same or different DB)
- Database connections remain open

Database Connections

Visual Query Builder

- ■Access data without knowing SQL
  - •Scroll through tables and fields
  - •Customize your query
- ■Built-in visualization tools
  - •Plotting and charting
  - •Creating HMTL reports
  - •Handling date strings
- ■Reuse SQL statements in your own program



Data source: ECDCMusic, AcuityDemo2, MicroArray, MS Access Database, Brain Microarray, CancerData

Tables: ClassificationErrors, DenverStudy, WisconsinStudy, DenverLargeRadius, DenverMalignant, Misclassified, WisconsinBenign

Fields: ID, PatientID, Malignant, Mean_radius, Mean_texture, Mean_perimeter, Mean_area

Advanced query options

● All

○ Distinct

SQL statement

```
SELECT ALL ID,Mean_radius FROM WisconsinStudy
```

MATLAB workspace variable: m

Data

```
Workspace variable     Size
bp                     1x1
bps                    1x10
colnames               1x33
dataColumns            1x1
dataRows               1x1
m                      569x
```

Charts: plot, plot3, polar, quiver, quiver3, ribbon, scatter, scatter3, semilogx, semilogy, stairs

X data: ID, Mean_radius

Y data: ID, Mean_radius

Z data: ID, Mean_radius

Color data: ID, Mean_radius

□ Aggregate column data

X labels: ID, Mean_radius, Use Field Names

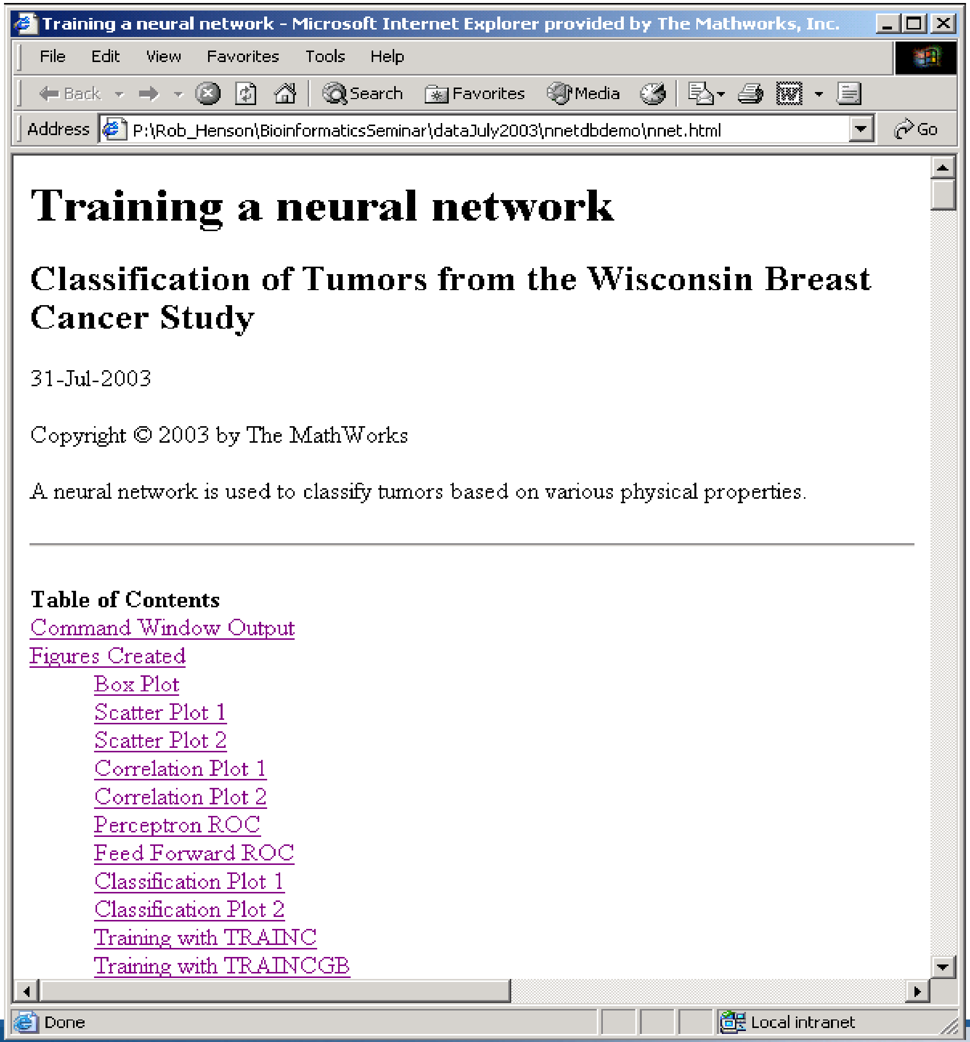
Y labels: ID, Mean_radius, Use Field Names

Z labels: ID, Mean_radius, Use Field Names

□ Show legend

Legend labels: ID, Mean_radius, Use X Data Field Names, Use Y Data Field Names

Customized Reports



Training a neural network - Microsoft Internet Explorer provided by The Mathworks, Inc.
File Edit View Favorites Tools Help
Back Search Favorites Media
Address P:\Rob_Henson\BioinformaticsSeminar\dataJuly2003\nnetdbdemo\nnet.html
Go
Done
Local intranet

Training a neural network

Classification of Tumors from the Wisconsin Breast Cancer Study

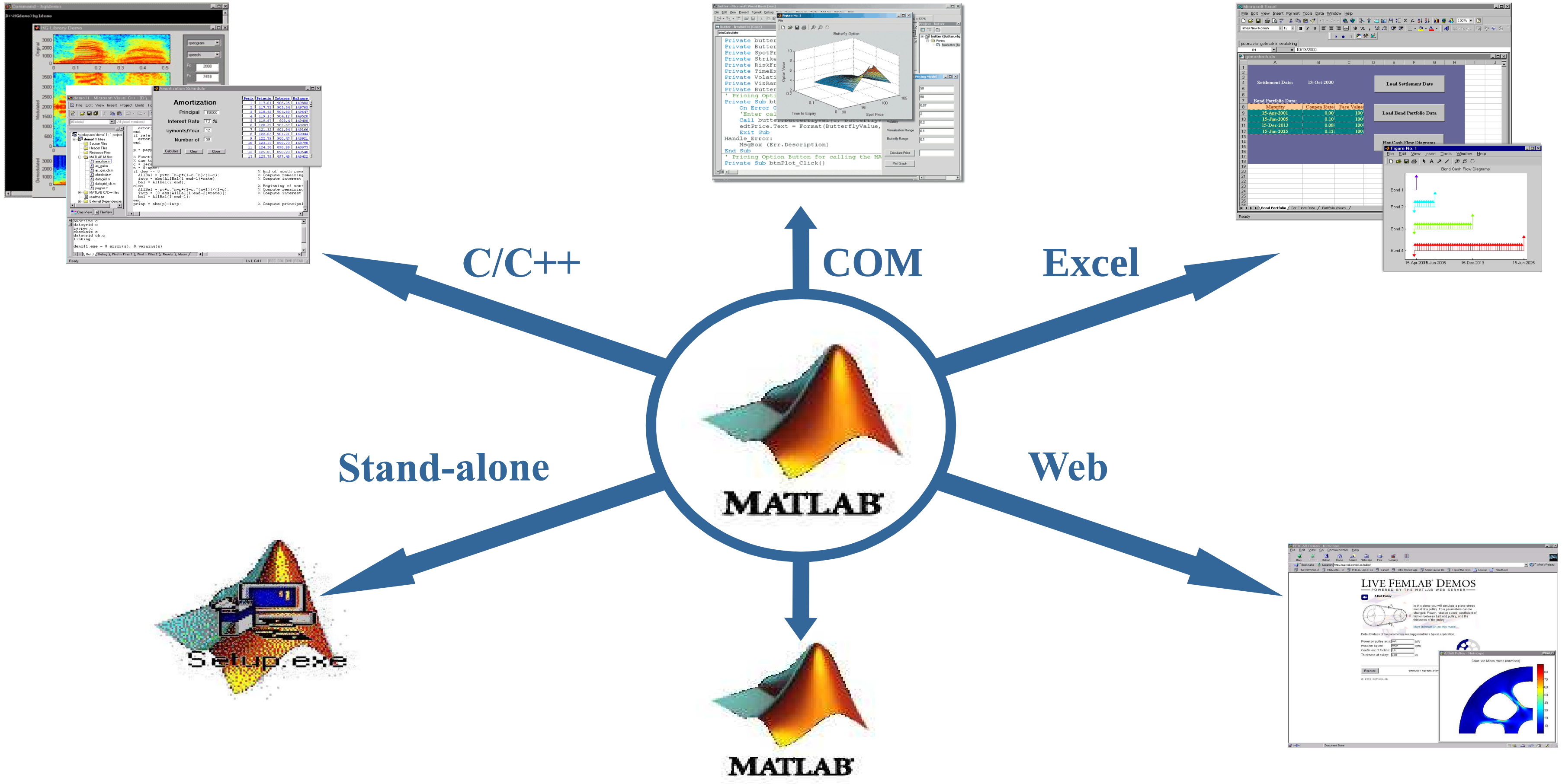
31-Jul-2003

Copyright © 2003 by The MathWorks

A neural network is used to classify tumors based on various physical properties.

Table of Contents

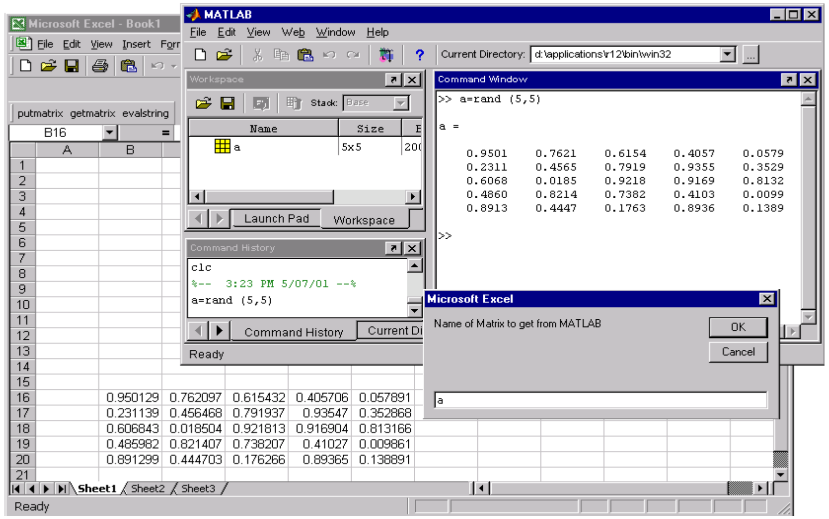
Deploying with MATLAB



Push Data into MATLAB

Data I/O

- Import Excel ranges into MATLAB
- Export MATLAB data into Excel ranges
- Evaluate MATLAB Statements in Excel



```
>> a=rand (5,5)

a =

    0.9501    0.7621    0.6154    0.4057    0.0579
    0.2311    0.4565    0.7919    0.9355    0.3529
    0.6068    0.0185    0.9218    0.9169    0.8132
    0.4860    0.8214    0.7382    0.4103    0.0099
    0.8913    0.4447    0.1763    0.8936    0.1389

>>
```

```
clc
%--  3:23 PM 5/07/01 --%
a=rand (5,5)
```

Microsoft Excel

Name of Matrix to get from MATLAB: a

OK | Cancel

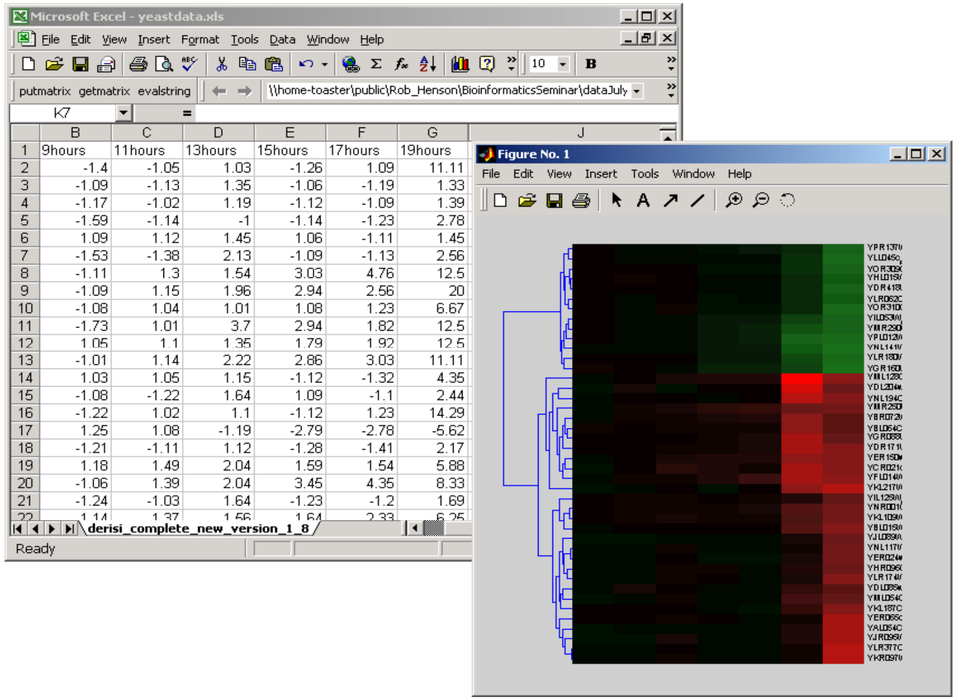
	A	B				
16		0.950129	0.762097	0.615432	0.405706	0.057891
17		0.231139	0.456468	0.791937	0.93547	0.352868
18		0.606843	0.018504	0.921813	0.916904	0.813166
19		0.485982	0.821407	0.738207	0.41027	0.009861
20		0.891299	0.444703	0.176266	0.89365	0.138891

Computational Engine for Excel

Spread Sheet Applications

- MATLAB Excel Link can be the computational engine behind your Excel applications

- Fast scalable solution


Microsoft Excel - yeastdata.xls
File Edit View Insert Format Tools Data Window Help
putmatrix getmatrix evalstring
\\home-toaster\public\Rob_Henson\BioinformaticsSeminar\dataJuly
K7
=
derisi_complete_new_version_1_8
Ready
Figure No. 1
File Edit View Insert Tools Window Help

	B	C	D	E	F	G
1	9hours	11hours	13hours	15hours	17hours	19hours
2	-1.4	-1.05	1.03	-1.26	1.09	11.11
3	-1.09	-1.13	1.35	-1.06	-1.19	1.33
4	-1.17	-1.02	1.19	-1.12	-1.09	1.39
5	-1.59	-1.14	-1	-1.14	-1.23	2.78
6	1.09	1.12	1.45	1.06	-1.11	1.45
7	-1.53	-1.38	2.13	-1.09	-1.13	2.56
8	-1.11	1.3	1.54	3.03	4.76	12.5
9	-1.09	1.15	1.96	2.94	2.56	20
10	-1.08	1.04	1.01	1.08	1.23	6.67
11	-1.73	1.01	3.7	2.94	1.82	12.5
12	1.05	1.1	1.35	1.79	1.92	12.5
13	-1.01	1.14	2.22	2.86	3.03	11.11
14	1.03	1.05	1.15	-1.12	-1.32	4.35
15	-1.08	-1.22	1.64	1.09	-1.1	2.44
16	-1.22	1.02	1.1	-1.12	1.23	14.29
17	1.25	1.08	-1.19	-2.79	-2.78	-5.62
18	-1.21	-1.11	1.12	-1.28	-1.41	2.17
19	1.18	1.49	2.04	1.59	1.54	5.88
20	-1.06	1.39	2.04	3.45	4.35	8.33
21	-1.24	-1.03	1.64	-1.23	-1.2	1.69
22	1.14	1.37	1.56	1.64	2.33	6.25

```
MLPutMatrix("data",B2:H43)
MLPutMatrix("Genes",A2:A43)
MLPutMatrix("TimeSteps",B1:H1)
MLEvalString("clustergram(data,'RowLabels',…
                            Genes,'ColLabels',TimeSteps)")
```

Summary

- Read and write to a database
  - Powerful math and data analysis
  - Generate custom reports
- Create standalone applications
- Easy integration with Excel
  - MATLAB as a computational engine
  - Create Excel Plug-ins in minutes

Industry Issues & Solutions

- Integrating tools from various programming languages is difficult, closed source tools are not customizable, and freeware is often not supported.
- There is no standard biological data format.
- Applications must be easily deployable within organizations.

- MATLAB is a supported, viewable source, user-friendly environment for data analysis across applications, algorithm development, and deployment.
- MATLAB and the Bioinformatics Toolbox provides file format support for common data sources (web-based, sequences, microarray, etc.).
- MATLAB’s deployment tools and user-interface design environment allow easy deployment of MATLAB based applications.

Further Information

- Product Information and Demos

  Trials and technical literature are available through the MathWorks. www.mathworks.com

- MATLAB Central
  - File exchange and newsgroup access for MATLAB and Simulink users
  - www.mathworks.com/matlabcentral
  - Access to comp.soft-sys.matlab

MATLAB® CENTRAL

file exchange and newsgroup access for
the MATLAB & Simulink user community

Visit www.mathworks.com

Free trials and technical literature are available through the MathWorks