Information Systems for Genetics

Erwin M. Bakker

ISLS 23-9 2004

Overview

- DIAL
- Sequence Alignment
  - BLAST, FASTA, etc.
- Genebrowsing
  - TIGR, MeV
  - Ensembl
  - http://www.genebrowser.com/

DIAL

CMSB

- CMSB (Centre for Medical Systems Biology, www.cmsb.nl) is one of the Genomics Centre of Excellence
  - Genomics for identifying hidden connections between diseases: Improving diagnosis, treatment and prevention of common diseases *such as Alzheimer's, cardiovascular disease, diabetes and rheumatism.*
- DIAL (Data Integration, Analysis and Logistics) The project works on the data of the experimental projects of the CMSB.

DIAL Institutes

- Leiden University Medical Center
- Leiden University
- VU University medical center
- Vrije Universiteit Amsterdam
- TNO Pharma Leiden
- Erasmus MC Rotterdam

DIAL Example Study Groups

- RotterdamStudy (ERGO: Hofman, van Duijn, a.o.)
  - population-based cohort study of 12,000 subjects aged 55+ years. Patients have been followed for over 10 years now.
- Grip Cohort Study (Rotterdam: van Duijn, Oostra)
  - population-based cohort study of 3 generation families (2500 subjects). They are screened for the presence of multiple diseases.
- Netherlands Twin Register (Boomsma VU Amsterdam)
  - number of twins (60,670) and siblings (3,175)

CMSB Center for Medical Systems Biology

Six Projects

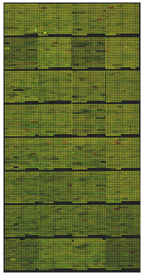
- Epidemiology: cohorts & genotyping
- Systems Biology: transcriptomics/arraying proteomics metabolomics
- Technology: magnetic resonance microscopy other imaging molecular interactions
- Model Systems: animal models (mouse, zebra fish etc).
- Clinical Applications : translation (cells, vaccines, viral, pharmaceutical)
- DIAL: Data Integration, Analysis and Logistics

DIAL Data Integration, Analysis and Logistics

- database/data mining group
- genomical statistics
- biomolecular bioinformatics
- computational systems biology

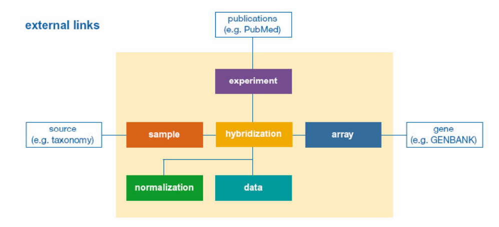
DIAL Micro array CGH

- CGH (Comparative Genomic Hybridization) micro array experiments
- Data collection
- Data processing (Normalization)
- Data storage
- Data selection
- Data integration



DIAL Micro Array CGH

MIAME minimum information about a micro array experiment



external links

publications (e.g. PubMed)

experiment

source (e.g. taxonomy)

sample

hybridization

array

gene (e.g. GENBANK)

normalization

data

Fig. 3 A schematic representation of six components of a microarray experiment.

DIAL Interfaces

DIAL Interface

- GenePix, ImageGene
- Normalization Routines
- Smoothing
- BAC (Bacterial Artificial Chromosomes) updates
- CGHAnalyzer, MeV
- Rosetta, Ensembl, etc.
- Data mining

Sequence Alignment

- sequence similarity <-> homology
- database queries; comparative genomics
- different possible alignments
- different quality of alignments

```
ACTGTGACCATATCG     ACTGTGACCATATCG
      ||X||  |      |||X|X
      ACTAT-T       ACTATT
```

Sequence Alignment

- different possible alignments
- different quality of alignments -> score/metric
- match/mismatch: PAM, BLOSUM matrices
- insertions and deletions =>gaps: gap-penalty, gap-extension-penalty

```
ACTGTGACCATATCG     ACTGTGACCATATCG
      ||X||  |      |||X|X
      ACTAT-T       ACTATT
```

Sequence Alignment

- what kind of alignments are valid
- scoring system for ranking the alignments
- the algorithm to find optimal or good alignments using the scoring mechanism
- statistical evaluation of the significance of an alignment score

Sequence Alignment

- There are ~2^{2n}/sqrt(pi.n) different alignments between to sequences of length n (too much)
- Dynamic Programming
  - O(nm) (n<=m) (time and space)
  - O(km) (differences are bounded by k <<n)
- Exclusion Algorithms: fast expected time
- FASTA
- BLAST
  - sequence alignment
  - Heuristics
- Hidden Markov Models (model a family of sequences)
- Chomsky Hierarchy
- Stochastic Context Free Grammars (structure prediction)

Exclusion Algorithms

- T text, P pattern
- **Partition** P in consecutive regions of length r
- **Search** T to find length-r intervals R that could be contained in an approximate occurrence of P (surviving intervals)
- **Check** for each surviving interval R if there is an approximate occurrence of P in a larger interval around R

It is expected that already a large part of T is excluded from the **check** phase so that only (sub) linear time is needed here.

FASTA

- A heuristic exclusion method using dynamic programming
- D.J. Lipman, W.R. Pearson Rapid and sensitive protein similarity searches. Science, 227:1435-41, 1985
- W.R. Pearson, D.J. Lipman Improved tools for biological sequence comparison. Proc. Natl. Academy Science, 85, 2444-48, 1988

FAST A (Fast All)

- query string S; text string T
- k: is the length of the hot-spots in the dynamic programming table (k=6 for DNA, k=2 for Protein)

1. find pairs (i,j) such that the substring of length k starting at position i in S exactly matches the substring of length k at position j in T (using lookup table), and then look for diagonals with many supporting word matches (sort the matches on the difference (i-j))
2. The best diagonals from 1) are pursued: extend with ungapped alignment to find maximal scoring ungapped regions
3. join ungapped regions with gaps and find highest scoring matches using dynamic programming

- (full local dynamic programing, if k=1)

BLAST (Basic Local Alignment Search Tool)

- Altschul et al., 1990
- O(mn) is too slow
- Probabilistic approach to searching
- True alignments will have short stretches of perfect match

BLAST (Basic Local Alignment Search Tool)

- make a list of words W in P (proteins: length 3; nucleic acids: 11) that match T with score > threshold
- scan T for occurrences of W
- if a hit is obtained extend the match in both directions (no gaps), stopping at maximum scoring extension

=> finds ungapped alignments only

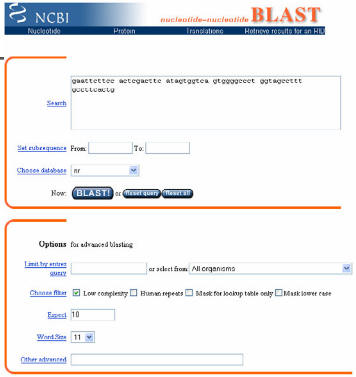
newer versions of Blast => gapped alignments

BLOSOM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

BLAST Queries

- NCBI http://www.ncbi.nlm.nih.gov/
- Ensembl http://www.ensembl.org/



Search: gaattcttcc actcgacttc atagtggtca gtggggccct ggtagccttt gccttcactg

Set subsequence From: To:

Choose database: nr

Now: BLAST! or Reset query Reset all

Options for advanced blasting

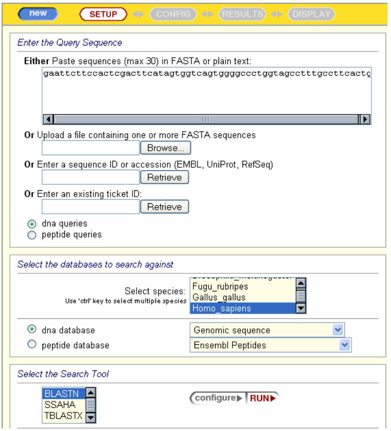
Limit by entrez query: or select from: All organisms

Choose filter: ☑ Low complexity ☐ Human repeats ☐ Mask for lookup table only ☐ Mask lower case

Expect: 10

Word Size: 11

Other advanced:



Enter the Query Sequence

Either Paste sequences (max 30) in FASTA or plain text:

gaattcttccactcgacttcatagtggtcagtggggccctggtagcctttgccttcactg

Or Upload a file containing one or more FASTA sequences Browse...

Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq) Retrieve

Or Enter an existing ticket ID: Retrieve

● dna queries
○ peptide queries

Select the databases to search against

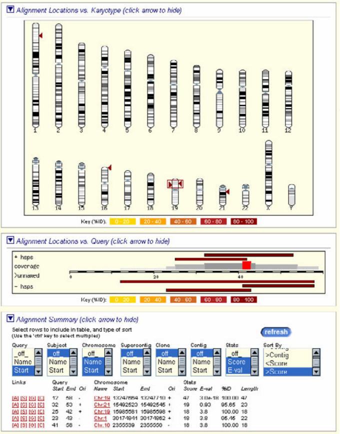
Select species: Fugu_rubripes / Gallus_gallus / Homo_sapiens
Use 'ctrl' key to select multiple species

● dna database: Genomic sequence
○ peptide database: Ensembl Peptides

Select the Search Tool

BLASTN / SSAHA / TBLASTX

configure ▸ RUN ▸


Alignment Locations vs. Karyotype (click arrow to hide)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Key (%ID): 0 - 20 20 - 40 40 - 60 60 - 80 80 - 100
Alignment Locations vs. Query (click arrow to hide)
+ hsps
coverage
>unnamed
0
20
40
- hsps
Key (%ID): 0 - 20 20 - 40 40 - 60 60 - 80 80 - 100
Alignment Summary (click arrow to hide)
refresh

Select rows to include in table, and type of sort
(Use the 'ctrl' key to select multiples)

Query: _off_ / Name / Start — Subject: off / Name / Start — Chromosome: _off_ / Name / Start — Supercontig: off / Name / Start — Clone: off / Name / Start — Contig: off / Name / Start — Stats: _off_ / Score / E-val — Sort By: >Contig / <Score / >Score

Links	Query			Chromosome				Stats			
	Start	End	Ori	Name	Start	End	Ori	Score	E-val	%ID	Length
[A] [S] [G] [C]	12	58	-	Chr:19	13247664	13247710	+	47	3.0e-18	100.00	47
[A] [S] [G] [C]	32	53	+	Chr:21	15492523	15492545	+	19	0.93	95.65	23
[A] [S] [G] [C]	25	42	+	Chr:19	15985581	15985598	+	18	3.8	100.00	18
[A] [S] [G] [C]	23	43	-	Chr:1	30174841	30174862	+	18	3.8	95.45	22
[A] [S] [G] [C]	41	58	-	Chr:10	2355539	2355550	-	18	3.8	100.00	18

BLAST Results

- NCBI http://www.ncbi.nlm.nih.gov/
  - Different types of BLAST
  - BLAST Results
- Ensembl http://www.ensembl.org/
  - BLAST Result
  - Detailed View

Genebrowsers and Experiment Visualization Tools

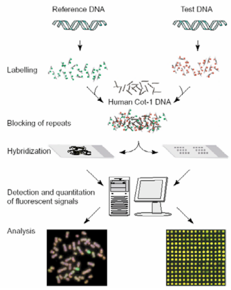
- TIGR TM4, MeV
  - FISH, Micro array CGH Experiments
- Ensembl
- http://www.genebrowser.com/

CGH Experiments

Determination of Genomic Imbalances by Genome-wide Screening Approaches

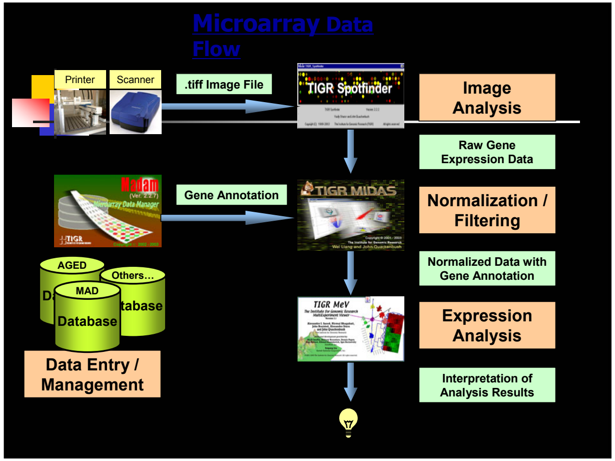
By

Károly Szuhai



CGH Experiments TIGR

- Spotfinder
- Madam
- Midas
- MeV (Demo)



Microarray Data Flow

Printer

Scanner

.tiff Image File

TIGR Spotfinder

Image Analysis

Raw Gene Expression Data

Madam

Gene Annotation

TIGR MIDAS

Normalization / Filtering

Normalized Data with Gene Annotation

AGED

Others…

MAD

Database

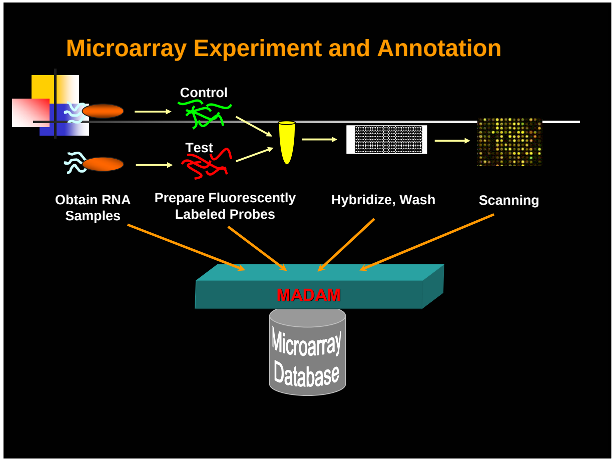
Database

Data Entry / Management

TIGR MeV

Expression Analysis

Interpretation of Analysis Results



Microarray Experiment and Annotation

Control

Test

Obtain RNA Samples

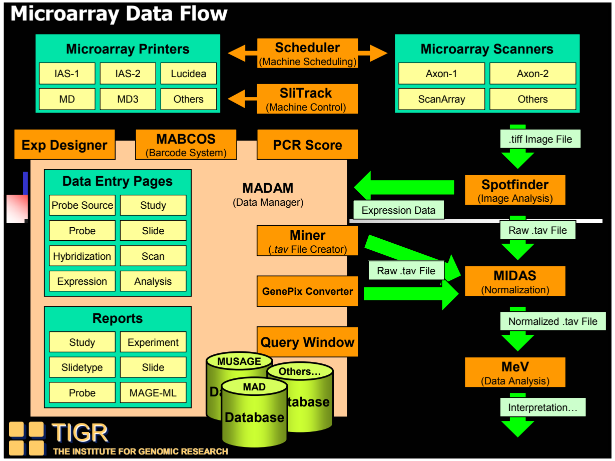
Prepare Fluorescently Labeled Probes

Hybridize, Wash

Scanning

MADAM

Microarray Database



Microarray Data Flow

Microarray Printers

IAS-1 | IAS-2 | Lucidea

MD | MD3 | Others

Scheduler (Machine Scheduling)

SliTrack (Machine Control)

Microarray Scanners

Axon-1 | Axon-2

ScanArray | Others

Exp Designer

MABCOS (Barcode System)

PCR Score

MADAM (Data Manager)

Data Entry Pages

Probe Source | Study

Probe | Slide

Hybridization | Scan

Expression | Analysis

Reports

Study | Experiment

Slidetype | Slide

Probe | MAGE-ML

Miner (*.tav* File Creator)

GenePix Converter

Query Window

MUSAGE

Others…

MAD

Database

Database

.tiff Image File

Spotfinder (Image Analysis)

Expression Data

Raw .tav File

Raw .tav File

MIDAS (Normalization)

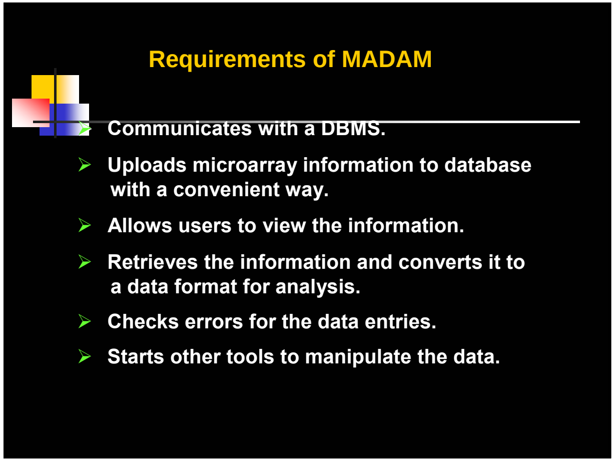
Normalized .tav File

MeV (Data Analysis)

Interpretation…

TIGR

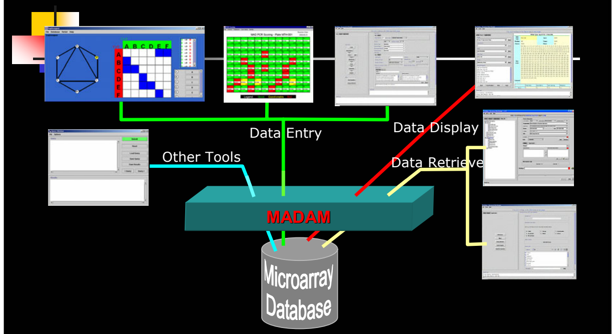
THE INSTITUTE FOR GENOMIC RESEARCH



Requirements of MADAM

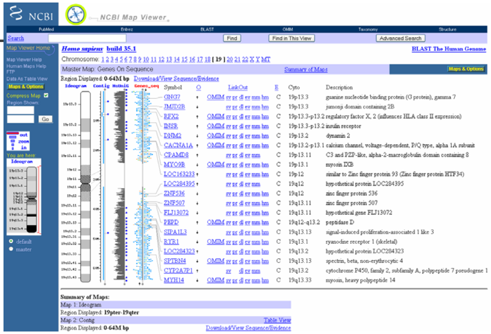
- Communicates with a DBMS.
- Uploads microarray information to database with a convenient way.
- Allows users to view the information.
- Retrieves the information and converts it to a data format for analysis.
- Checks errors for the data entries.
- Starts other tools to manipulate the data.

Functions of MADAM



Genebrowsers and Experiment Visualization Tools

- TIGR TM4, MeV
  - FISH, Micro array CGH Experiments
- Ensembl
- http://www.genebrowser.com/



Homo sapiens build 35.1

BLAST The Human Genome

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 [19] 20 21 22 X Y MT

Master Map: Genes On Sequence

Summary of Maps

Maps & Options

Region Displayed: **0-64M bp** Download/View Sequence/Evidence

Symbol	O	LinkOut	E	Cyto	Description
GNG7	↑	OMIM sv pr dl ev mm hm	C	19p13.3	guanine nucleotide binding protein (G protein), gamma 7
JMJD2B	↓	sv pr dl ev mm hm	C	19p13.3	jumonji domain containing 2B
RFX2	↑	OMIM sv pr dl ev mm hm	C	19p13.3-p13.2	regulatory factor X, 2 (influences HLA class II expression)
INSR	↑	OMIM sv pr dl ev mm hm	C	19p13.3-p13.2	insulin receptor
DNM2	↓	OMIM sv pr dl ev mm hm	C	19p13.2	dynamin 2
CACNA1A	↑	OMIM sv pr dl ev mm hm	C	19p13.2-p13.1	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
CPAMD8	↑	sv pr dl ev mm	C	19p13.11	C3 and PZP-like, alpha-2-macroglobulin domain containing 8
MYO9B	↓	OMIM sv pr dl ev mm hm	C	19p13.1	myosin IXB
LOC163233	↑	sv pr dl ev mm hm	C	19p12	similar to Zinc finger protein 93 (Zinc finger protein HTF34)
LOC284395	↑	sv pr dl ev mm	C	19q12	hypothetical protein LOC284395
ZNF536	↓	sv pr dl ev mm hm	C	19q12	zinc finger protein 536
ZNF507	↓	sv pr dl ev mm hm	C	19q13.11	zinc finger protein 507
FLJ13072	↓	sv pr dl ev mm hm	C	19q13.11	hypothetical gene FLJ13072
PEPD	↑	OMIM sv pr dl ev mm hm	C	19q12-q13.2	peptidase D
SIPA1L3	↓	sv pr dl ev mm hm	C	19q13.13	signal-induced proliferation-associated 1 like 3
RYR1	↓	OMIM sv pr dl ev mm hm	C	19q13.1	ryanodine receptor 1 (skeletal)
LOC284323	↑	sv pr dl ev mm hm	C	19q13.2	hypothetical protein LOC284323
SPTBN4	↓	OMIM sv pr dl ev mm hm	C	19q13.13	spectrin, beta, non-erythrocytic 4
CYP2A7P1	↑	sv dl ev mm	C	19q13.2	cytochrome P450, family 2, subfamily A, polypeptide 7 pseudogene 1
MYH14	↓	OMIM sv pr dl ev mm hm	C	19q13.33	myosin, heavy polypeptide 14

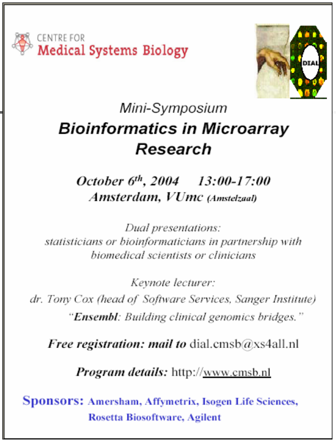
Summary of Maps:

Map 1: Ideogram

Region Displayed: **19pter-19qter**

Map 2: Contig Table View

Region Displayed: **0-64M bp** Download/View Sequence/Evidence



CENTRE FOR
Medical Systems Biology

Mini-Symposium

Bioinformatics in Microarray Research

October 6th, 2004 13:00-17:00
Amsterdam, VUmc (Amstelzaal)

Dual presentations:
statisticians or bioinformaticians in partnership with biomedical scientists or clinicians

Keynote lecturer:
dr. Tony Cox (head of Software Services, Sanger Institute)
*"**Ensembl**: Building clinical genomics bridges."*

Free registration: mail to dial.cmsb@xs4all.nl

Program details: http://www.cmsb.nl

Sponsors: Amersham, Affymetrix, Isogen Life Sciences, Rosetta Biosoftware, Agilent