

# Electronics and Informatics techniques for Genetic Analysis

## *Micrel Lab people involved*

Elisa Ficarra

Carlotta Guiducci

Daniele Masotti

Christine Nardini

Claudio Stagni Degli Esposti

## Collaborations

- Scanning Force Microscopy Lab at Biochemistry Department-UniBo
- STI Urbino
- DAUIN Politecnico di Torino
- STMicronics
- INFN Ferrara
- La Sapienza Roma
- Biochemistry Department c/o S.Luigi Hospital, University of Torino
- Computer Science of Stanford University

# Genetic Analysis

## *To Determine the sequence of a strand of DNA*

- Sequencing (DNA analysis)

## *To understand the molecular bases of the (human) phenotype*

- DNA structural properties analysis and DNA-protein interaction investigation
- Expression Profiling (RNA and protein analysis)
- Biochemical pathways defining
- Genotyping (Statistic of the presence of Single base mutations in a population)

## *To have information on the health of living being*

- Diagnostics
- Therapeutic treatments
- Drug Development

# Research themes

- Point-of-care Electronic Systems for Genetic Analysis

**Carlotta Guiducci - Claudio Stagni  
Degli Esposti**

- Computational Biology

**Elisa Ficarra - Daniele Masotti -  
Christine Nardini**

# Point-of-care Electronic Systems for Genetic Analysis

- ❖ Genetic Analysis with microfabricated sensors
- ❖ Electrical sensing of genetic-affinity reactions
- ❖ Optical detection

## Point-of-care Electronic Systems for Genetic Analysis

### Point-of-care diagnostics

Analytical testing performed outside the physical facilities of the clinical laboratories

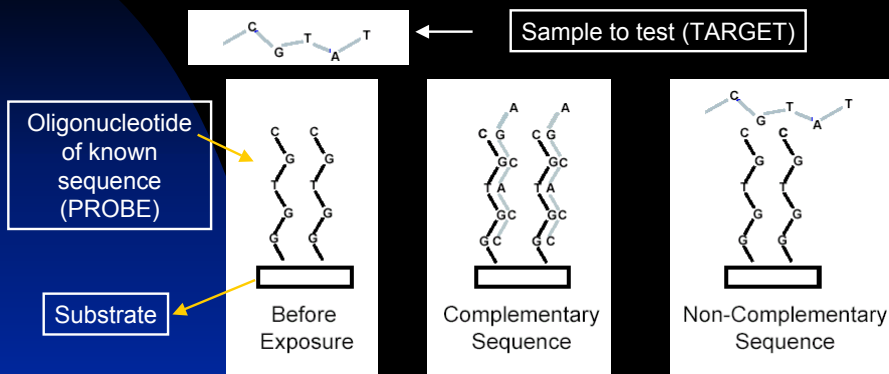
### Aim and scope of our research activity

*To enable the implementation of point-of-care genetic analysis for the detection of plant pathogens, of genetically modified organisms in foods, of marker proteins for pathologies, by developing technologies based on direct generation of electrical signals*

# Fundamental innovation in genetic analysis

- Possibility of attach, localize and/or address receptors onto a substrate in a very precise and dense way
- More simple efficient and precise analysis
- Micro-arrays: Microfabricated two-dimensional structures for parallel analysis
- Multi-site detection (more fast and parallel)
- Miniaturized devices (less sample quantity and reagent cost, mass production)

## Genetic Analysis with microfabricated sensors



### Know-how

Implementation of a sensing method or a transduction system

Systemic Chemistry  
Surface physical-chemistry  
Analytical chemistry  
Microfabrication technology

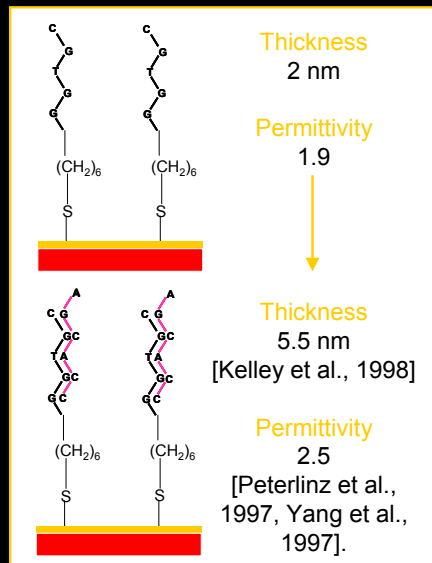
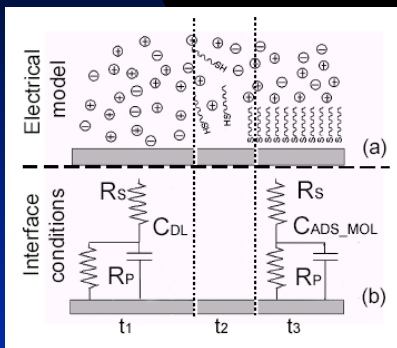
# Advantages of electrical methods

The measured signal is electrical

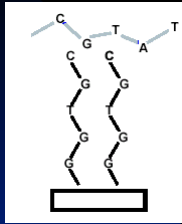
Signal measurement and processing can be integrated on the same chip

Fundamental step towards the development of lab-on-a-chip technology and point-of-care analysis

## A metal/solution interface as an electrical structure

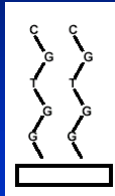


# Electrical sensing of genetic-affinity reactions

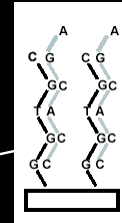
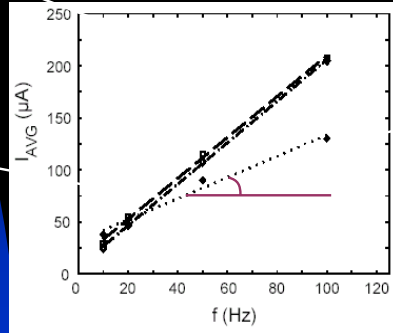


Non Complementary

Total capacitance variation: 52%



Probes

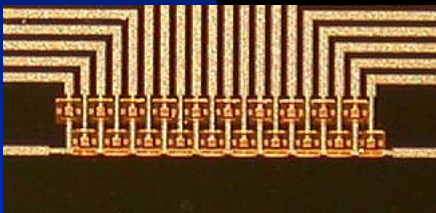


Complementary

Guiducci et al. *Biosensors and Bioelectronics* (2004)

# Microfabricated gold electrodes

STMicroelectronics  
Microelectrodes on silicon  
 $2 \times 10^3 \mu\text{m}^2$



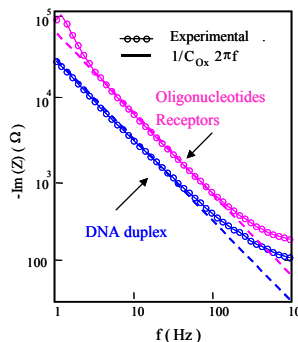
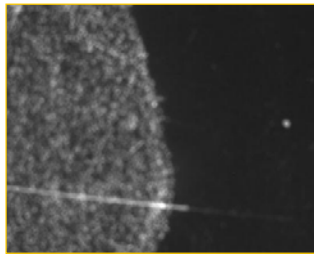
INFN Ferrara  
Microelectrodes on glass  
 $10^4 \mu\text{m}^2$



Stagni et al. et al. *Proc. of AISEM 2004*

# Low-cost materials for substrate

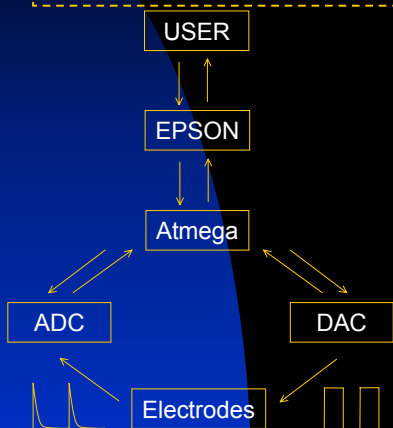
Analysis of new materials by means of Impedance Spectroscopy and Fluorescence imaging



Guiducci et al. *Proc. of Biosensors 2004*

# System on board for electrical DNA detection

$\mu$ -processors, DAC and ADC programmable by  $\mu$ -processor, all mounted on a board with microelectrodes



C++ programming

Interface and communication with DAC or ADC: RS232 serial

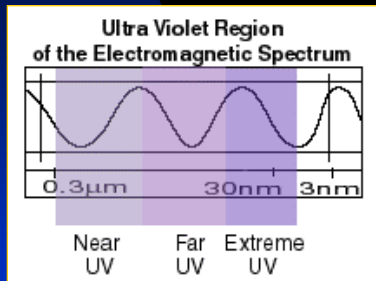
Sample rate: 60 Ksamples/sec with clock system at 8 MHz

Schematic Design based on datasheet and realization on board

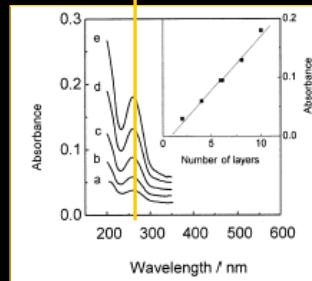
*N.B: in this case we have two  $\mu$ -controller, one for the interface with a display, the second for the measure*

# Optical detection, in progress...

*DNA detection by means of UV measurements with high sensitive integrated UV sensors*



Numbers of DNA layer



Luo et al. *Biophysical Chemistry* 2001

## Research themes

- Point-of-care Electronic Systems for Genetic Analysis

**Carlotta Guiducci - Claudio Stagni Degli Esposti**

- Computational Biology

**Elisa Ficarra – Daniele Masotti – Christine Nardini**

# Computational Biology

## Our Capabilities!

- Techniques for Automated Analysis of DNA Molecules in Atomic Force Microscope Images
- Clustering and Cluster Biological Evaluation of Gene Expression Data

## Works in Progress

- Extraction of Clinical Information from Gene Expression Data
- Modelling Gene Regulatory Networks
- siRNA Design for RNA silencing

## Techniques for Automated Analysis of DNA Molecules in Atomic Force Microscope Images

Development of Automated Algorithms for DNA Molecules Feature Analysis and Extraction

DNA Sizing and Molecular Profiles determination algorithm through a set of fully automated **Image Processing** steps

DNA Intrinsic Curvature profile computation using a fast heuristic technique → **Combinatorial Optimization Problem**

# Motivation

## Importance of DNA Sizing, Molecular Profile Determination and DNA Curvature Analysis

- ✓ Specific DNA target identification
- ✓ Physical genome maps and genotyping construction
- ✓ Transcription rules investigation
- ✓ Analysis of DNA secondary structure transitions
- ✓ DNA molecule structural properties investigation
- ✓ DNA-protein interaction analysis

# Atomic Force Microscopy (AFM)

## Characteristiques and Properties

- Low amount of DNA samples (vs. Gel electrophoresis)
- Direct visualization of DNA molecules ⇒ **lower processing time** (few minutes vs. 2 hours with gel electrophoresis )
- **High resolution** (from 2 to 20nm vs. Over 200nm with optical microscopy)
- **High signal-to-noise ratio**
- Direct visualization of DNA molecules **without contrast-enhancing agents**

# Objectives

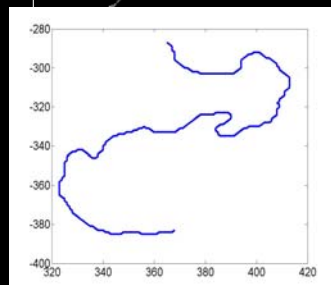
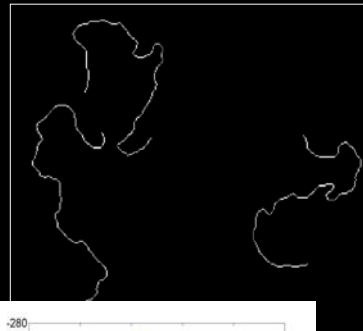
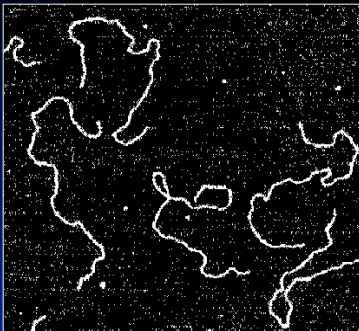
## Automated Algorithm for molecular profile determination and DNA sizing from AFM images

- High accuracy
- High robustness w.r.t. changes on DNA curvature profiles
- High speed
- DNA secondary structural transitions analysis

## Automated Algorithm for DNA intrinsic curvature profile computation from molecular profiles

- Automated Molecular Profile determination through Molecular Orientation detection

# DNA Sizing and Molecular Profile Determination Algorithm

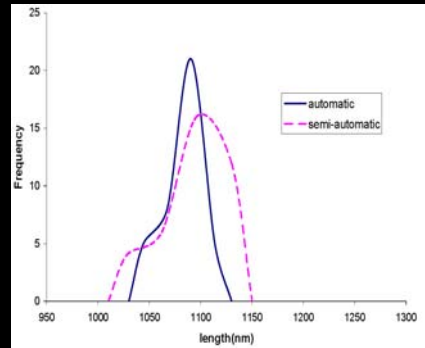
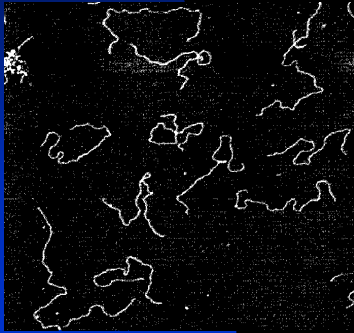


- Sequential Image Processing Steps
- Outputs:
  - DNA length calculation
  - Molecule Profile Extraction and Smoothing (for DNA curvature and flexibility analysis)

# Example of Experimental DNA Sizing Results

## Crithidia fasciculata AFM images

- ✓ Characterized by a very high curvature region → very irregular shapes
- ✓ DNA length computation gets harder because surrounding noise shadows DNA shapes



## DNA Curvature Model

- Highly asymmetric form factor
  - ◆ Molecules can be idealized as one-dimensional curved line
- **Curvature value**
  - ◆ Intrinsic curvature
    - Nucleotide sequence-dependent, static contributions
  - ◆ Flexibility
    - Susceptibility to thermal deformation, dynamic thermal contributions

$$C(n) = C_0(n) + f(n)$$

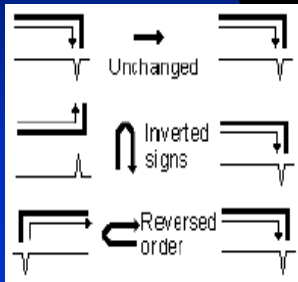
- **Filtering of Dynamic Contributions**

- Averaging along the chain on a significant population of molecules flexibility contribution is null and the average of sampled curvature is equal to intrinsic curvature

$$\langle C(n) \rangle = \langle C_0(n) \rangle + \langle f(n) \rangle = C_0(n)$$

# Curvature Reconstruction Algorithm

- Four different DNA adsorbing modality on AFM surface
  - Molecular face (two different ways due to Dna molecule planarity)
  - Direction of sampling (difference due to DNA molecule asymmetry)
- Four different curvature profile orientations → from an AFM image with  $m$  equal molecules,  $v$  curvature values are sampled at regular intervals along each chain. Since all these curvature vectors have the same dimension we can define a curvature matrix  $C(m \times v)$



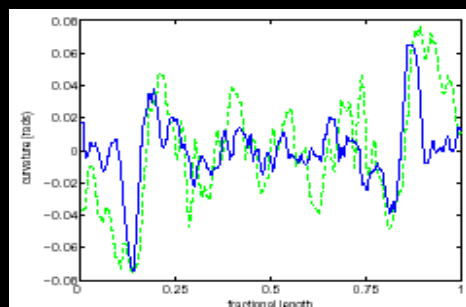
- Representation of molecules in the matrix with the same orientation to evaluate the curvature average on corresponding points of the molecule.
- The optimal configuration, all the molecules share the same orientation → minimal value of curvature variance for each point, i.e. the minimal column variance in matrix of curvature  $C$  → **Greedy Heuristic**

## Example of Experimental Curvature Profile Computation

EcoR V-EcoR V dimer intrinsic curvature profile

- Theoretical curvature peak of 0.08 rads
- Deviation of 8.44nm, that is 1.3% of molecule length in the location of the peaks w.r.t. the theoretical curvature profile
- Reconstructed intrinsic curvature profile well approximates the theoretical one with a standard deviation in the regions of the peaks of  $6.1E-3$

Figure: Theoretical (dashed plot) and EcoRV-EcoRV reconstructed intrinsic curvature profile (solid plot)



# Microarray Clustering

- Unsupervised learning technique
- In general, an NP-complete problem
- Examples
  - ◆ K-means, hierarchical, graph partitioning, self-organizing map,...
- Mostly approximate algorithms
- May lose global patterns
- High-dimensional data
  - ◆ More difficult for a cluster to form
  - ◆ Harder to find a cluster
- Two-way clustering
  - ◆ Cluster attributes as well
- Subspace clustering
  - ◆ Focus on subset of attributes

## pCluster<sup>1</sup> – metric and definition

- Two-way clustering algorithm
- Only 3 parameters are required:  
 $M_G, M_E, \delta$

$$pScore \left( \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix} \right) = \left| (x_1 - y_1) - (x_2 - y_2) \right|$$

	J		
	x1	x2	
I			
	y1	y2	

pCluster: cluster which elements all have pScore smaller than a threshold

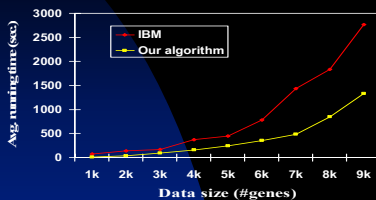
$$(I, J) pCluster \Leftrightarrow pScore \leq \delta, \forall x_i, i \in I, \forall y_j, j \in J$$

<sup>1</sup>Clustering by Pattern Similarity in Large Data Sets, H. Wang et al, SIGMOD 2002

# Enhanced pCluster Algorithm - flow

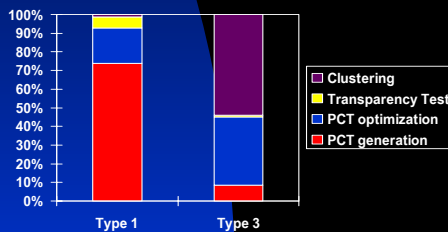
- Generation of  $PCT_{genes}$ : sets of all pCluster of size 2 genes by any number of experiments (and viceversa with  $PCT_{exp}$ )
- Test for *Well Shaped property*, if this holds the PCT already contains the final solution in it
- Depending on the former test, application of appropriate algorithm for clustering
- Finds ALL pClusters in the matrix

## Performances



← Comparison with original pCluster (on syntetic data set)

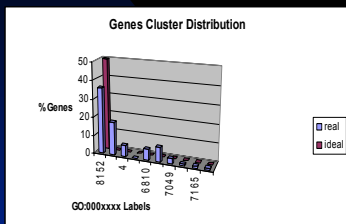
Number of clusters found (Real data set, yeast)



Running time breakdown

Case	Ours	IBM
1	5	5
2	21	21
3	102	102
4	167	167
5	65	65
6	372	207
7	572	210

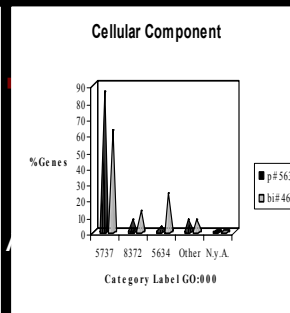
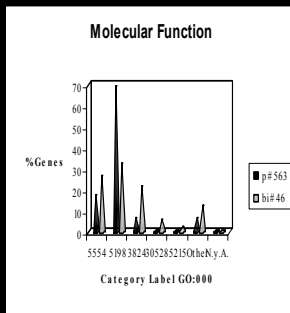
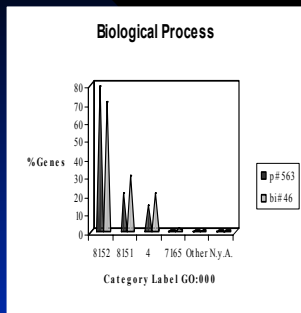
# Cluster Biological Validation



- Use of Gene Ontology (GO)
- Generation of cluster distributions: frequency of genes through GO categories
- Quantitative evaluation of cluster *purity*: peak value and coefficient of variation measure how close a cluster is to a discrete impulse representation (*highest purity*).

GOID	GO term	Frequency	Gene(s)
8152	Process: metabolism	36 out of 50 genes, 72%	RPS8A SSA3 RPL32 RPS9B DER1 YBP1 PBN1 HEX3 PRP9 RP
8151	Process: cell growth and/or maintenance	18 out of 50 genes, 36%	SSA3 VAC17 HEX3 MPH2 GSG1 NUM1 SEC7 SIR4 SBE2 ESC2
6350	Process: transcription	8 out of 50 genes, 16%	CTH1 SIR4 YAP6 ESC2 SSN2 RAD3 NUT1 SRB5
6810	Process: transport	6 out of 50 genes, 12%	SSA3 MPH2 GSG1 SEC7 VPS3 YEL006W
4	Process: biological_process unknown	6 out of 50 genes, 12%	YCR001W YDR109C FIN1 YER128W YGL085W YGL250W
7049	Process: cell cycle	3 out of 50 genes, 6%	GSG1 NUM1 RAD3
67	Process: DNA replication and chromosome cycle	1 out of 50 genes, 2%	NUM1
7165	Process: signal transduction	1 out of 50 genes, 2%	CDC43
6520	Process: amino acid metabolism	0 out of 50 genes, 0%	none
Other	Other	1 out of 50 genes, 2%	SNG1

## Results



	Biological Process		Molecular Function		Cellular Component	
	epC	C&C	epC	C&C	epC	C&C
peak	0.8	0.7	0.8	0.6	0.8	0.7
Coeff. Variat	169.3	139.8	180.6	90.3	186.5	122.8

Comparison between high overlapping cluster (C&C)<sup>2</sup> and enhanced pCluster (epC)

## Work in progress (Microarray Clustering)

- Clinical genomic: introduction of clinical information in the gene expression matrix
- Goal:  
Diagnose diseases with the accuracy of the genetic pattern through clinical information

## Work in Progress (Gene Networks)

- **siRNA Design for RNA interference to**
  - ◆ systematic analysis of gene expression and function
  - ◆ therapeutic gene silencing
- **Modelling Gene Regulatory Networks**
  - ◆ defining gene function
  - ◆ defining biochemical pathwaysthrough network mathematical models design and microarray screening of RNAi knockouts
- **Goals:**
  - Drug Development
  - Therapeutic treatment
    - ◆ Cancer
    - ◆ HIV
    - ◆ Viral infection
    - ◆ Parasitic infection