

- *Swiss Federal Research Station for Fruit-Growing,
Viticulture and Horticulture*
- *CH8820 Waedenswil*
- *Switzerland*
-
-
-
-
-
-
-



Towards The Random Chip



A Biologist's Approach to Bioinformatics

Juerg E. Frey

Needs of Agricultural Diagnostics

- Identification of Many Different Organisms/Taxa
 - Need for Multiplex Capacity
- Handling of Large Sample Sizes
 - Several Hundred Specimen rather than One Single Individual
- Ease of Use
- Reproducibility
- Economically Reasonable

Microarrays in European Agriculture



Aim

- Reduction in Complexity

| Taxon | Gene | Technique |
|------------------|---|---|
| Nematodes | ITS | PCR-RFLP |
| Insects | mitochondrial COI, COII, 16S rRNA; nuclear 18S rRNA, ITS | PCR-Sequencing, PCR-RFLP |
| Fish | mitochondrial cytb, COI | PCR-RFLP |
| Bacteria | 16S rRNA, 28S rRNA | PCR-Sequencing, PCR-RFLP, Hybridisation |
| Fungi | ITS, 28S rRNA, mt-LrRNA | PCR-Sequencing, PCR-RFLP, PCR- Hybridisation |
| Crops | Microsatellites | PCR |

Microarrays in European Agriculture



Aim

- Massively Parallel Identification
 - Many Species – One Assay
 - Expands Potential of Laboratories – Diversity and Throughput
- Standardisation / Harmonisation of Methods
 - Decrease in System Complexity
 - Increase in Quality and Reproducibility
 - Facilitates International Collaboration
- Economic Dimension

Microarrays in European Agriculture



Requirement for Implementation

- Knowledge of Among vs. Within-Species Genetic Variation

```
Consens  TTATCWAAGGAAGGAGCAGGAAACAGGATGAACAGTTTATCCACCTYTATCAACATTTATC
Fra.occi  ----A-----T-T-----C-----T-G-----Y-----
Ech.amer  ----A-----A-----YC-----
Thr.palm  ----AT-----T-----
Thr.taba  C-T-AT-----G-----G-----A-----T---W-G-----
Tae.pici  --AGA-----G-----C-C-----T-----
Her.femo  A---TTGG-----T--T-----G--G-----T--C-----AT-----T---C-
Hel.haem  A-GATGGG---T-----T-----T---C-T-----
Thr.angu  C-T-T-----T-----T---C-T-----
Par.drac  A-GCTCGGT--T---C-T--T-----TT-G--C-A-----C-----T-A--T-
Ana.obsc  --T-AGG-----T-----G-----T---C-T-----T-
```

„Intelligent Chips“ - Based on Specific Sequences

- **Mitochondrial Cytochrome Oxidase I Gene**
 - Many Copies per Cell
- **Conserved Primers for PCR** (polymerase chain reaction)
 - Can be Used in Many Species
- **C. 400 Base Pair Fragment**
 - Easy to PCR Amplify and Sequence
- **Arthropod Model Thrips (Thysanoptera: Thripidae)**

The Random Chip



Information on the Entire Genome

- **Use of Many Primers with Random Sequence**
 - No Need for Sequence Information > can be used for Any Species
 - Many Potential Targets per Genome
- **Labelling by Primer Extension**
 - Many Markers per Genome > High Resolution
- **Analysis on Microarray**
 - De-Multiplexing
- **Identification by Comparison to Database**
 - E.g., Clusteranalysis

Technology

- Pattern-based characterization
- Production of many labelled anonymous markers - not PCR-based
- > 40'000 labels of each marker
- Identification by clustering methods based on reference database - quality assessment
- Suitable for array/chip technology
- Chip composition can be optimized for targeted application

Areas of Application

Undescribed Genomes

- Molecular identification
 - Identification of plant varieties / animal species and/or strains
- Molecular markers for characteristics
 - pathogen and/or pesticide resistance genes
 - Quality characters - Marker-assisted breeding
- Genetic diversity (Biodiversity) and population genetics
- SNP identification and analysis

Random Chip



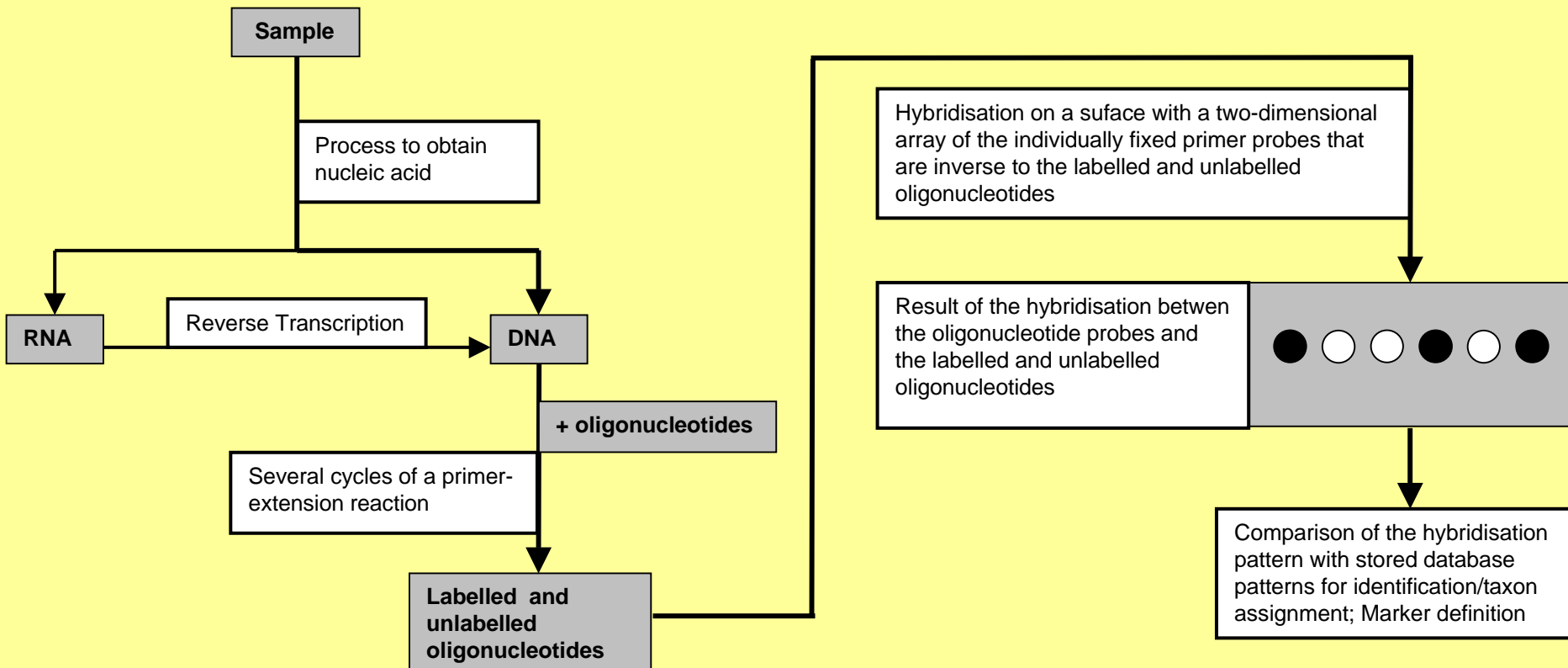
Strategy

- Use one or few Chip models
 - large production numbers and possibility of multiple use decrease costs
- Develop separate Databases for specified fields
 - increased speed and specificity
 - optimal marketing strategy

Random Chip



Flowchart



Oligo Construction

- Generate random number between 1 - 4
- Assign bases (1=A, 2=C, 3=G, 4=T)
- Do this until required length reached
- Do selection testing
 - Fixed %GC
 - Fixed Tm
 - No hairpin
 - No complement
 - Minimal similarity



Search Genome for Oligo Hits

- Open genome file
- Read string of oligo length
- Compare to oligo and inverted oligo
- Read next base, drop last base, compare again
- Do this until end of file, then for next genome
- Record for each oligo/inverted oligo number of hits
- Produce 1/0 table for cluster analysis

| | |
|----------|---|
| Ecoli | 001011000010001000001101110000100010100100101111000000100001000 |
| Staphaur | 010010011010001000001001000100010010100100010011000100000100000 |

Feasibility

Parameters for producing the computer-generated virtual bacteria strains. Average gene size is 1200 base pairs (bp). The length of both genomes had to be slightly adjusted (<0.02%) for computer analysis. Level of polymorphism: Genotype PGL: low, PGM: medium, PGH: high.

| | N | Percent | Mutations per gen | Average percentage of polymorphic nucleotide positions |
|--------------------|------|---------|-------------------|--|
| E. coli | | | | |
| Genotype PGL | 966 | 25,0 | 30 | 0,0250 |
| Genotype PGM | 1611 | 41,7 | 84 | 0,0700 |
| Genotype PGH | 1289 | 33,3 | 138 | 0,1150 |
| Total | 3866 | 100 | 88.5 | 0.0738 |
| B. subtilis | | | | |
| Genotype PGL | 874 | 24,9 | 30 | 0,0250 |
| Genotype PGM | 1487 | 42,3 | 84 | 0,0700 |
| Genotype PGH | 1152 | 32,8 | 138 | 0,1150 |
| Total | 3513 | 100 | 88.3 | 0.0736 |

Virtual Random Chip - medium density

- 768 Marker (Random Sequence)
- 2 fully sequenced Strains of 4 Microorganisms
 - **Escherichia coli K12 MG1655**
 - **Escherichia coli O157H7**
 - **Staphylococcus aureus Mu50**
 - **Staphylococcus aureus N315**
 - **Mycobacterium tuberculosis**
 - **Mycobacterium tuberculosis CDC1551**
 - **Neisseria meningitidis serogroup A strain Z2491**
 - **Neisseria meningitidis serogroup B strain MC58**
- All by 12.9.2001 publicly available sequences > 2 Mio bp

Virtual Random Chip

- Robustness Computer Simulation

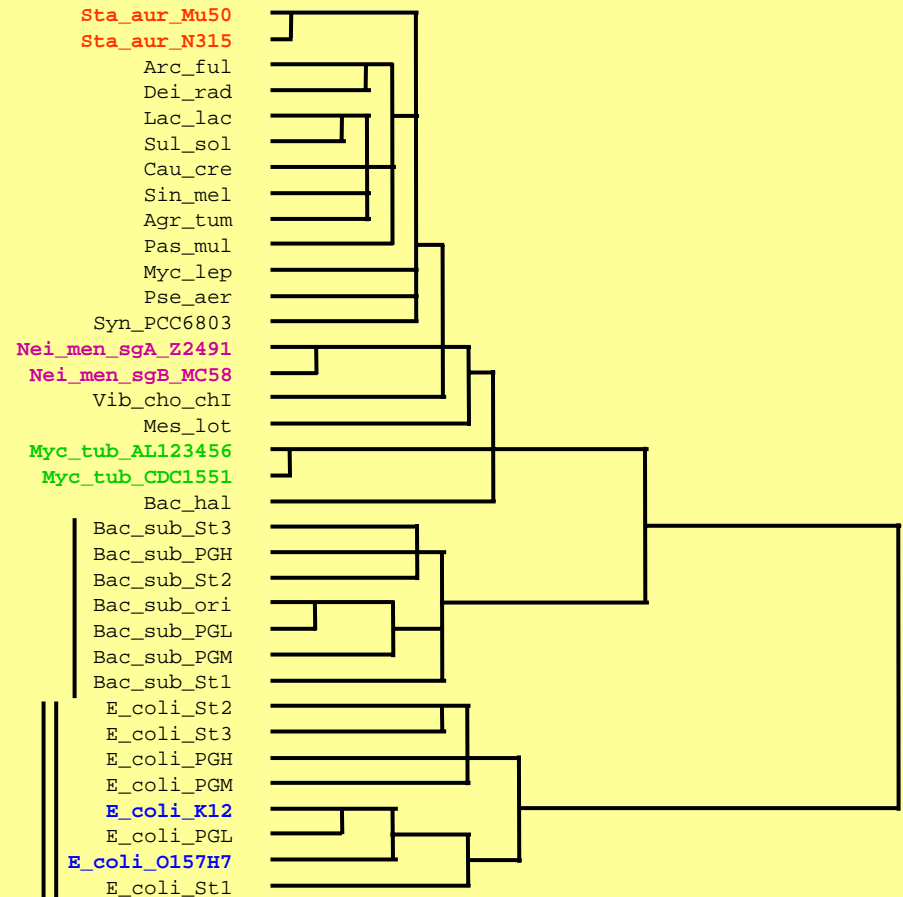
Result from 20 virtual chips with 768 markers each

- 100% correct group assignment of the sequenced strains
- 100% correct group assignment of the virtual strains (exception PGH)
- PGH in *E. coli* - 60% in the correct group
- PGH in *B. subtilis* - 10% in the correct group
- 61% of these false assignments clustered *B. subtilis* PGH with *B. halodurans* C125

Random Chip

Virtual Random Chip

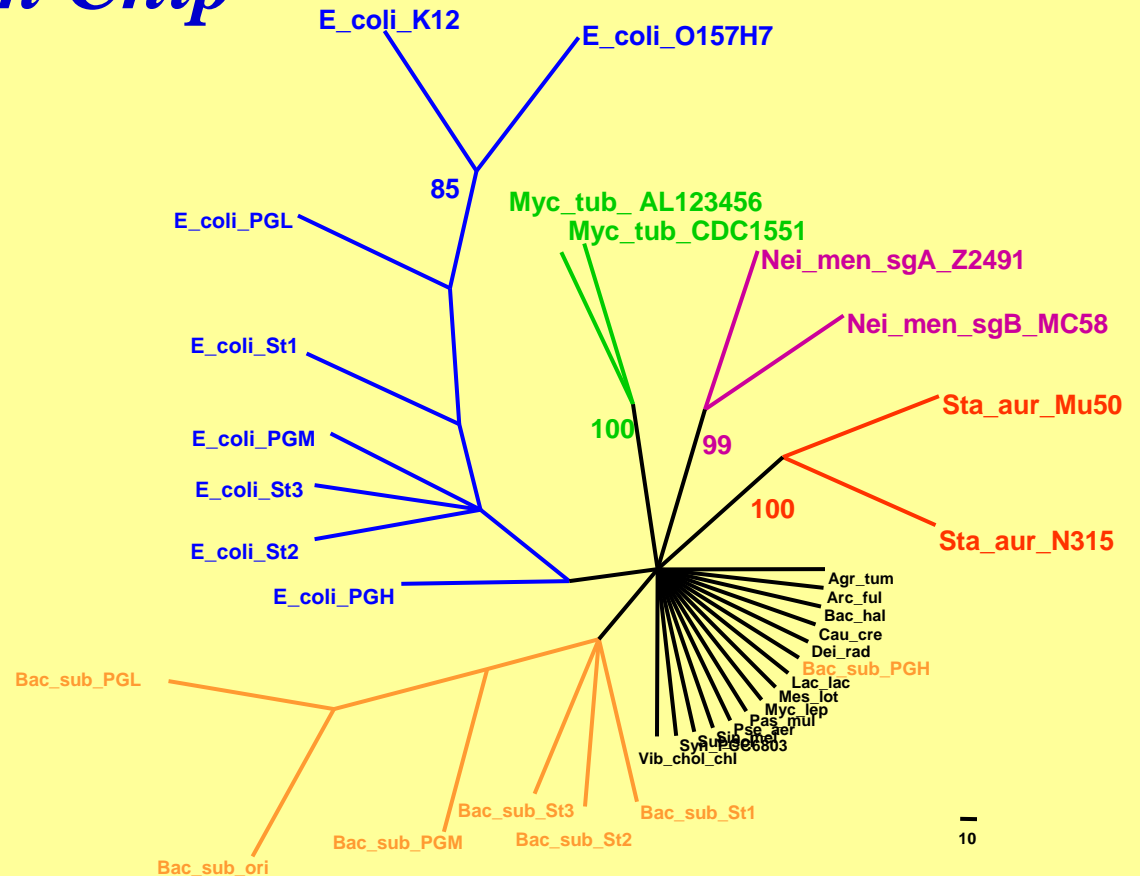
Dendrogram of a clusteranalysis of the data matrix (presence/absence) for 768 oligonucleotides with a random sequence. All until 14.9.2001 freely accessible fully sequenced genomes of at least 2 Mio bp in length were used. All strains of the same species and the computer-generated strains of *E. coli* and *B. subtilis* were correctly assigned to the respective group (expection PGH). Compared to the results based on a 10'000 feature chip, the resolution within groups is partly lost.



Random Chip

Virtual Random Chip - low Density

Computer Simulation



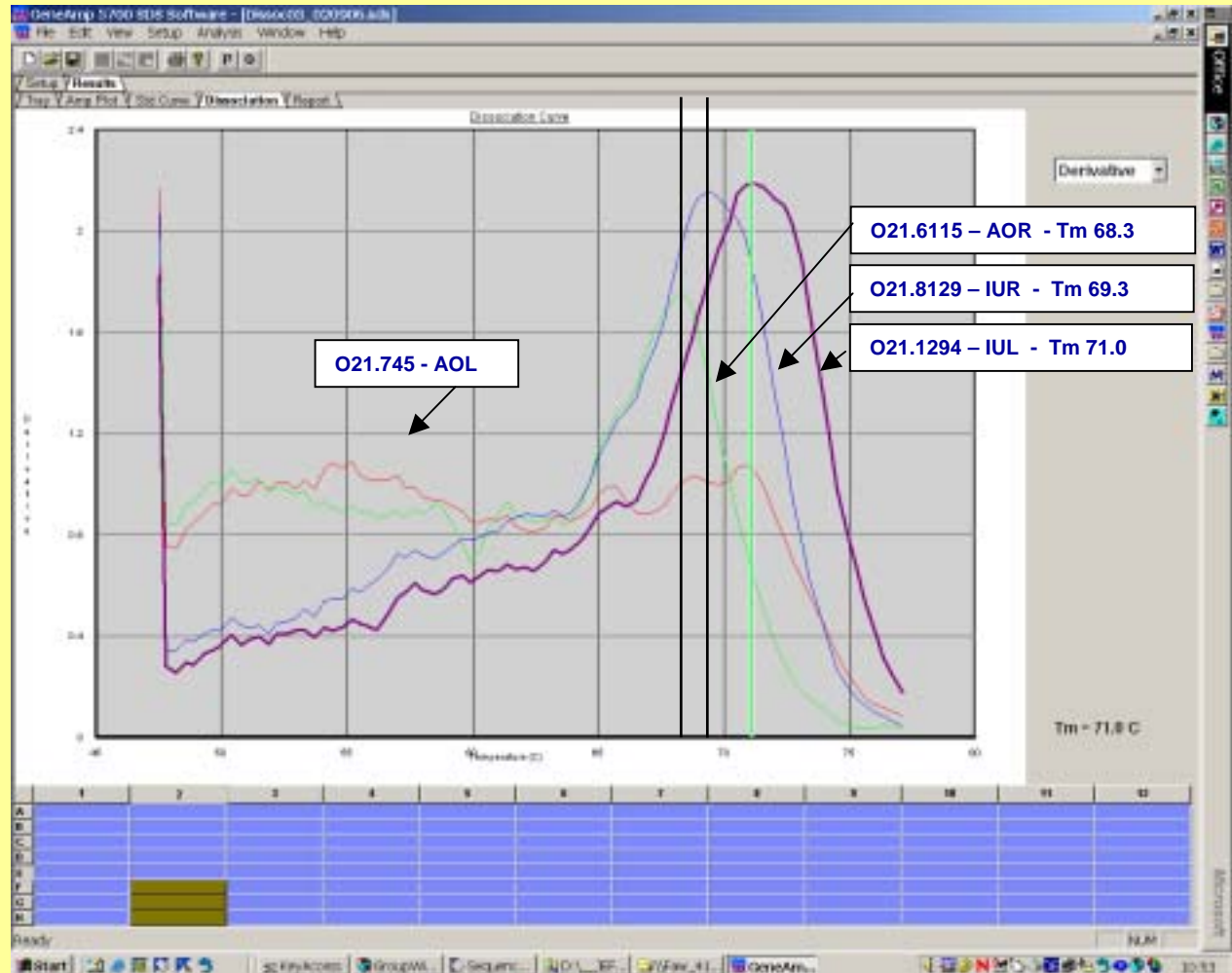
„Real World“ Problems

- Tm calculation
 - Tm calculations not accurate
 - Tm solution $\langle \rangle$ Tm surface
 - Development of new algorithm required
 - Empirical testing
 - Expensive – funding?

Random Chip



Tm calc = 55



„Real World“ Problems

- Required DNA/RNA amount is high
 - >0.5 ug – Preamplification necessary?
- Which technology?
 - Whole genome amplification
 - Uneven amplification
 - Rolling circle amplification under development at Amersham
 - Targeted genome amplification
 - May produce false positives
 - Seems very robust and reproducible

The Team



JEF

Beatrice Frey

Monika Pfunder

Franz Schwaller

Patrick Brunner

Apprentices