Propositional Logic for Efficient Microbial Community Assessment by DNA Microarray Data Analysis.

Faouzi JAZIRI, Céline RIBIÈRE, Cyrielle GASC, Nicolas PARISOT, Réjane BEUGNOT, Clémence DEFOIS, Owen K. APPADOO, Eric PEYRETAILLADE, David R.C. HILL and Pierre PEYRET.

July 8, 2015
I. Introduction/Context.

II. Proposed approach.

III. Results.

IV. Conclusions.
I. **Introduction/Context.**

II. Proposed approach.

III. Results.

IV. Conclusions.
✓ DNA microarrays: High-throughput molecular tool.

✓ Current format: up to several million probes.

✓ Allow studying the presence, or the expression levels of several thousands of genes, combining qualitative and quantitative aspects in only one experiment (Schena et al., 1995).

✓ Based on the ability of complementary strands of DNA to hybridize to one another in solution with high specificity.
The principle of DNA Microarrays

- Sample (nucleic acids: DNA or RNA)
- Labeled targets
- Probe-target hybridization
- Microarray scanning (intensity measurements: Scanner)
- Data Analysis
✓ **DNA microarray data analysis:** a crucial step for a successful microarray experiment.

✓ **High-density** formats: microarrays produce huge amount of data.

- The complexity of data analysis

- Important increase in computational capacity requirements of microarray data analysis algorithms.
Existing methods: several tools (Dudoit et al., 2003; Mehta and Rani, 2012; Koschmieder et al., 2012; Jaziri, 2014).

- dedicated to the transcriptomic studies of isolated organisms and thus are not suitable for evaluating metagenomic samples.
- don't take into account the quality of the probes used.
- not adapted to analyse microarrays developed using multiple probe selection tools (PhylArray, PhylGrid, HiSpOD, KASpOD, Metabolic Design, MetaExploArrays).
We proposed a parallel algorithm for DNA microarray data analysis:

- We use parallel computing and the concepts of propositional logic to determine the microbial composition of a hybridized biological sample.

- This software is well adapted to microarrays developed using multiple probe selection tools.
I. Introduction/Context.

II. Proposed approach.

III. Results.

IV. Conclusions.
✓ **Goals:**

- A software adapted to multiple probe selection tools (PhylArray, PhylGrid2.0, MetaExploArrays, KASpOD, HiSpOD).

- Determine the microbial composition of a hybridized biological sample.

- Take into account the specificity of all probes used on the microarray.

✓ **Proposed approach « PhylInterpret »:**

**Implementation:**

Tools used:

- Blastn (Altschul et al., 1990)
- zchaff (Zhang et al., 2001)
✓ **Algorithm: 2 main steps.**

1\textsuperscript{st} step: determining an initial list of positive groups.

- Calculating a minimum response value using all control probes:

\[ d(s, c) = \sqrt{(X_c - X_s)^2 + (Y_c - Y_s)^2} \]

where \((X_c, Y_c)\) are the coordinates of \(c\) and \((X_s, Y_s)\) are the coordinates of \(s\).
Analyse des résultats de biopuces

\[ \forall s \in S \text{ et } \forall c \in C, \text{ on définit par } P_{cs} \text{ le poids de la sonde contrôle } c \text{ relatif à la sonde } s \text{ tel que:} \]

\[ P_{cs} = f(d(s, c), d_{s_{max}}) = \frac{d(s, c)}{d_{s_{max}}} \]

Où \( d_{s_{max}} \) est la distance entre \( s \) et la plus distante sonde de contrôle négative.

\[ \forall s \in S \text{ et } \forall c \in C, \text{ on définit par } I_{s_{min}} \text{ la valeur minimale de réponse de } s \text{ tel que:} \]

\[ I_{s_{min}} = \frac{\sum_{c=0}^{n-1} (P_{cs} \times I_{c})}{n} \]

Où \( I_{c} \) est l'intensité d'hybridation de \( c \) et \( n \) le nombre total des sondes de contrôle négatives de la biopuce.
✓ Algorithm: 2 main steps.

1st step: determining an initial list of positive groups.

- Calculating a minimum response value using all control probes.
- Calculating a SNR value for each probe (« Signal to Noise Ratio »):

\[
\forall s \in S, \ SNR_s = \frac{I_s}{I_{S_{\text{min}}}}
\]
✓ **Algorithm:** 2 main steps.

1st step: determining an initial list of positive groups:

- Calculating a minimum response value using all control probes.
- Calculating a SNR value for each probe (« Signal to Noise Ratio »).
- Defining a preliminary list of positive groups.
✓ Algorithm: 2 main steps.

1st step: determining an initial list of positive groups:

- Calculating a minimum response value using all control probes.
- Calculating a SNR value for each probe (« Signal to Noise Ratio »).
- Defining a preliminary list of positive groups.

Positive response or cross-hybridization??
Algorithm: 2 main steps.

2nd step: determining the real composition of a hybridized sample: the specificity of the probes.

1. Specificity test: considerable computation time (up to several hours to process high-density microarrays): a parallel implementation using a computing cluster.

2. Formulate the problem using the concepts of propositional logic: SAT (Boolean Satisfiability) problem.

3. Use a SAT solver (zchaff (Zhang et al., 2001)) to identify all possible solutions to this problem: all possible lists of groups present in the hybridized sample.

The goal is to remove all groups that appear in the initial analysis (Step 1), due to the cross-hybridization of its probes.
I. Introduction/Context.

II. Proposed approach.

III. Results.

IV. Conclusions.
1. Results of the parallelization of the specificity test: a microarray composed of 54,129 25-mer probes (PhylOPDb: http://g2im.u-clermont1.fr/phylopdb/):

Speedup of 52x using 64 cores on a computing cluster.
2. **Results of the proposed method to determine the real prokaryotic composition of a hybridized biological sample:** we used a prokaryotic microarray composed of 19,874 25-mer probes targeting 2,069 genera (developed using PhylGrid).

- The hybridized sample is composed of a DNA mixtures of species from 31 prokaryotic genera.

- Taking into account the specificity of probes, PhylInterpret has detected 6 genera that appear mistakenly in the original list of results due to cross-hybridizations.

- Error rate: less than 2% of the genera targeted by the microarray.

The quality of the results generated by PhylInterpret + the quality of probes selected by PhylGrid.
Outline

I. Introduction/Context.

II. Proposed approach.

III. Results.

IV. Conclusions.
1. Conclusions

- PhylInterpret determine the microbial composition of a hybridized biological sample using parallel computing and the concepts of propositional logic.

- We presented the performance of PhylInterpret on real biological datasets.

- PhylInterpret is currently the only available algorithm well adapted to analyse the results of complex microarrays developed using multiple probe selection tools.

- PhylInterpret is the only available software that uses the specificity of probes to allow better biological interpretation of results.
2. **Prospects**

- More biological tests...
- Web interface ???
Thank you for your attention!