PhylochipAnalyzer - A Program for Analysing Hierarchical Probe-Sets

Katja Metfies[*], Philipp Borsutzki[‡], Christine Gescher[*], Linda K. Medlin[*], Stephan Frickenhaus[*,+]

[*]Alfred-Wegener-Institute for Polar and Marine Research, Am Handelshafen 12,
27570 Bremerhaven, Germany

[‡] Urnenfelderstraße 9, 85051 Ingolstadt, Germany

Keyword:

Phylochips, molecular probes, DNA-microarrays, analysis, hierarchical probes, software tool

[+]Corresponding author:

Stephan Frickenhaus, Alfred-Wegener-Institute for Polar and Marine Research,
Am Handelshafen 12, 27570 Bremerhaven
Tel. ++49471-4831 1179, Fax. ++49471-4831 1590, e-mail: Stephan.Frickenhaus@awi.de

Abstract

The recent introduction of phylochips that contain molecular probes facilitates environmental microbial identification in a single experiment without previous cultivation. A set of probes recognizing species at different taxonomic levels is denoted as a hierarchical set. Application of hierarchical probe sets on a DNA-microarray allows the assessment of biodiversity with different resolutions. It significantly increases the robustness of the results retrieved from phylochip experiments because of the possible consistency checks of hybridization across different taxonomic levels. Here, we present a computer program, PhylochipAnalyzer, for the hierarchy editing and the evaluation of phylochip data generated from hierarchical probe-sets.

Basic rationale: Recently, more and more publications describe the application of DNA microarrays for species identification (phylochips) from environmental samples (Guschin et al. 1997; Loy et al. 2002; Metfies and Medlin 2004; Call 2005, Medlin et al. 2006). Phylochips are DNA-microarrays containing molecular probes that bind to unique sequences in a target. The target sequence is usually part of marker genes, e.g., the ribosomal RNA gene. Ribosomal RNA-genes are particularly well suited for phylochip- and phylogenetic analysis, because they are universal, found in all cellular organisms, are of relatively large size; and contain both highly conserved and variable regions with no evidence for lateral gene transfer (Woese 1987). The large number of published 18S rDNA-sequences, (e.g., RDP, Maidak et al. 2001) makes it possible to design hierarchical probe sets that specifically target the 18S-rDNA from higher taxa down to species level (Lange et al. 1996; Guillou et al. 1999; Groben et al. 2004). Phylochips provide a promising tool to identify large numbers of microbial species in complex environmental samples quickly without a cultivation step. Our phylochip contains a hierarchical set of probes, which target phytoplankton species at different taxonomic levels (Metfies and Medlin 2004; Medlin et al. 2006). In a hierarchical probe-set, a target species is only considered present, if all hierarchical probes for each species result in a positive signal. Therefore, hierarchical probes add to the accuracy of molecular probe based identification approaches.

In spite of the growing number of applications for phylochips, they represent only a small proportion of all DNA-microarray related work. Most publications describe expression studies (e.g., Lockhart et al. 2000; Stoughton 2005; Rensink 2005; Csako 2006). Consequently, the majority of protocols are optimized for applications related to expression analysis. However, the application of phylochips for species identification in environmental samples presents technical challenges that are not encountered in gene expression studies of laboratory samples (Peplies et al. 2003; Call et al. 2005; Metfies et al. 2006). There are numerous commercial and non-commercial programs for the analysis of expression studies (e.g. Dondrup et al. 2003; Vaquerizas et al. 2005) but few programs exist for phylochip analysis. One example is the Unix-based program ChipChecker (Loy et al. 2002), which is dedicated to data interpretation from phylochips. It calibrates signal to noise ratios to a set threshold determined by the user and finds positive signals with respect to that threshold based on the fact that a positive signal can only be located where there is a fully complementary probe to its target. However, in a hierarchical probe set, a signal is only considered truly positive, if all probes in the hierarchy are positive. Therefore, the analysis of hierarchically organized phylochips requires an additional step in comparison to the functions provided by ChipChecker. The positive signals must be tested for their robustness in relation to the hierarchy on the phylochip. In summary, a program for the analysis of hierarchically organised phylochips has to provide an algorithm for the calculation of a signal to noise-value and a tool that allows to set positive signals in relation to the hierarchy inherent in the design of the probe-set. Here we present the program, PhylochipAnalyzer, that implements the calculation of signal to noise ratios and the evaluation of phylochip-data with respect to probe hierarchy.

Funcionality and Implementation aspects of the Program

PhylochipAnalyzer is a GUI-based Windows-program, developed under Borland-Delphi. The program combines two strongly interconnected functions: hierarchy editing and data analysis. The user starts editing interactively and graphically the hierarchy that is inherent in the chip/probe design process. Editing is started by loading a spot description file in GAL-format generated by the GenePix- software (Axon Instruments Inc., USA). A procedure to convert other formats is described in the software documentation. Spot entries are shifted manually so that a hierarchically structured tree-like layout appears, in correspondence to the hierarchical probe design of the chip seen in Fig. 2A, upper part. Probes must not be placed in a hierarchy at all, e.g., positive or negative controls should be placed as stand-alone, i.e., with no parent probes and no child probes. However, a positive control could be placed as the parent probe to all other. The hierarchy is then saved as an XML-file that is used later for data analysis. Whereas the XML-file stores the pure hierarchy information of the chip, spot-intensity data are read from files with externally defined format, such as tab-delimited tables. The user may include the probe sequence in the comment field. The hierarchy can be exported as a tree file in Newick-format.

The second mode of operation is for the analysis of processed scanner data, i.e., tables with data for foreground and background intensities of the individual spots. The presence or absence of a hybridization signal is checked by a threshold criterion. The foreground-background intensity contrasts are normalized with respect to intensities of the negative control spots (Loy et al. 2002). Here intensity data of multiple copies (blocks) of the spots on each chip are evaluated and means and standard deviations are computed. The results for the blocks on the chip are shown independently (Fig. 2A, bottom right) such that entire blocks can be excluded from the analysis. It is assumed that if some spots in a certain block are identified as outliers or if positive controls fail, the user should exclude the whole block from evaluation because of the questionable quality of hybridization. A false positive signal on a higher hierarchical level has consequences for the validity of lower levels, down to the species level: PhylochipAnalyzer marks all positive signals that are below the hierarchy level of a spot showing a negative signal, i.e., corrected lines are crossed out. Because a signal is marked positive when the majority of copies give signals above the threshold, a correction is always contradicting. The user should inspect whether the underlying probe is correctly designed or maybe placed in the wrong hierarchy level.

The user may export the evaluation results directly to an Excel-graph (Fig.2B) in which the signals are given as bars, labelled with the probe identifier. The size of a bar indicates the quality above the threshold, i.e., the longer the bar, the stronger the evidence for a positive signal. All data are shown with error bars of the mean due to the variance over the different blocks.

Validation

The PhylochipAnalyzer was used to analyse data retrieved from a hybridization of PCR-products of *Micromonas pusilla* 18S rDNA to a phylochip that contained 44 probes, including a hierarchical probe-set for the Prasinophyte genus *Micromonas*. The hierarchical probe-set consisted of six probes that bind, respectively, at the level of Kingdom (EUK 1209, EUK 328), Class (Chlo01, Chlo02), Clade or Order (Pras 04) and Genus (Micro01) to *Micromonas pusilla*. The additional probes on the chip identified other phytoplankton taxa, a negative control, and two positive controls. Fluorescence images of the hybridized phylochips were taken with the Genepix 4000B Scanner (Axon Instruments Inc. USA). The signal intensities were quantified using the GenePix 6.0 software (Axon Instruments Inc. USA). Raw data were saved as a GPR-file and imported to the PhylochipAnalyzer-program. The computation of the raw data with the PhylochipAnalyzer-program identified only positive signals for the perfectly matching probes. For those probes, a signal/noise ratio was calculated that was above the threshold. The complete hierarchical probe set resulted in positive signals, therefore the signal for Micro01 can be considered truly positive (see Fig. 1 and Fig. 2B).

Discussion

The program simplifies tremendously the time consuming tasks of data processing of results from hierarchical phylochips. This is from particular interest, if high-throughput data are analyzed. The program is flexible with respect to configuration because the user can influence the threshold criterion by modifying the code that is implemented as a Delphi-script. This allows arbitrary modifications of the basic formula of data processing. Other formats of intensity description can easily be converted into appropriate GAL-format. On screen, the user may change the threshold value (default 2) interactively for sensitivity studies and recalculation. The rather simple criterion for elimination of false positives could be extended towards more quantitative measures. We plan to extend the program for quantitative analysis, i.e., spots from higher hierarchical levels are expected to show stronger signals than the lower hierarchical spots because they target more individuals. Multi-chip comparative analysis (e.g., clustering) for time-series analysis is also a desirable feature. The proposed XML-format for hierarchy representation can be seen as a prototype for standardization in phylochip hierarchy description. It is now necessary to introduce community standards for the representation of both, chip description and data-processing details. For gene-expression analysis by means of DNA-microarrays guidelines already exist (Brazma et al. 2001). Standards for phylochip design and processing description are considered to be a prerequisite for permanent archiving of publication supplemental data accompanied by catalogues of metadata in repositories.

Acknowlegement

The work was partly funded by the EU-project FISH&CHIPS (GOCE-CT-2003-505491). We thank the group of F.O. Glöckner for providing a data file in Perkin-Elmer format to verify the applicability of the program.

Copyright, Download

The copyright is specified by the author of the software (PB). The use of PhylochipAnalyzer is free of charge. For software and supplemental material see http://www.awi.de/en/go/phylochipanalyzer .

References

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball WA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature Genetics 29, 365 - 371.

Call DR (2005) Challenges and opportunities for pathogen detection using DNA microarrays. Critical Reviews in Microbiology 31, 91-99.

Csako G (2006) Present and future of rapid and/or high-throughut methods for nulcleic acid testing. Clinica Chimica acta 363, 6-31.

Dondrup M, Goesmann A, Bartels D, Kalinowski J, Krause L, Linke B, Rupp O, Sczyrba A, Puhler A, Meyer F (2003) EMMA: a platform for consistent storage and efficient analysis of microarray data. J Biotechnol 106:135-46.

Groben R, John U, Eller G, Lange M. and Medlin, L.K. (2004) Using fluorescently- labelled rRNA probes for hierarchical estimation of phytoplankton diversity. Nova Hedwigia. 79, 313-320.

Guschin DY, Mobarry BK, Proudnikov D, Stahl DA, Rittmann BE, Mirzabekov AD (1997) Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. Appl. Environm. Microbiol. 63, 2397-2402.

Guillou L, Moon-van-der-Staay SY, Claustre H, Partensky F, Vaulot D (1999) Diversity and abundance of Bolidophyceae (Heterokonta) in two oceanic regions. Appl. Environ. Microbiol. 65, 4528-4536.

Lange M, Guillou L, Vaulot D, Simon N, Amann RI, Ludwig W, Medlin LK (1996) Identification of the class Prymnesiophyceae and the genus *Phaeocystis* with ribosomal RNA-target nucleic acid probes detected by flow cytometry. J. Phycol. 32, 858-868

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature 405, 827-836.

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J, Schleifer KH, Wagner M (2002). Oligonucleotide Microarray for 16S rRNA Gene-Based Detection of All Recognized Lineages of Sulfate-Reducing Prokaryotes in the Environment., Appl. Environm. Microbiol. 68, 5064-5081.

Maidak BL, Cole JR, Lilburn T G, Parker CTJ, Saxman PR, Farris RJ, Garrity, GM, Olson GJ, Schmidt TM, Tiedje JM (2001) The RDP-II (Ribosomal Database Project). Nucleic. Acids Res. 29, 173-174.

Medlin LK, Metfies K, Mehl H, Wiltshire K, Valentin K (2006) Picoeukaryotic plankton diversity at the Helgoland time series site as assessed by three molecular methods. Microbial Ecology, 52, 53-71.

Metfies K, Medlin L (2004). DNA Microchips for Phytoplankton: The Fluorescent Wave of the Future. Nova Hedwigia, 79, 321-327.

Peplies J, Glöckner FO, Amann R (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Appl. Environ. Microbiol., 69(3), 1397-407.

Stoughton RB (2005). Applications of DNA Microarrays in Biology. Annual Review of Biochemistry, 74, 53-82.

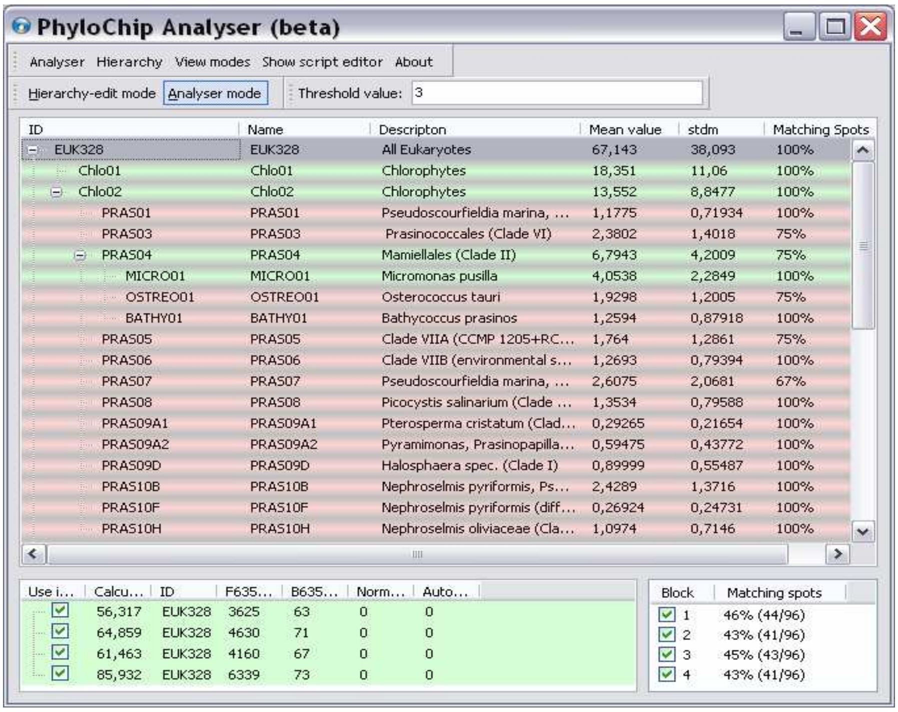
Rensink WA, Buell CR (2005) Microarray expression profiling resources for plant genomics. Trends in Plant Sciences 10, 603-609.

Vaquerizas JM, Conde L, Yankilevich P, Cabezon A, Minguez P, Diaz-Ulriarte R, Al-Shhrour F, Herrero J, Dopazo J (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Res. 33(Web Server issue):W616-20.

Woese C.R. (1987). Bacterial evolution. Microbiological Reviews. **51**, 221-271.

Figure 2:

A



PhyloChip Analyser (beta)

Analyser Hierarchy View modes Show script editor About

Hierarchy-edit mode | Analyser mode | Threshold value: 3

ID	Name	Descripton	Mean value	stdm	Matching Spots
⊟ EUK328	EUK328	All Eukaryotes	67,143	38,093	100%
Chlo01	Chlo01	Chlorophytes	18,351	11,06	100%
⊟ Chlo02	Chlo02	Chlorophytes	13,552	8,8477	100%
PRAS01	PRAS01	Pseudoscourfieldia marina, ...	1,1775	0,71934	100%
PRAS03	PRAS03	Prasinococcales (Clade VI)	2,3802	1,4018	75%
⊟ PRAS04	PRAS04	Mamiellales (Clade II)	6,7943	4,2009	75%
MICRO01	MICRO01	Micromonas pusilla	4,0538	2,2849	100%
OSTREO01	OSTREO01	Osterococcus tauri	1,9298	1,2005	75%
BATHY01	BATHY01	Bathycoccus prasinos	1,2594	0,87918	100%
PRAS05	PRAS05	Clade VIIA (CCMP 1205+RC...	1,764	1,2861	75%
PRAS06	PRAS06	Clade VIIB (environmental s...	1,2693	0,79394	100%
PRAS07	PRAS07	Pseudoscourfieldia marina, ...	2,6075	2,0681	67%
PRAS08	PRAS08	Picocystis salinarium (Clade ...	1,3534	0,79588	100%
PRAS09A1	PRAS09A1	Pterosperma cristatum (Clad...	0,29265	0,21654	100%
PRAS09A2	PRAS09A2	Pyramimonas, Prasinopapilla...	0,59475	0,43772	100%
PRAS09D	PRAS09D	Halosphaera spec. (Clade I)	0,89999	0,55487	100%
PRAS10B	PRAS10B	Nephroselmis pyriformis, Ps...	2,4289	1,3716	100%
PRAS10F	PRAS10F	Nephroselmis pyriformis (diff...	0,26924	0,24731	100%
PRAS10H	PRAS10H	Nephroselmis oliviaceae (Cla...	1,0974	0,7146	100%

Use i...	Calcu...	ID	F635...	B635...	Norm...	Auto...
☑	56,317	EUK328	3625	63	0	0
☑	64,859	EUK328	4630	71	0	0
☑	61,463	EUK328	4160	67	0	0
☑	85,932	EUK328	6339	73	0	0

Block	Matching spots
☑ 1	46% (44/96)
☑ 2	43% (41/96)
☑ 3	45% (43/96)
☑ 4	43% (41/96)

B

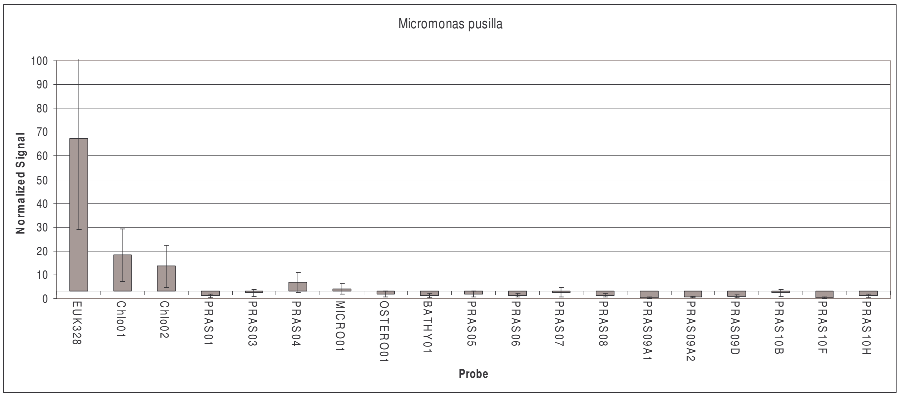


Figure 2. A: Screenshot of the analyser-mode. Any set of molecular probes can be organized as a user defined phylogenetic tree by a drag and drop function in editor-mode. The screenshot displays a tree of probes that bind to Prasinophytes at different hierarchical levels. The bottom part shows an individual probe result for the selected probe (EUK328, top part). **B**: Output of signal-noise values in graphical form.

Figure 1:

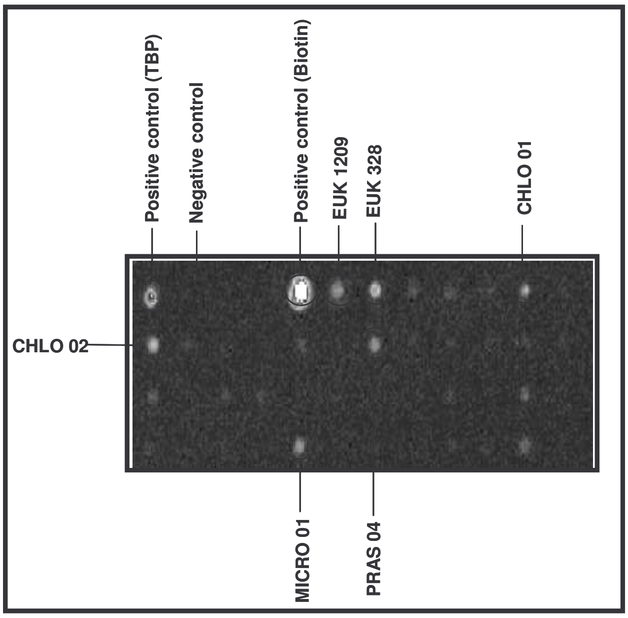


Figure 1: The 18S rDNA of *Micromonas pusilla* was hybridized to a set of 44 probes. The set of probes contained a hierarchical set that binds to the 18S rDNA of *M. pusilla* at four different taxonomic levels (EUK 1209, EUK 328, Chlo01, Chlo02, Pras04 and Micro01).