Q-440: Hybridization Efficiency of Probes Targeting 16S rRNA Genes Using the Affymetrix GeneChip Platform

T. Z. DeSantis[1,2] K. D. Hansen[3] E. L. Brodie[1] Y. M. Piceno[1] J. Bullard[3] P. Hu[1] G. L. Andersen[1,2]

[1]Ecology Department, Earh Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA [3]University of California, Berkeley, Berkeley, CA [2]Virtual Institute of Microbial Stress and Survival (VIMSS.lbl.gov)

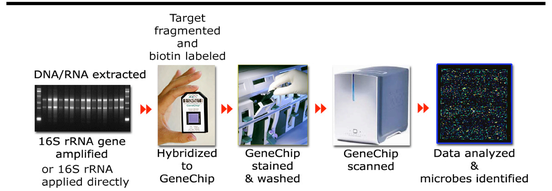
Abstract

Background: Detection of diverse 16S rRNA gene types in complex mixtures can be achieved using arrays of probes targeting specific sequences in 16S rRNA genes. Whereas probes for expression arrays are designed to leverage the diversity among various genes in one genome, 16S probes rely upon the diversity of the same gene found in many genomes. Also, expression arrays are validated by their accurate estimation of changes in analyte concentration, but 16S arrays are expected to provide definitive present-absent scoring of each prokaryotic taxa. The degree of uniqueness of a probe for a particular target species or other defined operational taxonomic unit will dictate its reliability but has yet to be quantified for prediction of hybridization accuracy. Methods: To obtain these metrics, amplicons of the16S rRNA gene from *Francisella tularensis* were fragmented, labeled and isothermally hybridized to replicate Affymetrix custom arrays containing 491,069 unique 25mer probes with various degrees of probe-target complimentarity, melting temperature, and secondary structure potential. Hybrid abundance at each probe location was determined by florescence intensity. Results: As expected, probes exactly complimentary to the target but with various sequence composition produced intensities ranging over 3 orders of magnitude, yet replicate probes on the same array produced a coefficient of variation under 10%. Although mismatching probes were able to capture target sequence, a general decrease in intensity was observed with probes divergent from the target. Conclusion: The data collected allows the development of a probabilistic model that aids in predicting the confidence that a probe's response is due to the presence of the corresponding target in solution.

Introduction

Challenge:
- Create universal 16S rRNA gene microarray.
- Limited sequence diversity among same gene across many genomes.
- Pick probes specific for each taxonomic cluster.
- How unique does a "specific" probe need to be for reliable detection?

Approach:
- One target gene was hybridized against an array of probes.
- Degree of similarity between probe and target was varied to characterize cross hybridization.
- Determine variation among redundant probes and replicate arrays.
- Find an affinity metric which best predicts a probe's response.
- Estimate probe failure rate.
- Estimate false positive rate where a false positive is any probe predicted to be negative yet experimentally was positive.
- Attempt to find a predictor or combination of predictors that minimize both the over-prediction of cross-hybridization and the number of false positives.



Methods

Probes:
491,069 unique 25mers, synthesized at 506,944 positions using Affymetrix high density platform.

2207 25mers tiled 3 or more times for redundancy.

Various degrees of probe-target complimentarity, melting temperature, and secondary structure potential.

Single Target Molecule:
1E9 copies (1.7 fmoles) of 1.5kb 16S rRNA gene from *Francisella tularensis*, were fragmented to ~100 bp, end-labeled and isothermally hybridized (48°C) to four replicate arrays.

After washing and staining, images were captured with fluorescence scanning and intensities were recorded in arbitrary units (a.u.).
5.9E9 copies (9.9 fmoles) of a pre-labeled control 25mer were added for image orientation (Oligo213).

F. tularensis target exactly complements 963 of the 491,069 unique 25mers represented at 1,826 of the 506,944 coordinates on the array.

Hybrid abundance at each probe location was determined by florescence intensity. No background subtraction.

Normalization - All intensities from each array were multiplied by a factor in order to produce a constant average intensity of all 963 25mers which exactly complement the F. tularensis 16S rRNA gene.

For each of the 491,069 unique 25mers, calculate:
- PM_mean - average of replicate spots
- MM_mean - average of all spots with central (position 13) mismatch relative to PM
- PM minus MM. If (PM-MM) < 1, then shift to 1.

Predictors:
Target-independent probe properties were measured: Percent G+C, perfect complement Melting Temperature by ThermAlign (Kaderali, 2005), Secondary Structure Potential by RNAfold(Zuker, 1981) assuming 48°C aqueous solution.

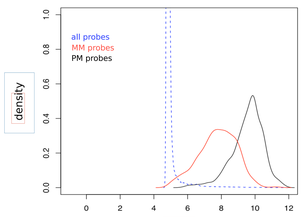
Target-dependent probe properties (affinities) were calculated: Alignment by BLAST (Altschul, 1990), perfect/imperfect complement Melting Temperature by ThermAlign, longest contiguous centrally-positioned nmer in common, percent of shared 9mers, positional-dependent nearest neighbor score (PDNN) modified from (Zhang, 2003).

Observe probe response in relation to probe attributes and probe-target attributes.

"Positive" response fixed at 256 a.u..

Results and Observations

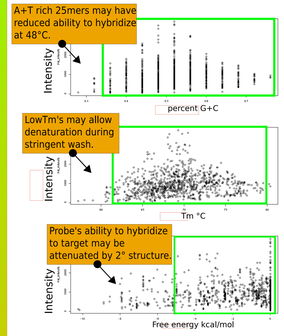
Probe response distribution



Precision in intensity measurements

	category	count	n observations per probe	mean coefficient of variation
inter-chip	Exact complement set	1,826	4	0.0964
	Near complement set	1,374	4	0.1162
	All probes	506,944	4	0.0775
intra-chip	25mers redundantly tiled	2,207	3 to 19	0.0964

Probe Characteristics and Filtering



Predictor-response analysis was limited to probes with G+C content between 35% and 70%, Tm between 62°C and 80°C, and free energy from secondary structure formation above -5kcal/mol.

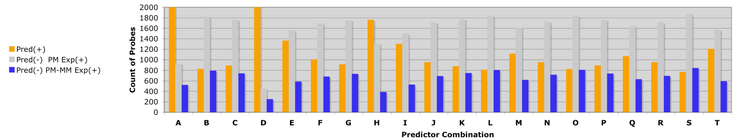
Additionally, any probe complementing Oligo213 along seven or more nucleotides was not included in further analysis. The concentration of Oligo213 is 6X that of the amplicon. Furthermore each Oligo213 25mer is covalently labeled during synthesis while, in contrast, the digested ~100mer amplicons are end-labeled with incomplete efficiency. The effective mean label per 25mer can be 30X greater for Oligo213 than the 16S rRNA gene amplicon.

372,003 unique 25mers met all conditions. Of these, 711 exactly matched the target (shown in green below).

Predictor Reliability

Predictor Combination

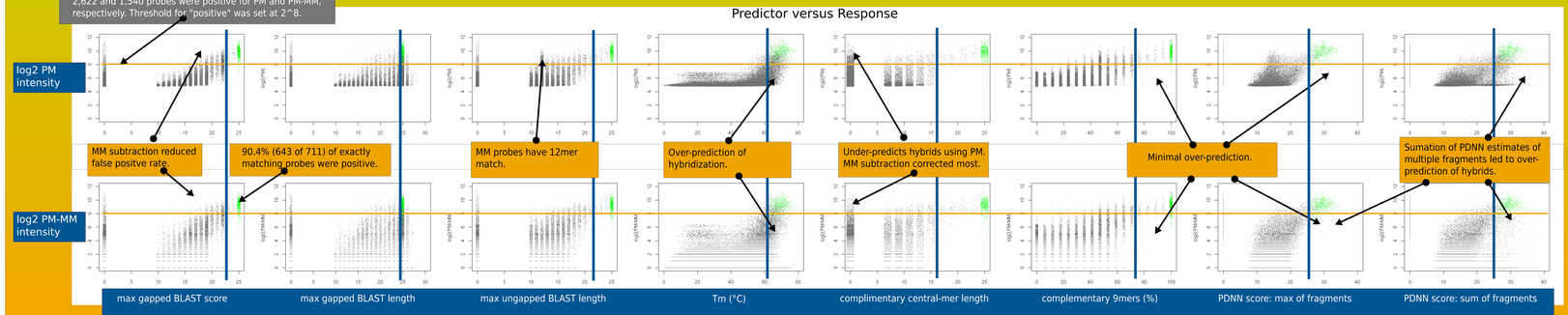
Predictors	Relative compute time	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
blast gapped score	+++++	22	23																		
blast ungapped length	+++++			22																	
Tm	+++++++++				62																
common central-mer length	+					15	17	19									17	17	19	19	15
common 9mers %	+								64	70	77	83	89				77	77	89	89	64
PDNN max	++++++													25	26	27					
Boolean:																	AND	OR	OR	AND	AND
Predicted(+) *cross-hybe breadth*		2464	827	893	6518	1367	1009	914	1764	1306	956	880	811	1119	956	824	896	1069	956	769	1209
Predicted(-) PM Experimental (+)		926	1822	1765	449	1562	1691	1753	1302	1497	1710	1773	1836	1613	1730	1839	1762	1639	1712	1877	1571
Predicted(-) PM-MM Experimental (+)		521	794	743	252	587	681	732	387	529	691	747	805	618	716	810	740	632	693	844	592



In all cases, MisMatch subtraction produced a lower count of false positives (Pred(-) Exp(+)) using a fixed intensity threshold (256 a.u.) for determining positives. Tm (combo D) and gapped BLAST score (combo A) allowed low numbers of false positives but vastly over-predicted cross-hybridizations. Scoring the 9mers in common (combos H-L) also provided a low rate of false positives and gave a 2 to 3 fold improvement in accuracy in predicting cross-hybridizations. Searching for contiguous centrally-positioned nmers within a probe found also in the target (combos E-G) was almost as reliable as matching all 9mers. Dependency on the maximum ungapped BLAST fragment length (combo C) with or without position dependent nearest-neighbor weighting (PDNN) (combos M-O) produced a balance between cross-hybridization prediction and false positives. Combining predictors can allow a slight decrease in false positives without extensive over-predictions of cross hybridizations.

Predictor versus Response

2,622 and 1,540 probes were positive for PM and PM-MM, respectively. Threshold for "positive" was set at 2^8.



Conclusions:
The intensity distributions of the Perfectly Matching (PM) probes and single base-pair MisMatching (MM) probes overlapped but were distinct from the background.

Intensity measurements were reproducible.

Broad ranges of probe GC content, Tm and secondary structure potential can be allowed in probe design. 9.6% probe failure rate indicates that multiple probes should be used for each target amplicon.

PM-MM scoring created less false positives compared to straight PM scoring.

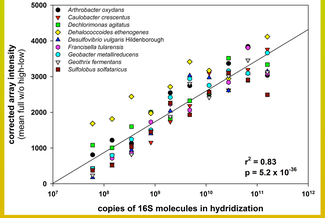
When multiple fragments from the target have ability to hybridize to same probe, the effect from the most stable hybrid is a better predictor of intensity than the sum of the fragments.

To achieve a low rate of unexpected (false) positives, a liberal cross-hybridization predictor is utilized. The converse is also true.

Using the central 17mer match as a predictor, 0.18% (681 of 372,003) of probes produced unexpectedly high hybridizations. Of the positives, 44% (681 of 1540) did not match a central 17mer to the target. This false positive rate is reduced to 25% (387of 1540), or even lower, when predicting with 9mers.

If array design does not include multiple probes for each taxa, probes must be choosen using a broad cross-hybe predictor to minimize unexpected positives.

Next.....
Determine probes that predictably hybridize through a range of concentrations in a complex mixture.



References:
- Altschul, 1990, J Mol Biol
- Kaderali, 2005, http://www.zaik.uni-koeln.de
- Liebich, 2006, AEM
- Zhang, 2003, Nat Biotech
- Zuker, 1981, NAR

Acknowledgements:
Microarray development and synthesis were funded in part by the Department of Homeland Security under grant number HSSCHQ04X00037. The computational infrastructure was provided in part by the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Program and the Natural and Accelerated Bioremediation Research Program. This work was performed under the auspices of the U.S. Department of energy the the University of California, Lawrence Berkeley National Laboratory, under contract no. DE-AC02-05CH11231.