BMC Bioinformatics

Methodology article

Mathematical design of prokaryotic clone-based microarrays

Bart Pieterse*[1,2,3], Elisabeth J Quirijns[4,5], Frank HJ Schuren[2] and Mariët J van der Werf[1,2]

Address: [1]Wageningen Centre for Food Sciences. Diedenweg 20, 6700 AN Wageningen, The Netherlands, [2]TNO Quality of Life. Utrechtseweg 48, 3700 AJ Zeist, The Netherlands, [3]BioDetection Systems, Kruislaan 406, 1098 SM, Amsterdam, The Netherlands, [4]Wageningen University and Research Centre, Systems and Control Group, Department of Agrotechnology and Food Sciences, Bornsesteeg 59, 6708 PD Wageningen, The Netherlands and [5]HAS Den Bosch, Onderwijsboulevard 221, 5200 MA, Den Bosch, The Netherlands

Email: Bart Pieterse* - Bart.Pieterse@bds.nl; Elisabeth J Quirijns - Pieterse-Quirijns@wanadoo.nl; Frank HJ Schuren - Schuren@voeding.tno.nl; Mariët J van der Werf - vanderwerf@voeding.tno.nl

* Corresponding author

Published: 28 September 2005

BMC Bioinformatics 2005, **6**:238 doi:10.1186/1471-2105-6-238

Received: 19 May 2005
Accepted: 28 September 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/238

Abstract

Background: Clone-based microarrays, on which each spot represents a random genomic fragment, are a good alternative to open reading frame-based microarrays, especially for microorganisms for which the complete genome sequence is not available. Since the generation of a genomic DNA library is a random process, it is beforehand uncertain which genes are represented. Nevertheless, the genome coverage of such an array, which depends on different variables like the insert size and the number of clones in the library, can be predicted by mathematical approaches. When applying the classical formulas that determine the probability that a certain sequence is represented in a DNA library at the nucleotide level, massive amounts of clones would be necessary to obtain a proper coverage of the genome.

Results: This paper describes the development of two complementary equations for determining the genome coverage at the gene level. The first equation predicts the fraction of genes that is represented on the array in a detectable way and cover at least a set part (the minimal insert coverage) of the genomic fragment by which these genes are represented. The higher this minimal insert coverage, the larger the chance that changes in expression of a specific gene can be detected and attributed to that gene. The second equation predicts the fraction of genes that is represented in spots on the array that only represent genes from a single transcription unit, which information can be interpreted in a quantitative way.

Conclusion: Validation of these equations shows that they form reliable tools supporting optimal design of prokaryotic clone-based microarrays.

Background

In the past decade, whole transcriptome comparison by microarray hybridizations has proven to be an effective tool for studying genome wide gene responses. The general approaches for the development of microarrays are based on the completely annotated genome sequence of an organism. Usually each spot on the array represents one open reading frame (ORF). Whereas this approach has clear advantages for strains for which the complete

annotated genome sequence is available, it is not applicable to strains for which this is not the case.

A method that allows for the rapid construction of microarrays for which the completely annotated genome sequence is not required is by the construction of a clone-based array. In this approach, a chromosomal DNA library is constructed from the strain of interest. From this library the genomic fragments, the inserts, are amplified from the clones by PCR with generic primers and spotted on the array-slide [1,2].

The major differences between ORF-based and clone-based arrays with respect to the data interpretation are that in case of clone-based arrays the differential signals can only be linked to a specific gene after the DNA fragment from the spot of interest on the array has been sequenced, and that it is beforehand uncertain whether a gene is represented on the array. Moreover, whereas ORF-based microarrays exclusively generate gene specific data, a differential signal within a spot on a clone-based array can originate from multiple genes on the insert that are not necessarily linked at the transcriptional level.

The extent of these limitations can be quantified by estimating the genome coverage by the spots present on the array. The standard formulas for estimating the genome coverage of a DNA library, the Clark-Carbon equation [3] and the Lander-Waterman equation [4], determine this coverage at the nucleotide level. In other words, they consider the genome as a set of nucleotides, which is useful when the library is to be used for genome sequencing. However, these formulas will overestimate the required number of clones for hybridization purposes. The reason for this is that for hybridization purposes small overlapping fragments that allow for specific binding of the labeled cDNA suffice. Akopyants *et al*. [5] developed an equation for the estimation of the fraction of genes that are at least partially represented. This formula is directly derived from classical probability calculations and contains the organism specific variables genome size and average gene size. Due to the fact that Akopyants *et al*. determine the genome coverage at the gene level, and consider a gene represented if a fragment is present that is large enough to hybridize to and large enough to identify the gene, the required number of clones to obtain a certain coverage is reduced.

A general drawback of these three formulas is that they give no insight into the fraction of genes for which specific data can be generated in a transcriptomics experiment. The data from a spot are considered specific if the expression ratios from the quantified signal from that spot can directly be related to the gene(s) represented by the spot. This is not the case if DNA from multiple (neighboring) transcription units is present in one spot, since it would be uncertain which gene is responsible for which part of the total signal from that spot.

In this paper, two formulas were developed that enable for mathematical predictions of genome coverage by a prokaryotic clone based-array at the gene level as a function of genome size, number of clones, insert size, and either the minimal part of the insert that is covered by the gene or the minimal overlap of the gene and the insert: the minimal insert coverage (MIC) equation, and the gene specific information (GSI) equation.

In order to develop equations that are applicable to a broad range of microorganisms, model datasets were generated for 15 prokaryotes originating from several genera (Table 1) that functioned as templates on which the MIC- and GSI-equations were fitted. The resulting formulas were validated on 10 other prokaryotic species.

Description of the developed equations

Minimal Insert Coverage (MIC)-equation

Since the generation of inserts for a genomic DNA library is a random process, a large part of the represented genes may be co-represented with other genes by one spot on the microarray. This complicates data interpretation since it introduces an uncertainty on which gene or genes are responsible for differential signals from these spots. The impact of differential expression of a specific gene on the observed difference of the signal from a spot will be larger when a larger part of the genome fragment in that spot is covered by that gene. Moreover, the larger the part of the insert that is covered by a specific gene, the larger the chance that differential signals for the spot can be attributed to that gene, and the higher the chance that differential expression levels from that gene result in a statistically significant differential signal on the array.

The MIC-equation anticipates to this effect by predicting the number of genes that are (at least partially) present on an insert *and* cover at least a predefined part of the insert (*DIC*). This predefined part is defined as a percentage of the total insert. E.g. if the insert size is 1000 base pairs and the predefined minimal insert coverage (*DIC*) is set at 50%, then at least 500 bp of that gene should be present on an insert to be considered as represented by the array. Genes smaller than the size of the predefined part of the insert, are considered as not represented on the array.

Gene Specific Information (GSI)-equation

Information on differential expression of a gene can only be quantitative and specific for that gene if it originates from a spot that only represents genes from a single transcription unit, assuming that all genes within one transcription unit are equally expressed. This was the

Table 1: Overview of prokaryotes from several genera with their genes/transcription unit-ratio. Microorganisms that were used for model development (M) or validation (V) of the MIC- and the GSI-equation are depicted in the list.

Genus	Organism	genes/TU (R)	Model (M) or validation (V) strain
Proteobacteria Gammaproteobacteria Enterobacteriales	*Escherichia coli* K-12 MG1655	1.6	
	Escherichia coli O157:H7 EDL933	1.6	
	Escherichia coli CFT073	1.6	M
	Salmonella typhi CT19	1.4	
	Salmonella typhimurium LT2	1.6	
	Yersinia pestis CO92	1.4	
	Shigella flexneri 2a str. 2457T	1.5	
	Buchnera aphidicola Sg	1.5	V
	Wigglesworthia glossinidia	1.5	
ProteobacteriaGammaproteobacteria Pasteurellales	*Haemophilus influenzae Rd*	1.7	
	Pasteurella multocida PM70	1.7	V
Proteobacteria Gammaproteobacteria Xanthomonadales	*Xylella fastidiosa* 9a5c	1.5	
	Xanthomonas campestris ATCC33913	1.5	V
Proteobacteria Gammaproteobacteria Vibrionales	*Vibrio cholerae* El Tor N16961	1.8	M
	Vibrio parahaemolyticus RIMD2210633	1.5	
	Vibrio vulnificus CMCP6	1.5	
Proteobacteria Gammaproteobacteria Pseudomonadales	*Pseudomonas aeruginosa* PA01	1.6	M
	Pseudomonas putida KT2440	1.6	
Proteobacteria Gammaproteobacteria Legionellales	*Coxiella burnetii* RSA 493	1.6	
Proteobacteria Betaproteobacteria	*Neisseria meningitidis* Z2491	1.6	M
	Ralstonia solanacearum GMI1000	1.6	
Proteobacteria Epsilonproteobacteria	*Helicobacter pylori* 26695	2.3	M
	Campylobacter jejuni NCTC11168	2.7	M
Proteobacteria Alphaproteobacteria	*Rickettsia prowazekii* Nadrid E	1.4	V
	Sinorhizobium meliloti 1021	1.5	
	Agrobacterium tumefaciens C58	1.5	
	Brucella suis 1330	1.5	
	Caulobacter crescentus	1.5	
Firmicutes Bacillales	*Bacillus subtilis* 168	1.6	M
	Oceanobacillus iheyensis HTE831	1.6	
	Staphylococcus aureus MW2	1.6	
	Listeria monocytogenes EGD-e	1.8	M
	Listeria innocua Clip11262	1.8	
Firmicutes Clostridia	*Clostridium acetobutylicum* ATCC824	1.6	
	Clostridium tetani E88	1.6	
	Thermoanaerobacter tengcongensis MB4T	2.0	
Firmicutes Lactobacillales	*Lactococcus lactis* IL1403	1.5	M
	Streptococcus agalactiae 2603	1.8	
	Streptococcus pneumoniae R6	1.8	
	Lactobacillus plantarum WCFS1	1.6	M
	Enterococcus faecalis V583	1.8	
Firmicutes Mollicutes	*Mycoplasma pneumoniae* M129	2.1	M
	Mycoplasma genitalium G37	3.1	V
	Mycoplasma penetrans HF-2	1.6	
	Ureaplasma urealyticum (serovar 3)	2.1	
Actinobacteria	*Mycobacterium tuberculosis* H37Rv	1.7	M
	Corynebacterium glutamicum ATCC 13032	1.5	
	Streptomyces coelicolor A3(2)	1.4	
	Tropheryma whipplei Twist	1.9	
	Bifidobacterium longum NCC2705	1.3	V
Fusobacteria	*Fusobacterium nucleatum* ATCC25586	2.0	V
Chlamydia	*Chlamydia trachomatis* (serovar D)	1.6	

Table 1: Overview of prokaryotes from several genera with their genes/transcription unit-ratio. Microorganisms that were used for model development (M) or validation (V) of the MIC- and the GSI-equation are depicted in the list. *(Continued)*

	Chlamydophila pneumoniae AR39	1.6	
Spirochete	*Borrelia burgdorferi* B31	1.8	
	Treponema pallidum Nichols	1.9	
	Leptospira interrogans 56601	1.5	
Bacteroid	*Bacteroides thetaiotaomicron* VPI-5482	1.8	M
Cyanobacteria	*Thermosynechococcus elongatus* BP-1	1.6	
	Nostoc sp. PCC 7120	1.2	
Green sulfur bacteria	*Chlorobium tepidum* TLS	1.6	
Deinococcus	*Deinococcus radiodurans* R1	1.5	V
Hyperthermophilic bacteria	*Aquifex aeolicus* VF5	2.1	
	Thermotoga maritima MSB8	3.0	V
Archae Euryarchaeota	*Methanococcus jannaschii* DSM2661	1.8	M
	Pyrococcus furiosus DSM3638	2.0	M
	Archaeoglobus fulgidus DSM4304	2.1	
	Thermoplasma acidophilum DSM1728	1.5	
	Methanosarcina acetivorans C2A	1.3	V
	Methanosarcina mazei Goe1	1.3	
	Pyrococcus abyssi	2.1	
Archae Crenarchaeota	*Aeropyrum pernix* K1	2.0	
	Sulfolobus solfotaricus P2	1.6	
	Pyrobaculum aerophilum IM2	1.7	

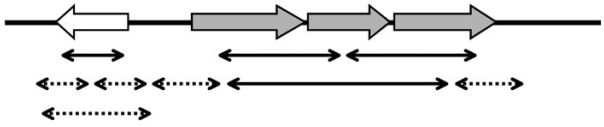

Figure 1
Schematic representation of the criteria that were applied to determine whether gene specific information is generated by a specific insert. The upper line represents a genome fragment in which the block arrows represent genes. Arrows with a gray filling belong to the same transcription unit. The thinner lines represent possible locations of the inserts. The dashed lines represent inserts for which no gene specific information can be generated, since they contain genomic material that possibly belongs to another transcription unit.

requirement that was set for a gene to be considered represented according to the gene specific information (GSI) equation. The criteria for spots that could generate gene specific information are visualized in Fig. 1. One of the variables in the GSI-equation, the minimal overlap (O_{mf}), allows one to set the minimal number of base pairs that are required for identification of a specific gene or transcript on an insert on the clone-based array.

Dataset preparation

Fifteen prokaryotes from various genera were selected as model species (Table 1). Genome data from these microorganisms were used for the generation of species-specific values for the expected fraction of represented genes as a function of the genome size (*GS*), number of clones (*N*), insert size (*IS*), and either *DIC* or O_{mf}. Coordinates from all annotated genes from these organisms were obtained from GenBank, and were used to determine the gene sizes. In addition, information was obtained on the start and stop coordinates from the transcription unit to which the gene belongs, and the position of the gene in this transcription unit. It was assumed that transcription units start at the first base pair of the first gene and finish at the last base pair of the last gene. This information was generated by the combination of intergenic region based transcription unit predictions, generated by Moreno-Hagelsieb and Collado-Vides [6], with gene coordinates from GenBank.

The genome size (*GS*) could be included as a fictitious variable in the datasets, since not the species-specific genome size, but the species-specific gene size distribution and genome organization in genes and transcription units were relevant.

It was assumed that each possible genome fragment of the size of the insert size (*IS*) has an equal chance of being represented. To achieve this, fragments should be generated by physical fragmentation, and not by the use of endonucleases.

Dataset preparation for the fitting procedure for the MIC-equation

For each model organism, the fraction of the represented genes was determined for multiple combinations of the number of clones (*N*), fictitious genome size (*GS*), the insert size (*IS*), and the minimal insert coverage (*DIC*) in

the ranges depicted in Table 2. In total, 140 different combinations of values for these variables were tested per strain. This was performed by first calculating the probability value of being represented per gene, and subsequently calculating the average of the probability values from all genes from the organism.

The following formulas were developed for the calculation of the probability value per gene:

$$O_{mv} = \frac{IS \cdot DIC}{100} \quad (1)$$

$$Gene > O_{mv} \Rightarrow p = 1 - \left(1 - \frac{Gene + 1 + IS - 2 \cdot O_{mv}}{GS}\right)^N \quad (2)$$

$$Gene < O_{mv} \Rightarrow p = 0 \quad (3)$$

Dataset preparation for the fittimg procedure for the GSI-equation

For each organism, the fraction of genes for which specific information could be generated was determined for 114 different combinations of the number of clones (*N*), fictitious genome size (*GS*), the insert size (*IS*), and the minimal required overlap (O_{mf}) in the ranges depicted in Table 2. The represented fraction was determined by taking the average of the probability values per gene. Formulas were developed that describe different situations with respect to the localization and organization of the gene of interest on the insert (eq 4 – 15)

Formulas that were developed to determine the probability value for genes that are transcribed into a single gene transcript:

$$Gene \geq IS \Rightarrow p = 1 - \left(1 - \frac{Gene + 1 - IS}{GS}\right)^N \quad (4)$$

$$Gene < IS \Rightarrow p = 0 \quad (5)$$

Formulas that were developed to determine the probability value for genes that are at the beginning of a transcription unit:

$$BP_e \leq IS - O_{mf} \Rightarrow O_e = IS - BP_e \quad (6)$$

$$BP_e > IS - O_{mf} \Rightarrow O_e = O_{mf} \quad (7)$$

$$BP_e + Gene > IS \Rightarrow p = 1 - \left(1 - \frac{Gene + 1 - O_e}{GS}\right)^N \quad (8)$$

$$BP_e + Gene < IS \Rightarrow p = 0 \quad (9)$$

Formulas that were developed to determine the probability value for genes that are flanked at both sides by other genes that belong to the same transcription unit:

Table 2: Overview of the variables that were used for the model datasets on which the MIC- and the GSI-equation are based. Multiple combinations of the mentioned values were applied.

Variable	Values
N	500; 1500; 2500; 3500; 4500; 5500; 6500; 7500; 8500; 9500
IS	100; 300; 500; 700; 900; 1100; 1300; 1500; 2100; 2700; 3000
GS	0.5; 1.5; 2.5; 3.5; 4.5; 5.5; 6.5; 7.5; 8.5; 9.5
O_{mf}	50; 100; 150; 200; 250; 300; 350; 400; 450
DIC	10; 20; 30; 40; 50; 60; 70; 80; 90

$$BP_b \leq IS - O_{mf} \Rightarrow O_b = IS - BP_b \quad (10)$$

$$BP_b > IS - O_{mf} \Rightarrow O_b = O_{mf} \quad (11)$$

$$BP_e \leq IS - O_{mf} \Rightarrow O_e = IS - BP_e \quad (12)$$

$$BP_e > IS - O_{mf} \Rightarrow O_e = O_{mf} \quad (13)$$

$$BP_b + BP_e + Gene > IS \Rightarrow p = 1 - \left(1 - \frac{Gene + 1 + IS - O_b - O_e}{GS}\right)^N \quad (14)$$

$$BP_b + BP_e + Gene > IS \Rightarrow p = 0 \quad (15)$$

Models and fits

The datasets with the expected fractions of represented genes for the various combinations of parameters as presented in the previous section functioned as template for the fitting of the predictive equation for *MIC* and *GSI*.

MIC equation

From equation 2, which was used to determine the probability value per gene, it became apparent that organism-dependent gene size distribution influenced the expected number of represented genes on a clone based array. These organism dependent differences were neglected for the preparation of the MIC equation, which proofed to be justified when validating the MIC-equation (see validation section).

A polynome was developed as MIC model. In the polynome all variables were present in first and second order and in cross terms between two variables. Because of a high expected correlation between *IS* and *DIC* (based on equation 2), this relation was extended with a second order term composed of *IS* and *DIC*, resulting in:

$$p_{MIC} = a + b_1 \cdot DIC + b_2 \cdot DIC^2 + c_1 \cdot N + c_2 \cdot N^2 + d_1 \cdot GS + d_2 \cdot GS^2 + e_1 \cdot IS + e_2 \cdot IS^2 + f \cdot DIC \cdot IS + g \cdot DIC \cdot N + h \cdot DIC \cdot GS + i \cdot IS \cdot N + j \cdot IS \cdot GS + k \cdot GS \cdot N + l \cdot (IS \cdot DIC)^2 \quad (16)$$

The model datasets for the 15 model species were used together in the regression procedure to estimate the parameters in the MIC model. Linear regression using a standard least squares algorithm (fminsearch) provided by Matlab (The MathWorks) was applied to search the parameters that minimize the sum of squares (SSQ) defined as:

$$SSQ = \sum(p_{MIC,\exp} - p_{MIC,\mathrm{mod}})^2 \quad (17)$$

The resulting parameters are presented in Table 3. The average absolute deviation of the MIC equation from the model dataset was 0.0517.

GSI equation

From the model datasets for the GSI equation it appeared that an organism dependent variable had a strong influence on the calculated number of represented genes (results not shown). Analysis revealed a positive correlation between the number of represented genes and the species-dependent average number of genes per transcription unit, *R*. *R* was determined by dividing the total number of genes (GenBank) by the total number of predicted transcription units [6] (Table 1).

Starting-point for the GSI model was a second order polynome for all variables, extended with the cross terms between two variables. A set of parameters was estimated for each individual model species (results not shown). Parts which appeared to contribute less than 1% to p_{GSI} were not included, which resulted in the following relation:

$$p_{GSI} = a + b_2 \cdot O_{mf}^2 + c_1 \cdot N + c_2 \cdot N^2 + d_1 \cdot GS + d_2 \cdot GS^2 + e_1 \cdot IS + e_2 \cdot IS^2 + f \cdot O_{mf} \cdot IS + h \cdot O_{mf} \cdot GS + i \cdot IS \cdot N + j \cdot IS \cdot GS + k \cdot GS \cdot N \quad (18)$$

For each prokaryote a set of parameters was obtained by minimizing the SSQ, equivalent to equation 17. The average absolute deviation of the GSI equation from the model datasets was 0.0258.

In order to obtain one generic equation for the organism specific relations for p_{GSI}, the species specific values for the parameters (a - k) in equation 18 were related to the species related variable *R* by a linear relation:

$$parameter(a\text{-}k) = q + r \cdot R \quad (19)$$

in which *R* is species specific (Table 1). Since no dependency of a with *R* could be established, a was set at the average of all individual a values: 0.544. With this value the polynome was fitted again, and the final relations between the other parameters and *R* were determined (Table 3).

Validations

In order to validate the MIC- and the GSI-equation, datasets were generated (as previously described in the "dataset preparation" section) for ten validation species (Table 1). Represented gene fractions were calculated per species for all possible combinations for the variables as presented in Table 4 and distracted from the values as they were predicted by MIC- and the GSI-equations 16 and 18, respectively. The distributions of the residuals, i.e. the difference between predicted and the calculated fraction, for both equations are presented as histograms in Figures 2a and 2b. The residual distributions of both the MIC- and the GSI-equation approach the normal distribution with a slight tendency to underestimate the fraction of represented genes (Fig. 2a and 2b). Moreover, in Table 5 the reliability of the formulas is depicted as the fraction of predictions that differ less than 0.01, 0.05 and 0.10 from the real values. It should be noted that the indicated reliabilities relate to the range of variables as depicted in Table 4.

Deviations between the predicted fractions by the MIC-equation and the true values as they were determined for the validation species are mainly to be attributed to species-specific gene size distribution. In order to obtain one generic equation, and based on the accuracy of the equation in its current form (Table 5), it was decided not to include a species-specific variable.

Prediction of the optimum value for the insert size (IS)

Whereas an increase in *N* will always have a positive contribution to the fraction of represented genes, and an increase in *GS*, O_{mf}, and *MIC* a negative contribution, there may be an optimum *IS* that depends on the values of the other variables. This optimum can be estimated by differentiation of equation 16 and 18 to *IS* (dp/d*IS*).

For the determination of the optimal value for *IS* for the MIC-approach this results in the following equation:

$$\frac{dp_{MIC}}{dIS} = e_1 + 2e_2 \cdot IS + f \cdot DIC + i \cdot N + j \cdot GS + 2 \cdot l \cdot DIC^2 \cdot IS = 0 \Rightarrow$$
$$IS_{MIC-opt} = \frac{-e_1 - f \cdot DIC - i \cdot N - j \cdot GS}{2e_2 + 2l(DIC)^2} \quad (20)$$

For the determination of the optimal value for *IS* for the GSI-approach the equation is as follows:

$$\frac{dp_{GSI}}{dIS} = e_1 + 2e_2 \cdot IS + f \cdot O + i \cdot N + j \cdot GS = 0 \Rightarrow$$
$$IS_{GSI-opt} = \frac{-e_1 - f \cdot O - i \cdot N - j \cdot GS}{2e_2} \quad (21)$$

If the indicated values for IS_{opt} are outside the range of 0 to 2000 bp (the range that was applied for validation of the models) no optimum can be identified within the

Table 3: Values for the parameters in the MIC- and the GSI-equation.

parameter	MIC equation	GSI equation *q*	GSI equation *r*
a	4.85E-01	0.544	0
b_1	2.54E-03	*	*
b_2	-1.51E-05	-4.26E-08	-3.05E-07
c_1	1.27E-04	6.13E-05	1.46E-05
c_2	-5.22E-09	0	-1.96E-09
d_1	-1.22E-01	-7.84E-02	-1.06E-02
d_2	3.42E-03	3.31E-03	3.23E-04
e_1	3.95E-04	-5.36E-04	2.08E-04
e_2	-9.57E-08	9.73E-08	-4.62E-08
f	-9.85E-06	1.69E-08	3.42E-08
g	-4.61E-07	*	*
h	3.25E-04	2.55E-06	4.12E-06
I	-1.69E-08	-2.22E-08	5.47E-09
j	2.01E-05	2.42E-05	-6.04E-06
k	2.26E-06	-1.76E-06	1.30E-06
l	2.60E-11	*	*

ad *: this parameter is not present in the GSI equation

Table 4: Overview of the variables and the values used for these variables that were used for the datasets that were used for the validation of the MIC- and the GSI-equation. All possible combinations of the mentioned values were tested.

Variable	Values
N	1000; 4000; 7000; 10000
IS	100; 500; 1000; 1500; 2000
GS	1; 3; 5; 7
DIC	25; 50; 75
O_{mf}	100; 200

boundaries of the model. In these cases small values of *IS* will give the best results.

Influence of the average number of genes per transcription unit (R) *on the predicted values*

From the input variables for the MIC and GSI formulas, *N*, *IS*, *DIC* and O_{mf} are user-defined, while *GS* and *R* have to be estimated for the specific organism. Whereas current techniques allow for rapid and accurate estimations of *GS* [7-9], the organism specific value for *R* is difficult to determine for species from which little sequence information is available.

R was determined for 73 prokaryotes from multiple genera, as previously described in the "models and fits" section (Table 1). For 61 of the 73 strains in this list, *R* was within the narrow range from 1.5 – 2.0. Moreover these data indicate that accurate estimations of *R* can be made, based on the genus of the organism, with an exception for the mollicutes, the hyperthermophilic bacteria and the euryarchaeota.

The effect of false estimations of *R* was studied by the generation of validation sets as defined in Table 4 with the exception that higher or lower values for *R* were applied. The resulting values from the GSI-equation were compared with the true values (Table 6). It appeared that an over- or underestimation of 0.2 on *R* had limited effects on the fraction of predictions that differ less than 0.1 from the real values from the validation dataset (0.90 vs. 0.95 for the exact value of *R*). While an overestimation of 0.3 still results in 88% of the predictions that differ less than 0.1 from the real value from the validation dataset. This percentage was 80% in case of an underestimation of the same size.

Application

As an example for the applicability of the developed equations, the effect of different combinations of the number of clones and insert size was determined for a prokaryote with a genome size of 4 Mbp and an estimated value for *R* of 1.8 using equations 16 and 18. The effect of multiple combination of *N* and *IS* on p_{MIC} was determined for minimal insert coverage (DIC) values of 25%, 50% and 75%. The results are depicted in the contour plots in figure 3a–3c. The predicted fractions of represented genes for which gene specific information could be generated (p_{GSI}) with a minimal overlap between the insert and the gene of 100 bp is depicted in figure 4.

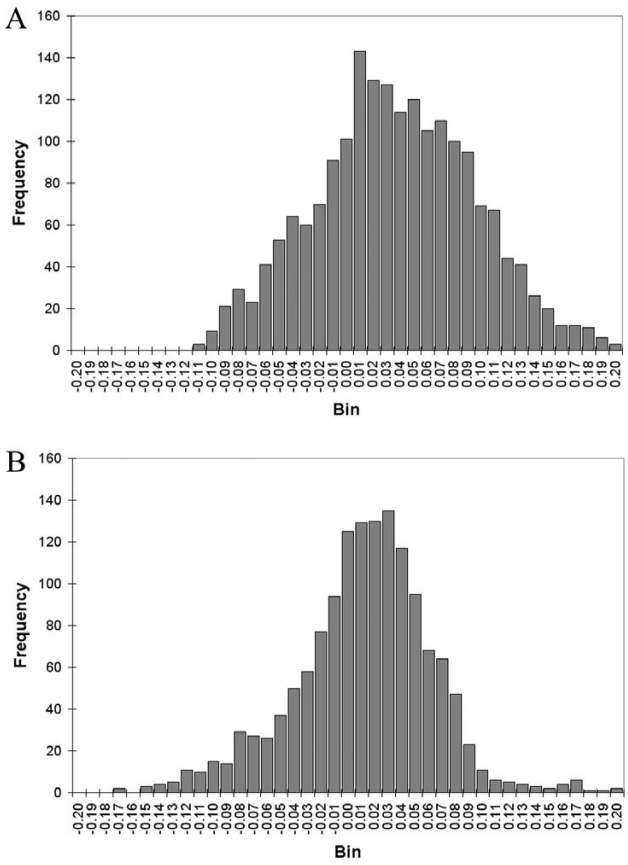

Figure 2
Histogram representations of the residuals from the validation of the MIC-equation (A) and the GSI-equation (B).

Table 5: Reliability of the MIC- and the GSI-equation, depicted as the fraction of predictions that differ less than 0.01, 0.05 or 0.10 from the real values, for the validation sets defined in Table 4.

Abs (Δ predicted vs. real)	Fraction for MIC-equation	Fraction for GSI-equation
< 0.01	0.19	0.24
< 0.05	0.58	0.73
< 0.10	0.87	0.95

Plots like those depicted in figures 3a–c and 4 can be used to determine the preferred combination of the number of spots on the array and the insert size. If for instance the number of spots would be limited to 6000, an insert size of approximately 800 bp would be optimal with respect to the fraction of genes that are represented with a minimal insert coverage of 25% (Fig. 3a). From equation 20 this optimum appears to be 803 bp. With this combination of array parameters the predicted fraction of genes that cover at least 25% of the insert (which equals 803 × 0.25 = 201 bp) is 0.75 (eq. 16). Meanwhile the predicted fraction of genes for which gene specific information can be generated is 0.49 (eq. 18). If the specificity of the data is considered to be more important than the amount of represented genes, it is preferable to have an optimum value for p_{MIC} for higher values of *DIC* (e.g. Fig. 3c) and a high value for p_{GSI} (Fig. 4). These requirements are best fulfilled by combinations with low values for the insert size.

Table 6: Effect of false estimations of *R* on the fraction of predictions that differ less than 0.01, 0.05 or 0.10 from the real values, for the validation set defined in Table 4.

Applied value for *R*	Abs (Δ predicted vs. real)	Fraction
R	< 0.01	0.24
	< 0.05	0.73
	< 0.10	0.95
R - 0.1	< 0.01	0.17
	< 0.05	0.65
	< 0.10	0.95
R + 0.1	< 0.01	0.24
	< 0.05	0.75
	< 0.10	0.94
R - 0.2	< 0.01	0.12
	< 0.05	0.55
	< 0.10	0.90
R + 0.2	< 0.01	0.23
	< 0.05	0.69
	< 0.10	0.91
R - 0.3	< 0.01	0.10
	< 0.05	0.38
	< 0.10	0.81
R + 0.3	< 0.01	0.21
	< 0.05	0.59
	< 0.10	0.88

A Microsoft Excel fill in-spreadsheet that allows for calculations of pGSI, pMIC, and the optimal values for the insert size, is available as additional file with this paper [see Additional file 1].

Discussion

Classical approaches for the construction of DNA libraries form a suitable base for the construction of clone-based microarrays. However, as the construction of these libraries is a random process, it is beforehand uncertain whether a gene or transcription unit will be uniquely represented on a separate insert on the array. Genome coverage by a DNA library is usually determined by calculating the expectation that each single nucleotide from that gene

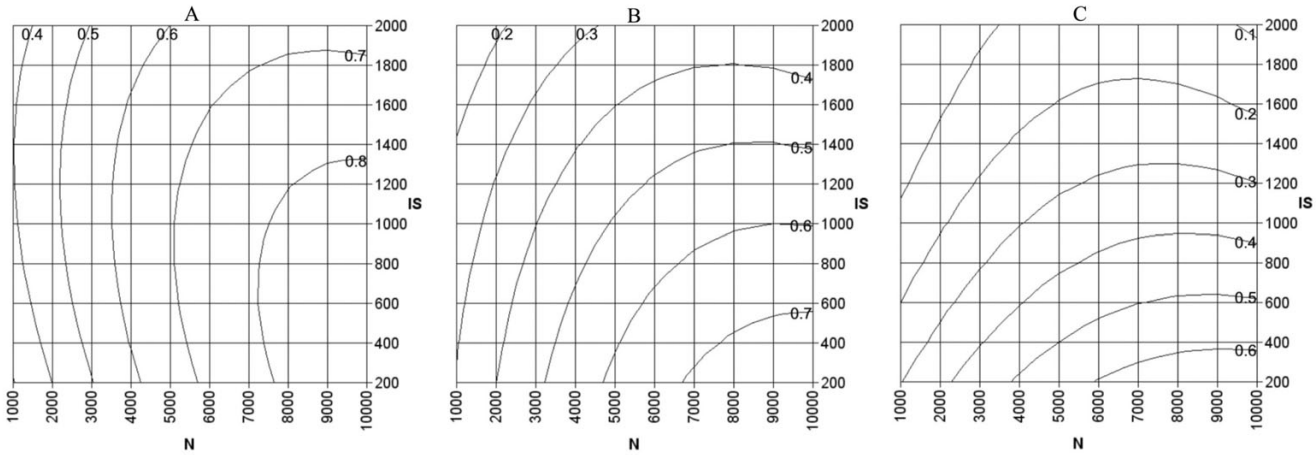

Figure 3
Contour plots of the predicted fractions of represented genes with a minimal insert coverage of 25% (A), 50% (B), or 75% (C) as a function of the number of clones (*N*) and the insert size (*IS*) for a prokaryote with a genome size of 4 Mbp. The predicted fractions are depicted in the plot on top of the lines by which they are represented.

is present [3,4]. These formulas will overestimate the number of clones required when the library is to be used for the construction of a microarray, since for this purpose partial representation of a gene is sufficient for hybridization.

To our knowledge, Akopyants *et al*. were the first to estimate genome coverage at the gene level [5]. They predicted the fraction of represented genes using equation 22:

$$p_{Akopyants} = 1 - \left(1 - \left(\frac{\text{average transcript size} + \text{insert size} - 2 \times \text{required overlap}}{\text{genome size}}\right)\right)^{\text{number of clones}} \quad (22)$$

An important variable in this formula is the average transcript size. However, use of this variable is not legitimate for this type of probability calculations since the average probability per gene (the required information) is not *per se* equal to the probability per average gene. When we validated the Akopyants formula on the same dataset that was applied for the validation of the MIC-equation, it appeared that 49% of the predictions deviated more than 0.1 from the real value (calculated as the average chance per gene), with a strong tendency to overestimation. The Akopyants formula therefore appears unreliable for calculating optimal library sizes

None of the previous formulas give insight in the fraction of genes for which gene specific information can be generated, while this is one of the most important features when one is interested in studying differential gene expression. The MIC-and GSI-equations that were developed in this study allow for good estimations of both the genome coverage at the gene level, and the fraction of genes for which gene specific transcription information can be generated.

Whereas the MIC-equation is rather straight-forward with respect to the input variables and interpretation, application of the GSI-equation requires the estimation of the average number of genes per transcription unit for an organism. Although a false estimation of this variable could lead to a wrong prediction of the represented fraction, Tables 1 and 6 indicate that this risk is limited.

The GSI-equation is partially based on operon predictions. For the development of the model and validation datasets we used log-likelihood based transcription unit predictions for adjacent pair of genes to be in the same operon [10]. This log-likelihood based prediction method is only applicable to organisms for which at least large parts of the genome have been sequenced, and will therefore not be useful when sequence data from array spots for which differential expression was identified, have to be interpreted. Nevertheless, good predictions can be made on whether or not genes that are co-represented in a single spot on the array belong to the same transcription unit. Strong indications can already be obtained from the physical organization of the DNA fragment of interest, like

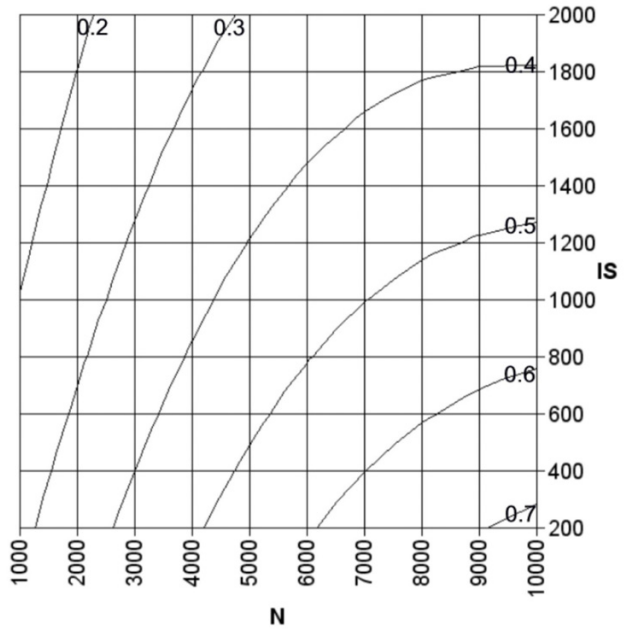

Figure 4
Contour plot of the predicted fraction of represented genes for which gene specific information could be generated as a function of the number of clones (*N*) and the insert size (*IS*) for a prokaryote with a genome size of 4 Mbp, an average number of genes per transcription unit (*R*) of 1.8, and a minimal overlap between the insert and the gene of 100 bp. The predicted fractions are depicted in the plot on top of the lines by which they are represented.

gene orientation and intergenic distance [6,11]. Other indications are the co-occurrence of genes with a joint function, and the conserved organization of homologous genes in other prokaryotes [11,12].

Conclusion

The MIC- and GSI-equations that were developed in this study were based on genomes from 15 prokaryotes from different genera, and validated on the genomes of 10 other prokaryotes. These validations show that these equations form reliable tools for optimal design of prokaryotic clone-based microarrays within the ranges that were tested (Table 4), and that they are applicable to a broad range of prokaryotes. Therefore, these equations form a good basis for the design of microarrays for prokaryotes from which the genome sequence is not available.

List of abbreviations

BP_b number of base pairs within the operon in front of the specific gene [bp]

BP_e number of base pairs within the operon behind the specific gene [bp]

DIC predefined minimal insert coverage, i.e. the minimal required representation of the gene on the insert [%]

Gene gene size [bp]

GS genome size [Mbp]

IS insert size [bp]

$IS_{opt\text{-}MIC}$ Optimal value of *IS* for the MIC equation [-]

$IS_{opt\text{-}GSI}$ Optimal value of *IS* for the GSI equation [-]

N number of clone [-]

O_b overlap of the fragment and the beginning of the gene [bp]

O_e overlap of fragment and the end of the gene [bp]

O_{mf} minimal required overlap of the fragment and the gene (fixed) [bp]

O_{mv} minimal required overlap of the fragment and the gene (variable) [bp]

p gene specific probability value [-]

p_{GSI} predicted fraction of specifically represented genes [-]

p_{MIC} predicted fraction of represented genes represented with a minimal insert coverage [-]

R average number of genes per transcription unit [-]

SSQ Residual Sum of Squares [-]

a-l parameters in MIC or GSI model [-]

Authors' contributions

Bart Pieterse is responsible for the original idea behind this work and performed the statistical and validation procedures. Elisabeth Quirijns developed the MIC- and GSI-equations and performed the fitting procedures. Frank Schuren provided input on the construction and application of clone based microarrays. Mariët van der Werf focused on the interpretability and applicability of

the developed equations. All authors read and approved the final manuscript.

Additional material

Additional File 1

The MIC- and GSI-equations (eq. 16 and 18), and the derived equations for prediction of the optimal values for IS *(eq. 20 and 21), are available as a Microsoft Excel fill in spreadsheet. This spreadsheet can also be applied for the generation of contour plots in which the represented gene fractions are depicted as a function of the number of clones and the insert size.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-6-238-S1.xls]

Acknowledgements

We would like to thank Rolf Boesten, Martien Caspers, Nicole van Luijk and Karin Overkamp for their critical remarks and useful suggestions.

References

1. Cho JC, Tiedje JM: **Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays.** *Appl Environ Microbiol* 2001, **67:**3677-3682.
2. Askenazi M, Driggers EM, Holtzman DA, Norman TC, Iverson S, Zimmer DP, Boers ME, Blomquist PR, Martinez EJ, Monreal AW, Feibelman TP, Mayorga ME, Maxon ME, Sykes K, Tobin JV, Cordero E, Salama SR, Trueheart J, Royer JC, Madden KT: **Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains.** *Nat Biotechnol* 2003, **21:**150-156.
3. Clark L, Carbon J: **A colony bank containing synthetic *Col* E1 hybrids representative of the entire *E. coli* genome.** *Cell* 1976, **9:**91-99.
4. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2:**231-239.
5. Akopyants NS, Clifton SW, Martin J, Pape D, Wylie T, Li L, Kissinger JC, Roos DS, Beverley SM: **A survey of the *Leishmania major* Friedlin strain V1 genome by shotgun sequencing: a resource for DNA microarrays and expression profiling.** *Mol Biochem Parasitol* 2001, **113:**337-340.
6. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002, **18(Suppl 1):**S329-336.
7. Chu G, Vollrath D, Davis RW: **Separation of large DNA molecules by contour-clamped homogeneous electric fields.** *Science* 1986, **234:**1582-1585.
8. Sun LV, Foster JM, Tzertzinis G, Ono M, Bandi C, Slatko BE, O'Neill SL: **Determination of Wolbachia genome size by pulsed-field gel electrophoresis.** *J Bacteriol* 2001, **183:**2219-2225.
9. Wilhelm J, Pingoud A, Hahn M: **Real-time PCR-based method for the estimation of genome sizes.** *Nucl Acids Res* 2003, **31(10):**e56.
10. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in *Escherichi coli*: Genomic analyses and predictions.** *Proc Nat Acad Sci* 2000, **97:**6652-6657.
11. Westover BP, Buhler JD, Sonnenburg JL, Gordon JI: **Operon prediction without a training set.** *Bioinformatics* 2005, **21:**880-888.
12. Ermolaeva MD, White O, Salzberg SL: **Prediction of operons in microbial genomes.** *Nucleic Acids Research* 2001, **29:**1216-1221.