Revista Latinoamericana de MICROBIOLOGÍA
ALAM

Vol. 48, No. 2
Abril - Junio. 2006
pp. 56 - 65

CONFERENCIAS PLENARIAS

In silico evaluation of a novel DNA chip based fingerprinting technology for viral identification

Alfonso Méndez-Tenorio,* Perla Flores-Cortés,* Armando Guerra-Trejo,** Hueman Jaimes-Díaz,* Emma Reyes-Rosales,* Arcadio Maldonado-Rodríguez,* Mercedes Espinosa-Lara,* Rogelio Maldonado-Rodríguez,* Loren Beattie Kenneth***

ABSTRACT. The identification of microorganisms by whole genome DNA fingerprinting was tested "in silico". 94 HPV genome sequences were submitted to virtual hybridization analysis on a DNA chip with 342 probes. This Universal Fingerprinting Chip (UFC) constitutes a representative set of probes of all the possible 8-mer sequences having at least two internal and non contiguous sequence differences between all them. A virtual hybridization analysis was performed in order to find the fingerprinting pattern that represents the signals produced for the hybridization of the probes allowing at most a single mismatch. All the fingerprints for each virus were compared against each other in order to obtain all the pairwise distances measures. A match-extension strategy was applied to identify only the shared signals corresponding to the hybridization of the probes with homologous sequences between two HPV genomes. A phylogenetic tree was constructed from the fingerprint distances using the Neighbor-Joining algorithm implemented in the program Phylip 3.61. This tree was compared with that produced from the alignment of whole genome HPV sequences calculated with the program Clustal_X 1.83. The similarities between both trees are suggesting that the UFC-8 is able to discriminate accurately between viral genomes. A fingerprint comparative analysis suggests that the UFC-8 can differentiate between HPV types and subtypes.

Key words: UFC, evaluation, virtual hybridization, in silico, HPV.

RESUMEN. En este trabajo se evaluó, "in silico", la identificación de organismos por medio de su huella genómica. Las secuencias genómicas de 94 HPVs se sometieron a hibridación virtual sobre un chip de DNA que contiene 342 sondas. Este Sensor Universal de Huella Genómica o UFC está integrado por un conjunto de sondas representativo de todas las secuencias posibles de 8 nucleótidos de longitud que contienen al menos dos diferencias, internas y espaciadas, entre todas ellas. El análisis de hibridación virtual permitió calcular las huellas genómicas que representan las señales producidas por la hibridación de las sondas permitiendo a lo mucho una base no apareada. Las huellas genómicas fueron comparadas entre sí para obtener mediciones de distancias entre todos los pares posibles. Se utilizó una técnica de extensión del alineamiento para considerar solo las señales compartidas por dos genomas correspondientes a la hibridación de las sondas contra sitios homólogos. Se construyó un árbol filogenético a partir de las distancias entre las huellas genómicas utilizando el algoritmo Neighbor-Joining del programa Phylip 3.61. Este árbol fue comparado con el obtenido a partir del alineamiento de los genomas completos de HPV obtenido con el programa Clustal_X 1.83. La similitud entre los árboles obtenidos por ambos métodos sugiere que el UFC-8 es capaz de discriminar con precisión los genomas virales. El análisis comparativo de las huellas genómicas indica que el UFC-8 es capaz de distinguir los tipos y subtipos de HPV.

Palabras clave: UFC, evaluación, hibridación virtual, in sílico, HPV.

INTRODUCTION

Bacterial identification commonly is based in growth conditions (Lachica, 1976; Brown and Walpole, 2001), biochemical and physiological properties (Toledo and Trabulsi, 1983; Edinger et al, 1985), staining characteristics (Zimmer et al, 1999) and immunological reactivity (Roach et al, 2006). More recently an increasing number of molecular technologies have been applied to bacterial identification (Fredicks et al, 2005). Such is the case of PCR (Hryniewiecki et al, 2002; Yang et al, 2002), real-time PCR (Wellinghausen et al, 2004), RFLP (Lu et al, 2000; Okhravi et al, 2000; Sakamoto et al, 2003), AFLP (Shou, et al, 2006) and pulsed field electrophoresis (Bautsch, 1994; Lorenz et al, 1997). These technologies give limited information on the genomic sequences of the organisms determined. In another more informative approach, the sequence of a single gene (16S rRNA) is searched (Keinanen-Toivola et al, 2006; Lau et al, 2006). Unfortunately this procedure sometimes is unable to distinguish between highly related organisms as happens with Bacillus thuringiensis, Bacillus cereus and Bacillus anthracis (Kaneko et al, 1978; Helgason et al, 2000; Han et al, 2006). In a more discriminatory approach, arrays of sequence derived probes have been used for identification of bacterial groups (Kim et al, 2005; Liu et al, 2005; Kelly et al, 2005; Francois et al, 2005).

* Laboratorio de Tecnología del DNA, Escuela Nacional de Ciencias Biológicas. IPN. Prol. de Carpio y Plan de Ayala s/n, Col. Santo Tomás, 11340, México D.F.
** Departamento de Biofísica, Escuela Nacional de Ciencias Biológicas. IPN. Prol. de Carpio y Plan de Ayala s/n, Col. Santo Tomás, 11340, México D.F.
*** Amerigenics, Inc. USA.

First version received: 05-05-06; first version revised: 22-05-06.
22-09-06; accepted: 23-09-06.

Recently our research group has selected a universal set of 13-mer probes aimed for bacterial identification, which was tested, by virtual hybridization, in 191 fully sequenced bacterial genomes (manuscript in preparation). The comparison of the 191 genomic fingerprints allowed us to construct a bacterial taxonomic tree. The distribution of bacterial strains in this tree had a significant number of differences with the tree based on the alignment of the aminoacid sequences from 55 ribosomal proteins (The Institute of Genomic Research, 2006). The discordances can be associated to the different type of sequences searched in each approach, since our analysis is done in the complete genome (having conserved and non conserved sequences) while the ribosomal proteins are codified by conserved sequences. This suggestion is supported by the observation that a third bacterial tree, made only with the conserved sequences contained in our fingerprints, showed numerous similarities with the tree obtained from ribosomal (conserved) sequences (data non published).

Another way to verify the reliability of the UFC is to search if it produces a DNA fingerprinting taxonomy (tree distribution) similar to that obtained by the sequence alignment approach when tested on the same conserved genes. This work reports the results obtained with this type of analysis, done in 94 Human Papillomavirus (HPV) strains.

MATERIAL AND METHODS

HPV databank. The genomic sequences from 94 HPV strains were obtained from GenBank. The list of accession numbers is included in Table 1. All the sequences were saved in FASTA format in a single folder.

UFC-8. A DNA chip named UFC-8, constituted by 342 probes, representing all possible 8-mer oligonucleotide sequences, having two sequence differences, internal and spaced, was designed. The Tm probe values vary from 32.6°C to 50.2°C. To draw the fingerprints the 342 probes were distributed in 17 columns of 20 probes each and 1 column with two probes, ordered from the top left corner to the bottom right corner by their increasing Tm values, which also corresponds to an increasing G+C content.

Virtual Hybridization (VH). A computer program, able to predict perfect and mismatched hybridizations, based in the determination of the stability of fully or partially complementary target-probe duplexes, was used to determine all the hybridizations occurring between each HPV genome and the UFC-8. The group of hybridization signals

Table 1. HPV strains and GenBank accession numbers.

HPV type	Accession #	HPV type	Accession #	HPV type	Accession #	HPV type	Accession #
RTRX7	U85660.1	18	NC_001357.1	42	NC_001534.1	70	NC_001711.1
1a (3-3)	U06714.1	19	NC_001581.1	44	NC_001689.1	71	NC_002644.1
1a	NC_001356.1	20	NC_001679.1	45	NC_001590.1	72	X94164.1
2a	NC_001352.1	21	NC_001680.1	47	NC_001530.1	73	X94165.1
3	NC_001588.1	22	NC_001681.1	48	NC_001690.1	74	NC_004501.1
4	NC_001457.1	23	NC_001682.1	49	NC_001591.1	75	Y15173.1
5	NC_001531.1	24	NC_001683.1	50	NC_001691.1	76	Y15174.1
5b	NC_001444.1	25	NC_001582.1	51	NC_001533.1	77	Y15175.1
6	NC_000904.1	26	NC_001583.1	52	NC_001592.1	82	NC_002172.1
6a	NC_001668.1	27	NC_001584.1	53	NC_001593.1	82sub IS39_AE2	AF293961.1
6b	NC_001355.1	28	NC_001684.1	54	NC_001676.1	83	NC_000856.1
7	NC_001595.1	29	NC_001685.1	55	NC_001692.1	84	NC_002676.1
8	NC_001532.1	30	NC_001585.1	56	NC_001594.1	85 candidate	AF131950.1
9	NC_001596.1	31	NC_001527.1	57	NC_001353.1	86	NC_003115.1
10	NC_001576.1	32	NC_001586.1	57b	U37537.1	87	NC_002627.2
11	NC_001525.1	33	NC_001528.1	58	NC_001443.1	89	NC_004103.1
12	NC_001577.1	34	NC_001587.1	59	NC_001635.1	90	NC_004104.1
13	NC_001349.1	35	NC_001529.1	60	NC_001693.1	91	NC_004085.1
14D	NC_001578.1	36	NC_001686.1	61	NC_001694.1	92	NC_004500.1
15	NC_001579.1	37	NC_001687.1	63	NC_001458.1	93	NC_005133.1
16	NC_001526.1	38	NC_001688.1	65	NC_001459.1	94	NC_005352.1
16 iso 16W12E	AF125673.1	39	NC_001535.1	66	NC_001695.1	96	NC_005134.2
16 variant	U89348.1	40	NC_001589.1	67	D21208.1		
17	NC_001580.1	41	NC_001354.1	69	NC_002171.1		

for each HPV strain corresponds to its DNA fingerprint. A preliminary VH was done, without considering Tm values, for perfectly matched target-probe sequences. Then a second, more appropriated, VH analysis was made with 1ºC UFC-8 probe subsets under conditions to permit the formation of single mismatched duplexes.

Fingerprinting pairwise distances. The Fingerprints for each HPV virus were compared against each other in order to obtain a table of distances for all the possible combinations of fingerprint pairs (pairwise distances). During this comparison an extended match strategy was applied to identify only the signals shared for two fingerprints that correspond to the hybridization against homologous sites. This strategy consist in extend the site of the hybridization with a shared probe until a length enough to exclude the possibility of finding other site in the genome by chance. For this purpose, the original target sequences of 8 bases recognized by each probe were extended 5 bases at each side in both targets. Therefore, a sequence of 18 bases is obtained for each hybridization site in each target. When two sequences of 18 bases, from different HPV types, share at least 16 matches, were considered as homologous signals and used to compute the distances between two fingerprints.

Fingerprinting tree. A phylogentic tree was built from the table of all pairwise fingerprint distances using Neighbor-Joining algorithm from the program Phylip 3.61 (Felsenstein, 2002).

HPV phylogenetic tree. A phylogenetic HPV whole genome tree was constructed as follows: Genomic sequences were aligned with the program Clustal_X 1.83 (Thompson et al, 1997). Then, the program MEGA3 (Kumar et al, 2004) was used to estimate the table of p-distances derived from the alignment, and the phylogenetic tree was constructed using the Neighbor-Joining algorithm (NJ).

Analysis. The distributions of viral strains in both trees were compared to search for the main similarities and differences. A comparison of fingerprint HPV pairs was done to determine the UFC power to discriminate between related HPV strains.

RESULTS

The probe sequences are property of Amerigenics, Inc, their commercial or research use requires a permit from the company. The main purpose of this work was to evaluate the capability of the UFC to discriminate between different organisms. The strategy was to compare the distribution of a group of organisms in a tree obtained from the distances derived of the alignment of the genomic sequences with that obtained in other tree calculated from the distances between genome fingerprinting with the UFC DNA chip.

A key difference between both approaches is the type of genome sequences analyzed. Universal DNA fingerprints are done by exploring the complete genomic information, which frequently contains not only the effect of mutations but also is affected by gene loss and lateral gene transfer (Bushman, 2002), while phylogenic studies are commonly performed on conserved sequences. HPV genomes, which are approximately 9,000 bp long, are mostly constituted by essential (conserved) genes. The HPV genome is integrated by 8 genes and a control region known as Long Control Region or LCR. There are 6 genes that are early expressed, which participate in the replication and the control of the cell cycle. Two more genes that are lately expressed (L1 and L2) form part of the virus protein envelope and they have interactions with the viral DNA and with the host cell. Moreover, the LCR participates in the control of the transcription (http://www.ircm.qc.ca/microsites/hpv/en/390.html).

The HPV genomic UFC-8 fingerprints were done by virtual hybridization (VH). Virtual hybridization is an algorithm able to predict the formation of complementary and imperfect target/probe duplexes during hybridization. The program calculates the stability of the duplexes, and selects those having higher stability than any chosen free energy cut-off value. The reliability of the VH algorithm was recently tested with the successful discrimination of Pseudomonas aeruginosa and closely related bacterial strains via hybridization fingerprinting using oligonucleotide microarrays (Reyes-Lopez, et al, 2003). It was clear from this work, that hybridization signals are obtained at more negative predicted free energy values (more stable). Moreover, other works have showed that, despite that the methods for predicting the secondary structure are not perfect yet; the design of microarrays is optimized by this type of predictions. These designs show better performance than those where this pre-evaluation was not done (Matveeva, 2003).

An UFC 8-mer constituted by 342 probes (Table 2), with two spaced and internal base differences between each other, was designed and used to simulate the hybridization against the 94 fully sequenced HPV types contained in GenBank up to date. The Tm for this UFC varies from 37 °C to 54 °C whereas the free energy varies from –12.21 to –7.18 Kcal/mol.

A first analysis was done to establish the hybridization conditions giving an appropriated number of hybridization signals as to discriminate between the viral strains, and simultaneously, to permit the identification of the target sequences involved in the duplexes formed. It is expected to get good discrimination when the number of hybridization signals in all the strains varies between 20 and 80% of the set of probes. By other side, the identification of target sequences decreases with the number of mismatches allowed in the duplexes formed. Therefore, it is desirable to perform the hybridizations under conditions

Table 2. Probe sequences and Tm values.

Sequence	Tm	Sequence	Tm	Sequence	Tm	Sequence	Tm	Sequence	Tm	Sequence	Tm	Sequence	Tm
CTATGCTA	32.64	AGAACTTC	34.72	AGATACGT	36.96	AGACATCC	39.59	TACTTGCG	43.13	ACTGCACT	45.68	GTGAACGC	47.77
AATAGGAC	32.79	AGTCCTTA	34.78	TCTTGCTA	37.27	GTCTCCAT	39.59	GCCACTAG	43.16	CAGCTGTG	45.71	TGCTCTCG	47.84
GATTAGGT	32.79	CGTTATCT	34.84	TAAGCAGA	37.27	AAGCTCAA	39.74	GCTGTAGG	43.16	GTACCCGA	45.76	AGGGCATG	47.92
TAATGAGG	32.84	GTGAAGAT	34.89	CAACTTGT	37.32	TAGTCCCT	39.76	GAATGCCA	43.24	CCGAGACT	45.81	TGTCTGCC	48.03
TTATCAGG	32.84	ACAGAATC	34.89	CCAAGTAC	37.36	AGGCATAG	39.77	GCAACAGA	43.34	CTCCGACT	45.81	GTCAGGCA	48.03
CCTCAATA	32.84	ATCTGTTC	34.89	GATTCAGG	37.38	GAAGTTCG	39.87	TCATGGGT	43.5	TCGGAGTC	45.91	GCAGACCA	48.03
GGATGTAT	32.97	CTGCTAAT	35.07	AATGGTCT	37.52	AGCATGAA	39.92	CTGATCGG	43.61	CCCAGTGA	45.97	CGGTCGAT	48.07
GTTCTACT	33.08	CATCATGA	35.11	CATGTACC	37.53	TTGAGCAT	39.92	TCGACTGA	43.66	TGGGTCAG	45.97	GTCCACGT	48.29
CTTACTGA	33.13	TACAGACT	35.23	ATACTCCC	37.55	CAACGTAC	39.99	CGAGAGTG	43.71	CAGGGTGA	45.97	CCTCGCAT	48.57
CAGAGTAA	33.13	TCATTAGC	35.24	TATCCGTT	37.57	ACGAAAGT	40.19	CGAGTGAG	43.71	ATCCGCTT	46.13	ATCTGGCG	48.57
CTGACTTA	33.13	TATGATGC	35.42	TACCGAAT	37.57	CCGTCTTA	40.21	GTCACAGG	43.77	CTGCTTCG	46.17	CTCTGCGT	48.58
TTAGACTG	33.13	AGTTTCTG	35.55	CAGTTCTC	37.58	CTAGGGTC	40.22	CTGTCACC	43.77	CTTCGCTG	46.17	CGTCAGCT	48.58
CACTTAGA	33.13	CAAGACTT	35.55	CAGAAGTC	37.58	ACTTGAGG	40.24	CAGGTCAC	43.77	ACGTTTGC	46.22	ACGAGCAG	48.58
CTACTCAA	33.13	AAAGTCTG	35.55	ATTGAAGC	37.63	ACGTATCC	40.35	CTGTGGAC	43.77	GCTACCCA	46.23	GCCATGGT	48.78
CACTATCA	33.31	ACTTTCAG	35.55	AGAAATGC	37.63	TTCGACAA	40.36	GTAGCTCG	43.96	TGGGTAGC	46.23	ACTGCCAC	48.79
TGTGATAG	33.31	AAGATCCT	35.62	GCTCATTT	37.63	AGGCTAAC	40.52	TAACACGC	43.98	TGTAGCCC	46.23	ACACTGGC	48.79
GCATATTC	33.36	TAAACGAG	35.65	AGCAATTC	37.63	GCTAAGGT	40.52	GTAGGCAC	44.02	TAGCACCC	46.23	ACTCCGGA	49.07
TCGATATC	33.41	TTTACTCG	35.65	GAATAGCC	37.66	AAGCAACT	40.54	GGCAGATC	44.09	GCAAGGTG	46.31	GCTGCTCA	49.16
GTTAGCTA	33.44	TGAAACTC	35.72	ACTTCTGT	37.74	TGCCTAAG	40.55	CTCAGCAG	44.1	ATGGCGAT	46.31	AACGGCTC	49.22
GTAGCATA	33.62	CAACATCT	35.73	GTGAGATG	37.75	CTGTGAGT	40.58	CTTGCACA	44.13	TGCAACCT	46.36	GGCGAGTT	49.22
CTACCATT	33.64	ATCACTTG	35.73	ACTAGGTC	37.77	TGCAGAAA	40.71	AGACCCTC	44.22	AGTGGCAA	46.36	AAGGACGC	49.22
ATAACCTG	33.64	GATAGTGG	35.79	ACTACGAA	37.78	TATCTCGC	40.86	CTTAGCGG	44.46	TCAGCCTC	46.37	CTTCGGCA	49.23
CCTACAAT	33.64	TAATCACG	35.82	ATTCGAGA	37.81	AGATGAGC	40.9	GGGCTAGA	44.51	TACCGTGG	46.56	TGAAGGCG	49.23
ACTAATGG	33.64	TGTAAGGA	35.84	GCAATGAT	37.81	GTGTACAC	40.92	GAGACGTG	44.54	AACGTCGA	46.61	CAGCGGAA	49.23
CATTCTTG	33.7	TCACCTAA	35.84	TAGCCAAT	37.87	GAGCAAAC	41.05	CGTCACTC	44.54	ACTCGGAC	46.67	GCGATCGA	49.24
CATATGGT	33.82	AGCTTTAC	35.85	GCATCTAC	38.03	GCTTGTTC	41.05	CCATAGCG	44.63	CCACTCGA	46.69	ATGCCGTC	49.39
GACAAATC	33.83	TAAGCTTG	35.89	TAGCAGTT	38.08	TCGGTAAC	41.11	TCGCTTTG	44.65	CTCGTCCA	46.69	ACCGCATC	49.39
GTGATTTC	33.83	TGATCAAC	35.89	TAAGTGCT	38.08	GGGAGTAC	41.13	GGCTGAAG	44.69	CGCGTATG	46.87	ACTGTCGC	49.4
TATTCCAC	33.83	GATGTCAA	35.89	CATCGTTT	38.16	AACTGGTC	41.16	ATTCACGC	44.8	TATCGGCC	46.9	TCGCACAG	49.41
GGAATGTA	33.83	TATCCACA	36.02	CACGAATT	38.16	AGTCCAAC	41.16	GCGATGTT	44.8	GATTGGCG	46.93	TCACTGCG	49.41
TGGTAATC	33.83	GTAATGCT	36.03	ATGGTAGG	38.41	AGGTTGAC	41.16	ACGCATTC	44.8	TCTACCGC	46.94	CATCGCCA	49.41
TTCATACC	33.83	GATCAGAG	36.04	CCCTATGT	38.41	TGAACCAG	41.19	ATGCCTGA	44.8	TACGAGGC	46.94	TCAGGGCT	49.63
CTCTTGAT	33.92	ATGTCTCT	36.07	CATTTGCT	38.45	AAATCGCT	41.34	CCAGATGC	44.87	TAGGCGAC	46.94	AGGCCTCA	49.63
CATTAAGC	33.99	AGTCTCAT	36.07	GCATAAGG	38.4	CGTTTCAC	41.47	ACGGTTCT	44.97	ATCCAGCC	47.11	TCGGGACA	49.96
GCTTAATG	33.99	TAGAGCTT	36.24	ATGGAACA	38.55	GACTGTGT	41.48	CTCGGTTG	45.04	GCGACATG	47.14	CGTGCCTT	49.98
CGAGATAA	34.07	TATGCATG	36.25	TCAATTGC	38.6	TTGCAAGT	41.51	AGGACTGG	45.05	ACCAGAGC	47.15	CCAAGCGT	49.98
TTAGTACG	34.07	GACAGTTT	36.5	TGACAACT	38.74	AGGGTACA	41.61	TGTTCCGA	45.1	GGCTGTCT	47.15	AACAGGCG	49.98
TAGTAACG	34.07	CAACTGAA	36.54	TACCTCAC	38.76	TCTTCAGC	41.66	TCTGACCC	45.15	CTCCTGCA	47.17	GCGTGGAA	50.05
ACATACAG	34.1	GTTTACGA	36.58	TAACGACA	38.76	TGGACTCT	41.82	GTCTGGGA	45.15	ATGACGGG	47.42	TGCGAACC	50.05
TTAGTGAC	34.1	GACGTAAA	36.58	ACATCGAT	38.79	CAATCGGA	41.96	GAGGTCCA	45.15	ACGATGGG	47.42	TTGCGTCC	50.05
TGAGTTAC	34.1	ACCTTGTA	36.67	TCGAATCA	38.79	GATCCCTG	41.98	CACCGATG	45.21	CCGGTGAT	47.42	AGCACACG	50.15
TAGTGTTC	34.1	GTCAATGT	36.68	TCGTAGAG	38.84	CGTCGTAT	42.04	GCCTTACG	45.29	TAGCCTGC	47.42	GCGCGATA	50.2
TATGGAGA	34.12	TTGGATCT	36.68	TAGAGTCG	38.84	ATCGTGTC	42.23	AGTTGGCT	45.42	CGACAGGT	47.45		
TCCTGATA	34.12	TCTGGAAT	36.68	TGGCATTA	38.89	ATACGGGA	42.42	CATACGCC	45.46	GGTGTCGA	47.54		
CAATATGC	34.16	CACATTGA	36.72	ATCTAGGC	38.93	CCGGATAC	42.67	AGGAGCTC	45.46	GCGGGTAT	47.66		
AAGATAGC	34.23	CACTACTG	36.82	CAATACCG	38.96	GGAGATCG	42.84	GAAACGCA	45.49	CGGCCATA	47.68		
TGTGTATC	34.28	CAGTACAG	36.82	GAACAGTG	39.26	CGAGGATC	42.84	GCAGGAAC	45.54	TATGGCCG	47.68		
TGTATGTC	34.28	TAAAGCAC	36.84	GGTACTGT	39.55	ACAAGTCG	42.84	GAAGCCAC	45.54	ACGCCTAC	47.69		
AGCTAGTA	34.54	GCTGTTTA	36.84	AGTACCAC	39.55	GACACTCC	42.99	GCCAACTC	45.54	TGCTACGG	47.71		
CGTAATTG	34.58	GCTTACAA	36.84	GATTGCTC	39.57	CTTGCCAT	43.12	TGCTCCAA	45.55	CGCAGGTA	47.71		

giving enough number of signals and simultaneously keeping a minimum or avoiding the formation of mismatches in the duplexes.

Table 3 shows the median number of matched and mismatched hybridization signals predicted by the 94 HPV genomes listed in table 1 on the UFC-8 by virtual hybridization. This table also shows the minimal and maximal free energy in all the target/probe duplexes formed, with 0, 1, 2, 3, or 4 mismatches, respectively. It can be seen that in average only 38 perfect matched duplexes are formed with each HPV, which represents only 11.1% of the 342 probes. This proportion of signals is too low to permit good discrimination between related viral strains with the UFC-8. Additionally, when the formation of all the single mismatched target/probe duplexes was allowed, 336 out of 342 probes yield duplexes in each HPV strain on average, corresponding to 98.2% of the probes. This is also undesirable because there is a small margin of variation to discriminate between the HPV strains. This data also shows that there are more single mismatched hybridizations with higher stability (-11.22 Kcal/mol) than perfect matched duplexes (-7.18 Kcal/mol). Similarly, some double mismatched hybridizations are more stable (-10.12 Kcal/mol) than some other matched (-7.18 Kcal/mol) and single mismatched (-2.20 Kcal/mol) duplexes. By other side there is only one sequence fully complementary to each probe, while there are 24 (3 for each one of the 8 probe positions) possible target sequences able to form single mismatched duplexes with each probe. Under these circumstances it was decided to test the number of hybridization signals obtained when using a cut-off free energy value that permits the formation of matched and single mismatched duplexes while avoids the formation of two or more mismatches in the duplexes. To avoid false hybridizations the UFC-8 was divided in seventeen probe subsets, having a variation of 1ºC in the Tm values for each probe subset, and a specific cut-off free energy value was used for each probe subset. Under these conditions the average of the hybridization signals increased to 81 (Table 4) for each HPV genome, which corresponds to 24 % of the probes and seems enough to discriminate between the HPV strains. Therefore these conditions were used to obtain the HPV genomic fingerprints.

Table 3. Virtual hybridization analysis of UFC-8 *vs* HPV genome sequences.

Number of mismatches	Average number of probes[1]	ΔG° min (kcal/mol)	ΔG° max (kcal/mol)
0	38	-12.21	-7.18
1	336	-11.22	-2.20
2	342	-10.12	-0.36
3	342	-8.82	0.00
4	342	-6.65	0.00

[1] Corresponds to the number of probes of the UFC that bind to any target.

To improve the estimation of phylogenic relationships between the HPV strains only the hybridization signals contained in their fingerprint and corresponding to highly conserved (extended shared signals) sequences were consider. For this purpose, the 8 bases long sequence of each hybridization signal shared by two HPV strains was located in its respective genome. Then the 10 flanking bases (5 from each side) were added to have a sequence of 18 bases long existing at the respective sites in each HPV strain. The length of this extended section was estimated using the Karlin and Altschul statistics (Karlin, 1993). According with this statistical method, when comparing two locally aligned sections and using a score of +1 for each match, a score of –2 for each mismatch, K = 0.621 and λ = 1.33 are used, the probability to find an alignment score S by chance is calculated by:

$$P(S \geq \chi) = 1 - e^{-Kmne-\lambda s}$$

where m and n represent the length of the target and probe sequences respectively, K and λ are the Karlin and Altschul parameters which depend on the kind of sequences to be compared (DNA or Protein) and the score for matches and mismatches. The score S is calculated by:

Table 4. Division of the UFC 8 in subsets with defined ranges of stability, cut-off values and average number of signals predicted by virtual hybridization against HPV sequences allowing 1 mistmatch at most.

Set	ΔG° cut-off	Average number of signals
A	-7.00	8
B	-7.13	9
C	-7.39	7
D	-7.54	10
E	-7.72	9
F	-7.95	6
G	-8.09	4
H	-8.32	6
I	-8.55	3
J	-8.73	1
K	-8.89	1
L	-9.05	3
M	-9.31	3
N	-9.50	4
O	-9.67	2
P	-9.93	2
Q	-10.15	3
Total	-	81

$$S = (\textit{probe length} + \textit{total extension} - \textit{mismatches}) - (2 \times \textit{mismatches})$$

Where, *total extension* is the sum of the left and right extensions (the programs uses both extensions of the same length). If a probe length is 8, and a total extension of 10 nucleotides is done allowing only 2 mismatches (threshold = 16), the score is (8+10-2) - (2*2) = 12 and the probability for finding such score by chance with m = 9,000 and n = 18 is 0.0117 (1.1%). Therefore such score is not easily found by chance and it is expected that the distances between organisms based in such score have a better correlation with the *real* distances between the sequences.

The total number of extended shared signals was used to calculate the extended shared scores between pairs of HPV strains. As an example, the analysis of shared extended matches in the fingerprints of the two more common HPV high risk types, related with cervical cancer, was performed (Table 5). HPV16 and HPV 18 gave 83 and 76 hybridization signals, respectively. 121 of these signals were different and 38 shared. The number of shared extended signals was 14 and the distance score was 0.2620.

In order to verify if the phylogenetic distances between sequences were correctly assigned with the fingerprint analysis, the fingerprint distances were compared with those obtained from the alignment of the genome sequences. A total of 94 HPV genome sequences were aligned with the program Clustal_X 1.83. Then the program MEGA3 was used to estimate the table of distances from the alignment. The distance between two genomic sequences was calculated as a p-distance which is defined as:

$$p = \frac{\textit{number of differences}}{\textit{length of the alignment}}$$

Table 5. Distance scores in three different pairs of HPV types. The distance scores for HPV 16 with other three HPV types located at different positions in the three were calculated. HPV 35 is located next to HPV 16, HPV 18 is located in another cluster near to the HPV 16 and, HPV 77 is distantly located from HPV 16. Numbers in parenthesis correspond to HPV types.

HPV types	(16 and 18)	(16 and 35)	(16 and 77)
Hybridization signals	83(16) 76(18)	83(16) 87(35)	83(16) 91(77)
Different signals	121	123	137
Signals shared	38	47	37
Extended shared signals	14	25	13
Extended match score	0.2620	0.1953	0.2860

The trees calculated from the alignments and fingerprints are shown in Figure 2. Both trees were calculated from the distance data with the Neighbor Joining (NJ) algorithm using Phylip 3.6. Trees derived by using other distance-based algorithms as the UPGMA were topologically similar to Neighbor-Joining trees (data not shown). Although there are notable differences between the fingerprint and alignment trees, they show strong similarities which are indicated by the main five groups and by the clustering of high risk HPV types. To explain the differences between the trees, it must be considered that global multiple alignments obtained by the Clustal_X program are not optimal. Clustal_X uses a heuristic method for multiple alignments, which is prone to errors especially for divergent sequences. Errors are propagated during the alignment and the most distant sequences can show a considerable high number of errors in their alignment (Mount, 2001). The extended match score approach can be considered as a method that uses local alignments to derive the phylogenic distances. It is known that local alignments provide more reliable information about similarity between sequences than global alignments (Durbin et al, 1998). Therefore this example illustrates how the Karlin and Altschul (1993) statistics can be conveniently used to estimate extension and threshold values for this phylogenetic approach.

The strategy to asses the UFC-8 capability to distinguish between different strains by DNA fingerprinting was to compare the similarity in the distribution of HPV types in the tree produced by the classical sequence alignments (Fig. 1 B) with that produced by this new (shared extended signals contained in the fingerprint) procedure (Fig. 1 A). It is clear that similar viral distribution is obtained in both trees, the main differences are that group 2 in the B tree, seems divided in two parts, 2a and 2b, in the A tree. However just by turning the 2a subgroup to the right they become similar. The same situation happened with groups 3, 4 and 5 (Fig. 1B) which are just inverted in the other tree (Fig. 1A). Besides high risk HPV types were placed in the same HPV cluster (group 3) in both trees, which agrees with the viral location in previous HPV classifications (http://www.stgen.lanl.gov/stgen/virus/hpv/compendium/htdocs/COMPENDIUM_PDF/94PDF/3/MakePart3.pdf). Therefore, these results suggest that the UFC-8 is able to discriminate reliably between all these HPV types.

According to **Fowlie and Schmidt** (1998) a good analytical tool for diagnostic purposes should be able to discriminate between close and distantly related organisms. Therefore, the UFC-8 capability to discriminate between differently related HPV types was tested. The genomic fingerprints of tree pairs of HPV types, having different degree of relatedness, were compared. HPV16 was compared

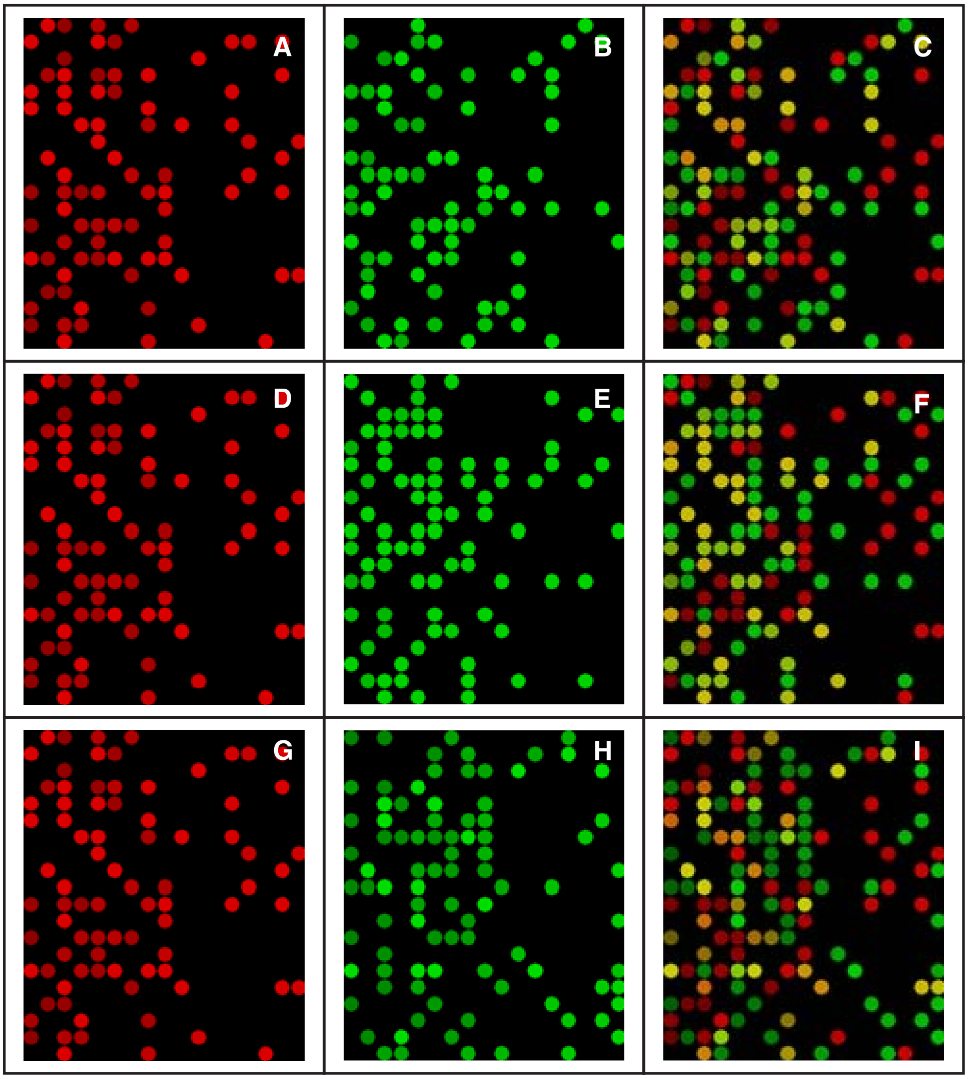

Figure 1. HPV Virtual hybridization fingerprints with the UFC-8. *Virtual hybridizations were done under conditions for allowing the formation of single mismatched duplexes. The hybridization signals for the reference HPV 16 strain are shown in red. For test HPV strains (18, 35 or 77) the signals are shown in green. In the comparison of reference and test fingerprints the shared signals are shown in yellow, whereas particular signals are shown in red for reference and in green for test strains respectively. A) HPV 16, B) HPV 18, C) HPV 16 compared with HPV 18, D) HPV 16, E) HPV 35, F) HPV 16 compared with HPV 35, G) HPV 16, H) HPV 77, I) HPV 16 compared with HPV 77.*

with HPV35, HPV18 and HPV77 (Table 5). HPV16 and HPV35 types are highly related since they were placed next to each other in group 3. HPV16 and HPV18 are less related, since they were placed in different subgroups from group 3, while HPV16 and HPV77 are distantly related HPV types, because they belong to the 3 and 5 groups, respectively.

Figure 2 shows the isolated and overlapped fingerprints for these pairs of strains before to the selection of shared extended signals. Table 5 shows the data corresponding to the fingerprint of these pairs of HPV types. It also includes the number of shared extended signals and their respective scores. There can be seen that the extended shared distance scores increases (0.1953, 0.2620 and 0.2860 in the 16-35, 66-18 and 16-77 pairs of HPV types, respectively) with the lowering in the degree of relatedness. These results suggest that the UFC-8 is able to discriminate between HPV types which have at least 10% of differences between them (Heinzel et al, 1995).

To test if the UFC-8 is able to discriminate between HPV subtypes, which have from 1 to 10% of sequence differences between them (Heinzel et al, 1995), a similar fingerprint statistical analysis was done on HPV 6 subtypes. Due to the higher degree of relatedness HPV 6, 6a and 6b fingerprints gave the same number of hybridization shared signals before and after the sequence extended shared procedure (Table 6), showing strong similarity scores, 0.0163, 0.0096 and 0.0158 for the 6-6b, 6-6a and 6a-6b pairs of HPV subtypes. The most related (0.0096 extended shared distance score) HPV pair (6-6a) has six differences (75-74 + 79-74) between them. The 6a-6b HPV pair has a 0.0158 extended shared distance score and shows 10 differences

(79-73 + 77-73) between them. And the 6-6b HPV pair has an extended shared distance score of 0.0163 showing 10 differences (75-71 + 77-71) between them. The different score for the 6a-6b and 6-6b HPV pairs, which have the same number (10) of differences between them, is due to the slightly lower amount of different signals (81 in the 6-6b HPV pair versus 83 in the 6a -6b HPV pair). Similar results were obtained in the comparison of HPV 16 with its variants (16variant and 16iso16W12E) (data not shown). All this data suggest that the UFC-8 is able, under appropriated hybridization conditions, to reliably discriminate between different HPV subtypes by genomic fingerprinting, under the conditions tested.

All the previous data suggest that the UFC-8 is a device able to discriminate between all the HPV types and subtypes. In our knowledge, even when there are many very good HPV diagnostic procedures (Sandri et al, 2006; Coutlee et al, 2006;) there is no other method able to make the discrimination of all the HPV types and subtypes simultaneously in a single assay.

A big difference between the design of UFC-8 and other oligonucleotide arrays made for the identification of organisms is the type of sequences selected for probes. In the UFC-8 the set of probes represents all the possible 8-mer sequences with G+C content between 35 to 65% of G+C, with properties appropriated for hybridization (for example avoiding sequence repeats), and with optimized discriminatory properties, such that when a given target sequence forms a stable duplex with one probe, the same sequence is unable to react with any of the other probes. In other DNA chip designs, such as that from Yang et al (2006), the sequences of the types of organisms to be identified are used as source of information to locate the specific sequences contained in them, which are then used as reference for the design of the probes. Therefore most designs are useful only for the specific type of organisms corresponding to the probes, and additional DNA chips will be needed for other viruses, while the UFC-8 which has been evaluated "in silico" for HPV and HIV (data not shown) is potentially able to differentiate between all classes of viruses.

Table 6. Distance scores in three highly related pairs of HPV subtypes. The distance scores and particular or shared number of hybridization signals for HPV 6 and other two HPV 6 subtypes, located in the same branch in the three, are shown. Numbers between parenthesis correspond to HPV subtypes.

HPV types	(6 and 6b)	(6 and 6a)	(6a and 6b)
Hybridization signals	75(6) 77(6b)	75(6) 79(6a)	79(6a) 77(6b)
Different signals	81	80	83
Signals shared	71	74	73
Extended shared signals	71	74	73
Extended match score	0.0163	0.0096	0.0158

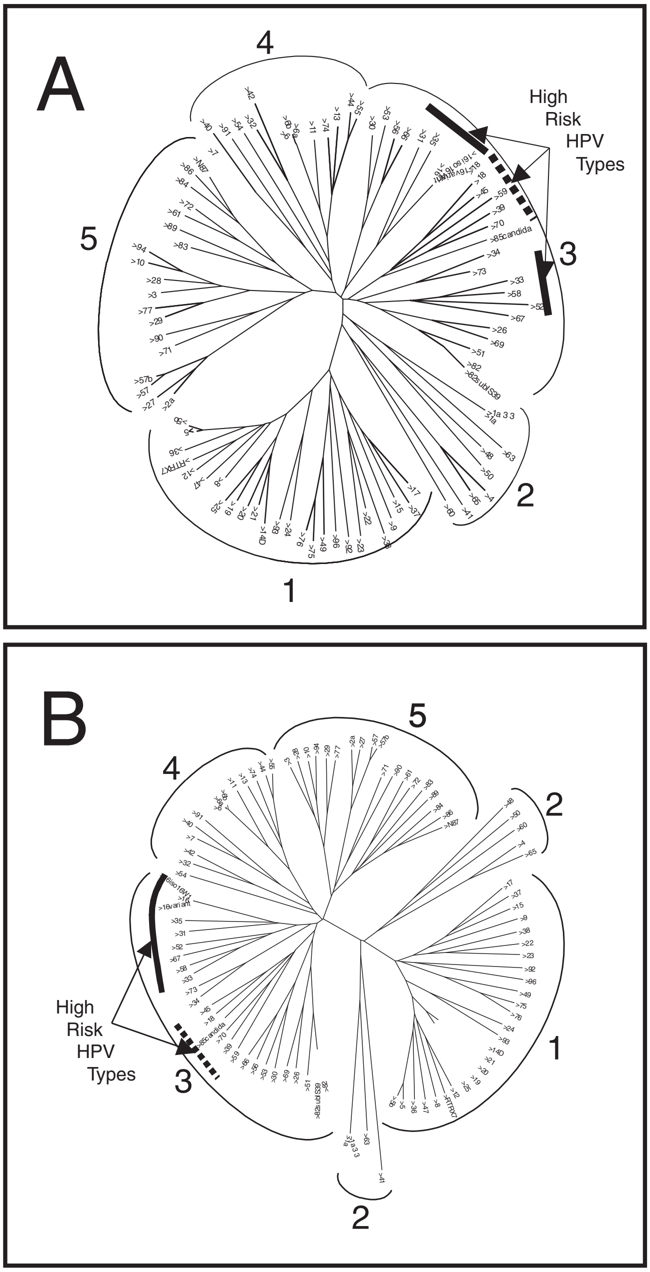

Figure 2. HPV trees produced by fingerprinting and genomic alignment. *A) Neighbor-Joining tree calculated from the distances between the fingerprints obtained with the hybridization of the UFC-8 and 94 HPV genomic sequences considering only the homologous signals. B) Neighbor-Joining tree calculated from the alignment of 94 HPV genomic sequences. The main HPV clusters are indicated by the numered regions from 1 to 5. HPV groups of High Risk for producing cervical cancer are indicated with arrows.*

When there is interest only in a given class of viruses, a UFC-8 probe subset can be selected for this purpose. This type of probe arrays has been denominated as Cluster Associated Fingerprinting Chips.

Preliminary analyses of the distribution of the sites which were recognized with probes contained in the UFC suggest that they are distributed uniformly along the HPV genome, hybridizing with all the genes. Of special interest are the probes binding specifically with High Risk viruses because they could be recongnizing gene sequences related with this property. However a more detailed analysis on this topic is required.

Even with all this "in silico" support, it is necessary to perform experimental tests to confirm these results. Also, when necessary it will be convenient to add more probes to the UFC-8 to make the identification of minor sequence variations, by example for diagnostics of HPV variants, because HPV variants have less than 1% of sequence variations between them.

REFERENCES

1. Bautsch W. 1994. Bacterial genome mapping by two-dimensional pulsed-field gel electrophoresis (2D-PFGE). Mol Biotechnol. 2(1):29-44.
2. Brown DF, Walpole E. 2001. Evaluation of the Mastalex latex agglutination test for methicillin resistance in *Staphylococcus aureus* grown on different screening media. J Antimicrob Chemother. 47(2):187-189.
3. Coutlee F, Rouleau D, Petignat P, Ghattas G, Kornegay JR, Schlag P, Boyle S, Hankins C, Vezina S, Cote P, Macleod J, Voyer H, Forest P, Walmsley S, Franco E. 2006. Enhanced Detection and Typing of Human Papillomavirus (HPV) DNA in Anogenital Samples with PGMY Primers and the Linear Array HPV Genotyping Test. J Clin Microbiol. 44(6):1998-2006.
4. Durbin R., S. R. Eddy, A. Krogh, G. Mitchison. (1998): Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, London, England
5. Edinger RC, Migneault PC, Nolte FS. 1985. Supplementary rapid biochemical test panel for the API 20E bacterial identification system. J Clin Microbiol. 22(6):1063-1065.
6. Felsenstein, J. 2002. PHYLIP (Phylogeny Inference Package) version 3.6a3. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
7. Francois P, Charbonnier Y, Jacquet J, Utinger D, Bento M, Lew D, Kresbach GM, Ehrat M, Schlegel W, Schrenzel J. 2005. Rapid bacterial identification using evanescent-waveguide oligonucleotide microarray classification. J Microbiol Methods. [Epub ahead of print].
8. Fredricks DN, Fiedler TL, Marrazzo JM. 2005. Molecular identification of bacteria associated with bacterial vaginosis. N Engl J Med. 353(18):1899-911.
9. Han CS, Xie G, Challacombe JF, Altherr MR, Bhotika SS, Bruce D, Campbell CS, Campbell ML, Chen J, Chertkov O, Cleland C, Dimitrijevic M, Doggett NA, Fawcett JJ, Glavina T, Goodwin LA, Hill KK, Hitchcock P, Jackson PJ, Keim P, Kewalramani AR, Longmire J, Lucas S, Malfatti S, McMurry K, Meincke LJ, Misra M, Moseman BL, Mundt M, Munk AC, Okinaka RT, Parson-Quintana B, Reilly LP, Richardson P, Robinson DL, Rubin E, Saunders E, Tapia R, Tesmer JG, Thayer N, Thompson LS, Tice H, Ticknor LO, Wills PL, Brettin TS, Gilna P. 2006. Pathogenomic Sequence Analysis of *Bacillus cereus* and *Bacillus thuringiensis* Isolates Closely Related to *Bacillus anthracis*. J Bacteriol. 188(9):3382-3390.
10. Heinzel PA, SY Chan, LHo, M O'Connor, P Balaram, MS Campo, K Fujinaga, N Kiviat, J Kuypers, and H Pfister. 1995. Variation of human papillomavirus type 6 (HPV-6) and HPV-11 genomes sampled throughout the world. J Clin Microbiol. 33(7): 1746–1754.
11. Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, Hegna I, Kolsto. 2000. *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*—one species on the basis of genetic evidence. Appl Environ Microbiol. 66(6):2627-2630.
12. Hryniewiecki T, Gzyl A, Augustynowicz E, Rawczynska-Englert I. 2002. Development of broad-range polymerase chain reaction (PCR) bacterial identification in diagnosis of infective endocarditis. J Heart Valve Dis. 11(6):870-874.
13. http://www.ircm.qc.ca/microsites/hpv/en/390.html
14. http://www.stgen.lanl.gov/stgen/virus/hpv/compendium/htdocs/COMPENDIUM_PDF/94PDF/3/MakePart3.pdf
15. Kaneko T, Nozaki R, Aizawa K. 1978. Deoxyribonucleic acid relatedness between *Bacillus anthracis*, *Bacillus cereus* and *Bacillus thuringiensis*. Microbiol Immunol. 22(10):639-641.
16. Karlin, S. y S. F. Altschul (1993): "Applications and statistics for multiple high-scoring segments in molecular sequences". Proc Natl Acad Sci USA 90(12): 5873-5877.
17. Keinanen-Toivola MM, Revetta RP, Santo Domingo JW. 2006. Identification of active bacterial communities in a model drinking water biofilm system using 16S rRNA-based clone libraries. FEMS Microbiol Lett. 257(2):182-188.
18. Kelly JJ, Siripong S, McCormack J, Janus LR, Urakawa H, El Fantroussi S, Noble PA, Sappelsa L, Rittmann BE, Stahl DA. 2005. DNA microarray detection of nitrifying bacterial 16S rRNA in wastewater treatment plant samples. Water Res. 39(14):3229-3238.
19. Kim BC, Park JH, Gu MB. 2005. Multiple and simultaneous detection of specific bacteria in enriched bacterial communities using a DNA microarray chip with randomly generated genomic DNA probes. Anal Chem. 15;77(8):2311-2317.
20. Kumar S, Tamura K, Nei M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5(2):150-63.
21. Lachica RV. 1976. Simplified thermonuclease test for rapid identification of Staphylococcus aureus recovered on agar media. Appl Environ Microbiol. 32(4):633-634.
22. Lau SK, Ng KH, Woo PC, Yip KT, Fung AM, Woo GK, Chan KM, Que TL, Yuen KY. 2006. Usefulness of the MicroSeq 500 16S rDNA bacterial identification system for identification of anaerobic Gram positive bacilli isolated from blood cultures. J Clin Pathol. 59(2):219-222.
23. Liu Y, Han JX, Huang HY, Zhu B. 2005. Development and evaluation of 16S rDNA microarray for detecting bacterial pathogens in cerebrospinal fluid. Exp Biol Med (Maywood). 230(8):587-591.
24. Lorenz E, Leeton S, Owen RJ. 1997. A simple method for sizing large fragments of bacterial DNA separated by PFGE. Comput Appl Biosci. 13(4):485-486.
25. Lu JJ, Perng CL, Lee SY, Wan CC. 2000. Use of PCR with universal primers and restriction endonuclease digestions for detection and identification of common bacterial pathogens in cerebrospinal fluid. J Clin Microbiol. 38(6):2076-2080.
26. Matveeva, O.V., Shabalina S. A., Nemtsov V. A., Tsodikov A. D., Gesteland R. F. Atkins J. F. (2003): Thermodynamic calculations and statistical correlations for oligo-probes design, Nucleic Acids Research, 31, 4211-4217.

27. Mount D. W. (2001): Bioinformatics. Cold Spring Harbor Laboratory Press, New York, USA.
28. Okhravi N, Adamson P, Matheson MM, Towler HM, Lightman S. 2000. PCR-RFLP-mediated detection and speciation of bacterial species causing endophthalmitis. Invest Ophthalmol Vis Sci. 41(6):1438-1447.
29. Reyes-Lopez, M.A., Mendez-Tenorio, A., Maldonado-Rodriguez, R., Doktycz, M., Fleming, J.T., Beattie, K. L. (2003): Fingerprinting of prokaryotic 16S rRNA genes using oligodeoxyribonucleotide microarrays and virtual hybridization, Nucleic Acids Research. 31: 779-789.
30 Roach JC, Levett PN, Lavoie MC. 2006 Identification of Streptococcus iniae by commercial bacterial identification systems. J Microbiol Methods. 6; [Epub ahead of print].
31. Sakamoto M, Takeuchi Y, Umeda M, Ishikawa I, Benno Y. 2003. Application of terminal RFLP analysis to characterize oral bacterial flora in saliva of healthy subjects and patients with periodontitis. J Med Microbiol. 52(Pt 1):79-89.
32. Sandri MT, Lentati P, Benini E, Dell'orto P, Zorzino L, Carozzi FM, Maisonneuve P, Passerini R, Salvatici M, Casadio C, Boveri S, Sideri M. 2006. Comparison of the Digene HC2 Assay and the Roche AMPLICOR Human Papillomavirus (HPV) Test for Detection of High-Risk HPV Genotypes in Cervical Samples. J Clin Microbiol. 44(6):2141-6.
33. Shou SY, Feng ZZ, Miao LX, Miu FB. 2006. Identification of AFLP markers linked to bacterial wilt resistance gene in tomato. Yi Chuan. 28(2):195-199.
34. The Institute of Genomic Research. 2006. *Comprehensive microbial resource (database of prokaryotic genomes). http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi*.
35. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25(24):4876-82.
36. Toledo MR, Trabulsi LR. 1983. Correlation between biochemical and serological characteristics of *Escherichia coli* and results of the Sereny test. J Clin Microbiol. 17(3):419-421.
37. Wellinghausen N, Wirths B, Franz AR, Karolyi L, Marre R, Reischl U. 2004. Algorithm for the identification of bacterial pathogens in positive blood cultures by real-time LightCycler polymerase chain reaction (PCR) with sequence-specific probes. Diagn Microbiol Infect Dis. 48(4):229-241.
38. Yang G, Liang CH, Cui JH, Chen S. 2006. The development and clinical application of papillomavirus genotyping by DNA chip. Zhonghua Liu Xing Bing Xue Za Zh;27(1):47-49.
39. Yang S, Lin S, Kelen GD, Quinn TC, Dick JD, Gaydos CA, Rothman RE. 2002. Quantitative multiprobe PCR assay for simultaneous detection and identification to species level of bacterial pathogens. J Clin Microbiol. 40(9):3449-3454.
40. Zimmer K, Drager KG, Klawonn W, Hess RG. 1999. Contribution to the diagnosis of Johne's disease in cattle. Comparative studies on the validity of Ziehl-Neelsen staining, faecal culture and a commercially available DNA-Probe test in detecting Mycobacterium paratuberculosis in faeces from cattle. Zentralbl Veterinarmed B. (2):137-140.

Correspondence to:

Maldonado-Rodríguez Rogelio
Cerrada Merced de las Huertas 28,
México 11420 D.F.,
Phone-Fax 57296000, ext. 62322.
E-mail: romaldodr@hotmail.com