Evaluación de los coeficientes Jaccard, Dice, Jeffrey´s X y Ochiai en la genotipificación de *Leuconostoc* spp.

Padilla-Frausto, J.J.[1]*, Ceja-Farias, T.K. [1], Cobo-Jauregui, J. J. [1] y Navarro-Villarruel, C.L.[1]
[1]División de Desarrollo Bio-Tecnológico, Centro Universitario de la Ciénega, Universidad de Guadalajara, Av. Universidad, No.1115, Col. Lindavista, 47810, Ocotlán, Jalisco. México. Tel.: +52(392)9259400 Ext. 48357
Correo: j.padilla@academicos.udg.mx

Palabras clave: Coeficientes de similitud, Genotipificación, PFGE-RAPD-Perfil plasmídico, *Leuconostoc* spp..

Introducción

Los estudios nacionales e internacionales de vigilancia sanitaria de los alimentos emplean métodos como la electroforesis en gel por campos pulsados (PFGE), amplificación polimórfica aleatorizada de ADN (RAPD) y perfiles de restricción plasmídicos, entre otros, para la genotipificación de cepas microbianas, con el objetivo de identificar clonas relacionadas cronológica y espacialmente. La PFGE y los perfiles de restricción plasmídicos se utilizan ampliamente para el intercambio de datos de identificación clonal entre laboratorios de investigación, debido a su alta reproducibilidad interlaboratorios [1], alto poder discriminatorio y adecuada relación costo-beneficio [2]. La RAPD resulta eficiente en la subclasificación de cepas en el laboratorio, sin embargo, su reproducibilidad y poder discriminatorio, es variable y dependiente del número y secuencia de los iniciadores arbitrarios y las condiciones de amplificación [3]. Una enorme variedad de cepas bacterianas se han analizado empleando los criterios propuestos por Tenover *et al.,* (1995) [4] para la cuantificación de diferencias entre dos patrones de bandas. Sin embargo, el empleo de este criterio está limitado a la comparación de un pequeño número de cepas [5]. Por lo que, cuando se incrementa en número de cepas a clasificar, no se brinda una adecuada clasificación debido a que las definiciones son arbitrarias. En este sentido, los coeficientes de similitud basados en el análisis del número de bandas, como el coeficiente de Dice o Jaccard, se utilizan luego para la construcción de dendrogramas, lo que proporciona una evaluación cuantitativa de la similitud de deformaciones y la asignación del genotípo se basa en la definición de un valor de enlace umbral, por debajo del cual las cepas se asignan al mismo grupo. Normalmente, esto se realiza de forma empírica mediante la inspección del dendrograma de análisis de conglomerados jerárquico que contiene las cepas de interés [2, 3]. Sin embargo, en una colección de cepas cuyo contenido de integrantes siempre es variable (tendiente a incrementarse continuamente) la eficiente organización jerárquica por este criterio se ve limitada, ya que el algoritmo límite debe reorganizarse cada vez que una nueva cepa se añade a la colección [6]. Una posible solución a la inestabilidad de la clasificación es generar una base de datos que colecte los patrones de bandas ya clasificados y en la misma, calcular continuamente el valor de similitud. Sin embargo, esta solución requiere determinar cuál es el coeficiente de similitud de patrones que mejor reproduce la clasificación de referencia. En consecuencia, el objetivo de este trabajo fue evaluar la exactitud de los coeficientes de similitud; Jaccard, Dice, Jeffrey´s X y Ochiai para clasificar patrones de bandas, con la finalidad de emplearlos en una clasificación automática de una colección de cepas bacterianas y comparar su eficiencia frente a la asociación visual de Tenover. En este documento se pretende destacar el valor crítico de las grandes colecciones que se han clasificado de forma visual y de la importancia de la implementación de la automatización informatizada en el laboratorio de referenciación genotípica de cepas bacterianas.

Metodología

Se emplearon tres colecciones de patrones de bandas: una que contenía 87 patrones de bandas generados por PFGE, otra de 112 patrones de RAPD y finalmente, una más de 23 patrones de restricción de plásmidos, resultado de la tipificación de cepas de *Leuconostoc* spp.

Clasificación visual de grupos (CVGTenover):

Los patrones de bandas obtenidos por los tres métodos de genotipificación fueron clasificados en tipos y subtipos mediante una inspección visual empleando los criterios establecidos por Tenover *et al.,* (1995); que establecen que dos cepas se consideran clonas si tienen una coincidencia exacta de los patrones de bandas y se consideran como cepas posiblemente relacionadas si tienen un máximo de seis bandas diferentes en

ambos carriles. Se identificaron mediante letras en mayúscula los grupos principales de asociación (≤ seis bandas diferentes) y subtipos: cepas con mayor similitud (Indistinguibles y ≤ tres bandas diferentes) con letras y números (ej., A1 y A2).

Base de datos de los patrones de bandas:
Las imágenes del gel obtenidas mediante un fotodocumentador (Kodak®, Edas® 290) en formato de negativo de 8 bits en escala de grises en formato TIF fueron analizadas empleando el sistema Bionumerics® ver. 3.0 e importados en un sistema binario en una base de datos de Excel® (2003). Se emplearon marcadores de 100pb y 1kpb como estándar para la normalización y calibración de cada gel. Todos los patrones de bandas fueron analizados mediante detección automática y fueron verificadas visualmente. Este último paso es de suma importancia, ya que la intensidad y/o las irregularidades de gel a gel causan errores en la asignación automática de la banda. Para los patrones generados por PFGE sólo las bandas que se encontraban entre 14 a 400 kpb se consideraron en este estudio. De igual manera, para RAPD y perfil de restricción plasmídica los segmentos y fragmentos superiores a los 200 pb.

Evaluación de los coeficientes de similitud entre patrones de bandas:
En la base de datos de Excel® (2003) se generaron matrices de comparación de similitud de patrones de bandas mediante el empleo de los cuatro diferentes coeficientes de similitud; Jaccard, Dice, Jeffrey´s X, y Ochiai (Tabla 1), generando grupos de asociación empleando diferentes límites de tolerancia en la posición de la banda en el gel (1, 2.5 y 5 %).

Tabla 1. Coeficientes de similitud ($S_{i,j}$) entre dos patrones de bandas, *i* y *j*

Coeficiente	Formula
Jaccard	$\frac{n_{i,j}}{n_{i,j} + n_i + n_j}$
Dice	$\frac{2n_{i,j}}{2n_{i,j} + n_i + n_j}$
Jeffrey´s X	$\frac{n_{i,j}}{N_i} + \frac{n_{i,j}}{N_j}$
Ochiai	$\frac{n_{i,j}}{\sqrt{[(n_{i,j} + n_i)*(n_{i,j} + n_j)]}}$

Donde:
$S_{i,j}$ describe el coeficiente de similitud;
$n_{i,j}$: Número de bandas que coinciden en los dos patrones.
n_i , n_j: Número de bandas sólo presentes en el patrón *i* o en el *j*.
N_i, N_i: Número total de bandas en el patrón *i* o en el *j*.

Curvas de análisis ROC:
Se utilizaron curvas de Característica Operativa Relativa (ROC, *receiver operating characteristic*, por sus siglas en inglés) para seleccionar él o los coeficientes posiblemente óptimos para la clasificación de patrones de bandas obtenidos por PFGE, RAPD y restricción de plásmidos. En este estudio se compararon los grupos formados por cada coeficiente, según límite de tolerancia en la posición de las bandas en el gel, identificándose los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP), o falsos negativos (FN). La exactitud de la clasificación de cada coeficiente fue entonces medido mediante el trazado de curvas de análisis ROC que relaciona el umbral de sensibilidad o la tasa de VP, en comparación con la tasa de FP (1- especificidad) (Tabla 2), para cada colección de patrones de bandas. Se calculó además, el área bajo la curva (ABC) que es la probabilidad de clasificar correctamente un par de patrones de bandas, seleccionados al azar de la colección. Un valor de ABC de 1, identifica un desempeño perfecto en la clasificación, es decir, que el resultado correcto será obtenido el 100% de las veces que sea evaluado.

Parámetro	Formula
Sensibilidad o tasa de verdaderos positivos	$\frac{VP}{VP + FN}$
Especificidad	$\frac{VN}{VN + FP}$
1 – Especificidad o tasa de falsos positivos	$\frac{FP}{VN + FP}$

Tabla 2. Parámetros de la curva de análisis ROC

Resultados y discusión

Con la finalidad de lograr una cabal comparación entre la agrupación visual de Tenover y las medidas alternativas de la similitud, se construyeron inicialmente las gráficas de CVGTenover (Figuras 1) que muestran la clasificación de las cepas en los tipos y subtipos, junto con los patrones normalizados. Con dicho estudio fue posible identificar 13 grupos tipo en los patrones generados por PFGE y 10 tipos más, tanto en los obtenidos por RAPD, como en los generados por restricción plasmídica.

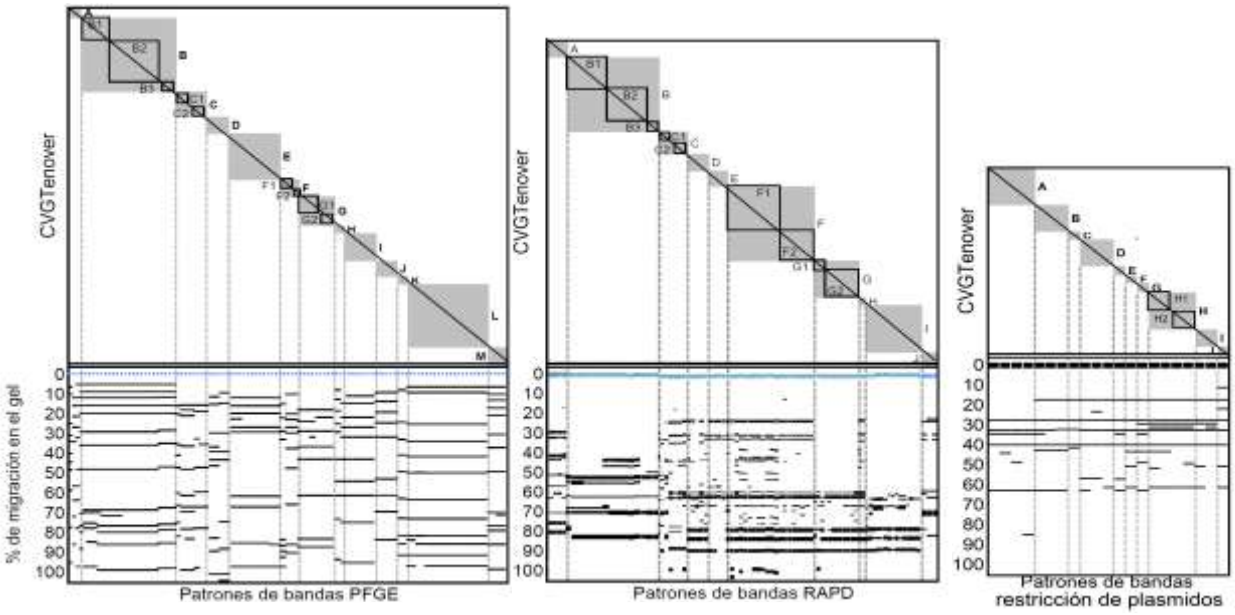

Figura 1. Representación de los grupos de CVGTenover según técnica.

En segundo lugar, se construyeron curvas ROC para cada coeficiente de similitud genética empleando diferentes valores de tolerancia de posición de las bandas en el gel, con la finalidad de determinar el coeficiente más exacto en la discriminación de patrones de bandas (Figura 2). Es importante mencionar que el análisis de curvas ROC constituye uno de los mejores métodos estadísticos para determinar la exactitud diagnóstica de dos o más métodos, siendo utilizadas con tres propósitos específicos: i) determinar el punto de corte de una escala continua en el que se alcanza la sensibilidad y especificidad más alta, ii) evaluar la capacidad discriminativa de la prueba, es decir en este contexto, su capacidad de diferenciar entre un patrón de bandas y otro, y iii) comparar la capacidad discriminativa de dos o más métodos que expresan sus resultados como escalas continuas [7]. Una prueba con discriminación perfecta, sin solapamiento de resultados, tiene una curva ROC que pasa por la esquina superior izquierda, donde la sensibilidad y la especificidad toman los valores máximos (S+E=1). Mediante el cálculo del ABC, fue posible identificar que el coeficiente Jaccard con tolerancia de 1% es la más sensible y específico para evaluar los patrones de bandas generados por PFGE, mientras que el coeficiente Dice con tolerancia de 5% resulto el más exacto en la comparación de patrones de bandas obtenidos mediante RAPD y restricción plasmídica.

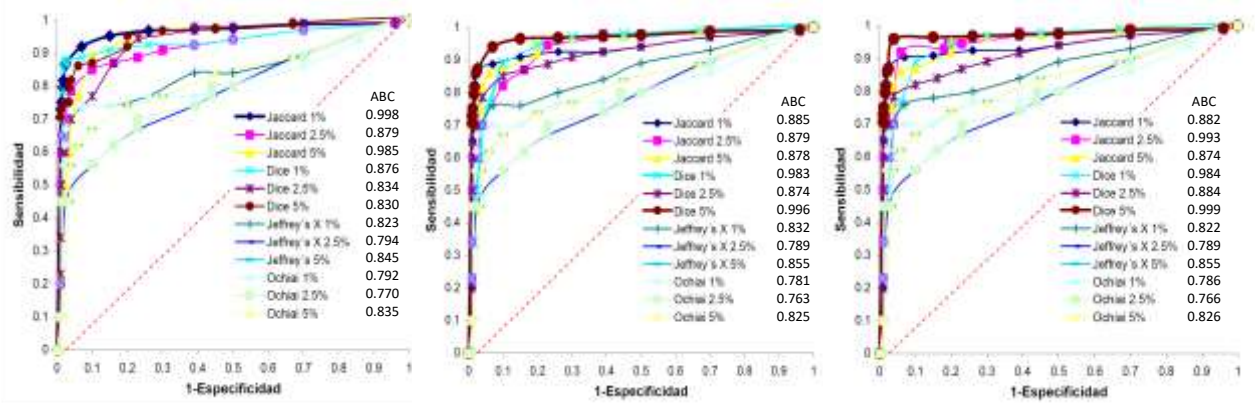

Figura 2. Curvas ROC para cada coeficiente de similitud, según tolerancia de migración de bandas en el gel

Kilic et al., (2010) [8] empleó una tolerancia de 5% en la posición de las bandas en el gel para comparar patrones de bandas obtenidos por restricción de plásmidos de *Salmonella* spp., reportando una eficiente clasificación de las cepas. Así mismo, la base de datos PulseNet, recomienda el empleo de 1% de tolerancia para la evaluación mediante PFGE.

Finalmente, en la Tabla 3 se muestra la correlación entre los grupos formados mediante la CVGTenover y los encontrados empleando los coeficientes que resultaron ser más exactos según técnica de tipificación.

Tabla 3. Correlación* entre los grupos de genotípos formados según técnica de tipificación.

Técnica	Observación	*r**
PFGE	CVGTenover *vs.* Jaccard 1%	0.8719
RAPD	CVGTenover *vs.* Dice 5%	0.5632
Restricción plasmidica	CVGTenover *vs.* Dice 5%	0.5402

* *r*: índice de correlación de Pearson (alfa=0.05); Si *r* =1, existe una correlación positiva perfecta

Conclusiones

La parametrización óptima e identificación de los coeficientes de similitud idóneos para la comparación de patrones de banda, abre la posibilidad de llevar a cabo de manera satisfactoria la automatización del análisis de los datos gráficos generados mediante PFGE, RAPD y restricción de plásmidos. Finalmente, dado que no se observó una adecuada correlación entre la clasificación visual y la estadística, se sugiere limitar el empleo de la CVGTenover a colecciones reducidas de cepas.

Bibliografía

1. Murchan, S., M. E. Kaufmann, R. Coombes, & B. Cookson. (2003). Harmonization of PFGE protocols for epidemiological typing of strains. *J. Clin. Microbiol.* **41:**1574–1585.
2. Van Belkum, A., M. Struelens, A. de Visser, & M. Tibayrenc. (2001). Role of evolutionary genetics, and microbiol epidemiology. *Clin. Microbiol. Rev.* **14:**547–560.
3. Struelens, M. J. (1998). Molecular epidemiologic typing systems of bacterial pathogens: current issues and perspectives. *Memorias Inst Oswaldo Cruz* **93***, 581–585.*
4. Tenover, F., Goering, P., H. Persing, & B. Swaminathan. (1995). Interpreting DNA restriction patterns produced by PFGE. *J. Clin. Microbiol.* **33**:2233–2239.
5. Brueggemann, A. B., D. T. Griffiths, D. W. Crook, and B. G. Spratt. (2003). Clonal relationships between the clone-specific differences. *J. Infect. Dis.* **187:**1424–1432.
6. Padilla-Frausto, J. J., Cepeda-Márquez, L. G., Salgado, L. M., Iturriaga, M. H., & Arvizu-Medrano, S. M. (2015). Detection and genotyping of *Leuconostoc* spp. in a sausage processing plant. *Journal of food protection*, ***78***(12), 2170-2176.
7. Cerda, J. & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. *Revista chilena de infectología*, **29**(2), 138-141.
8. Kilic, A., Bedir, O., Kocak, N., & A. Basustaoglu, C. (2010). Analysis of an Outbreak of *Salmonella* by Repetitive-Sequence-Based PCR and PFGE. *Inter. Med.* **49**: 31-36.