DIVERSIDAD GENÓMICA ENTRE AISLAMIENTOS URUGUAYOS DE *MAGNAPORTHE GRISEA*: DESARROLLO DE MODELOS DE CLASIFICACIÓN BASADOS EN AFLPS PARA RECONOCIMIENTO DE LINAJES

Fabián Capdevielle[1/], Victoria Bonnecarrère[1/], Silvia Garaycochea [2/], Stella Avila[3/]

INTRODUCCIÓN

Los marcadores de tipo AFLP constituyen una tecnología apropiada para la exploración de un gran número de loci distribuidos en el genoma de un organismo, sin requerir información a priori sobre la secuencia de los mismos. La tecnología AFLP combina la especificidad, resolución y poder de muestreo de las técnicas de digestión del ADN que utilizan enzimas de restricción con la velocidad y practicidad de la detección de polimorfismos mediante la amplificación de fragmentos de ADN via reacción en cadena de polimerasa (PCR). Los polimorfismos detectados dependen de la relación entre el número de nucleotidos selectivos utilizados en las reacciones de PCR y la complejidad del genoma del organismo. Para genomas menos complejos, como los de microorganismos, se utiliza un menor número de nucleotidos selectivos para obtener una cantidad mayor de fragmentos visualizables.

En forma preliminar se utilizó en 2003 la técnica de AFLP ("Amplified Fragment Lenght Polimorphisms") para identificar polimorfismos a nivel molecular entre aislamientos de *M. grisea* (Capdevielle, F., Federici, MT., Solares, E., Branda, A., Avila, S. 2003. Molecular strategies for characterization of fungal isolates from Uruguayan rice fields. Proceedings 3rd International Temperate Rice Conference, Uruguay). Posteriormente en 2004 se extendió este estudio sobre diversidad genética a un conjunto de aislamientos colectados a partir de lesiones sobre plantas de diferentes genotipos en campos experimentales y comerciales de Uruguay, que constituyen una colección de trabajo utilizable para modelar los cambios ocurridos en la población del agente causal del "bruzzone" entre 1995 y 2003.

[1/] Unidad de Biotecnología, INIA Las Brujas;
[2/] Estudiante Tesis Fac. de Ciencias, UdelaR;
[3/] INIA Treinta y Tres

Las tecnologías modernas de la información, basadas en disponibilidad de sistemas de computación intensiva, han permitido ajustar nuevas herramientas para la colección, transferencia, almacenamiento y combinación de datos fenotípicos y moleculares provenientes de diferentes poblaciones De acuerdo a este enfoque, consideramos que la información molecular sobre aislamientos de *M. grisea*, generada a través del proyecto FONTAGRO, permitiría la aplicación de un conjunto de procedimientos analíticos aplicados en minería de datos ("data mining" : DM). El proceso de DM, que también se conoce como descubrimiento del conocimiento en bases de datos (KDD), consiste en la extracción de información implícita, previamente desconocida y potencialmente útil a partir de bases de datos (Capdevielle, 2004. KDD: IT approach and models, NSF Pan-American Advanced Studies Institute Proceedings, Uruguay). En este caso se aplicarán sobre un tipo de información molecular (perfiles AFLP de cada aislamiento) y estarán orientados a inferir y validar patrones que identifiquen grupos de aislamientos en base a su linaje, cultivar afectado y época de muestreo.

METODOLOGÍA UTILIZADA

Entre las aplicaciones más importantes de los enfoques de DM se encuentra la clasificación de casos en dos o más clases previamente etiquetadas (predefinidas), desarrollando funciones y algoritmos para asignar de forma óptima nuevos objetos a las clases predefinidas. La clasificación de datos es el proceso que identifica las propiedades comunes a un conjunto de objetos en una base de datos y utiliza esas propiedades para construir un modelo de asignación a clases diferentes (Johnson and Wichern 1998). Para construir este modelo de clasificación, una base de datos de referencia (E) es tratada como el conjunto de entrenamiento ("training set")

para desarrollar algoritmos que permitan asignar objetos a clases predefinidas.

Cada caso incluido en el "training set" contiene información para el mismo conjunto de atributos (en este caso corresponden a loci AFLP) que otros casos disponibles en una base de datos relevante (W), y adicionalmente cada caso en el "training set" tiene una identidad de clase asociada (por ejemplo, linaje establecido mediante MGR, cultivar infectado, año de colecta). El objeto de esta clasificación es analizar los datos del conjunto de entrenamiento y desarrollar una descripción precisa o un modelo para cada clase utilizando las variables disponibles (Hand, 1997). Dichas descripciones son entonces utilizadas para clasificar casos adicionales o futuros en las bases de datos relevantes (W) o para desarrollar mejores modelos de clasificación para diferentes clases en las bases de datos.

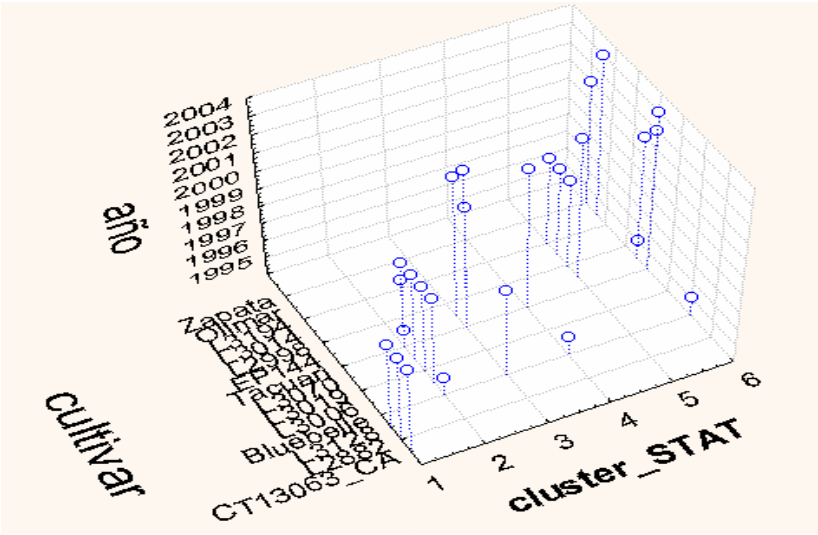
Considerando que el genoma haploide de *M. grisea* tiene un tamaño aproximado de 38 Mbp, se utilizaron oligonucleótidos no selectivos (EcoRI+0 y MseI+0) para amplificar secuencias a partir del ADN cortado con las enzimas EcoRI y MseI, seguido por una etapa de amplificación selectiva con oligonuclótidos marcados (EcoRI+2 y MseI +2, en diferentes combinaciones), lo que permitió identificar más de 160 fragmentos polimórficos en el genoma de los aislamientos analizados.

Los marcadores AFLP representarían un sistema de muestreo de la diversidad genómica de *M. grisea* que podría utilizarse para inferir patrones con valor predictivo respecto a diferentes subconjuntos dentro de la estructura poblacional del patógeno en Uruguay. En base a la información molecular obtenida se ajustó un procedimiento de análisis discriminante basado en métodos no paramétricos ("K-Nearest Neighbor") como herramienta para construir modelos predictivos de pertenencia a grupos de referencia (linajes, etc.), utilizando algoritmos de validación cruzada ("leave-one-out crossvalidation" proc discrim, SAS Institute) para evaluar la efectividad de la clasificación a partir de diferentes muestras de observaciones para las cuales la asignación de linajes es conocida.

RESULTADOS

Diversidad genómica detectada mediante marcadores AFLP

Utilizando una matriz de datos definida para presencia / ausencia de marcadores AFLP se agruparon los aislamientos en 6 grupos (procedimiento K-means-Statistica, variables = 168 loci AFLP), donde las distancias relativas entre aislamientos respecto al "centro" de cada grupo son minimizadas en comparación con las distancias relativas entre grupos.



Los grupos identificados en este estudio presentan las siguientes características en cuanto a sus componentes:

Grupo 1, incluye aislamientos colectados sobre hojas de diferentes líneas en 1999 (L2882, L3128,CT13063-CA).

Grupo 2, incluye aislamientos colectados sobre hojas en Bluebelle (1995) y diferentes líneas en 1999 (L3006, L3019,L3070), así como aislamientos colectados sobre cuello de panoja y hojas en INIA Tacuarí (1995, 1998, y 1999).

Grupo 3, incluye aislamientos colectados sobre hojas en Bluebelle (1995) y sobre tallos en INIA Tacuarí (2001), y mayoritariamente aislamientos colectados sobre lesiones del cuello de la panoja en EP144 (2002) e INIA Tacuarí (1 aislamiento en 2002 y 4 aislamientos en 2003).

Grupo 4, incluye aislamientos colectados sobre lesiones del cuello de la panoja en Bluebelle (1995) e INIA Tacuarí (2002).

Grupo 5, incluye tres aislamientos colectados en 1999 sobre lesiones en hojas de diferentes líneas (L 2998, L3014, L3194), y 1 aislamiento colectado sobre lesiones del cuello de la panoja en EP144 (2002).

Grupo 6, incluye aislamientos colectados sobre lesiones del cuello de la panoja en Bluebelle (1995), EP144 (1995 y 2001), INIA Zapata (2001), INIA Tacuarí (2002 y 2003), e INIA Olimar (2003).

En las dos últimas zafras INIA Tacuarí ha pasado a ser el cultivar más afectado por este patógeno, lo cual pone de manifiesto una importante diferencia respecto de zafras anteriores en las cuales el cultivar más afectado era El Paso 144. Estas observaciones pueden interpretarse en función de los cambios ocurridos en el comportamiento de la población del hongo causante del "bruzzone" (*M. grisea*) durante el período de muestreo considerado. De acuerdo a los resultados obtenidos en este estudio, INIA Tacuarí sería susceptible a aislamientos de los grupos 2, 3, 4 y 6, mientras EP 144 sería susceptible a aislamientos de los grupos 3, 5 y 6. Los aislamientos comprendidos en los grupos 3 y 6 son relativamente escasos antes de 2001, aumentando rápidamente su proporción en los últimos años. Los orígenes de estos grupos se podrían identificar en aislamientos que se encontraban presentes desde 1995 (grupo 6) y desde 1999 (grupo 3), y que representaron una pequeña proporción (5%-10%) del total de aislamientos realizados antes del año 2001.

A partir de 2001 los aislamientos que afectaron a INIA Tacuarí corresponderían mayoritariamente al grupo 3, lo cual coincide con una expansión del área afectada por la enfermedad que se presentaría casi endémica en algunas regiones y extendiendo su presencia más al sur de la zona Este del país, donde anteriormente era prácticamente inexistente. Durante 2003 la mayoría de los aislamientos de *M. grisea* fueron colectados sobre INIA Tacuarí, distribuidos entre el grupo 3 (60 %) y el grupo 6 (40%). Como hipótesis a considerar para las orientaciones futuras del programa de mejoramiento genético podemos suponer que los aislamientos relacionados con el grupo 6 representarían un mayor riesgo desde el punto de vista de la propagación del patógeno a diferentes áreas de cultivo, debido a que aparentemente tienen el potencial de afectar a la mayoría de los cultivares (Bluebelle, EP144, INIA Tacuarí, INIA Zapata, INIA Olimar).

La diferenciación genética entre los 6 grupos de aislamientos se estimó a través de la comparación del índice de fijación de Wright (*Fst*), calculado utilizando software AFLP-Surv 1.0, con la distribución de todos los posibles valores de Fst obtenidos mediante permutaciones de la base de datos. El valor de *Fst* calculado (0,4405) fué significativamente superior ($p<0.001$) al esperado de acuerdo a la distribución ad-hoc de posibles valores de *Fst*; esto permitió rechazar la hipótesis nula de que no existen diferencias entre diferentes poblaciones componentes de la estructura genética de M. grisea. Por lo tanto se puede afirmar que los grupos identificados presentan diferencias a nivel genómico que son detectadas mediante modificaciones en

los patrones de marcadores AFLP. Aplicando análisis discriminante con selección de variables (proc stepdisc, SAS Institute) fue posible seleccionar un número reducido de loci AFLP (B6, H24, B2, B11, C9, D13, B12, A11, B8, B3, H3, C15, B5) que maximizan la diferenciación entre los 6 grupos identificados dentro de la estructura poblacional de los aislamientos analizados. Estos marcadores AFLP se incorporaron en un modelo de clasificación (algoritmo k-NN), alcanzando un porcentaje de clasificación correcta superior al 99% en pruebas de validación cruzada que estiman su potencial para asignar nuevos aislamientos de *M. grisea* dentro de cada grupo.

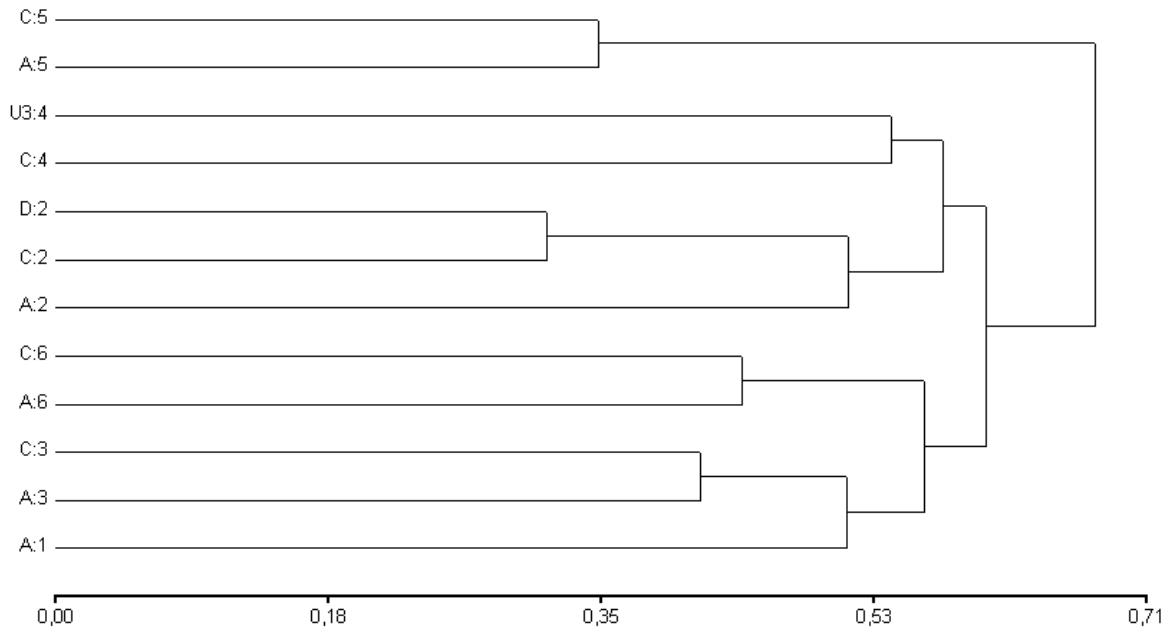
A partir de estos resultados se sugiere la utilización de la información de AFLP para trazar el posible origen y dispersión de aislamientos de *M. grisea* recolectados sobre cultivares y líneas experimentales a través de una serie de años en diferentes localidades de Uruguay.

MODELOS DE CLASIFICACIÓN BASADOS EN AFLPS

Los aislamientos agrupados mediante información de marcadores AFLP (Cluster_stat) para los que se dispone de información sobre el correspondiente linaje (MGR_obs) fueron agrupados utilizando el algoritmo UPGMA en función de ambos criterios. Los resultados obtenidos indican que no existe una relación directa entre la pertenencia a un determinado grupo inferido mediante marcadores AFLP y la pertenencia a un linaje particular; esto sería esperable si la distribución genómica de los marcadores AFLP resulta más amplia que la correspondiente a aquellos loci explorados mediante la técnica MGR. Como se observa en el gráfico anterior, los aislamientos correspondientes a los linajes A y C pueden subdividirse en diferentes grupos en función del grado de similaridad estimado a partir de la matriz de marcadores AFLP mediante agrupamiento no jerárquico (“k-means”).

Por lo tanto no resultaría efectivo buscar una correlación directa entre los agrupamientos generados con diferentes técnicas cuya cobertura genómica es diferente, debiendo explorarse el uso de algoritmos bioinformáticos que permitan analizar la información de marcadores AFLP e inferir patrones aplicables a la clasificación de aislamientos en sus correspondientes linajes. A efectos de establecer el valor informativo de los marcadores AFLP respecto a la asignación de aislamientos a grupos previamente identificados con metodologías de referencia, se estableció un modelo de clasificación (proc discrim - algoritmo k-NN, SAS Institute) que permite clasificar los aislamientos en cada uno de los linajes predefinidos, tomando como referencia un conjunto de aislamientos para los que se dispone de información sobre perfiles de AFLPs y sus correspondientes linajes establecidos mediante MGR.

Agrupamiento de linajes y conglomerados basados en AFLPs



La precisión del algoritmo de clasificación utilizado (k-NN) fué evaluada mediante una validación cruzada total del set de datos (*n* aislamientos), con respecto a la probabilidad de error en la ubicación de cada uno de los *n* aislamientos en linajes predefinidos ("missclassification") utilizando un set de *n-1* aislamientos para predecir la ubicación de cada uno de los restantes aislamientos ("leave-one-out crossvalidation"). Los aislamientos utilizados como referencia (pertenecientes a los linajes A, C, D y U3) fueron correctamente clasificados en la totalidad de los casos, por lo que se plantea en forma hipotética la asignación de los restantes aislamientos a los linajes A (15 aislamientos sobre líneas experimentales y cultivares), C (12 aislamientos sobre líneas experimentales y cultivares, con predominio de INIA Tacuarí), D (2 aislamientos sobre INIA Zapata e INIA Tacuarí), y U3 (1 aislamiento sobre Bluebelle).

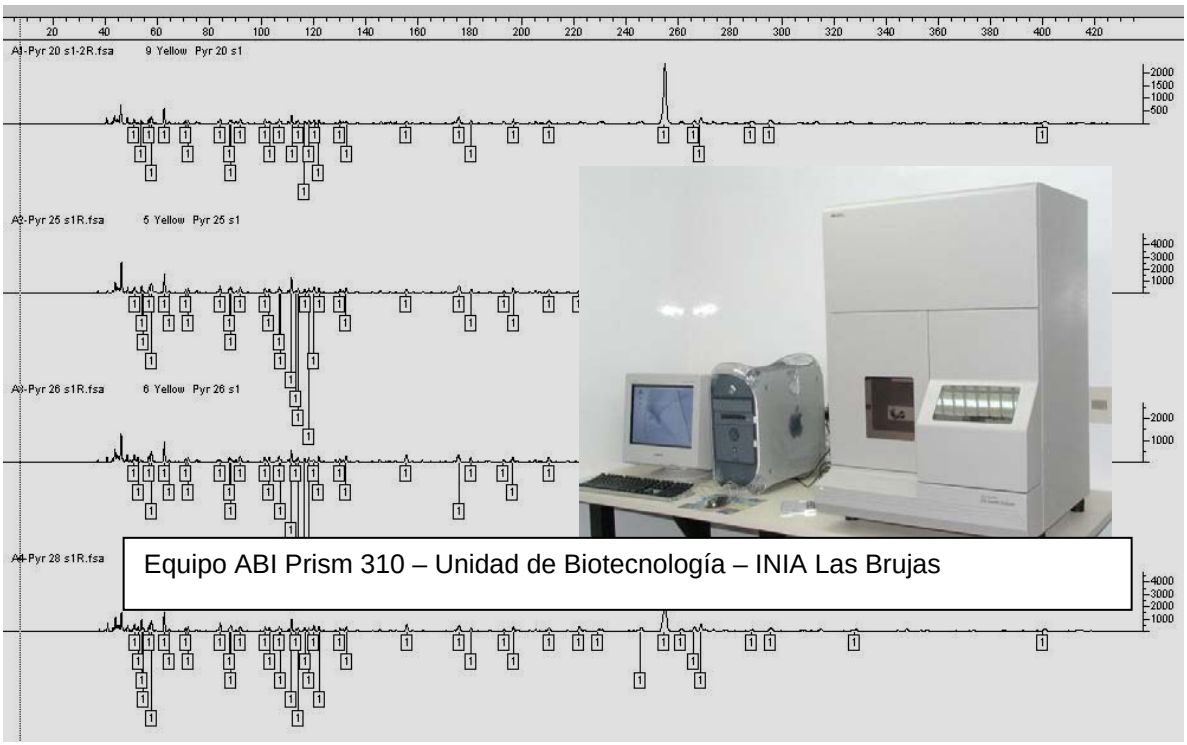
Cuadro 1. Porcentaje de aislamientos clasificados correctamente utilizando marcadores seleccionados mediante análisis discriminante (proc stepdisc, SAS Institute, V8)

	Número de Marcadores incluídos en modelo k-NN													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**15**	**16**	**20**
MGR_obs	25	83	80	80	82	89	89	93	93	95	100	100	100	100
Cluster_stat	17	33	50	50	83	98	98	98	98	98	100	100	100	100
Cultivar	0	0	24	29	36	43	57	57	73	74	77	85	99	99.5

En el Cuadro 1 se describen ejemplos del valor discriminativo de los modelos de clasificación basados en AFLPs evaluados mediante validación cruzada respecto al % de aislamientos correctamente asignados a las diferentes clases (cultivar, agrupamiento de aislamientos basado en k-means, ó linaje MGR al que pertenece el aislamiento) consideradas en este estudio. En todos los casos se obtuvo un alto porcentaje de clasificación correcta (superior al 99 %) incluyendo menos de 20 marcadores AFLP seleccionados dentro de la matriz de información molecular en un procedimiento de análisis discriminante no paramétrico ("K-Nearest Neighbor"), donde se utiliza la

información de los perfiles AFLP de cada aislamiento para establecer las distancias entre todos los pares de aislamientos posibles y se asignan los aislamientos en estudio a la clase a la que pertenece el aislamiento más similar (k=1) en el espacio multivariado definido por la matriz de AFLPs disponible.

De acuerdo a este experimento, la clasificación basada en 11 marcadores AFLP permitiría asignar correctamente todos los aislamientos en sus respectivos linajes, dado un número suficiente de aislamientos con información de linajes (MGR_obs) que serán utilizados como referencia para ajustar un modelo k-NN aplicable a cada colección de aislamientos. Considerando la necesidad de mantener un sistema actualizado de genotipado, aplicable en caracterización de los aislamientos utilizados para selección del germoplasma de arroz y en monitoreo de nuevos aislamientos, hemos incorporado la utilización del kit AFLP para microorganismos (Applied Biosystems) dentro de la plataforma genómica-bioinformática que está siendo implementada por la Unidad de Biotecnología en apoyo a proyectos de investigación del Programa Arroz.



Equipo ABI Prism 310 – Unidad de Biotecnología – INIA Las Brujas