UNIVERSIDAD: Universidad Nacional de La Plata

NUCLEO DISCIPLINARIO: Matemática Aplicada

Título del trabajo: **CÓMO ESTIMAR LA CAPACIDAD PREDICTIVA DE UN MODELO QUÍMICO-MATEMÁTICO: UN PROBLEMA DE MUESTREO**

AUTOR(es): Alan Talevi, Carolina L. Bellera, Eduardo A. Castro, Luis E. Bruno-Blanch

CORREOS ELECTRÓNICOS DE LOS AUTORES: atalevi@biol.unlp.edu.ar, cbellera@biol.unlp.edu.ar, castro@quimica.unlp.edu.ar, lbb@biol.unlp.edu.ar

PALABRAS CLAVES: análisis discriminante – validación – partición óptima
PALAVRAS CHAVES: análise discriminante - validação - divisória óptima

INTRODUCCIÓN

Los estudios QSAR/QSPR (Relaciones cuantitativas estructura-actividad/estructura propiedad) constituyen un área de la Química y la Fisicoquímica de notable expansión en las últimas décadas. Se fundamentan en el uso del conocimiento fisicoquímico, herramientas estadísticas y tecnología computacional para establecer relaciones cuantitativas entre una actividad biológica o una propiedad fisicoquímica determinada y la estructura molecular de un grupo de compuestos químicos [1]. A los fines del modelado matemático de la actividad o propiedad estudiada, diversos aspectos de la estructura molecular (por ejemplo, la simetría de la misma, su grado de complejidad, su distribución electrónica, la presencia de ciertos grupos químicos determinados) se traducen en variables numéricas independientes discretas o continuas llamadas descriptores moleculares; luego se establece la contribución cuantitativa de cada descriptor molecular a la actividad o propiedad de interés (por ejemplo, una actividad farmacológica dada). La Teoría QSAR/QSPR trata, en otras palabras, de cuantificar la manera en que una característica de la estructura molecular (nivel atómico) se traduce o refleja en una propiedad o actividad observada a nivel macroscópico. Es decir, la magnitud de cualquier propiedad observada a nivel macroscópico para un conjunto de compuestos químicos (el conjunto de entrenamiento) puede expresarse como una función de un conjunto de descriptores:

Propiedad observada microscópicamente = f (descriptor 1, descriptor 2, …, descriptor n)

Ya las primeras investigaciones en el campo de los estudios QSAR/QSPR demostraron que para modelar propiedades biológicas (tales como una actividad biológica o propiedades vinculadas a fenómenos de transporte) en general debe recurrirse a modelos multivariable que incluyan descriptores correspondientes a distintos aspectos de la estructura molecular, e incluso pueden requerirse modelos no lineales para una adecuada descripción o caracterización de la propiedad estudiada [2,3]. En ocasiones, entonces, la función anterior consistirá simplemente en una combinación lineal de un conjunto de descriptores; en caso de comportamientos observados más complejos deberá recurrirse a funciones/modelos no lineales.

Una vez desarrollado el modelo QSPR, el mismo puede aplicarse, con ciertas consideraciones, en la predicción de la propiedad estudiada para compuestos no incluidos en el conjunto de entrenamiento. Estas consideraciones se refieren, fundamentalmente, al **dominio de**

aplicabilidad del modelo y a la necesidad de validar el modelo adecuadamente para descartar *overfitting* o **sobreajuste**. Para entender ambos conceptos puede recurrirse a la Teoría de muestreo.

El conjunto de compuestos químicos (casos) a partir de los cuales se deriva el modelo QSPR puede entenderse como una muestra a partir de la cual el modelador realiza una inferencia determinada que luego extrapolará a una población dada (el conjunto de compuestos cuyo valor de la propiedad estudiada habrá de predecirse). Como en general el conjunto de entrenamiento se diseña retrospectivamente, buscando en literatura estructuras químicas cuya propiedad haya sido medida/determinada, el concepto de dominio de aplicabilidad del modelo se refiere a delimitar o definir, a partir de la muestra, la "población" de estructuras ajenas al conjunto de entrenamiento dentro de la que las predicciones del modelo generado serán confiables. En otras palabras, definir el dominio de aplicabilidad equivale a contestar la pregunta: ¿de qué población de compuestos químicos es representativo mi conjunto de entrenamiento? Las predicciones son confiables solamente dentro de una población de estructuras que comparte ciertas características estructurales de la muestra; la extrapolación fuera de esa población da lugar a predicciones de valor dudoso [4].

El **sobrejauste**, por otro lado, es un comportamiento indeseado y se refiere a que el modelo QSPR refleje o capture características muestrales que no estén presentes en la población general en la que se aplicará el modelo, accidentes estructurales que a nivel poblacional no se vinculan con la propiedad estudiada. Puesto de otro modo, sobreajustar consiste en aumentar la capacidad del modelo de explicar el comportamiento del conjunto de entrenamiento (muestra) en detrimento de su capacidad de generalización, su capacidad predictiva sobre un conjunto independiente de casos [5]. La figura 1 grafica la problemática del sobreajuste. Imaginemos que los puntos en verde representan las observaciones correspondientes a la muestra (el conjunto de entrenamiento); los puntos anaranjados corresponden a casos externos al conjunto de entrenamiento, es decir, cuatro casos de la población general no utilizados para el modelado. Si ajustamos los datos de la muestra a través de la relación polinomial representada mediante la curva verde, el modelo ajustará, *para la muestra*, perfectamente a los datos observados; los residuales para el conjunto de entrenamiento serán en todos los casos 0. Sin embargo, los residuales para los cuatro puntos independientes del conjunto de entrenamiento serán en general menores si el conjunto de entrenamiento se describe mediante un modelo lineal, más simple (la recta). En el ejemplo, entonces, si se ajustaran las observaciones mediante el modelo

representado por la curva de color verde se sacrificaría capacidad predictiva sobre los casos anaranjados, pertenecientes a la población. Esto no supone ninguna utilidad, ya que el valor de la propiedad para los compuestos del conjunto de entrenamiento ya se conoce, y el objetivo del modelado es predecir el valor de la propiedad para casos en los cuales el mismo se desconoce.

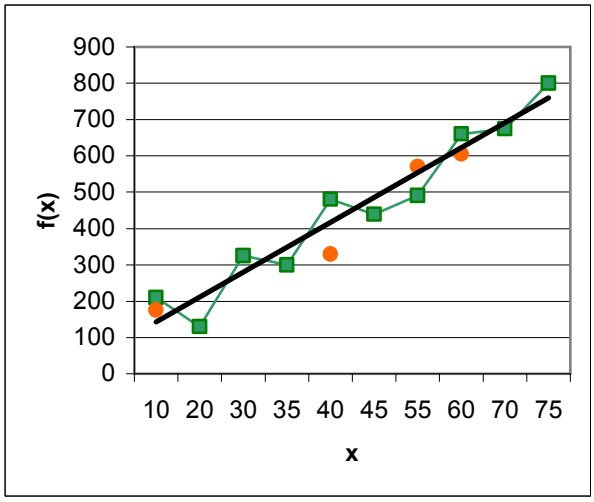

Figura 1. Ejemplo de sobreajuste.

Para escapar al sobreajuste deberá evitarse la inclusión de predictores innecesarios en el modelo (dado que conforme aumenta el número de descriptores incorporados mejora el ajuste del modelo a los datos observados para el conjunto de entrenamiento, pero más allá de un número crítico y en principio desconocido de descriptores se incurrirá en sobreajuste). Por otro lado, deberán evitarse sesgos en el diseño del conjunto de entrenamiento: idealmente, las distribuciones de las características estructurales y del valor de la propiedad estudiada en la muestra debería reflejar las de la población, y el número de casos considerado en el conjunto de entrenamiento debería ser estadísticamente significativo. Como se mencionó, el conjunto de entrenamiento suele recolectarse de manera retrospectiva, y lamentablemente el número de compuestos químicos ha los que se les ha medido o se les puede medir la propiedad que se desea estudiar es usualmente limitado.

A fin de evitar el sobreajuste se suele recurrir a técnicas de validación. Dentro de las técnicas de validación más utilizadas encontramos la validación cruzada interna Leave-Group-Out (LGO) o Leave-n-out y la validación externa [6]. La validación cruzada interna LGO consiste en, una vez generado el modelo, remover aleatoriamente un número n de casos del conjunto de entrenamiento, generar un nuevo modelo con los mismos descriptores del modelo original y utilizar este segundo modelo generado para predecir el valor de la propiedad estudiada para los

casos removidos. Esta mecánica se repite por lo menos hasta haber removido una vez cada uno de los casos del conjunto de entrenamiento. El procedimiento se esquematiza en la fig. 2.

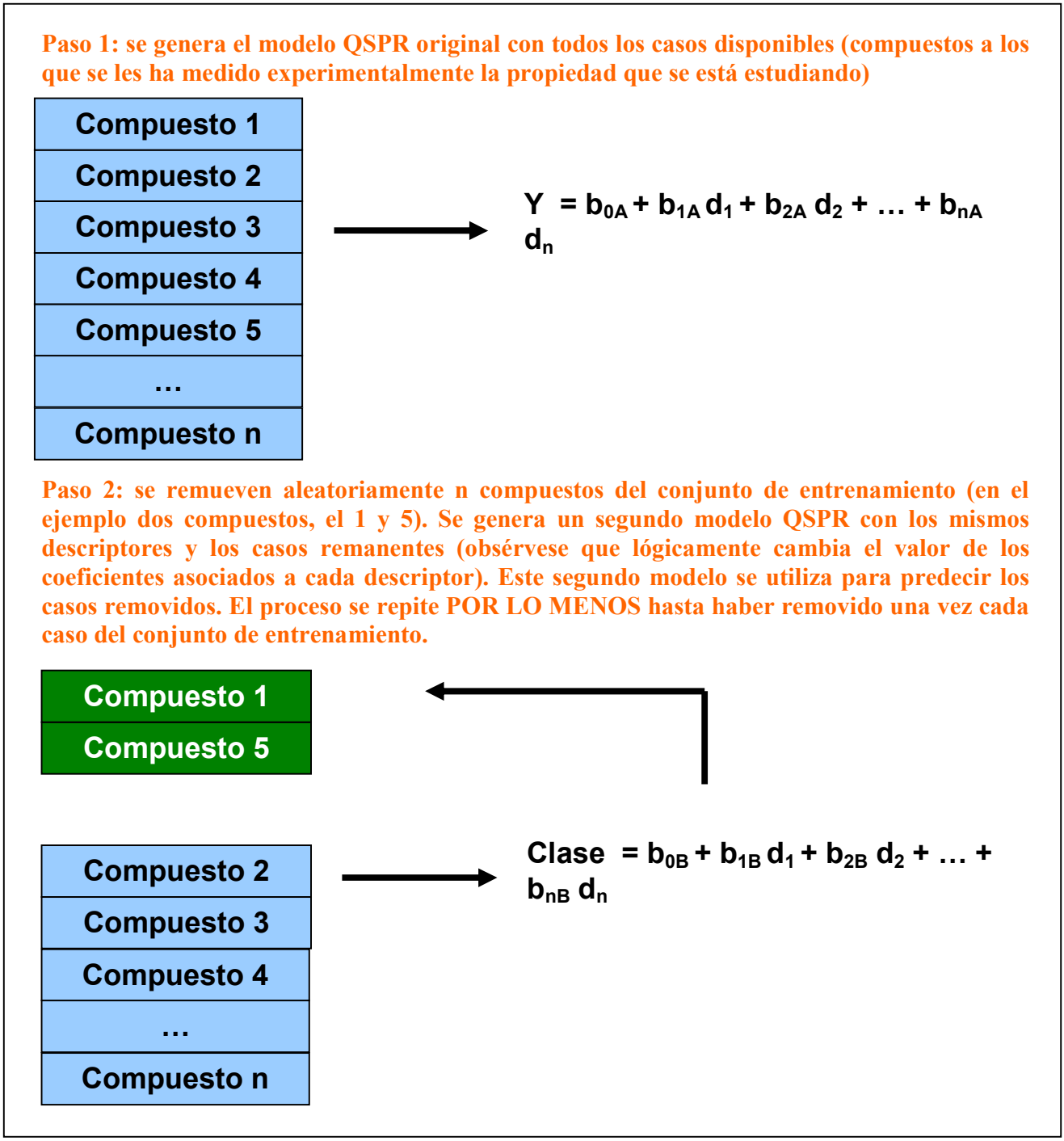

Fig. 2. Representación esquemática del procedimiento de la validación interna cruzada LGO.

El parámetro que mayormente se utiliza para evaluar la capacidad predictiva del modelo mediante esta técnica es el q^2:

$$q^2 = 1 - \frac{\sum_{i=1}^{n} \left(Yobs,i - Ypre,i \right)^2}{\sum_{i=1}^{n} \left(Yobs,i - Ymed \right)^2}$$

donde $Y_{pre,i}$ corresponde al valor predicho para cada caso i cuando es removido del conjunto de entrenamiento, $Y_{obs,i}$ corresponde al valor observado/medido experimentalmente de la propiedad estudiada para cada caso i del conjunto de entrenamiento y Y_{med} corresponde al valor medio de la propiedad para los n casos que componen el conjunto de entrenamiento. Se ha observado, sin embargo, que el q^2 tiende a sobreestimar la capacidad predictiva del modelo sobre la población [7,8]. Se asume que el desempeño del modelo en la predicción de la propiedad para un conjunto de prueba independiente es un mucho mejor estimativo de la capacidad predictiva del modelo. En esto consiste, justamente, la validación externa: trabajar con un conjunto de entrenamiento más pequeño a fin de reservar parte de los casos disponibles (la muestra *hold-out* o *split-out*) como conjunto de prueba independiente. Típicamente la parte de la muestra que se reserva consiste en un 10 a 20% de los casos totales [6]. Hawkins ha sugerido recurrir a la validación externa si y sólo si se dispone de un número de casos grande para el conjunto de entrenamiento (varios cientos de casos, como mínimo) a fin de no resignar información importante para la etapa de modelado en pos de una mejor validación cuando el número de casos está restringido [5].

OBJETIVOS

El presente trabajo consiste en utilizar una muestra de tamaño intermedio (160 compuestos) en la generación de modelos QSPR para predecir la permeabilidad intestinal de fármacos. La propiedad elegida es meramente anecdótica, ya que en este trabajo se desea estudiar, específicamente: 1) si efectivamente la validación externa es el método de validación más confiable para estimar la generalizabilidad del modelo (capacidad predictiva sobre la población de interés); 2) cómo debe particionarse el conjunto de casos disponible en un conjunto de entrenamiento y un conjunto de prueba a fin de tener menor probabilidad de sobreajuste. De todos modos, vale destacar que el %PI es una propiedad de sumo interés para definir si es posible administrar un fármaco por vía oral y que el mismo llegue cuantitativamente a circulación sanguínea.

MATERIALES Y MÉTODOS

En primer lugar se reunió el conjunto de casos totales, consistente en 160 fármacos con permeabilidad intestinal (%PI) conocida (la lista de casos no se muestra por cuestiones de espacio, pero está disponible para quien desee consultarla). Los valores de %PI de los compuestos se obtuvieron de publicaciones científicas especializadas y fueron chequeados exhaustivamente a través de los servicios HSDB (www. toxnet.nlm.nih.gov) y Pubchem (www.pubchem.ncbi.nlm.nih.gov), dependientes del Instituto Nacional de Salud de Estados Unidos. Dado que los datos de permeabilidad intestinal presentan gran variabilidad, se recurrió como técnica de modelado al Análisis Discriminante a fin de derivar modelos QSAR capaces de clasificar los compuestos utilizados en cuatro categorías, según su %PI: categoría 1) compuestos con %PI ≤ 20%; categoría 2) compuestos con 20 < %PI ≤ 50%; categoría 3) compuestos con 50 < %PI ≤ 80% y; categoría 4) compuestos con %PI > 80%. Se utilizó para ello el módulo General Discriminant Analysis Modeling del programa Statistica 7.0 (Statsoft Inc, 2004). No se consideraron compuestos que cayeran en distintas categorías según las distintas fuentes bibliográficas consultadas; sólo se tuvieron en cuenta compuestos absorbidos pasivamente a través de la membrana gastrointestinal (transporte pasivo transcelular y paracelular). La distribución de los 160 compuestos entre las 4 categorías fue de 47 compuestos en categoría 1; 26 compuestos en categoría 2; 40 compuestos en categoría 3 y; 47 compuestos en categoría 4.

Se procedió entonces a realizar distintas particiones del conjunto de casos en conjunto de entrenamiento y conjunto de prueba: partición 120/40; partición 80/80; partición 40/120. Es decir, en el primer caso se utilizan 120 casos de los 160 disponibles para entrenar el modelo, y 40 casos para validarlo mediante validación externa; en el segundo caso, 80 casos para el entrenamiento y 80 casos para la validación; en el tercer caso, se consideran sólo 40 casos para el entrenamiento y los 120 restantes para la validación. Para realizar las particiones se recurrió a un muestreo aleatorio sistemático. Los compuestos de cada categoría fueron ordenados por orden alfabético y luego numerados según ese orden. Para la partición 120/40 se reservaron para la validación externa los compuestos numerados como múltiplos de 4, para la partición 80/80 los compuestos pares, etc. Vale aclarar que los nombres de los compuestos son nombres de fantasía y que su orden alfabético no guarda ninguna relación con su estructura química; no existe por lo tanto relación alguna entre el orden de los compuestos y su estructura química que afecte la aleatoriedad del muestreo sistemático propuesto.

Se utilizaron como conjuntos de descriptores (posibles variables independientes del modelo) dos subconjuntos de descriptores moleculares del programa Dragon (Milano Chemometrics,

2003): el primero de ellos formado por índices de información, índices de carga de Gálvez, fragmentos centrados en átomos y frecuencias de grupos funcionales; el segundo formado por descriptores constitucionales, descriptores topológicos, índices de información e índices de carga de Gálvez. La elección de estos subconjuntos de descriptores para el modelado está vinculada a la experiencia del grupo de trabajo con el uso del programa Dragon y la eficiencia de ciertos grupos de descriptores para establecer relaciones QSPR. Se trabajó únicamente con grupos de descriptores que pueden calcularse a partir de una representación de la molécula en el plano, a fin de independizarnos de la conformación tridimensional de las moléculas. Se procedió a generar, mediante metodología Stepwise Forward, funciones discriminantes de hasta 10 descriptores, evitando la inclusión de descriptores con p valor asociado mayor a 0.05.

RESULTADOS

Las Figs. 3a y 3b presentan los resultados del análisis realizado a partir de los modelos derivados de cada uno de los dos subconjuntos de posibles predictores descritos en materiales y métodos. Se presentan el % global de buenas clasificaciones de los compuestos en las 4 categorías de %PI consideradas para los conjuntos de entrenamiento y de prueba definidos por cada una de las particiones consideradas.

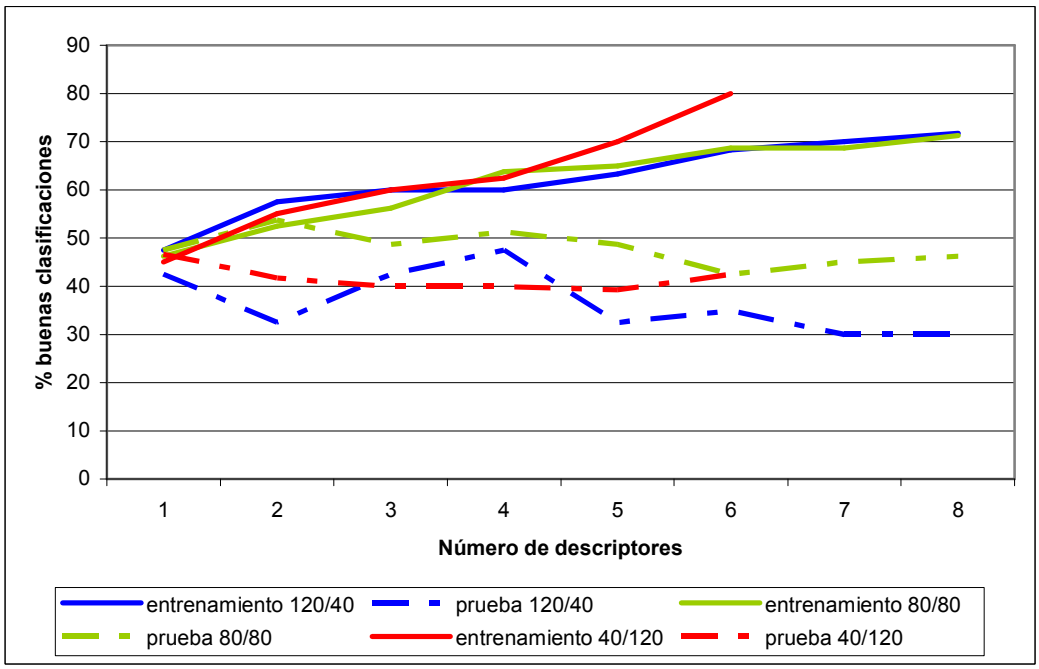

Fig. 3a. Desempeño de los modelos derivados del primer subconjunto de descriptores de Dragon, en la clasificación de los conjuntos de entrenamiento y prueba para cada una de las particiones consideradas.

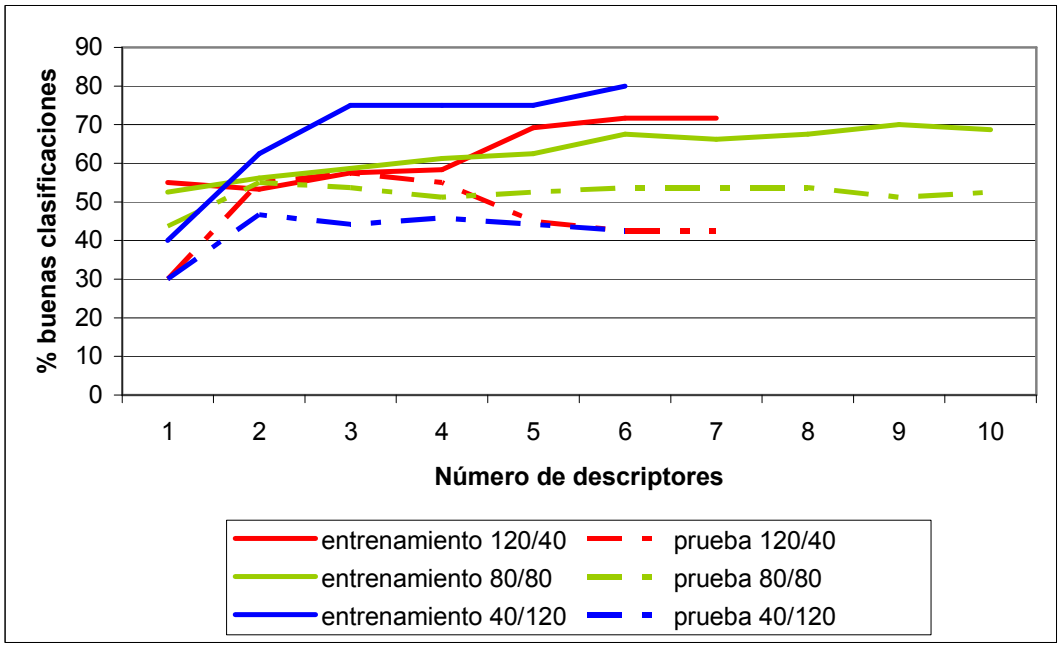

Fig. 3b. Desempeño de los modelos derivados del segundo subconjunto de descriptores de Dragon, en la clasificación de los conjuntos de entrenamiento y prueba para cada una de las particiones consideradas.

DISCUSIÓN

En las Figs. 3a y 3b puede observarse que la divergencia entre los %s de buenas clasificaciones en los conjuntos de entrenamiento y de prueba se acentúan en los casos de particiones 120/40 y 40/120. En el caso de un modelo de aplicación totalmente general (modelo ideal en el que la capacidad de describir/caracterizar la propiedad química para la muestra –el conjunto de entrenamiento- es idéntica a la capacidad de descripción de la población) debería observarse que la curva del gráfico correspondiente al conjunto de entrenamiento se superpone perfectamente con la curva correspondiente al conjunto de prueba. Consistentemente, ambas curvas, para todas las particiones, comienzan convergiendo (cuando el número de predictores es pequeño no existe sobreajuste, y no hay diferencia entre cada par de curvas) pero divergen conforme se agregan más descriptores al modelo (el punto en el que comienzan a divergir corresponde al posible –pero incierto- comienzo del sobreajuste). Sin embargo, la diferencia entre ambas curvas es menos acentuada para el caso de la partición 80/80: existe sobreajuste, pero no tan notorio como en las dos particiones restantes. ¿Significa esto que es más probable incorporar particularidades de la muestra que no se encuentran en la población al modelo cuando la relación (número de casos en el conjunto de entrenamiento/número de casos en el conjunto de prueba) es alta? No necesariamente; simplemente, cuando el número de compuestos del conjunto de entrenamiento es mucho mayor que el del conjunto de prueba, es

más probable encontrar elementos en el conjunto de entrenamiento ausentes en el conjunto de prueba pero que sin embargo podrían estar presentes en la población general: no hay garantías de que esos elementos presentes en el conjunto de entrenamiento y ausentes en el de prueba se encuentren o no se encuentren en la población general, y por lo tanto los resultados de nuestro análisis indican que la validación externa, cuando el conjunto de prueba es muy pequeño en relación al de entrenamiento, es una herramienta bastante pobre para estimar la capacidad predictiva de un modelo. Del mismo modo en que se ha demostrado que la validación interna cruzada LGO tiende a sobreestimar la capacidad predictiva [7,8], la validación externa bien podría estar subestimándola.

Las notables divergencias de la partición 40/120 pueden explicarse por el número pequeño de compuestos en el conjunto de entrenamiento, que no alcanzan para generar un modelo de aplicación general.

El % de buenas clasificaciones de los modelos (entre un 50 y un 70% dependiendo del número de descriptores incorporados) no es bajo si se considera que en el análisis discriminante se consideró que la probabilidad a priori de pertenecer a una categoría era la misma para las cuatro categorías (25%). Si bien esto no es cierto para el conjunto de entrenamiento (ya que sabemos de antemano que el número de compuestos en cada categoría no es el mismo) preferimos utilizar esta hipótesis para generar el modelo pensando en la aplicación del mismo a la población general (nada nos indica que en la población general hay más compuestos químicos con alta permeabilidad intestinal que con baja permeabilidad). Las diferencias en la distribución de %PI en el conjunto de 160 casos considerado responden únicamente a que hay pocos datos confiables en literatura de fármacos de permeabilidad intestinal intermedia, abundando en cambio los datos de fármacos de alta y baja permeabilidad (categorías 1 y 4). Utilizar distintas probabilidades a priori de pertenencia a una categoría determinada en base a la distribución de %PI en el conjunto de casos nos hubiera permitido lograr mayores porcentajes de buenas clasificaciones, pero sacrificando capacidad predictiva sobre compuestos ajenos al conjunto de entrenamiento. Hemos descartado esta opción para favorecer la aplicabilidad general de los modelos obtenidos. El porcentaje de referencia correspondiente a una clasificación al azar con el que deberíamos comparar el desempeño del modelo es, en caso de una distribución uniforme de los compuestos en todas las categorías, del 25%.

CONCLUSIÓNES

Los resultados obtenidos sugieren dos conclusiones de considerable importancia en el campo de los estudios QSPR/QSAR.

Por un lado, cuando el número de compuestos utilizado en el conjunto de entrenamiento es mucho mayor que el número de compuestos en el conjunto de prueba, la validación externa parece subestimar la capacidad predictiva real del modelo. Probablemente, el desempeño real del modelo se encuentre en tal caso en algún punto intermedio entre los resultados de la validación interna cruzada y la validación externa. Cuando el número de compuestos en los conjuntos de entrenamiento y de prueba es el mismo, el desempeño del modelo sobre el conjunto de prueba es más estable independientemente del número de predictores (hay que incorporar mayor cantidad de descriptores al modelo para que se observe sobreajuste en la validación externa). Esto implica que la práctica general de reservar una *hold out* sample del 10 al 20% del número total de casos disponibles nos lleva a subestimar la capacidad predictiva del modelo obtenido: puede o no haber sobreajuste, pero en el caso de no haberlo la validación externa podría conducirnos a descartar erróneamente modelos válidos. Una partición más equilibrada del número de casos entre el conjunto de entrenamiento y el conjunto de pruebas puede llevarnos a un modelo de menor desempeño; sin embargo, en ese caso, hay mayor probabilidad de que la capacidad predictiva estimada mediante validación externa se aproxime a la real, esto es, el modelador tiene mayor certidumbre respecto a la capacidad predictiva del modelo.

En segundo lugar, los gráficos de divergencia entre el desempeño del modelo sobre el conjunto de entrenamiento y el conjunto de prueba pueden utilizarse como herramienta para establecer a partir de qué número de descriptores hay probabilidad cierta de sobreajuste, y seleccionar el número de descriptores óptimo de un modelo.

Conviene ampliar este trabajo considerando otras propiedades moleculares, otros conjuntos de entrenamiento y otras particiones del conjunto de casos a fin de verificar la validez general de los resultados obtenidos.

AGRADECIMIENTOS: A. Talevi y E.A. Castro agradecen al CONICET. L. Bruno-Blanch agradece a la Universidad Nacional de La Plata (Incentivos UNLP) y a ANPCyTn (PICT 11985).

BIBLIOGRAFÍA

1. Dudek AZ, Arodz T, Gálvez J. Comput. Chem. High Throughput Screen (2006), 9, 213-28. **2.** Hansch C, Muir RM, Fujita T, Maloney P, Geiger E. J Am Chem Soc (1963), 85, 2817-24. **3.** Leo A, Panthanickal A, Hansch C, Theiss J, Shimkin M, Andrews AW. J Med Chem (1981), 24, 859-64. **4.** Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. ATLA (2005), 33, 445-59. **5.** Hawkins DM. J Chem Inf Comput Sci (2001), 41, 1218-27. **6.** Yasri A, Hartsough D. J. Chem. Inf. Comput. Sci. (2001), 41, 1218-27. **7.** Golbraikh A, Tropsha A. J. Mol. Graph Model. (2002), 20, 269-76. **8.** Schüürman G, Ebert R, Chen J, Wang B, Kühne R. J. Chem. Inf. Model. (2008), 48, 2140-5