ISSN 0121–1129

Análisis del proceso de minería de datos sobre la base de datos Bioinformática de segmentos de la proteína p53, asociada a la actividad cancerígena*

Data Mining Process Analysis, on a Bioinformatics Database about p53 Protein's Segments, Related to Carcinogenic Activity

Fecha de recepción: 3 de septiembre de 2010
Fecha de aprobación: 27 de noviembre de 2010

Alejandro Hadad**,
Franco Simonetti***

Resumen

Se estudió la utilización de estrategias para afrontar el problema del desbalanceo y la alta dimensionalidad de los registros que habitualmente forman parte de las bases de datos en el área bioinformática. Se tomó como caso de estudio la base de datos de segmentos de la proteína p53; sobre dicha base se construyen modelos con el fin de identificar si corresponden a patrones activos o inactivos. El problema del desbalanceo se abordó a través de una red neuronal no supervisada, y el de la selección de variables para reducir la alta dimensionalidad, a partir de una combinación de métodos con diferentes enfoques. Experimentos preliminares del modelo propuesto en datos estándar muestran resultados promisorios.

Palabras clave: Desbalanceo, Selección de variables, Bioinformática.

Abstract

It studies the strategies' used to address the problem of records imbalance and high dimensionality, which in the bioinformatics field's databases, are a known characteristic. The study case is a p53 protein segments' database. On this basis some models are built to identify whether they correspond to active or inactive patterns. The imbalance problem is addressed through some unsupervised neural network and the variables selection to reduce the high dimensionality by combining methods with different approaches. Preliminary experiments using the proposed model over standard data, shows promising results.

Key words: Imbalances, Variables Selection, Bioinformatics.

* Este trabajo forma parte de las actividades de formación del Grupo de Estudio y Análisis de Base de Datos en Bioinformática, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina.

** Facultad de Ingeniería, Univ. Nacional de Entre Ríos. Oro Verde, Entre Ríos, Argentina. hadad@santafe-conicet.gov.ar
*** Facultad de Ingeniería, Univ. Nacional de Entre Ríos. Oro Verde, Entre Ríos, Argentina.

I. Introducción

A. Descripción del problema

Actualmente, si se quiere abordar una tarea de análisis sobre las bases de datos en bioinformática, mediante un proceso de Minería de Datos, estas presentan varias dificultades, referidas fundamentalmente al fuerte desbalanceo en el número de registros asociados a una clase o comportamiento o patrón respecto de los demás patrones, y a la alta dimensionalidad de dichos registros. El problema del desbalanceo de datos es relativamente nuevo en la literatura de aprendizaje automático y minería de datos; sin embargo, es un tema de creciente interés en dicha comunidad, debido a sus efectos sobre los resultados obtenidos y al número de aplicaciones en donde se puede encontrar esta situación. Un conjunto de datos desbalanceados se puede definir como aquellos que presentan una desproporción notable en el número de instancias pertenecientes a cada clase; ello provoca un sesgo en el desempeño de los clasificadores estándares hacia el reconocimiento de las clases más numerosas, en detrimento de las más raras [1].

Entre las aplicaciones donde se puede observar prevalencia de datos desbalanceados se pueden citar, entre otras: detección de fraude e intrusión, manejo de riesgo, clasificación de texto, detección de fallas en procesos industriales y diagnóstico y monitoreo médico [2]. Para hacer las cosas más difíciles, en muchas de estas aplicaciones las clases más raras son justamente las que interesa especialmente reconocer. En la literatura se pueden encontrar varios métodos para tratar el problema de aprendizaje automático de clasificadores utilizando datos desbalanceados, sin embargo, este problema permanece abierto. Entre las estrategias propuestas se pueden distinguir dos enfoques: en el primero se opta por la asignación de un costo diferencial a las instancias de entrenamiento según las frecuencias de clases, mientras que en el segundo se remuestrea el conjunto de datos originales, ya sea agregando casos sintéticos o repetidos de la clase minoritaria o submuestreando las clases mayoritarias [3].

Por otro lado, el problema de la alta dimensionalidad es una problemática abordada desde hace tiempo a través de métodos paramétricos y no paramétricos, basados en enfoques estadísticos y de teoría de la información, generalmente con el objetivo de seleccionar las variables más relevantes. Este problema se ve agravado porque la complejidad de los procedimientos de aprendizaje y de los clasificadores en sí mismos depende de dos factores primarios: el volumen de muestras de aprendizaje y la dimensionalidad del espacio de representación.

En este trabajo se emplea un enfoque de remuestreo, y se propone un modelo para la selección de ejemplos de las clases mayoritarias, basado en el agrupamiento no supervisado de dichas clases. Se aplica el método propuesto a conjuntos de datos desbalanceados y luego se evalúa la selección de variables a través de 5 criterios de manera combinada.

B. Caso de estudio

La meta del proceso es modelar la actividad transcriptacional de la proteína p53 (activa vs. inactiva) basada en datos extraídos de simulaciones biofísicas. Los modelos biofísicos de las proteínas p53 mutantes poseen características de rendimiento que se pueden utilizar para predecir la actividad transcripcional de p53. Todas las etiquetas de clase están determinadas a través de estudios in vivo.

El cáncer es causado por la acumulación de mutaciones genéticas en dos vías regulatorias críticas: el crecimiento celular y la muerte celular programada (apoptosis); defectos en alguna de estas vías regulatorias pueden resultar en una proliferación celular descontrolada. En la aparición del cáncer se han implicado mutaciones en dos amplias clases de genes: los protooncogenes y los genes supresores de tumores; los protooncogenes son activados para volverse oncogenes mediante mutaciones que los hacen excesivamente activos en la promoción del crecimiento; en cambio, los genes supresores de tumores normalmente inhiben la proliferación celular excesiva, por lo que si sufren una mutación o una deleción aumentará la probabilidad de que se produzca un tumor.

Las proteínas supresoras de tumores, como la p53, normalmente desencadenan la apoptosis de las células afectadas y destruyen el tumor. La proteína p53 ejerce su capacidad supresora de tumores principalmente como un factor de transcripción que induce la detención del ciclo celular, la apoptosis, la reparación de ADN o senescencia; es activada por una complicada serie de modificaciones postraduccionales, lo cual induce la expresión de otros genes con función supresora de tumores. Además, p53 puede translocarse a la mitocondria en respuesta al daño al ADN y causar la liberación del citocromo c, lo cual activa la vía apoptótica.

Mutaciones que interrumpen cualquiera de estos mecanismos son causantes de cáncer en humanos. La base de datos TP53, de la Agencia Internacional de Investigación del Cáncer (IARC), contiene 26.597 mutaciones de p53 encontradas en cánceres de pacientes humanos, de las cuales, el 73,6% (19.579) se deben al cambio de un único aminoácido en el dominio de unión al ADN. Una de las grandes metas de la terapia anticáncer es poder rescatar estos mutantes, estabilizando la conformación wild-type y activando, por ende, la apoptosis en las células cancerosas, reduciendo o matando el tumor. Se han identificado varias moléculas que podrían servir a modo de droga para lograr este efecto, pero su mecanismo de acción y su rango de actividad son desconocidos.

Sin embargo, algunos mutantes de p53 causantes de cáncer han sido rescatados in vitro por mutaciones intragénicas supresoras, lo cual anula los efectos funcionales de las mutaciones, produciendo un doble mutante p53, cuya actividad wild-type ha sido restaurada. Infortunadamente, la evaluación in vitro de todas las combinaciones de mutaciones posibles para determinar si devuelven o no la funcionalidad a las p53 mutantes es inviable, debido al tiempo que conlleva y a los costos. Sería muy conveniente tener un modelo computacional que lleve a cabo experimentos virtuales en los mutantes y que pueda reducir a un número razonable la lista de posibles mutaciones que rescaten la funcionalidad de la p53, para su ensayo en el laboratorio. Un clasificador que realice esta tarea necesita de un gran número de datos para su entrenamiento, así como una elección apropiada de las características más informativas para lograr una predicción más precisa.

El conjunto de datos elegido para esta tarea posee características derivadas de modelos atómicos modelados por homología y de simulaciones de dinámica molecular. Si bien la estructura completa de la p53 wild-type es desconocida, se conoce la estructura cristalográfica del dominio principal de ella, lo cual ha hecho posible la construcción de los modelos por homología. Las características extraídas de los modelos atómicos corresponden a [11]:

- **Secuencia genómica (1D).** Contiene información acerca de la ubicación de la mutación y del residuo cambiado. Se extrajeron 247 características de este tipo por mutante.

- **Mapas de superficie (2D).** Los mapas de superficie bidimensionales contienen propiedades de la superficie de la proteína, tales como potencial electrostático, capacidad de aceptar/donar H y otras propiedades estéricas.

- **Mapas de distancia (3D).** Los mapas de distancia 3D fueron construidos a partir de los desplazamientos espaciales de los residuos de p53 mutante relativos a la p53 wild-type; esto refleja los cambios estructurales inducidos por la mutación.

- **Simulación (4D).** Las características 4D se obtuvieron al analizar el cambio de la estructura tridimensional a lo largo de un periodo simulado, donde se hace variar la temperatura, obteniendo así datos sobre la termoestabilidad de la proteína.

Para el trabajo se dispusieron de 16.772 registros de segmentos ejemplos (activos e inactivos), de los cuales el 0,85% corresponde a segmentos activos, y el resto a inactivos; por otro lado, cada segmento consta de 5408 variables o características. Estas cifras permiten cuantificar la magnitud del desbalanceo y de la alta dimensionalidad para este problema.

II. Metodología propuesta

A. Desbalanceo

Al evaluar las opciones existentes para resolver el problema descrito, se descartó en primer lugar la alternativa de llevar el número de elementos de cada clase al de la clase más numerosa, clonando ejemplos de las clases más raras, por dos razones: una de orden teórico y otra de orden práctico. La primera razón está vinculada con el problema de sobreajuste; si se duplican ejemplos de clases minoritarias, dichos ejemplos tendrán mayor frecuencia que los de la clase mayoritaria, aumentando la probabilidad de aprender detalles de estas instancias que perjudiquen la generalización; por otro lado, se podrían generar versiones ruidosas de los ejemplos originales, y así evitar el perjuicio antes mencionado; sin embargo, para hacerlo se debería contar con algún modelo que permita controlar las características de las modificaciones impuestas a los ejemplos originales para garantizar que los nuevos casos sigan perteneciendo a la clase original. Finalmente, la razón de origen práctico tiene que ver con el número de instancias de las clases más numerosas, que suele ser muy elevado; si se llevaran todos los conjuntos de instancias por clase a dicha cantidad, el número final de ejemplos haría mucho más lento el proceso de entrenamiento, empleando los algoritmos de clasificación estándar.

El modelo de remuestreo de clases propuesto en este trabajo consiste en elegir, de los conjuntos de datos cuyas clases son más numerosas, un subconjunto significativo de ejemplares, de tal manera que estos permitan representar las diferentes variantes de los elementos de dichas clases. Para ello, una vez determinado el número N de ejemplos por utilizar en el problema, se realiza un tratamiento diferencial de los conjuntos de instancias por clase, ya sea que tengan más o menos ejemplares que N. Para las clases con menos de N individuos, se clonan las instancias originales considerando en dicho proceso como heurístico minimizar el número de casos repetidos en el conjunto final. Si bien al operar de esta manera se generarán casos duplicados, el número por lo general será menor al que se tendría si el valor N fuera igual al número de instancias de la clase mayoritaria.

Para las clases cuya cantidad de ejemplares es superior a N se entrena un mapa autoorganizado específico para cada clase solo con instancias de dicha clase, con el objetivo de encontrar subgrupos de ejemplos dentro de cada clase y emplear representantes de cada uno de ellos en el conjunto final. Una vez que se tienen los mapas entrenados con todos los datos disponibles para cada clase mayoritaria, para cada ejemplo de entrada se identifica la neurona ganadora, obteniéndose un conjunto de ejemplos asociados a cada neurona. Con esta información se genera un ranking de neuronas de acuerdo con el número de ejemplares que tengan asociados. Para la selección de las instancias de las clases mayoritarias se recorren las neuronas del mapa tantas veces como fuera necesario según el N elegido, y en orden descendente de acuerdo con dicho ranking. Cada vez que se considera una neurona se elige al azar un ejemplo que tenga asociado; el ejemplo seleccionado se agrega al conjunto definitivo de instancias para dicha clase. Al recorrer de la forma mencionada la lista de neuronas del mapa se da prioridad en la selección a las que presentan mayor cantidad de ejemplos asociados, asegurando que si el N es pequeño, se logre representar la mayor cantidad de los casos originales. Con el procedimiento descrito se logra obtener un conjunto de patrones con igual frecuencia de casos por clase. En este caso, como punto de partida para los métodos de selección se submuestreó la clase mayoritaria, a fin de obtener las misma cantidad de ejemplos que la clase minoritaria.

B. Selección

Para el proceso de selección de características, a fin de contemplar la mayor parte de los enfoques, se utilizó la asistencia de diversos métodos que utilizan criterios independientes para la clasificación binaria [4]-[7]:

Prueba T: Se hace una prueba T de Student, donde se asume una distribución normal para cada característica.

Entropía: Se calcula la entropía relativa, también conocida como la divergencia de Kullback-Leibler.

Método de Bhattacharyya: Mide la similitud entre dos distribuciones de probabilidad, junto con el coeficiente de Bhattachary; ya que es una medida del grado de superposición entre las dos poblaciones, se utiliza para medir la separabilidad de clases.

Curva ROC: Se calcula el área bajo la curva ROC; permite decidir cuáles de un conjunto de instancias están en un grupo o en otro.

Test de Wilcoxon: Es una prueba no paramétrica para comparar la mediana de dos muestras relacionadas; en este caso se utiliza para verificar la separabilidad de clases según el parámetro elegido.

Luego de aplicados cada uno de los métodos se realizó un proceso de búsqueda exhaustiva para identificar las características presentes entre las posiciones más relevantes del ordenamiento generado por cada método. Se conformaron conjuntos de variables seleccionadas por 2, 3, 4 y los 5 métodos, a fin de comparar qué estrategia puede ser más conveniente para este conjunto de datos.

Para evaluar la calidad de los conjuntos de patrones seleccionados con la metodología propuesta se utilizó un clasificador basado en máquinas de soporte vectorial (SVM: Support Vector Machines) con núcleo Gaussiano [8]. Se evaluaron diferentes configuraciones del clasificador, analizando valores de costo (c) y gamma (g), con el objetivo de encontrar los parámetros que mostraran el mejor desempeño para los distintos conjuntos de entrenamiento. Los valores empleados en dicha evaluación fueron 0.001, 0.01, 0.1, 1, 10, 100 y 1000 para ambos parámetros. Los entrenamientos se realizaron incluyendo una evaluación mediante validación cruzada de 2 conjuntos, a fin de poder comparar con trabajos de referencia [9]-[11] que particionan los datos de entrenamiento y testeo de esa manera.

En dichos trabajos de referencia, para realizar un submuestreo se determina el mutante (cada registro) más informativo, estimando su impacto en la precisión del clasificador; en este enfoque se supone que el mutante es activo, se construye el clasificador y se determina su precisión; luego se supone al mutante inactivo y se repite. El máximo incremento en el coeficiente de correlación de validación cruzada para el mutante es llamado índice "curiosidad", y se calcula utilizando las estadísticas sobre verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos [11]. Para la selección de variables en [11], primero estudian el efecto de considerar las diferentes variables biofísicas 1D, 2D, 3D y 4D de los modelos por separado; luego de generadas estas características, utilizan un algoritmo basado en información mutua [12] para elegir las características más informativas; finalmente, para la tarea de clasificación, el autor del trabajo de referencia utilizó una SVM para las variables 1-4D; luego, un clasificador compuesto de Naive Bayes reúne las estadísticas para cada uno de los componentes clasificados anteriormente para determinar la probabilidad para cada clasificador de que haya predicho correctamente un mutante. En todos los casos obtienen medidas de performance en torno al 70%.

En el método propuesto se utiliza una red neuronal no supervisada, aprovechando su heurística de aprendizaje sin dejar de lado ningún registro; luego, en el proceso de selección de variables se utilizan criterios de diferente naturaleza para obtener las características más relevantes, sin considerar el origen biofísico de estas (1D, 2D, 3D y 4D). Dichos criterios se evalúan de manera combinada a fin de explorar sus mejores combinaciones. Utilizamos el algoritmo SVM directamente sobre las variables seleccionadas, en lugar de un clasificador de Naive Bayes sobre las salidas de los SVM de cada tipo de variable (1-4D).

III. Resultados

A continuación se presentan las tablas resultantes de los ensayos realizados para las diferentes combinaciones de características, con la precisión de clasificación para cada combinación de modelo.

Tabla 1

Modelos de SVM con las variables seleccionadas por los 5 criterios (11 características)

c/g	0.001	0.01	0.1	1	10	1000	1000
0.001	53,497	53,497	59,441	58,042	51,049	49,301	49,65
0.01	53,497	53,497	59,441	58,042	51,049	49,301	49,65
0.1	53,497	56,993	67,483	61,888	51,049	49,301	49,65
1	56,993	69,58	72,028	74,825	58,392	51,049	49,65
10	68,182	70,28	72,727	73,776	61,538	52,448	49,301
100	69,93	72,028	75,175	67,133	64,685	52,448	49,301
1000	72,378	74,476	71,329	62,587	64,685	52,448	49,301

Tabla 2

Modelos de SVM con las variables seleccionadas por 4 criterios (72 características)

c/g	0.001	0.01	0.1	1	10	1000	1000
0.001	67,832	66,084	53,497	51,399	51,049	56,294	49,65
0.01	67,832	66,084	53,497	51,399	51,049	56,294	49,65
0.1	86,014	77,622	53,497	51,399	51,049	56,294	49,65
1	91,259	84,965	77,972	57,343	51,399	56,294	49,65
10	94,755	85,664	76,224	57,343	51,399	56,294	49,65
100	93,357	82,867	75,874	57,343	51,399	56,294	49,65
1000	90,559	82,867	75,874	57,343	51,399	56,294	49,65

Tabla 3

Modelos de SVM con las variables seleccionadas por 3 criterios (353 características)

c/g	0.001	0.01	0.1	1	10	1000	1000
c/g	66,084	62,587	54,545	51,399	61,538	50,699	50
0.001	66,084	62,587	54,545	51,399	61,538	50,699	50
0.01	80,769	70,28	54,545	51,399	61,538	50,699	50
0.1	90,559	80,07	68,881	51,049	50,699	50,699	50
1	91,958	81,119	68,182	52,448	50,35	50,699	50
10	90,909	79,021	68,182	52,448	50,35	50,699	50
100	87,762	79,021	68,182	52,448	50,35	50,699	50

Tabla 4

Modelos de SVM con las variables seleccionadas por 2 criterios (885 características)

c/g	0.001	0.01	0.1	1	10	1000	1000
0.001	59,091	53,846	50,35	62,587	51,399	50	50
0.01	59,091	53,846	50,35	62,587	51,399	50	50
0.1	70,629	53,846	50,35	50,35	51,399	50	50
1	82,168	66,084	51,049	50,35	51,399	50	50
10	83,916	66,783	51,399	50,35	51,399	50	50
100	83,566	66,783	51,399	50,35	51,399	50	50
1000	83,566	66,783	51,399	50,35	51,399	50	50

Tabla 5

Modelos de SVM con todas las variables presentes

c/g	0.001	0.01	0.1	1	10	1000	1000
0.001	51,049	50,699	56,643	50	50	50	50
0.01	51,049	50,699	56,643	50	50	50	50
0.1	51,049	50,699	56,643	50	50	50	50
1	61,538	50,699	56,643	50	50	50	50
10	61,888	50,699	56,643	50	50	50	50
100	61,888	50,699	56,643	50	50	50	50
1000	61,888	50,699	56,643	50	50	50	50

Observando las cinco tablas se aprecia que las mejores performance del clasificador se dan para la combinación de parámetros c=10 y g=0.001. En función de estos resultados, en la Fig. 1 se observa el mejor rendimiento del clasificador en función del número de criterios considerados.

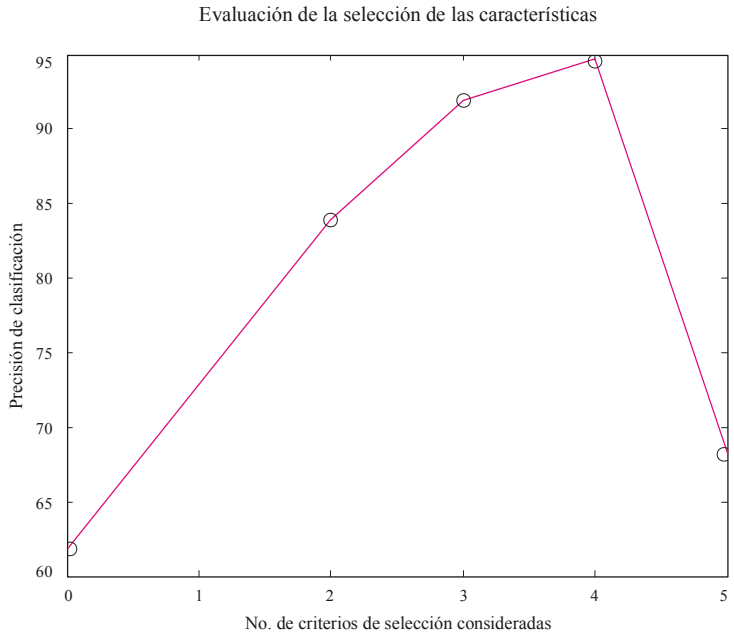

Fig. 1.

En este caso se observa la mejor performance al considerar 4 criterios de selección de variables y un incremento en la performance del clasificador, salvo para el caso de considerar los 5 criterios.

IV. Conclusiones

Se abordaron dos problemáticas presentes en muchos casos de análisis de datos, particularmente exigente en el área de bioinformática. Se construyó un modelo heurístico basado en una red neuronal autoorganizada, el cual permitió realizar un submuestreo aprovechando las características de representación de dichas redes, generando un subconjunto de datos más pequeño sobre el cual realizar el análisis.

Considerar varios métodos de selección de variables y procesar sus resultados de manera combinada nos permitió cuantificar, para este problema, hasta qué punto se puede evaluar el contenido de información de una variable a fin de seleccionarla. La baja performance del conjunto de variables que respondió a los 5 métodos se debe probablemente a lo exigente de esta condición, que fue cumplida por solo 11 variables. Igualmente, es interesante observar la fuerte reducción de 5408 a 72 variables para el reconocimiento de segmentos activos e inactivos.
También se observa que para las combinaciones de 2, 3 y 4 criterios de selección de variables se obtienen performances superiores al 70%, indicado en los trabajos de referencia, lo que a priori muestra la ventaja que presenta la utilización de criterios combinados para este problema. En este sentido, se evaluará como trabajo futuro esta condición, realizando variaciones controladas sobre el grado de desbalanceo a fin de conocer si la estructura del problema permite seguir realizando una identificación aceptable con las combinaciones de variables obtenidas.

Referencias

[1] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Handling imbalanced datasets: A review". GESTS International Transactions on Computer Science and Engineering, Vol. 30, No. 1, pp. 25-36. 2006.

[2] N. Chawla, N. Japkowicz, A. Kolcz, *Editorial*: *Special Issue on Learning from Imbalanced Data Sets*. Sigkdd Explorations, Vol. 6, No. 1, pp. 1-6. 2004.

[3] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis and P. Pintelas, "A Wrapper for Reweighting Training Instances for Handling Imbalanced Data Sets". In *IFIP International Federation for Information Processing*, Vol. 247, Artificial Intelligence and Innovations 2007: From Theory to Applications, eds. Boukis, C, Pnevmatikakis, L., Polymenakos, L., (Boston: Springer), pp. 29-36. 2007.

[4] AP. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms". *Pattern Recognition*, No. 30, pp. 1145-1159. 1997.

[5] D. J. Hand & R. J. Till, "A simple generalization of the area under the ROC curve to multiple class classification problems". *Machine Learning*, No. 45, pp. 171-186. 2001.

[6] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions". *Bulletin of the Calcutta Mathematical Society,* No. 35, pp. 99-109.

[7] F. Wilcoxon, "Individual Comparisons by Ranking Methods." *Biometrics* No. 1, pp. 80-83. 1945.

[8] Cortes and Vapnik. *Support Vector Networks*. *Machine Learning*. Springer Netherlands. ISSN 0885-6125 (Print) 1573-0565 (Online), Vol. 20, No. 3, pp. 273-297. DOI 10.1007/BF00994018. 1995.

[9] S. A. Danziger, R. Baronio, L. Ho, L. Hall, K. Salmon, G.W. Hatfield, P. Kaiser and R. H. Lathrop, "Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning, PLOS". *Computational Biology*, Vol. 5, No. 9, e1000498. 2009.

[10] S. A. Danziger, J. Zeng, Y. Wang, R. K. Brachmann and R. H. Lathrop, "Choosing Where to Look Next in a Mutation Sequence Space: Active Learning of Informative p53 Cancer Rescue Mutants", *Bioinformatics*, Vol. 23, No. 13, pp. 104-114. 2007.

[11] S. A. Danziger, S. J. Swamidass, J. Zeng, L. R. Dearth, Q. Lu, J. H. Chen, J. Cheng, V.P. Hoang, H. Saigo, R. Luo, P. Baldi, R. K. Brachmann and R. H. Lathrop, "Functional Census of Mutation Sequence Spaces: The Example of p53 Cancer Rescue Mutants", *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, No. 3, pp. 114-125.

[12] E. G. Miller, *Mutual Information From MathWorld—A Wolfram Web Resource*, created by Eric W. Weisstein. Disponible en: http://mathworld.wolfram.com/MutualInformation.html