Identificación de sitios en proteínas usando máquinas con vectores de soporte

Jaime Leonardo Bobadilla [1], Tobías Mojica Ph.D [2] y Luis Fernando Niño Ph. D.[3*]

[1 - 3] Departamento de Ingeniería de Sistemas, Universidad Nacional de Colombia. [2]Instituto de Genética, Universidad Nacional de Colombia, Bogotá.

Recibido: 01-07-03; Aceptado: 12-10-03

RESUMEN

Ante el incremento creciente de estructuras tridimensionales (3D) de proteínas determinadas por rayos X y tecnologías de NMR, así como de estructuras obtenidas mediante métodos computacionales, resulta necesaria la utilización de métodos automatizados para obtener anotaciones iniciales. Hemos desarrollado un nuevo método para reconocer sitios en estructuras tridimensionales de proteínas. Este método está basado en un algoritmo previamente informado para crear descripciones de microambientes proteicos, utilizando propiedades físicas y químicas muy específicas. El método de reconocimiento tiene 3 entradas: 1. Un juego de sitios que comparten alguna función estructural o funcional; 2. Un juego de sitios que no comparten funciones estructurales o funcionales; 3. Un sólo sitio para análisis. Una máquina clasificadora con vector de soporte utiliza detalles del vector, donde cada componente representa una propiedad en volumen dado. La validación contra tests independientes muestra que esta prueba de reconocimiento tiene una alta sensibilidad y especificidad. También describimos los resultados de examinar 4 proteínas de unión a calcio (y con el calcio removido) utilizando una rejilla tridimensional de puntos de prueba en un espacio de 1.25Ao. Nuestros resultados muestran que descripciones basadas en propiedades con máquinas de soporte de vectores pueden ser utilizadas para el reconocimiento de sitios de proteínas en estructuras no anotadas.

Palabras claves: Aprendizaje de máquinas, anotaciones, estructura de proteínas, sitios, algoritmos, no-sitios.

ABSTRACT

Sites Identification in Proteins, using machines with support vectors

The increasing amount of protein three-dimensional (3D) structures determined by x-ray and NMR technologies as well as structures predicted by computational methods results in the need for automated methods to provide initial annotations. We have developed a new method for recognizing sites in three-dimensional protein structures. Our method is based on a previously reported algorithm for creating descriptions of protein microenvironments using physical and chemical properties at multiple levels of detail. The recognition method takes three inputs: 1. a set of sites that share some structural or functional role, 2.a set of control non-sites that lack this role, and 3. a single query site. A support vector machine classifier is built using feature vectors where each component represents a property in a given volume. Validation against an independent test shows that this recognition approach has high sensitivity and specificity. We also describe the results of scanning four calcium binding proteins (with the calcium removed) using a three dimensional grid of probe points at 1.25Å spacing. Our results show that property based descriptions along with support vector machines can be used for recognizing protein sites in un-annotated structures.

Keywords: Machine learning, annoted, protein structure, sites, algorithm, non-sites.

* Correspondencia: nova@unicolmayor.edu.co

Introducción

Es indudable la gran expectativa y el seguimiento que han tenido los proyectos de secuenciamiento de genomas, proyectos de genómica estructural y aplicaciones en biotecnología, entre otros, los cuales tienen y tendrán una repercusión a nivel no solamente científico sino también económico, industrial y social.

Entramos en una era donde la secuencia del genoma humano y genomas de docenas de organismos han sido completadas. La comunidad biomédica y biológica está poniendo su atención en la proteómica, es decir, en el estudio de las proteínas productos de los genes secuenciados en un genoma. Es evidente que muchos en la comunidad científica están pidiendo un proyecto del proteoma humano, análogo al proyecto del genoma humano (1), pero que es, sin lugar a dudas, mucho más complejo (2).

El número de estructuras establecidas (3) y predichas por computador (4) se incrementa rápidamente; existe entonces una necesidad significativa de métodos automáticos que puedan colaborar en anotaciones bioquímicas y que puedan ser usados para realizar predicciones a escala genómica de sitios importantes en proteínas.

Bioinformática

La Bioinformática es un campo interdisciplinario que se encuentra en la intersección entre las Ciencias de la Vida y de la Información; proporciona las herramientas y recursos necesarios para favorecer la investigación en muchas áreas de la biología. En algunos problemas, la bioinformática es la única herramienta de estudio y, por lo tanto, es crucial para el avance y la investigación; de hecho, es una de las herramientas fundamentales para el secuenciamiento de los genomas. El genoma humano secuenciado por Celera es un genoma bioinformático.

Este campo interdisciplinario, que ya se considera no solamente un subconjunto de las ciencias de la computación o de la biología sino que es una disciplina independiente, comprende la investigación y desarrollo de herramientas útiles para llegar a entender el flujo de información desde los genes a las estructuras moleculares, a su función bioquímica, a su conducta biológica y, finalmente, a su influencia en las enfermedades.

El flujo de información que estudia la bioinformática es el flujo de información del DNA a la función biológica (5), el cual se refiere al dogma central de la biología: el DNA se transcribe en RNA, el RNA se traduce a proteína y las proteínas tienen funciones que realizan los procesos biológicos celulares. Los enfoques de informática para estudiar este flujo incluyen métodos para búsqueda de genes (6, 7), predicción de estructura tridimensional (8-12) y modelado de redes metabólicas.

La fisiología y el comportamiento de un organismo está dictado por sus genes, los cuales en un nivel muy básico pueden ser vistos como sitios de almacenamiento digital de información; es así como la biología es una ciencia de procesamiento de información (13).

Estructura de proteínas

La síntesis de todas las proteínas celulares está codificada por los genomas. Actualmente tenemos disponibles las secuencias de más de 150 genomas celulares incluyendo varios de organismos multicelulares. Se vislumbra una nueva etapa en la investigación biológica, que emana naturalmente de la genómica o el estudio de los genomas y que incluye la caracterización de la expresión de las proteínas codificadas por un genoma y el establecimiento de sus propiedades funcionales y estructurales.

Los genomas definen el contenido de información de los organismos y, por lo tanto, definen la tipología del organismo. La secuencia del genoma no dice cómo funciona un organismo; para tener respuestas a la pregunta de cómo funciona el organismo es necesario estudiar las proteínas. Una de las mayores áreas de investigación biológica en la actualidad, es cómo proteínas construidas de solo 20 aminoácidos diferentes, realizan gran variedad de tareas.

Las proteínas son cadenas de monómeros de aminoácidos sin ramificaciones. La forma tridimensional particular de las proteínas: 1) se origina en la secuencia de aminoácidos, 2) ocurre post-traduccionalmente, 3) Está dada por interacciones no covalentes entre regiones de la secuencia de aminoácidos (14). Solamente cuando la proteína está en su estructura tridimensional correcta es capaz de funcionar eficientemente. Un paso importante para la anotación de las proteínas es recono-

cer sitios funcionales en regiones locales tridimensionales con roles funcionales especiales y ciertas características conservadas como sitios activos, sitios de unión y sitios de soporte de la estructura (15).

Los sitios funcionales dan información valiosa acerca de la función de la proteína. Algunos sitios funcionales tienen una relación directa con la estructura primaria o secuencia de aminoácidos y pueden, por tanto, ser reconocidos usando métodos de búsqueda de motivos como Pfam (16). Sin embargo no todos los sitios funcionales tienen una relación directa con la secuencia.

Muchos sitios se componen de aminoácidos que están lejos unos de otros en la secuencia, pero cerca en el espacio tridimensional; éstos pueden no tener características definidas a nivel de secuencia pero pueden ser reconocidos utilizando un método basado en estructura. El trabajo realizado se enfoca en el reconocimiento de sitios que requieren información de la estructura terciaria. El objetivo de este trabajo es desarrollar un método que pueda reconocer diferentes sitios funcionales y que sea aplicable a sitios con o sin conservación de residuos o geometría.

Un procedimiento estadístico para caracterizar un sitio y el ambiente que lo rodea ha sido reportado anteriormente (17). Basado en éste se ha desarrollado un nuevo sistema que utiliza máquinas con vectores de soporte. En este artículo mostramos que podemos reconocer sitios de unión de calcio con sensibilidad y especificidad más alta que la informada en la literatura, usando máquinas con vectores de soporte.

Máquinas con vectores de soporte

La resolución de problemas por medio de aprendizaje de máquinas se adapta bien en áreas donde hay una gran cantidad de datos pero poca teoría, y este es el caso de la bioinformática (18). Los métodos de aprendizaje de máquina pueden ser, desde un punto de vista general, divididos en aprendizaje supervisado y no supervisado. Se dice que el aprendizaje es supervisado cuando a un algoritmo de aprendizaje se le da un conjunto de ejemplos junto a la clase a la que pertenecen y se prueba en un conjunto de datos en los que no se conocen las clases a las que pertenecen; para el caso de la identificación de sitios en proteínas, tenemos un aprendizaje supervisado, ya que contamos con una serie de sitios previamente identificados por métodos de cristalografía.

Dentro de los métodos de aprendizaje de máquina tenemos a las máquinas con vectores de soporte, las cuales son modelos de entrenamiento supervisado utilizados en problemas de clasificación binaria, fácilmente extensibles a modelos de clasificación múltiple.

Su funcionamiento se basa inicialmente en la transición del problema original a uno de mayor dimensión generalmente mediante una transformación con funciones no lineales (19). Dadas ciertas funciones especiales para la transformación que deben cumplir ciertas propiedades matemáticas (continua, integrable, acotada), se puede demostrar que en el nuevo espacio de clasificación, las categorías por discriminar tendrán mayor probabilidad de ser linealmente separables.

Con la transformación del espacio de entrada a uno de mayor dimensionalidad se convierte el problema de clasificación en la determinación del hiperplano de separación óptima, tomando como criterio de evaluación la calidad de la separación medida a partir de las distancias mínimas entre los datos clasificados y el hiperplano de separación (margen de separación); este último se considera normalizado para los datos fronterizos (vectores de soporte) donde se cumplen las siguientes proposiciones:

$w_0^T x_i + b_0 \geq 1$ para $d_i = +1$

$w_0^T x_i + b_0 \leq -1$ para $d_i = -1$

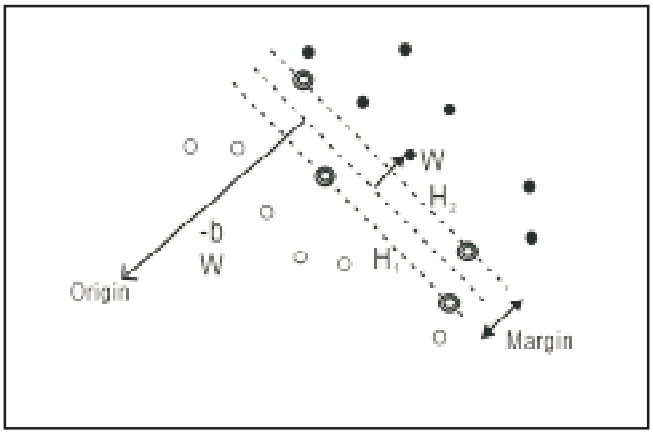

Figura 1 . Hiperplano para la separación. Los vectores de soporte están encerrados en círculo.

Correspondencia: nova@unicolmayor.edu.co

El problema de entrenamiento, dadas las poblaciones de ejemplos positivos y negativos, se puede ver como un problema de optimización restringida de una función convexa, planteando la ecuación lagrangiana donde se representa el margen de separación con base en la norma del vector de pesos, y se incluyen términos de la calidad de separación para cada vector de ejemplo (ponderados por los multiplicadores de Lagrange), teniéndose adicionalmente que satisfacer la condición de optimalidad.

$$J(\mathbf{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i\left[d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right]$$

Con base en esto se plantea el sistema dual y se soluciona por métodos de solución de optimización cuadrática, obteniendo los multiplicadores de Lagrange, de los cuales se deduce el vector de pesos del hiperplano óptimo de separación y con éste se deducen los valores de traslación del hiperplano, con el cual se pueden clasificar los nuevos casos de análisis.

Para el caso de que los datos no sean separables se puede plantear un modelo similar, que además, debe incluir variables de holgura que cuantifican los errores de clasificación para cada uno de los datos y se adicionan con coeficientes negativos (penalización) a la función objetivo del problema de programación cuadrática planteado.

Métodos

Se consideran sitios de unión de calcio las regiones esféricas de 7Å centradas sobre iones determinados por cristalografía. Los no-sitios son utilizados como controles explícitos y son regiones esféricas de 7Å en regiones esféricas en la superficie o al interior de proteínas que no se unan al calcio.

Se decide si una región (una esfera de 7Å alrededor de un sitio de prueba) es o no un sitio de calcio construyendo un clasificador con sitios conocidos de unión de calcio y no-sitios conocidos. Un diagrama esquemático del sistema se muestra en la Figura 2. El objetivo es clasificar la región y decidir, basados en la máquina con vectores de soporte, si se trata o no de un sitio de unión de calcio.

Se comenzó recolectando un conjunto de sitios de unión al calcio y de no-sitios de una versión local de la base de datos de estructura terciaria de proteínas PDB. Los sitios y no-sitios fueron divididos en volúmenes espaciales, los cuales son esferas concéntricas con un radio de 1Å. Para cada sitio y no-sitio el sistema calcula el conteo de cada propiedad en cada uno de los volúmenes espaciales. Luego toma un vector con todas las características en cada uno de los volúmenes y se entrena una máquina con vectores de soporte con ejemplos negativos y positivos. Una lista completa de las características usadas para el clasificador se provee en la Tabla 1.

CATEGORÍA	PROPIEDADES
Atómicas	Distribución de átomos de algunos elementos básicos
Grupos químicos	carboxilo, hidroxilo, amino, etc.
Residuos	de los diferentes tipos de aminoácidos
Motivos de estructura secundaria	hélice-a, hoja plegada-b, giros
Motivos de estructura súper-secundaria	bridge, bend, 3-helix
Biofísicas	movilidad, factor –B, volumen de Van der Waals, etc.

Tabla 1. Características usadas para el entrenamiento

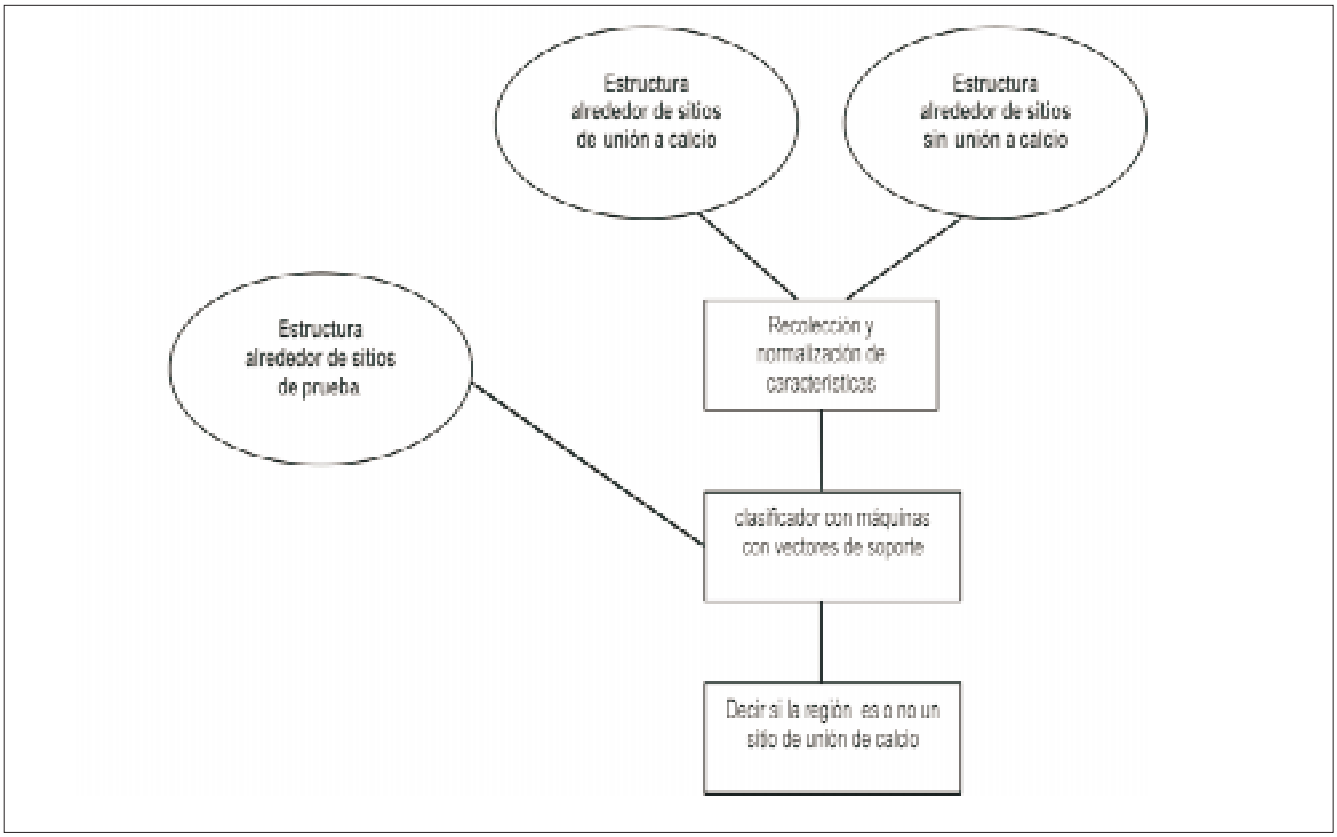

Figura 2. Esquema del sistema para la identificación de sitios.

El conteo de las propiedades de los sitios y los no-sitios puede ser usado para formar un modelo cuantitativo. Cuando se da una nueva región en una estructura dada, se divide de nuevo la estructura en conchas concéntricas y se aplica el sistema para calcular la presencia de la propiedad en los volúmenes espaciales. El clasificador construido toma el vector con las características y decide si se trata o no de un sitio de calcio.

Experimentos

Para el entrenamiento, usamos características derivadas de 68 sitios de unión al calcio y 120 no-sitios. Escogimos tres conjuntos de prueba independientes no usados previamente en el análisis con proteínas diferentes provenientes de organismos diversos. Para examinar el sistema a gran escala, en miles de regiones de prueba en una situación realista, se buscaron sitios en 4 proteínas con sitios de unión a calcio que no tenían ninguna relación y que no fueron usadas en el entrenamiento.

Para cada estructura de prueba, se definió una rejilla de 1.25Å. El clasificador con máquinas con vectores de soporte fue aplicado a cada punto de la rejilla para mirar si era o no un sitio de unión de calcio.

Puntos de la rejilla que estuvieron cerca al margen de clasificación fueron marcados como sitios potenciales de unión de calcio. Los puntos más cercanos son mostrados gráficamente en un visualizador y su ubicación fue comparada con los sitios reales de unión de calcio, como se muestra en las figuras 3 y 4.

Resultados

Para evaluar el desempeño del algoritmo de reconocimiento usamos dos medidas: sensibilidad (habilidad para reconocer un sitio de unión a calcio) y

especificidad (habilidad para reconocer un sitio que no se une al calcio) definidos de la siguiente manera:

$$\text{especificidad} = \frac{TN}{TN + FN}$$

$$\text{sensibilidad} = \frac{TP}{TP + FP}$$

Donde TP es el número de verdaderos positivos, FP es el número de falsos positivos, TN es el número de verdaderos negativos y FN es el número de falsos negativos.

La sensibilidad y especificidad en el reconocimiento de sitios de unión de calcio en el conjunto de prueba se muestra en la tabla 2. Las estructuras y los sitios de calcio potenciales encontrados por el método se encuentran en las figuras 3 y 4.

Número Test	Sitios	No sitios	Especificidad	Sensibilidad
1	47	49	100%	92%
2	33	30	100%	90%
3	193	126	100%	93%

Tabla 2. Resultado de los experimentos.

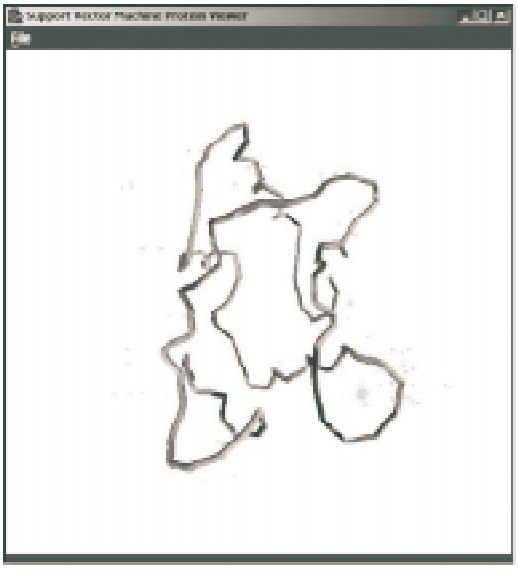

Figura 3. Búsqueda de sitios de calcio en la proteína con código PDB 3IOB

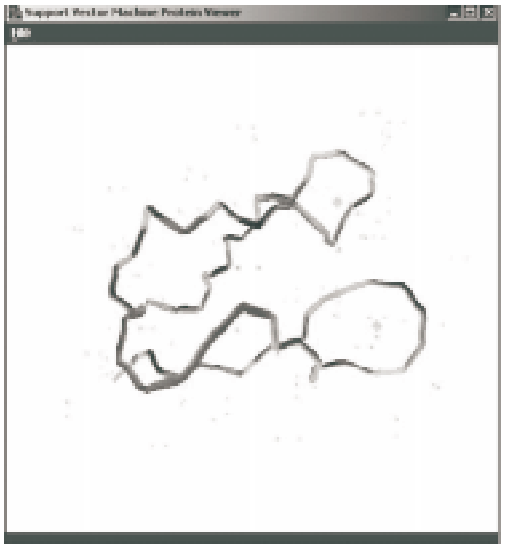

Figura 4. Búsqueda de sitios de calcio en la proteína con código PDB 3PAL.

Discusión

El método de reconocimiento presenta sensibilidad por encima del 90% y especificidad del 100%, lo cual muestra que el método es robusto y preciso. El desempeño del método de búsqueda en proteínas es promisorio.

Cada proteína requirió evaluación de mas de 15.000 puntos de prueba y condujo a un número pequeño de puntos candidatos. Para las cuatro proteínas el método reconoció los sitios de unión de calcio.

Los sitios candidatos estuvieron a una distancia de 1Å. El enfoque es, en principio, general y estamos tratando de crear clasificadores para otros sitios importantes en proteínas. También se trabaja para mejorar la eficiencia del código de búsqueda de sitios, para permitir anotación automática de estructura a escala genómica.

REFERENCIAS

1. Mojica T, Estrada L. Acerca del genoma humano. Agronomía Colombiana;27:7-12
2. Workshop Report National Research Council Steering Committee: George L. Kenyon, (Chair). Defining the Mandate of Proteomics in the Post-Genomics Era.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research 2000;28:235-42.

4. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated anealing and Bayesian scoring functions. J Mol Biol 1997;268:209-25.

5. Altman RB, Klein TE. Challenges for Biomedical Informatics and Pharmacogenomics. Annu Rev Pharmacol Toxicol 2002;42:113-33.

6. Koza JR. Evolution of a Computer Program for Classifying Protein Segments as Transmembrane Domains Using Genetic Programming. Proc of ISMB-94 1994:244-52.

7. Bryant SH, Altschul SF. Statistics of Sequence-structure Threading. Current Opinion in Structural Biology 1995;5:236-44.

8. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209-25.

9. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting Function: From Genes to Genomes and Back. J Mol Biol 1998;283:707-25.

10. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugent CW, Furey TS, et al. Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. PNAS 2000;97(1):262-7.

11. Koza JR. Evolution of a Computer Program for Classifying Protein Segments as Transmembrane Domains Using Genetic Programming. Proc of ISMB-94 1994:244-52.

12. Lathrop RH. The Protein Threading Problem with Sequence Amino Acid Interaction Preferences is NP-Complete. Protein Engineering 1994;7:9:1059-68.

13. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular Biology of the Cell. 3rd ed. New York and London: Garland Publishing; c1994.

14. Richards FM. Calculation of Molecular Volumes and Areas for Structures of Known Geometry. Methods in Enzymology 1996;115:440-64.

15. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The Pfam Protein Families Database. Nucleic Acids Research 2002;30(1):276-80.

16. Bagley SC, Altman RB. Characterizing the Microenvironment Surrounding Protein Sites. Protein Science 1995;4:622-35.

17. Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach. Cambridge, MA: MIT Press; 1998.

18. Burges CC. A Tutorial on Support Vector Machines for Pattern Recognition. In «Data Mining and Knowledge Discovery», 1998.

19. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, et al. The Protein Data Bank and the challenge of structural genomics. Nature Structural Biology 2000;7(11):957-9.