Identificación de sitios en proteínas usando métodos de aprendizaje de máquina

Leonardo Bobadilla*
Fernando Niño** Tobías Mojica***

Finding protein sites using machine learning methods

RESUMEN

Con el crecimiento de las bases de datos de estructuras tridimensionales determinadas por rayos-x, NMR (resonancia magnética nuclear) y de estructuras predichas por computador, se deriva la necesidad de sistemas automáticos que provean anotaciones iniciales. Se ha desarrollado un nuevo método para reconocer sitios en estructura terciaria de proteínas. El método propuesto se basa en un algoritmo previamente reportado para crear descripciones de microambientes en proteínas usando propiedades físicas y químicas con varios niveles de detalle. El método de reconocimiento toma tres entradas: 1. Un conjunto de sitios que comparte un rol funcional o estructural; 2. Un conjunto de no sitios que no tienen este rol; y 3. un sitio del cual se ignora si tiene la característica buscada o no. Se construyó un clasificador con máquinas con vectores de soporte usan vectores de características en que cada componente representa una propiedad en un volumen dado. La validación contra un conjunto de prueba independiente muestra que este enfoque tiene alta sensibilidad y especificidad. También se describen los resultados de escanear cuatro proteínas con sitios de unión a calcio (con el calcio removido) usando una rejilla tridimensional de puntos de prueba separada a 1.25 ángstroms. El sistema encuentra los sitios en las proteínas ubicando puntos en los sitios de unión o cerca de estos. Los resultados muestran que pueden usarse descripciones de propiedades junto con máquinas de soporte para reconocer sitios en proteínas no anotadas.

PALABRAS CLAVE

Bioinformática, dogma central de la Biología, aprendizaje de máquina, estructura terciaria de proteínas, máquinas con vectores de soporte

ABSTRACT

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 nonsites 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 set 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 1.25 angstrom spacing. The system finds the sites in the proteins giving points at or near the binding sites. Our results show that property based descriptions along with support vector machines can be used for recognizing protein sites in unannotated structures.

KEY WORDS

Bioinformatics, machine learning, support vector machines, protein tertiary structure

* Departamento de Ingeniería de Sistemas e Industrial de la Universidad Nacional de Colombia. jlbobadillam@unal.edu.co
** Ph.D. Departamento de Ingeniería de Sistemas e Industrial de la Universidad Nacional de Colombia. fninov@unal.edu.co
*** Ph.D. Instituto de Genética de la Universidad Nacional de Colombia. tobiasmojica@unal.edu.co

Introducción

Es indudable la gran expectativa y el gran seguimiento que han tenido los proyectos de secuenciamiento de genomas, proyectos de genómica estructural, aplicaciones en biotecnología, entre otros, los cuales tienen y tendrán repercusión científica, económica, industrial y social.

Entramos en una era en que se ha completado la secuencia del genoma humano y los genomas de docenas de organismos. La comunidad biomédica y biológica pone su atención en la proteómica, es decir, estudio de las proteínas productos de los genes secuenciados en un genoma. Es evidente que en la comunidad científica, muchos quieren un proyecto del proteoma humano análogo al proyecto del genoma humano[1], sin lugar a dudas mucho más complejo[2].

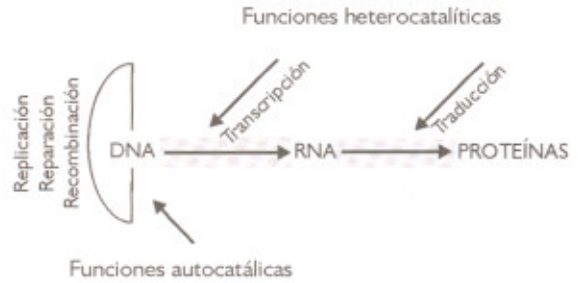
El número de estructuras establecidas [3] y predichas por computador [4] se incrementa rápidamente. Existe entonces una necesidad significativa de crear métodos automáticos que puedan colaborar en anotaciones manuales y que puedan ser usados para realizar predicciones a escala genómica de sitios importantes en proteínas.

Bioinformática

La Bioinformática, un campo interdisciplinario situado en la intersección de las ciencias de la vida y de la información, proporciona las herramientas y los 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 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 también una disciplina independiente, comprende la investigación y el 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.



DOGMA CENTRAL DE LA BIOLOGÍA

Figura 1. Dogma central de la Biología.

La fisiología y el comportamiento de un organismo está dictado por sus genes, los cuales pueden ser vistos, en un nivel básico como sitios de almacenamiento digital de información. Por esta razón 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. Hoy día tenemos disponibles las secuencias de más de 150 genomas celulares, incluidos varios organismos multicelulares. Se vislumbra una nueva etapa en la investigación biológica, que emana naturalmente de la geonómica o el estudio de los genomas, e 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 tanto, definen la tipología del organismo. La secuencia del genoma no dice cómo funciona un organismo. Para tener respuestas a la pregun-

ta de cómo funciona el organismo, es necesario estudiar las proteínas.

Para dar una idea de la complejidad del problema del estudio de las proteínas, consideremos que el genoma de levaduras tiene, calculadas por bioinformática y por otros medios, 6225 proteínas. Por anotación bioinformática se calcula que el 17% de esas proteínas está involucrado en el metabolismo general de la célula de levadura; el 30%, en organización celular y biogénesis de organelos y membranas; y el 10%, en transporte molecular[13]

Las proteínas son lo que se podría llamar los arquitectos de la vida pues son cruciales en los procesos celulares de los seres vivos. Las proteínas están implicadas en la catálisis de las reacciones químicas celulares, el transporte de moléculas, la transducción de señales, la segregación del material genético, la producción y el manejo de la energía. El programa celular vital necesita el trabajo coordinado de muchos tipos diferentes de proteínas. La mayor parte del peso seco de una célula está hecho de proteínas. Parece un hecho repetitivo, pero tendremos que entender las proteínas antes de que podamos entender la célula.

Una de las mayores áreas de investigación biológica hoy en día es saber cómo proteínas constituidas por sólo 20 aminoácidos realizan la gran variedad de tareas que cumplen. 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 postraduccionalmente, 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, su conformación es capaz de funcionar eficientemente. Un concepto clave para entender cómo funcionan las proteínas es que la función se deriva de la estructura tridimensional y la estructura tridimensional está especificada por la secuencia de aminoácidos [15].

Un paso importante para la anotación de las proteínas es reconocer 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.

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 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.

Algunos sitios se componen de aminoácidos que distan mucho entre sí en la secuencia, pero cerca en el espacio tridimensional. Éstos pueden no tener características definidas en la secuencia, pero pueden ser reconocidos utilizando un método basado en la 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 residuos o geometría conservada.

Anteriormente se ha reportado un procedimiento estadístico para caracterizar un sitio y el ambiente que lo rodea [17]. Basado en este método, 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 alta usando máquinas con vectores de soporte.

Máquinas con vectores de soporte

La resolución de problemas por medio de aprendizaje de máquina se adapta bien en áreas donde hay una gran cantidad de datos pero poca teoría; éste es 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.

Entre los métodos de aprendizaje de máquina están 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 [20].

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), en los que se cumplen las siguientes proposiciones:

$$w_0^T x_i + b_0 \geq 1 \text{ para } d_i = +1$$
$$w_0^T x_i + b_0 \leq -1 \text{ para } d_i = -1$$

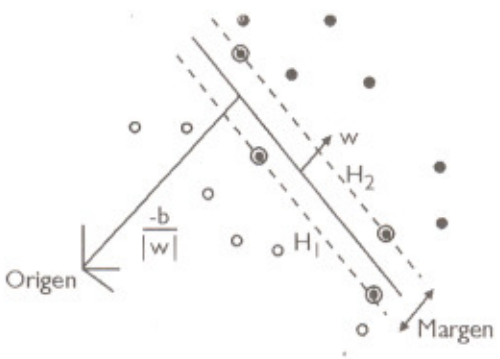


Figura 2. Hiperplano para la separación. Los vectores de soporte están encerrados en círculo.

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 del lagrangiano 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). Además se tiene que satisfacer la condición de optimalidad.

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

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

Para el caso que los datos no sean separables, se puede plantear un modelo similar que 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.

Con base en el conjunto de funciones (kernel de producto interior) que se utilice para efectuar la transformación del problema original, se pueden tener diferentes tipos de máquinas con vectores de soporte

- kernel polinomial:

$$K(x, x_i) = (1 + x^T x_i)^p$$

- kernel de base radial

$$K(x, x_i) = \exp\left(-\frac{1}{2\sigma^2} \| x - x_i \|^2 \right)$$

Métodos

Se consideran sitios de unión de calcio como regiones esféricas de 7 ángstroms (Å) centradas en iones determinados por cristalografía. Los no sitios son regiones esféricas de 7 Å en la superficie o al interior de proteínas que no se unen al calcio, y se utilizan como controles explícitos.

Se decide si una región (una esfera de 7 Å 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 3. 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 PDB [21].

Los sitios y no sitios son divididos en volúmenes espaciales, las cuales son esferas concéntricas con un ancho de 1 Å. 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 entrena una máquina con vectores de soporte con ejemplos negativos y positivos. Una lista de completa de las características usadas para el clasificador se provee en la tabla 1.

Tabla 1. Características usadas para el entrenamiento de la máquina con vectores de soporte

Categoría	Propiedades
Atómicas	Distribución de átomos de algunos elementos básicos
Grupos químicos	carboxilo, hidroxilo, amino, etc
Residuos	Diferentes tipos de aminoácidos
Motivos de estructura secundaria	hélice-α, hoja plegada-β, giros
Motivos de estructura súper-secundaria	bridge, bend, 3-helix
Biofísicas	movilidad, factor –B, volumen de Van der Waals, etc

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.

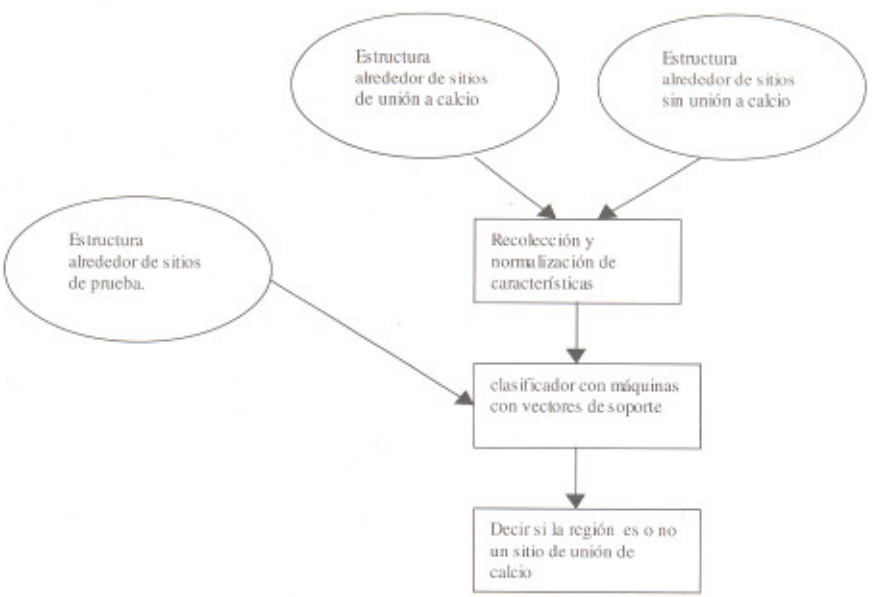


Figura 3. Esquema del sistema para la identificación de sitios.

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. A cada punto de la rejilla se aplicó el clasificador con máquinas con vectores de soporte para determinar 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 mas cercanos se muestran en un visualizador y su ubicación se comparó con los sitios reales de unión de calcio, como se muestra en las figuras 4 y 5.

Resultados

Para evaluar el desempeño del algoritmo de reconocimiento usamos dos medidas: sensibilidad (capacidad para reconocer un sitio de unión a calcio) y especificidad (capacidad para reconocer un sitio que no se une al calcio) definidos de la siguiente manera:

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

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

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 hallados por el método aparecen en las figuras 4 y 5.

Tabla 2. Resultado de los experimentos

Test número	Sitios	No sitios	Especificidad	Sensibilidad
1	47	40	100%	93%
2	33	30	100%	90%
3	193	126	100%	93%

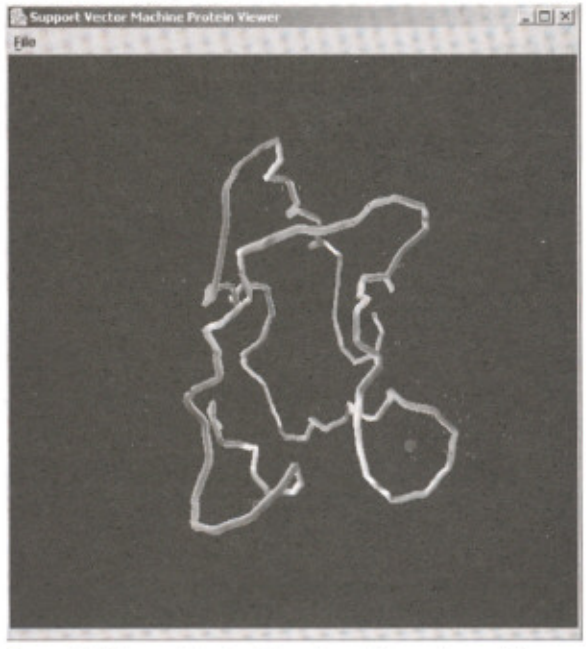


Figura 4. Búsqueda de sitios de calcio en la proteína con código PDB 3IOB.

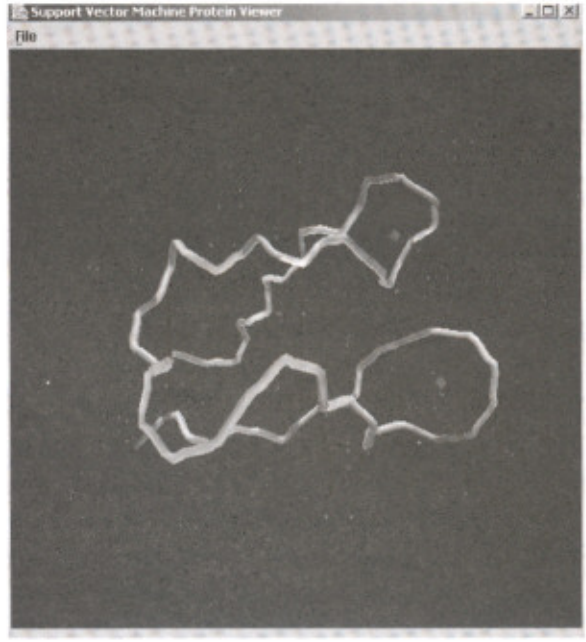


Figura 5. Búsqueda de sitios de calcio en la proteína con código PDB 3PAL.

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 más de 15000 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.

Muchos sitios candidatos estuvieron a una distancia de 1 Å. El enfoque es, en principio, general y en trabajo futuro se planea desarrollar 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] Acerca del genoma humano. Tobías Mojica, Luzardo Estrada. *Agronomía Colombiana*, Vol. 27, pp. 7-12 .

[2] Defining the Mandate of Proteomics in the Post-Genomics Era: Workshop Report National Research Council Steering Committee. George L. Kenyon (Chair).

[3] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28 pp. 235-242.

[4] Simons, K.T.; Kooperberg, C.; Huang, E. and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated anealing and Bayesian scoring functions. *J. Mol. Biol.* 268: 209-225.

[5] Altman, R. B. & T. E. Klein. Challenges for Biomedical Informatics and Pharmacogenomics. Annu. Rev. Pharmacol. Toxicol 2002. 42:113-33

[6] Bryant, S.H. and Altschul, S.F. (1995). Statistics of Sequence-structure Threading. *Current Opinion in Structural Biology*, 5, pp. 236-244.

[7] Simons, K.T.; Kooperberg, C.; Huang, E. and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated anealing and Bayesian scoring functions. *J. Mol. Biol.* 268: 209-225.

[8] Bork, P.; Dandekar, T.; Diaz-Lazcoz, Y.; Eisenhaber, F.; Huynen, M. and Yuan, Y. (1998). Predicting Function: From Genes to Genomes and Back. *J. Mol. Biol.*, 283: 707-725.

[9] Brown, M.P.S.; Grundy, W.N.; Lin, D.; Cristianini, N.; Sugent, C.W.; Furey, T.S.; Ares Jr., M..; and Haussler, D. (2000). Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. *PNAS*, 97(1): 262-267.

[10] Lathrop, R.H. (1994). The Protein Threading Problem with Sequence Amino Acid Interaction Preferences is NP-Complete. *Protein Engineering*, 7:9: 1059-1068.

[11] What is bioinformatics? An introduction and overview Nicholas M Luscombe, Dov Greenbaum & Mark Gerstein.

[12] *Molecular Biology of the Cell.* (1994c). 3rd ed. Alberts, Bruce; Bray, Dennis; Lewis, Julian; Raff, Martin; Roberts, Keith; Watson, James D. New York and London: Garland Publishing.

[13] Richards, F.M. (1996). Calculation of Molecular Volumes and Areas for Structures of Known Geometry. *Methods in Enzymology*, 115: 440-464.

[14] Bateman, A.; Birney, E.; Cerruti, L.; Durbin, R.; Etwiller, L.; Eddy, S.R.; Griffiths-Jones, S.; Howe, K.L.; Marshall, M.; Sonnhammer, E.L. (2002).The Pfam Protein Families Database. *Nucleic Acids Research* 30(1):276-280.

[15] Bagley, S.C. and Altman, R.B. (1995). Characterizing the Microenvironment Surrounding Protein Sites. *Protein Science*, 4: 622-635.

[16] Baldi, P. and Brunak, S. (1998). Bioinformatics: The Machine Learning Cambridge, MA: Approach. MIT Press.

[17] Burges, C. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. In: *Data Mining and Knowledge Discovery.*

[18] Simon S. Haykin. Neural Networks: A Comprehensive Foundation (2nd Edition).

[19] Berman, H.M.; Bhat, T.N.; Bourne, P.E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology* 7 (11): 957-959.