USO DE MODELOS ESCONDIDOS DE MARKOV EN BIOLOGÍA MOLECULAR COMPUTACIONAL

Fecha de recepción: 23-07-2008, aprobación: 09-09-2009

Carlos David Seligmann Trujillo

RESUMEN

Este documento pretende hacer una revisión del uso de los modelos escondidos de Markov en el análisis de secuencias biológicas y específicamente de cómo éstos son empleados en la biología molecular computacional.

PALABRAS CLAVE

Modelos escondidos de Markov, métodos de predicción, estructura secundaria, proteínas, predicción de genes.

ABSTRACT

This document revisits the use of Hidden Markov Models, and how this methodology is used in the analysis of biological sequences, specifically, how the Hidden Markov models are used in bioinformatics or computational molecular biology.

KEYWORDS

Hidden Markov Models, Prediction Methods, Secondary Structure, Proteins, Gene Prediction, Bioinformatics.

Este documento hace una revisión del uso de los modelos escondidos de Markov (HMM) en biología molecular computacional. Para lograr este fin, el texto está dividido en cinco partes. En las secciones I, II y III se da un marco teórico suficiente para entender la complejidad de los problemas biológicos que se pueden abordar con esta metodología. La sección IV ilustra los modelos escondidos de Markov como metodología, introduce los algoritmos usados en la utilización de estos modelos y da un ejemplo de su uso. La sección V hace un listado del uso de los modelos escondidos de Markov desde la bioinformática.

Los HMMs son un método estadístico usado en el reconocimiento de patrones (Krogh, 1998, Rabiner,1989), en especial en aplicaciones como reconocimiento de voz (Rabiner, 1989), aplicaciones bioinformáticas como la predicción de la estructura secundaria de proteínas a partir de la estructura primaria (Krogh, 1998., Mukherjee, 2005 & Thorvaldsen, 2005) y la predicción de genes eucariotas (Tomas, 2005).

DOGMA CENTRAL DE LA BIOLOGÍA MOLECULAR (PROCARIOTAS)

El dogma central de la biología molecular (figura 1) hace referencia al flujo de información dentro de la célula. La información celular está contenida en el ADN. El dogma señala que un gen (secuencia de ADN) produce una única molécula de ARN que a la vez determina la secuencia de la proteína. La molécula de ADN está compuesta de nucleótidos que pueden ser cuatro: A,T,C,G. La molécula de ARN es también una secuencia de nucleótidos: A,U,C,G. El paso del ADN a ARN es llamado transcripción. Básicamente se tiene la misma secuencia de nucleótidos cambiando el nucleótido T por U. Las proteínas son secuencias de aminoácidos (combinaciones de los veinte posibles aminoácidos). El paso del ARN a proteína (traducción) se realiza con un mapeo de información de tres letras de ARN a una letra de proteína (Birney, 2001) Dicho mapeo se da a través del llamado código genético (figura 2). La secuencia lineal de aminoácidos de la proteína (estructura primaria) determina la estructura proteica y la estructura proteica determina la función (Bergeron, 2002).

A diferencia de los organismos procariotas, donde se tiene que un gen (secuencia de ADN) produce una molécula (secuencia) de ARN y ésta una proteína, en organismos eucariotas un gen tiene secciones llamadas intrones y otras llamadas exones.

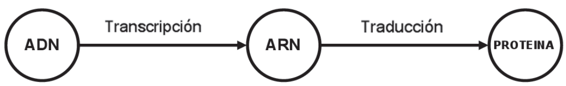

Figura 1. El dogma central de la biología molecular (modelo general).

De esta forma, al estudiar la secuencia generada a través de la representación de las moléculas mediante cadenas de residuos (nucleótidos en el caso del ADN y el ARN o aminoácidos en el caso de proteínas) es posible encontrar relaciones entre ellos que permitan establecer las propiedades biológicas de las moléculas asociadas (Rivas & Eddy, 1999; Searls, 1992). Uno de los problemas más importantes en biología molecular es el relacionado con la identificación de la función molecular (Baldi, 2001). El dogma central mostrado en la figura 1 es válido en organismos procariotas (sin núcleo) pero no necesariamente en el caso de organismos eucariotas (con núcleo celular).

RESEÑA DE AUTOR

Carlos David Seligmann Trujillo
(Politécnico Grancolombiano)
(dseligma@poligran.edu.co)
Biólogo e Ingeniero de Sistemas de la Universidad de los Andes, Bogotá, Colombia; MSc. Bioinformatics, DCU-Ireland. Se ha desempeñado como profesor de tiempo completo en el Politécnico Grancolombiano desde enero de 2007, en las asignaturas de Programación de Computadores, Pensamiento Algorítmico e Introducción a la Bioinformática, entre otras. Sus intereses académicos son la bioinformática, la biología computacional y la computación bioinspirada.

DOGMA CENTRAL DE LA BIOLOGÍA MOLECULAR (EUCARIOTAS)

A diferencia de los organismos procariotas, donde se tiene que un gen (secuencia de ADN) produce una molécula (secuencia) de ARN y ésta una proteína, en organismos eucariotas un gen tiene secciones llamadas intrones y otras llamadas exones. Sólo las secuencias correspondientes a exones contienen información útil para la elaboración de proteínas por esta razón, los intrones son cortados y desechados y los exones pegados unos detrás de otros donde forman una molécula de ARNm. Este proceso se llama *splicing* (Birney, 2001). La información de la molécula de ARNm es traducida en forma de proteína de forma similar a la usada por los organismos procariotas.

El dogma ampliado se muestra en la figura 3.

2a letra en el codon

1a letra en el codon	U	C	A	G	3a letra en el codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figura 2. Código genético.

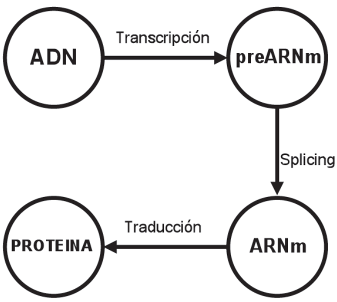

Figura 3. El dogma central de la biología molecular (ampliado).

ESTRUCTURAS PROTEICAS

Se distinguen cuatro tipos de estructura proteica: primaria, secundaria terciaria y cuaternaria. Cada una depende de la estructura anterior.

ESTRUCTURA PRIMARIA DE PROTEÍNAS

La estructura primaria de una proteína está dada por la secuencia lineal de aminoácidos es decir, el número de aminoácidos presentes y el orden en que están enlazados. Los aminoácidos estándar son 20, A, R, N, D, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V,[1] cada uno con diferentes propiedades biofísicas y bioquímicas. Una proteína es una macromolécula es decir, una molécula de gran tamaño (más de cincuenta aminoácidos sin límite de tamaño). Estamos hablando entonces de secuencias de más de cincuenta aminoácidos, donde se tiene que los veinte aminoácidos citados anteriormente no tienen restricciones de repetición. (Murzin, 1995).

Se distinguen cuatro tipos de estructura proteica: primaria, secundaria terciaria y cuaternaria. Cada una depende de la estructura anterior.

ESTRUCTURA SECUNDARIA DE PROTEÍNAS (TOOZE, 1999)

La estructura secundaria es la forma tridimensional general de segmentos locales de proteínas. Está caracterizada por arreglos de aminoácidos altamente regulares y repetitivos estabilizados por puentes de hidrógeno. No se describen las posiciones atómicas específicas que son consideradas como una estructura terciaria. Los tipos de estructuras secundarias más conocidas son las α-hélices: y las b-láminas (Eisenberg, 2003) sin embargo, se han descrito más estructuras secundarias2:

α-hélices: los aminoácidos en este tipo de estructura secundaria α están dispuestos en una estructura helicoidal dextrógira, con unos 3.6 aminoácidos por cada vuelta. Cada aminoácido supone un giro de unos 100° en la hélice. Casi no hay espacio libre

1. Alanina, arginina, asparagina, ácido aspártico, cisteína, ácido glutámico, glutamina, glicina, histidina, isoleucina, leucina, lisina, metionina, fenilalanina, prolina, serina, treonina, triptófano, tirosina y valina, respectivamente.
2. Bucles, giros, β-hélices, α-láminas, etcétera.

dentro de la hélice es decir, la hélice está estrechamente empaquetada. Todas las cadenas laterales de los aminoácidos están dispuestas hacia el exterior de la hélice. El grupo N-H del aminoácido (n) puede establecer un enlace de hidrógeno con el grupo C=O del aminoácido (n+4). De esta forma, cada aminoácido (n) de la hélice forma dos puentes de hidrógeno con su enlace peptídico y el enlace peptídico del aminoácido en (n+4) y en (n-4). En total son siete enlaces de hidrógeno por vuelta. Esto estabiliza enormemente la hélice (Murzin, 1988).

β-láminas: se forman por el posicionamiento paralelo (o antiparalelo) de dos cadenas de aminoácidos dentro de la misma proteína en el que los grupos N-H de una de las cadenas forman enlaces de hidrógeno con los grupos C=O de la opuesta. Es una estructura muy estable que puede resultar de una ruptura de los enlaces de hidrógeno durante la formación de la a-hélice. Los grupos R (radical) de esta estructura están posicionados sobre y bajo el plano de las láminas. Estos radicales no deben ser muy grandes, ni crear un impedimento estérico, ya que se vería afectada la estructura de la lámina (Eisenberg, 2003).

Es importante aclarar que como los aminoácidos son bioquímicamente diferentes entre sí, cada uno confiere a la cadena polipeptídica características particulares es decir, una cadena polipeptídica rica en aminoácidos hidrófilos se dobla de una forma, mientras que si es rica en aminoácidos hidrófobos se dobla de otra.

MODELOS ESCONDIDOS DE MARKOV

Un modelo escondido de Markov es un modelo estadístico que asume un proceso de Markov de parámetros desconocidos (Brown University, 2008). Un proceso de Markov, también llamado cadena de Markov (Mukherjee, 2005) es una serie de eventos en la cual la probabilidad de que ocurra uno depende del evento inmediatamente anterior. En un proceso de Markov existe el concepto de memoria. Los procesos de Markov "recuerdan" el evento inmediatamente anterior y esto condiciona las posibilidades de los eventos futuros. Esta dependencia del evento anterior distingue los procesos de Markov de las series de eventos independientes (como lanzar un dado varias veces) (Wikipedia, 2008).

En un proceso de Markov se tiene entonces un conjunto de estados posibles y un conjunto de probabilidades para pasar de un estado actual a otro futuro.

En este punto es preciso dar un ejemplo de un HMM, para lo cual se ha modificado y complementado un ejemplo tomado de (University of Leeds, 2007):

Suponga que se tiene el problema de definir el estado del clima y que los estados posibles son: *soleado, lluvioso, y nublado.*

Si el clima hoy es *soleado,* hay tres posibilidades para el clima de mañana, cada una con una probabilidad diferente: soleado (0.5), lluvioso (0.1) y nublado (0.4).

Si el clima hoy es *lluvioso,* hay tres posibilidades para el clima de mañana, cada una con una probabilidad diferente: soleado (0.2), lluvioso (0.5) y nublado (0.3).

Si el clima hoy es *nublado,* hay tres posibilidades para el clima de mañana, cada una con una probabilidad diferente: soleado (0.25),lluvioso (0.25) y nublado (0.5).

El proceso de Markov está definido como un grafo y se muestra en la figura 4.

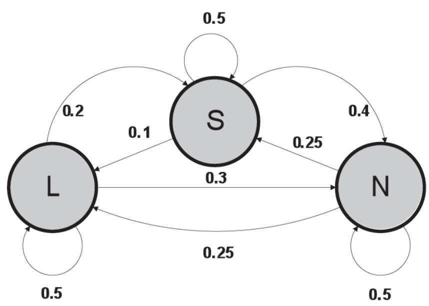

Figura. 4. Grafo que describe el proceso de Markov. Estados: S=soleado, L=lluvioso. N=nublado

Un modelo escondido de Markov es un proceso de Markov en el cual, además de los estados del sistema (que están ocultos), se tienen unos estados observables (que no son los estados del sistema (Mukherjee, 2005; Rabiner, 1986; Thorvaldsen, 2005). Los estados observables dan cierta información acerca de los estados del sistema. En este punto es necesario extender el ejemplo dado (University of Leeds, 2007).

Suponga que un ermitaño que vive en una cueva desea conocer el clima, pero por definición no saldrá a verificarlo. Sin embargo, él cuenta con un dispositivo de alta tecnología para inferir el estado del clima: una soga. El ermitaño revisa el nivel de humedad de la soga e infiere (con algún grado de probabilidad -o incertidumbre-) el estado del clima. Para esto, el ermitaño clasifica las posibles observaciones realizadas sobre la soga en cuatro estados: seco, escurrido, húmedo y empapado (o si se quiere nivel de humedad 0, 1, 2, 3). Estas observaciones le dan cierta información probabilística acerca del estado del clima (University of Leeds, 2007).

En la figura 5 se puede ver el grafo ampliado, conteniendo los estados escondidos y los estados observables. Las conexiones entre los estados escondidos y los estados observables representan la probabilidad de generar un estado observado dado que el proceso de Markov se encuentra en un estado escondido en particular.

Este tipo de procesos se forma como un modelo escondido de Markov donde existe un proceso de Markov (escondido) y unos estados observables que se relacionan de alguna forma con los estados escondidos (Rabiner, 1989; Thorvaldsen, 2005, University of Leeds, 2007).

Suponga que un ermitaño que vive en una cueva desea conocer el clima, pero por definición no saldrá a verificarlo. Sin embargo, él cuenta con un dispositivo de alta tecnología para inferir el estado del clima: una soga. El ermitaño revisa el nivel de humedad de la soga e infiere (con algún grado de probabilidad -o incertidumbre-) el estado del clima.

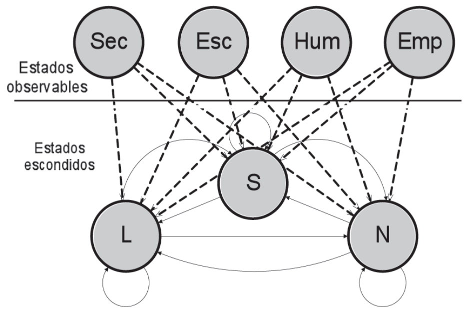

Figura 5. HMM que describe el ejemplo del clima. Muestra los estados observables (Sec=seco, Esc=escurrido, Hum=húmedo, Emp=empapado) y los estados escondidos (modelados por un proceso de Markov). S=soleado, L=lluvioso. N=nublado

Debería ser claro que se trata de una probabilidad condicional y que la suma de todas las probabilidades condicionales para llegar a un estado observable en particular debe ser igual a 1.

$$P(obs|\,sol) + P(obs|\,lluv) + P(obs|\,nub) = 1 \quad (1)$$

Para un modelo escondido de Markov se tiene un proceso de Markov y una matriz de probabilidades de los estados observables dados los estados escondidos. Esta matriz es llamada la matriz de emisión (Rabiner, 1989) o confusión (University of Leeds, 2007). Para el ejemplo del clima se muestra la matriz de emisión en la figura 6.

$$\begin{array}{r} \\ soleado \\ CLIMA\ lluvioso \\ nublado \end{array}\begin{array}{c} SOGA \\ \begin{array}{cccc} \text{sec} & esc & hum & emp \end{array} \\ \begin{pmatrix} 0.6 & 0.25 & 0.1 & 0.05 \\ 0.05 & 0.15 & 0.3 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \end{array}$$

Figura 6. Matriz de Emisión para el ejemplo del clima. Muestra los estados observables y los estados escondidos (modelados por un proceso de Markov). Note que la suma de probabilidades en cada fila es 1.

Se puede definir formalmente y finalmente un HMM (Λ) como una tripla (∏, A, B) donde: (Rabiner, 1989. Thorvaldsen, 2005. University of Leeds, 2007. Wikipedia, 2008).

- $\Pi = (\pi_i)$ Vector de probabilidades de estado iniciales
- $A = (a_i)$ Matriz de transición de estados; $P(x_{i_t} \mid x_{j_{t-1}})$
- $B = (b_{ij})$ Matriz de emisión; $P(y_i \mid x_j)$

Los HMMs asumen que cada probabilidad en la matriz de la transición de estados y en la matriz de la emisión es independiente del tiempo es decir, las matrices no cambian en el tiempo, pero en la práctica ésta es una de las suposiciones menos realistas de modelos de Markov acerca de procesos verdaderos.

Son tres los problemas clásicos tratados en el contexto de los HMMs (Eddy, 1996., Mukherjee, 2005., Rabiner, 1986; Rabiner, 1989; Thorvaldsen, 2005. University of Leeds, 2007). A continuación se nombran tales problemas y se muestra su solución.

EVALUACIÓN O CALIFICACIÓN

Dado un HMM (Λ) y una secuencia de observaciones $O = o_1, o_2, \ldots, o_t$ se debe calcular la probabilidad de que las observaciones sean generadas por el modelo $P(O|\Lambda)$. Si se tienen varios modelos se evalúa cuál es el que tiene mayor probabilidad de acuerdo a las observaciones (Mukherjee, 2005; Rabiner, 1986; Thorvaldsen, 2005).

DECODIFICACIÓN O ALINEAMIENTO

Dado un HMM (Λ) y una secuencia de observaciones $O = o_1, o_2, \ldots, o_t$ la pregunta a resolver es: ¿cuál es la secuencia de estados (escondidos) más probable (óptima) que produce las observaciones? Se deben conocer los estados escondidos –aunque no la secuencia de estados escondidos (University of Leeds, 2007)–.

APRENDIZAJE O ENTRENAMIENTO

Este problema es el más difícil de los tres y consiste en determinar un método para ajustar los parámetros de un HMM L = (P, A, B) de tal manera que se maximice la probabilidad de la secuencia de observaciones $O = o_1, o_2, \ldots, o_t$ dado este modelo. Es decir, maximizar $P(O|\Lambda)$ (Rabiner, 1989).

MODELOS ESCONDIDOS DE MARKOV EN EL ANÁLISIS DE SECUENCIAS BIOLÓGICAS

La utilización de modelos escondidos de Markov en los estudios bioinformáticos es ubicua sin embargo, su uso clásicamente se da en la solución de los siguientes problemas (Birney, 2001):

- Predicción de genes (discriminación de intrones y exones).
- Doblamiento de proteínas (predicción de la estructura secundaria a partir de la primaria).

PREDICCIÓN DE GENES

Se plantea el HMM de la siguiente manera:

La secuencia de ADN genómico –antes del *splicing*– es la entrada de este proceso donde los nucleótidos son los estados observados. Los estados escondidos son: o bien intrones o bien exones.

La salida es un árbol de intrones y exones sobre la secuencia original. De ahí se puede obtener el gen predicho.

Este modelo se ilustra en la figura 7 (Tomas, 2005).

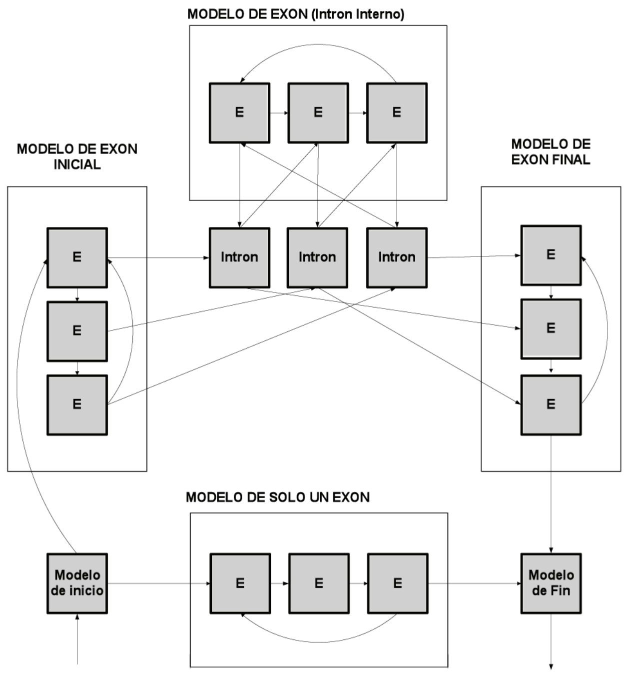

Figura 7. Modelo escondido de Markov para la detección de un gen sobre la cadena principal de ADN genómico. Los modelos de inicio y de fin corresponden, respectivamente, a codones conocidos de inicio y de fin de la trascripción

PREDICCIÓN DE ESTRUCTURA MOLECULAR

Se puede plantear acá el HMM de la siguiente forma:

La secuencia lineal de aminoácidos (estructura primaria) es la entrada del sistema. Los estados escondidos serían las diferentes estructuras secundarias (lámina beta, alfa hélice, etcétera), los estados observados: sub secuencias de aminoácidos de la estructura primaria.

La salida es un árbol de estructuras secundarias sobre la secuencia original. De ahí se puede obtener la estructura predicha (Asai, Hayamizu, & Handa, 1993). Este modelo se ilustra en la figura 8.

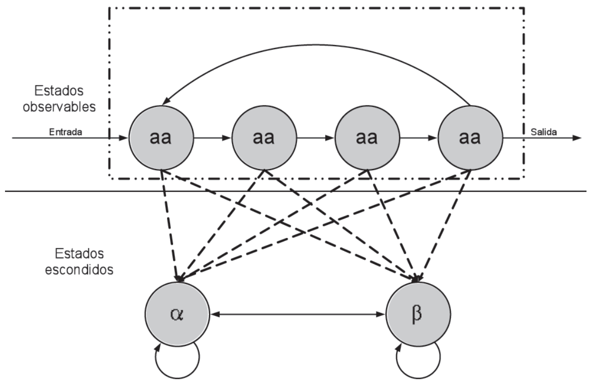

Figura 8. Un modelo escondido de Markov simplificado para la predicción de la estructura secundaria (en este caso sólo α-hélices y b-láminas) de un polipéptido.

Por otro lado, se ha desarrollado un HMM equivalente a un análisis de perfiles para investigar familias de proteínas (Birney, 2001., Krogh, Brown, Mian, Sjolander; & Haussler, 1994). El análisis de perfiles ofrece una forma de representar el "perfil consenso" de aminoácidos para un conjunto de secuencias proteicas (o ADN) pertenecientes a la misma familia. Un perfil consenso (HMM) puede hacerse para buscar, dentro de una base datos, otros miembros de una familia.

La principal característica de los perfiles HMM es que tratan los gaps de una forma sistemática (Krogh, 1998, University of Leeds, 2007).

BIBLIOGRAFÍA

- Asai, K., Hayamizu, S., & Handa, K. I. (1993). Prediction of Protein Secondary Structure by the Hidden Markov Model. *Comput. Appl. Biosci.*, *9(2),* 141-146.
- Baldi, P. B., S. (2001). *The Machine Learning Approach. Bioinformatics.* Boston: MIT Press.
- Bergeron, B. (2002.). *Applied Bioinformatics Computing: An Introduction*. New Jersey: Prentice Hall.
- Birney E. (2001). Hidden Markov Models in Biological Sequence *Analysis. IBM Journal of Research and Development, Vol. 45 (NO. 3/4),* 449-454.
- Bonneau, R., Baliga, N., Deutsch, E., Shannon, P., & Hood, L. (2004). Comprehensive de novo structure prediction in a systems-biology context for the *archaea Halobacterium sp. NRC-1. Genome Biology, 5(8),* R52.
- Brown University. (2008) Hidden Markov Models. Recuperado el 25 de abril de 2008, de http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/HiddenMarkovModels.html
- Eddy, S. (1996). Hidden Markov Models. *Current Opinion in Structural Biology, 6(3),* 361-365.
- Eisenberg, D. (2003). The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences* USA, 100, 11207-11210.
- Krogh, A. (1998). An Introduction to Hidden Markov Models for Biological Sequences. *Computational Methods in Molecular Biology.* Elsevier, NewYork pp. 45-63.

- Krogh, A. (1998). An Introduction to Hidden Markov Models for Biological Sequences. In S. Salzberg, Searls, D. and Kasif, S. (Ed.), Computational Methods in Molecular Biology (pp. 45-63): Elsevier.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol., 235,* 1501-1531.
- Mukherjee, S. , Mitra S. (2005). Hidden Markov Models, Grammars, And Biology: A Tutorial. *Journal of Bioinformatics and Computational Biology, 3(2),* 491-526.
- Murzin, A., Finkelstein, AV. (1988). General architecture of the α-helical globule. *Journal of Molecular Biology, 204*, 749-769.
- Murzin A. G., B. S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol., 247,* 536-540.
- Rabiner, L. J., B. (1986). An introduction to Hidden Markov Models. *ASSP Magazine also IEEE Signal Processing Magazine, 3(1 part 1),* 4-16.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE, 77,* 257-286.
- Rivas, E., & Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol., 285,* 2053-2068.
- Searls, D. (1992). The Computational linguistics of biological sequences. Artificial Intelligence and Molecular Biology, AAAI Press, pp. 47-120
- Thorvaldsen, S. (2005). A Tutorial On Markov Models Based On Mendel's Classical Experiments. *Journal of Bioinformatics & Computational Biology, 3(6),* 1441-1460.
- Tomas, V. (2005). Enhancements to Hidden Markov Models for Gene Finding and Other Biological Applications. University of Waterloo, Waterloo.
- Tooze, C. B. a. J. (1999*). Introduction to Protein Structure.* New York, NY: Garland Publishing.
- University of Leeds. Hidden Markov Model Tutorial. Recuperado el 25 de abril de 2007, http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html
- Wikipedia. (2008). Hidden Markov model. Retrieved. Recuperado el 26 de abril de 2008, de http://en.wikipedia.org/w/index.php?title=Hidden_Markov_model&oldid=209207033; Page Version ID: 209207033.