Detección de los puntos de unión de los factores de transcripción mediante el análisis de la variabilidad de la información mutua cruzada

Joan Maynou, Montserrat Vallverdú, Francesc Clarià, Joan-Josep Gallardo-Chacón, Pere Caminal y Alexandre Perera

Abstract— **La detección de secuencias de regulación es un reto importante en biología computacional. Concretamente, el proceso de síntesis de una proteína empieza con la unión de un factor de transcripción a la correspondiente secuencia de unión. Un mismo factor de transcripción puede unirse a diferentes secuencias de unión. Esta variabilidad intrínseca encontrada en las secuencias de unión aumenta la dificultad de su detección mediante algoritmos computacionales. En este trabajo, se propone un método para la detección de secuencias de unión basado en la correlación entre las correspondientes posiciones de la secuencia mediante medidas de la teoría de la información. La eficiencia del método está presentada mediante las curvas ROC (Receiver Operating Characteristic) para la detección de los diferentes factores de transcripción utilizados del organismo *Saccharomyces cerevisiae*. Finalmente, se comparan los resultados con otro método de detección de motivos, Motif Discovery Scan (MDscan).**

I. Motivación

Cada célula de un organismo contiene la información necesaria para la regulación de cualquier proceso biológico. Para su supervivencia, es necesario un control muy estricto de las redes de expresión génica (proliferación celular y diferenciación tisular) y en el tiempo (respuesta a estímulo) [1]. Durante el proceso de transcripción del gen, la información genética es transferida del ácido desoxirribonucleico (DNA) a ácido ribonucleico mensajero (mRNA). El primer proceso de control de la transcripción corresponde a la asociación de proteínas específicas con su secuencia de unión diana en el DNA. Además, la proteína específica se une a otros factores de modulación y al enzima RNA polimerasa. Estas proteínas, que actuan en regiones reguladoras génicas, se conocen como factores de transcripción. En organismos eucariotas, la transcripción empieza reclutando

RNA polimerasa para diferentes proteínas. Dichas proteínas reconocen señales específicas en la región previa al gen conocida como promotor. Una de estas señales es una secuencia de nucleótidos que contiene la información necesaria para iniciar el proceso de transcripción. Una vez sintetizado el mRNA, se traduce a una secuencia de aminoácidos, mediante el proceso conocido como traducción. Los polipéptidos originados forman las proteínas estructurales y enzimas que controlan todo proceso metabólico de las células.

Un factor de transcripción tiene la capacidad de unirse a diferentes posiciones y secuencias a lo largo del genoma. Debido a esta variabilidad, es difícil establecer una secuencia consensus para la detección de las secuencias de unión [2]. Consecuentemente, cualquier método de detección de secuencias de unión debe considerar la variabilidad de estas. Debido a esta problemática, se han desarrollado diferentes métodos de detección de patrones en secuencias de DNA. Los métodos de detección de patrones más relevantes son los métodos probabilísticos basados en Position Weigth Matrices, PWM, (e.g. MDscan, método basado en la enumeración de palabras combinadas y PWM [3]). Una PWM es una matriz de pesos donde cada fila corresponde a un símbolo del alfabeto y cada columna a una posición de la secuencia de unión. Según el peso considerado, hay diferentes tipos de PWMs [4]: matrices frecuenciales que contienen la frecuencia absoluta o relativa de un nucleótido en cada posición del motivo y matrices de pesos correspondientes al logaritmo de la razón de verosimilitud.

La teoría de la información ha sido utilizada en genética para visualizar y caracterizar la información de un conjunto de secuencias [5], [6]. Los detectores publicados basados en la entropía de Rényi miden el contenido de información total en una secuencia de unión [7]. Estos trabajos previos consideran que las posiciones de la secuencia de unión son independientes entre si. , otros estudios sugieren que la covarianza entre posiciones puede indicar posibles interacciones entre DNA y los factores de transcripción. La covarianza puede ser medida mediante el análisis de la correlación entre diferentes posiciones de la secuencia de unión. En este trabajo, se propone un detector de motivos aplicado a las secuencias de unión de los factores de transcripción utilizando una medida diferencial basada en la información mutua. A partir de un conjunto de secuencias alineadas con conocimiento de unión, se analiza la variación de la información total cruzada cuando la secuencia candidata es incluida en el conjunto.

Este trabajo ha sido parcialmente financiado por la CICYT TEC2007-63637/TCM del Ministerio de Ciencia y Tecnología, así como por el programa Ramón y Cajal del Ministerio de Educación y Ciencia. El CIBER de Bioingeniería, Biomateriales y Nanomedicina es una iniciativa del ISCIII.

J. Maynou, M. Vallverdú, P. Caminal y A. Perera pertenecen al Dep. ESAII, Centre Recerca en Enginyeria Biomèdica (CREB), Universitat Politècnica de Catalunya (UPC), Barcelona, Gargallo, 5, 08028 Barcelona, España. http://www.creb.upc.es, http://www.upc.edu. e-mail: joan.maynou, montserrat.vallverdu, pere.caminal, alexandre.perera@upc.edu

J.J. Gallardo pertenece al CIBER de Bioingeniería, Biomateriales y Nanomedicina. http://www.isciii.es/htdocs/redes/ciber.jsp e-mail:joan.josep.gallardo@upc.edu

F.Clarià pertenece al Dep. Informática y Ingeniería Industrial, Universitat de Lleida, LLeida, España. e-mail:Claria@eup.udll.es

TABLE I
FACTORES DE TRANSCRIPCIÓN ANALIZADOS

Organismo	Factores Transcripción	Nucleótidos	Sec. alineadas
S. cerevisiae	*MCM1*	38	16
S. cerevisiae	*ABF1*	37	22

II. MATERIALES Y MÉTODOS

A. Método

El método propuesto empieza con una matriz de secuencias alineadas con evidencia de unión. Cualquier secuencia candidata añadida a la matriz de entrenamiento causará una variación en la información mutua del conjunto de secuencias alineadas. La detección de una secuencia de unión se considerará dependiendo del cambio en la matriz de información mutua cuando la secuencia candidata es añadida al conjunto de secuencias alineadas. Para una secuencia aleatoria, la correlación entre posiciones del conjunto de secuencias alineadas decrecerá. En cambio, para una secuencia de unión la información mutua total del conjunto de secuencias alineadas no sé modificará de forma significativa. Por lo tanto, esta medida permite construir un detector basado en la dependencia entre las posiciones de la secuencia de unión. La validación del detector se ha realizado mediante "Leave one-out cross-validation". Cada secuencia individual es utilizada como una secuencia de validación. El clasificador está construido con las $n-1$ secuencias restantes como conjunto de entrenamiento. Los resultados han sido obtenidos para una secuencia candidata generada aleatoriamente con *1000* nucleótidos, testeada sucesivamente para cada una de las secuencias de la matriz de entrenamiento.

B. Información Mutua Cruzada entre Posiciones

La información mutua es una cantidad que mide la dependencia entre dos variables. Dado dos variables aleatorias discretas, X y Y, con N posibles estados $(X_1, X_2, \cdots, X_N)$ y $(Y_1, Y_2, \cdots, Y_N)$, la información mutua se define como,

$$I(X;Y) = \sum_N \sum_N p(X,Y) log_2 \left(\frac{p(X,Y)}{p(x)p(y)} \right) = H(X) + H(Y) + H(X,Y) \quad (1)$$

donde $H(X)$ y $H(Y)$ son las entropías marginales, y $H(X,Y)$ es la entropía conjunta de X y Y. La medida de información mutua es simétrica y no-negativa. $I(X;Y) = 0$ sólo si las dos variables (X, Y) son estadísticamente independientes. En una secuencia de DNA, las variables X y Y corresponden a los nucleótidos en dos posiciones diferentes. La probabilidad se calcula mediante la estimación de la frecuencia a partir de la matriz de entrenamiento. La medida de la información mutua cruzada entre posiciones (PCMI) permite estudiar la dependencia entre posiciones no adyacentes, proporcionando información sobre la correlación de los nucleótidos.

C. Base de datos

El algoritmo desarrollado requiere un grupo de secuencias de nucleótidos alineados con evidencia de unión. Estas secuencias provienen del organismo *Saccharomyces cerevisiae* el cual fue el primer organismo eucariota genotipado. Este organismo contiene aproximadamente 16 millones de nucleótidos distribuidos entre 16 cromosomas. Se ha considerado los factores de transcripción *MCM1* y *ABF1*, resumidos en la tabla I. El conjunto de datos se han obtenido de la base de datos *TRANSFAC*, `http://www.genregulation.com/pub/databases.html`, mediante una librería propia en R para la extracción automática de secuencias de DNA a partir de una palabra clave. Finalmente, estas secuencias han sido alineadas mediante *MUSCLE*, para poder obtener los diferentes nucleótidos involucrados en cada posición.

D. Detección de motivos

A partir de la matriz de secuencias alineadas, se realiza una medida de la correlación entre las posiciones mediante la información mutua. Los valores de la información mutua para posiciones muy correlacionadas son cercanos a H_i (entropía de Shannon para la posición i de la matriz de entrenamiento). En cambio, en posiciones no correlacionadas la información mutua toma valores cercanos a 0. Utilizando esta propiedad, el algoritmo desarrollado realiza una comparación entre la información mutua de la matriz de entrenamiento y la información mutua de la matriz de entrenamiento cuando la secuencia candidata es añadida al conjunto. Se define un conjunto de funciones para avaluar la variación en la información mutua. Las funciones consideradas son las siguientes,

$$Difference = \left(\sum \gamma \right)^{-1} \quad (2)$$

$$Power = \left[\sum MI_{matrix} \gamma \right]^{-1/2} \quad (3)$$

$$Normalization = \left[\frac{\left(\sum MI_{matrix} \gamma \beta^{-1} \right)}{max(H)} \right]^{-1} \quad (4)$$

donde, γ y β son

$$\gamma = |MI_{matrix} - MI_{matrix+seq}| \quad (5)$$

$$\beta = |MI_{matrix} + MI_{matrix+seq}| \quad (6)$$

donde, MI_{matrix} es la matriz de información mutua del conjunto de secuencias alineadas y $MI_{matrix+seq}$ es la información mutua entre las posiciones de la matriz de entrenamiento cuando la secuencia candidata es añadida. La variación en la matriz de información del conjunto de entrenamiento, cuando se añade la secuencia candidata, se evalua mediante el análisis de la variabilidad de la información mutua cruzada entre posiciones. Para una secuencia aleatoria, la dependencia entre posiciones decrece y, por consiguiente, decrece la matriz de información mutua. En cambio, para una secuencia de unión, no se modifica de forma significativa la información del conjunto de entrenamiento. Por lo tanto, $\gamma|_{random} > \gamma|_{binding}$ y $\beta|_{binding} > \beta|_{random}$. De esta forma, se puede definir un detector que discrimine entre una secuencia aletoria y una secuencia de unión. El método

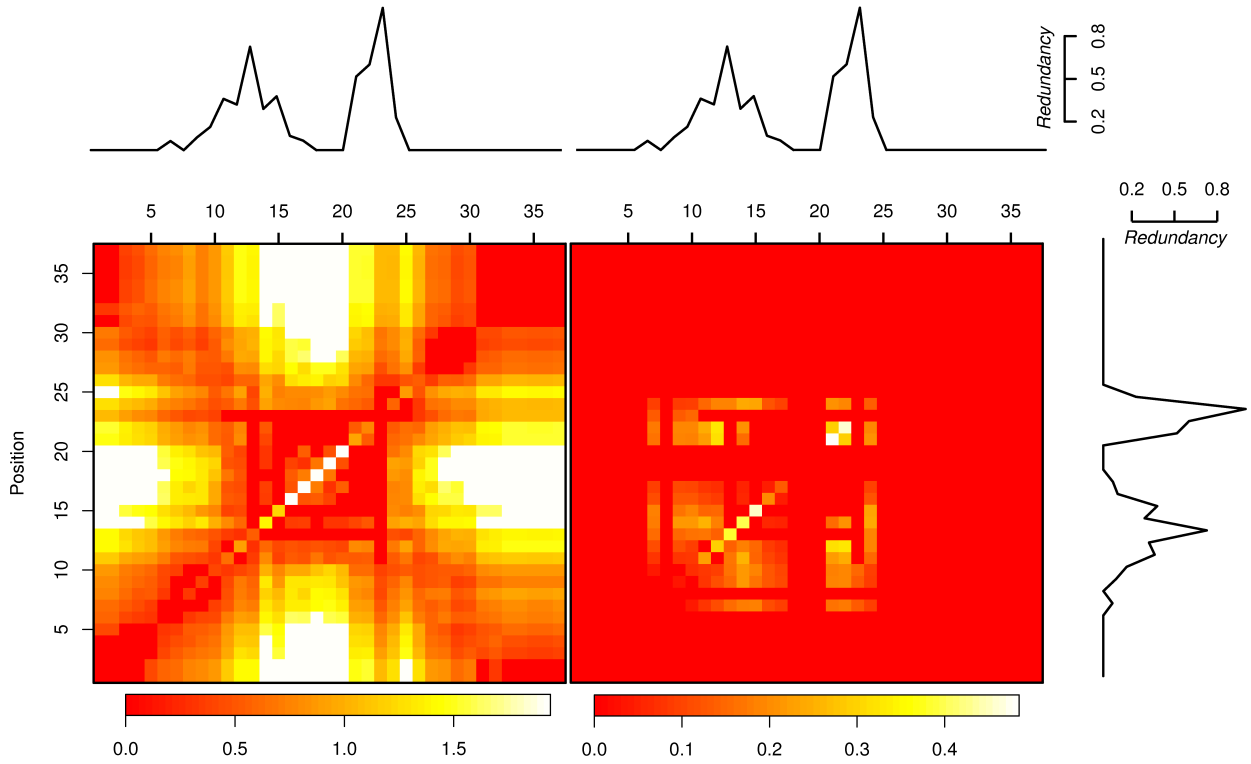

Fig. 1. (Izquierda) Heatmap de información mutua entre las posiciones de la secuencia de unión para ABF1. El perfil de redundancia se visualiza en la parte superior; (Derecha) Producto entre la matriz de información mutua ponderada por el producto exterior del perfil de redundancia.

desarrollado, basado en el criterio definido anteriormente, es el siguiente:

1) Para cada posición de la matriz de entrenamiento, se estima la probabilidad y la probabilidad conjunta correspondiente a cada nucleótido. Se considera la ausencia de símbolo como un multi-estado de nucleótidos con probabilidad igual a la frecuencia de cada nucleótido según el organismo.
2) La entropía marginal y conjunta es calculada a partir de PWM, corregiendo el efecto de muestra finita [8].
3) La información mutua es calculada a partir de la entropía marginal y conjunta.
4) Los puntos 1, 2 y 3 se repiten añadiendo a la matriz entrenamiento una nueva secuencia candidata.
5) Para cada matriz de información mutua obtenida, se calcula una magnitud escalar para cada una de las funciones definidas en (2), (3), (4), (5) y (6).

III. Resultados

En la figura 1 (izquierda) se muestra la información mutua entre las posiciones de la secuencia de unión y la variabilidad intrínseca de cada posición para el factor de transcripción ABF1. La medida de redundancia es una entropía normalizada que compara la entropía de la variable respecto a su máximo valor teórico, según $R = 1 - H/H_{max}$ [7], [9]. La medida de redundancia proporciona información sobre la varianza observada en una posición del conjunto de secuencias alineadas. Poca varianza corresponde a valores altos de redundancia. De hecho, la medida de redudancia muestra como una posición, en particular, ha estado *conservada* en un conjunto de secuencias. Por otra parte, en la figura 1 (derecha) se visualiza el producto entre la matriz de la información mutua y el producto exterior del perfil de redundancia. Esta medida ayuda a determinar la correlación entre posiciones de la secuencia de unión que desarrollan un papel importante en la unión des de un punto de vista conservativo. En este sentido, esta gráfica no muestra las posiciones conservadas entre diferentes secuencias de unión, pero sí la *correlación* entre las posiciones que han estado conservadas en un número de secuencias de unión de ejemplo. Cuando esta medida es positiva, se considera que existe dependencia entre posiciones. El detector propuesto en este trabajo evalua la perturbación en esta matriz para analizar si la *correlación conservada* es destruida cuando se añade la secuencia candidata al conjunto de secuencias alineadas.

La realización del detector para los factores de transcripción ABF1 y MCM1 y para las diferentes funciones definidas, se visualizan mediante las correspondientes curvas ROC, Receiver Operating Characteristic, tal como se muestra en las figuras 2 y 3, respectivemente. El mejor sistema de aprendizaje será aquel que produzca una mayor área bajo la superficie convexa (AUC). La realización del detector basado en información mutua ha sido comparado con el detector

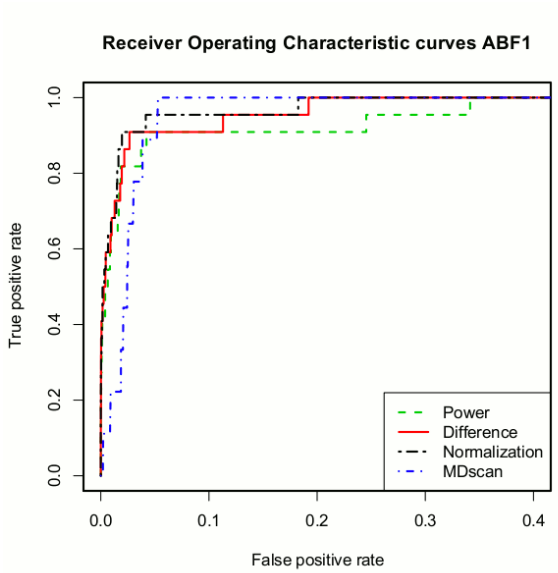

Fig. 2. Curvas ROC para ABF1 según diferentes funcionales

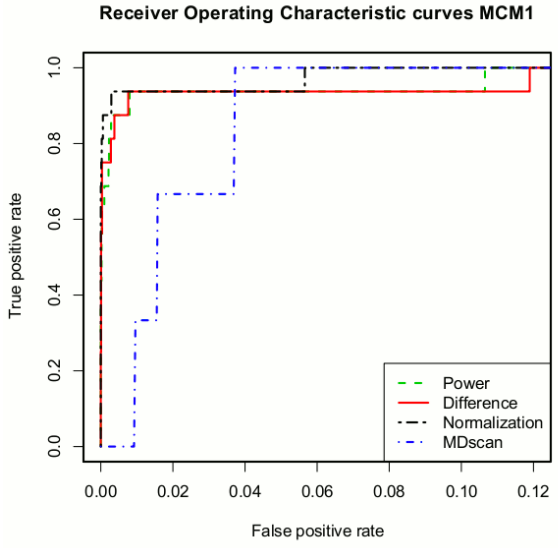

Fig. 3. Curvas ROC para MCM1 según diferentes funcionales

MDscan [3].

En la tabla II, se puede observar que el detector tiene diferentes comportamientos según el funcional utilizado. Además, el valor de la área bajo la superficie convexa mediante el método de información mutua es más cercano a 1 que mediante MDscan, excepto para la función *power* en ABF1. Por lo tanto, asumir la dependencia entre posicionesa través de la medida de información mutua ayuda a mejorar los resultados obtenidos mediante MDscan para estos ejemplos.

En las figuras 2 y 3 se observa como el número de verdaderos positivos, TP, y falsos positivos, FP, depende de la secuencia de unión del factor de transcripción y del

TABLE II
AUC

	MCM1	ABF1
Difference	0.9916	0.9801
Power	0.9923	0.9650
Normalization	0.9962	0.9848
MDscan	0.9793	0.9754

funcional considerado (e.g. dado un número de verdaderos positivos, el número de falsos positivos varia según la función considerada). El mejor funcional puede ser seleccionado para el uso final a partir del criterio de coste establecido: pérdida de clasificación de los verdaderos positivos y área máxima bajo la superficie convexa.

IV. Conclusiones

En este trabajo, hemos presentado una metodología para detectar las secuencias de unión de los factores de transcripción (TFBS). Este método esta basado en la variación de la información mutua cruzada entre posiciones a partir de un conjunto de secuencias de unión. El algoritmo propuesto ha sido aplicado en la detección de los factores de transcripción *ABF1* y *MCM1* a partir de una secuencia genómica aleatoria. La información mutua proporciona información adicional relacionada con el proceso de unión, como la correlación entre posiciones de la secuencia de unión. El método propuesto se comporta mejor que MDscan, método basado en la enumeración de palabras combinadas y PWM, en el caso de la clasificación de una secuencia de unión contra una secuencia generada aleatoriamente. Futuros trabajos extenderán el estudio de la dependencia entre las posiciones en la secuencia de unión mediante divergencias paramétricas.

References

[1] D. Latchman, *Eukaryotic Transcription Factors*, 5th ed. Academic Press, 2007.

[2] R. Mutihac, A. Cicuttin, and R. Mutihac, "Entropic approach to information coding in DNA molecules," *Materials Science & Engineering C*, vol. 18, no. 1-2, pp. 51–60, 2001.

[3] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments." *Nat Biotechnol*, vol. 20, no. 8, pp. 835–839, Aug 2002. [Online]. Available: http://dx.doi.org/10.1038/nbt717

[4] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements." *Nat Rev Genet*, vol. 5, no. 4, pp. 276–287, Apr 2004. [Online]. Available: http://dx.doi.org/10.1038/nrg1315

[5] T. D. Schneider, "Information content of individual genetic sequences." *J Theor Biol*, vol. 189, no. 4, pp. 427–441, Dec 1997. [Online]. Available: http://dx.doi.org/10.1006/jtbi.1997.0540

[6] G. D. Stormo, "Dna binding sites: representation and discovery." *Bioinformatics*, vol. 16, no. 1, pp. 16–23, Jan 2000.

[7] J. Maynou, M. Vallverdu, F. Claria, A. Perera, and P. Caminal, "Detection of transcription factor binding sites using r&yi entropy," in *Proc. 8th IEEE International Conference on BioInformatics and BioEngineering BIBE 2008*, 8–10 Oct. 2008, pp. 1–5.

[8] T. D. Schneider, G. Stormo, L. Gold, and A. Ehrenfeuch, "The information content of binding sites on nucleotide sequences," *J Mol Biol*, vol. 188, pp. 415–431, Nov 1986.

[9] A. Perera, M. Vallverdu, F. Claria, J. M. Soria, and P. Caminal, "Dna binding site characterization by means of rényi entropy measures on nucleotide transitions." *IEEE Trans Nanobioscience*, vol. 7, no. 2, pp. 133–141, Jun 2008. [Online]. Available: http://dx.doi.org/10.1109/TNB.2008.2000744