Predicción de sitios de unión de Factores de Transcripción utilizando información estructural

Vladimir Espinosa Angarica[1,2], Abel González Pérez[3], Ana Tereza Ribeiro de Vasconçelos[4], Julio Collado-Vides[2], Bruno Contreras-Moreira[2, 5]

[1]Departamento de Bioquímica y Biología Molecular y Celular, Universidad de Zaragoza, España; [2]Programa de Genómica Computacional, Centro de Ciencias Genómicas, UNAM, Cuernavaca, México; [3]Centro Nacional de Bioinformática, Ciudad de la Habana, Cuba; [4]Laboratorio Nacional de Computación Científica, Petrópolis–Río de Janeiro, Brasil; [5]Estación Experimental *Aula Dei*, Zaragoza, España

Contactos: vespinosa@gmail.com; bcontreras@eead.csic.es

INTRODUCCIÓN

El reconocimiento de los elementos de regulación en *cis* por los Factores de Transcripción (TF) es uno de los procesos determinantes de la regulación de la expresión coordinada de los genes en los organismos vivos. Es por esto que en los últimos años se ha dedicado especial interés al estudio de los mecanismos que determinan la afinidad de los TFs por sus secuencias de unión, con el objetivo de desarrollar metodologías computacionales de predicción de sitios en secuencias genómicas. La mayor parte de las aproximaciones utilizadas para predecir sitios de unión de TF están basadas en modelos estadísticos generados a partir de alineamientos múltiples de los sitios conocidos caracterizados experimentalmente. Estos modelos solo tienen en cuenta la información secuencial sobre la estructura primaria de los sitios reconocidos por un determinado TF; sin embargo el reconocimiento de un sitio de unión en un contexto genómico está mediado por interacciones que se establecen a nivel tridimensional entre determinantes atómicos situados en la región de contacto del TF y el sitio de unión.

En este trabajo describimos una nueva metodología de predicción de sitios de unión de TF basada en información estructural de la interacción proteína–DNA a nivel atómico. Nuestro método combina la información relacionada con la lectura directa del DNA, relacionada con la interacción específica entre átomos específicos de la proteína y el sitio de unión, y la lectura indirecta determinada por la deformación del DNA provocada por la unión del TF. Estas contribuciones estructurales son sintetizadas en un único modelo estadístico que puede ser utilizado para estimar la fracción de energía relacionada con el reconocimiento específico del complejo TF–DNA. Se generaron modelos para 11 TF bacterianos de diferentes familias estructurales y fueron utilizados para identificar sitios para estos TF en secuencias genómicas. La fortaleza predictiva de nuestro método fue comprobada para sitios conocidos determinados experimentalmente y predicciones obtenidas por otras metodologías y estos resultados demuestran la importancia de la utilización de información estructural para predecir sitios de unión de TF en secuencias genómicas.

MÉTODOS

• Se construyeron matrices de frecuencias de contactos atómicos para interacciones por puente de hidrógeno e interacciones hidrofóbicas a partir de un set no redundante de 237 complejos proteína–DNA extraídos del PDB. Estas matrices de preferencias de interacción a nivel atómico fueron convertidas en las correspondientes matrices de peso (PWM) utilizando la siguiente fórmula, en las cuales se representa la contribución energética (*i.e.* expresada en valores de probabilidades) de cada determinante atómico de las proteínas a interactuar con otro grupo determinado del sitio de unión[1].

$$W_{ij} = \ln \frac{\left(n_{ij} + p_{ab}\right) / \left(N_a + 1\right)}{p_{ab}}$$

• Se desarrolló una herramienta computacional (DNAPROT)[2, 3] para hacer "*threading*" de una determinada secuencia de DNA en la estructura 3D de un complejo proteína–DNA y estimar la energía de unión luego de la sustitución. En la estimación de la energía de unión tuvimos en cuenta la contribución de la lectura directa del DNA (*i.e.* relacionada con la interacción específica entre átomos específicos de la proteína y el sitio de unión), que fue estimada utilizando las PWM descritas arriba; y la lectura indirecta (*i.e.* relativa a la deformación del DNA provocada por la unión del TF) y que fue estimada utilizando el programa X3DNA[4]. Esta información fue sintetizada en nuestro modelo estadístico utilizando la siguiente fórmula la cual permite generar una nueva PWM con información de ambos tipos de lectura del DNA.

$$W_{bj} = \ln\left(\frac{\left(prob_{bj} + k_w\right) / \left(Np_j + 4k_w\right)}{Bf_b}\right)$$

• Los valores de *score* generados por nuestro método al hacer *threading* de una secuencia determinada en un complejo TF–DNA es un estimador de la energía estática de unión según la siguiente formula[5]

$$\Delta E_s = -RT \times Score_s$$

RESULTADOS

a) Estos resultados demuestran que la exclusión de un número considerable de las entradas utilizadas en la extracción de las reglas de preferencia atómica, no afecta significativamente la capacidad de discriminación de las matrices obtenidas, lo cual se puede inferir de la buena superposición de las distribuciones de *scores* de las poblaciones de entradas de los sets de *bootstrap* así como las excluidas de los mismos, Figura 1 **A)**. Esto significa que las preferencias de contactos obtenidas en el proceso de construcción de las matrices son significativas desde el punto de vista biológico y no son un artefacto producido por la redundancia del set de entrenamiento ni por el sobreentrenamiento.

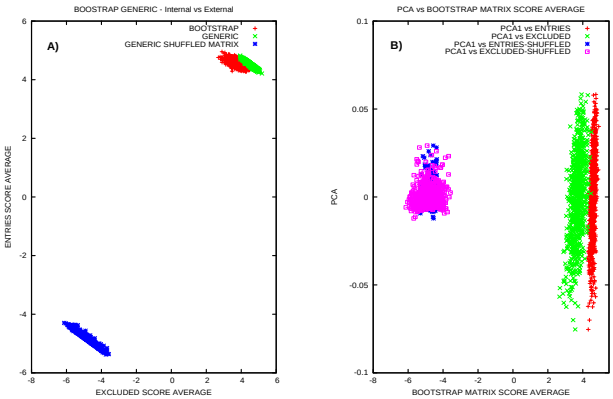

Figura 1: Análisis de "*bootstrap*" del set de entrenamiento utilizado en la construcción de las matrices de preferencia de contactos atómicos proteína–DNA. A partir del set de entrenamiento inicial se generaron 1000 subconjuntos de entrenamiento excluyendo aleatoriamente el 30% de las entradas iniciales. Para cada ensayo de *bootstrap* se generaron un par de PWM las cuales fueron utilizadas para evaluar las entradas que contribuyeron a su generación y a las que fueron excluidas. En el gráfico **A)** se muestra la evaluación de las poblaciones de *bootstrap* con las matrices genéricas, de *bootstrap* y una matriz de campos aleatorizados. En **B)** se muestra el análisis de componentes principales (PCA) de estos datos

b) Tal y como ha sido descrito en estudios previos[6, 7], las distribuciones de scores obtenidas con nuestra metodología para cada uno de los 11 TFs durante un *scanning* de regiones sin sentido biológico y sitios conocidos, son distribuciones normales, tal y como se observa en la Figura 1 para el caso del TF CRP. En todos los casos la distribución de scores correspondiente a los sitios reconocidos constitutivamente por el TF se encuentra desplazada hacia el extremo derecho, coincidiendo con los scores más altos. Para todos los TF estas distribuciones comienzan a separarse a partir de los scores positivos, donde es posible establecer valores de corte para identificar sitios en secuencias genómicas

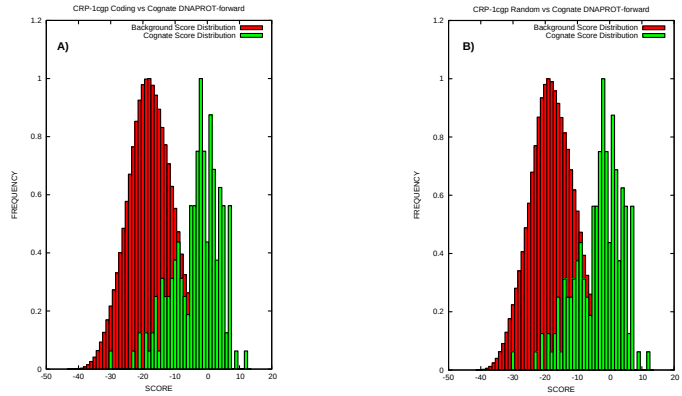

Figura 2: Distribución de frecuencias de los *scores* durante un *scanning* de un set de secuencias "*non–sense*" y un grupo de sitios de unión de un TF. Un set de secuencias donde no se deben encontrar sitios de unión funcionales para los TF –*i.e.* non–sense– así como el grupo de sitios reconocidos por un TF y caracterizados experimentalmente fueron *scaneados* con nuestro modelo. En el gráfico de arriba **A)** se realizó utilizando como secuencias de *background* las regiones codificantes del genoma y en **B)** se utilizó un set de secuencias generadas aleatoriamente a partir de las frecuencias AT:CG del genoma de *E. coli* utilizando el modelo estadístico obtenido para el factor de transcripción CRP

c) Los resultados obtenidos en los ensayos de identificación de sitios conocidos para todos los TF, como se puede ver en la Figura 3 para los ejemplos específicos de PhoB, NarL y CRP, corresponden a valores muy buenos de razones de sensibilidad/especificidad. Esto está relacionado con una buena fortaleza predictiva de nuestro método, el cual permite identificar las secuencias que comprobadamente se unen a un determinado TF con un buen compromiso entre los Falsos Negativos/Falsos Positivos

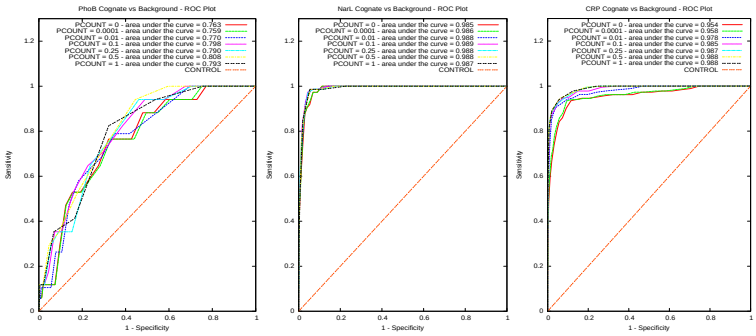

Figura 3: Curvas ROC de la identificación de los sitios conocidos para PhoB, NarL y CRP en un set de secuencias aleatorias respectivamente. Utilizando los modelos estadísticos correspondientes a cada uno de los TF se *scanearon* los sets de secuencias conocidas y random para valores crecientes de *score* y se calcularon los valores de sensibilidad y especificidad.

DISCUSIÓN

- Uno de los problemas al trabajar con subconjuntos del PDB está relacionado con el sesgo de esta base de datos hacia proteínas con plegamientos específicos y a una gran cantidad de información redundante. Nuestros resultados demuestran que las matrices de interacción atómicas obtenidas contienen información relevante desde el punto de vista biológico sobre las propensidades de interacción de determinantes específicos de las proteínas y el DNA. La Figura 1 **A)** demuestra que aún con la exclusión de hasta un 30% de las entradas del set de entrenamiento inicial, las matrices generadas retienen la capacidad de evaluar las energías de unión en las entradas de los conjuntos de *bootstrap* y las entradas excluidas del análisis. El análisis de componentes principales (PCA) permitió demostrar la existencia de relaciones en los valores de scores obtenidos para las poblaciones de complejos proteína–DNA de los subconjuntos incluidos y excluidos de la generación de las matrices
- Nuestro método reportó buenos resultados en ensayos de identificación de sitios conocidos en conjuntos de secuencias genómicas y aleatorias. Las distribuciones de score obtenidas, Figura 2, correspondieron a distribuciones con marcada tendencia a separarse para valores altos de score. Esta propiedad esta relacionada con la capacidad de discriminación del TF (señal/ruido) durante el proceso de reconocimiento de los sitios en el contexto genómico
- Los valores de Sensibilidad/Especificidad obtenidos por nuestro método para la identificación de sitios conocidos para TF en secuencias genómicas sin sentido biológico, son superiores a los obtenidos en otros trabajos en los que se utilizaron PWM para hacer las predicciones. En estos reportes los valores de Sensibilidad/Especificidad para estos TFs oscilan entre 74.4%/79.6% para CRP, 53.8%/97.8& para NarL y 50%/99.6% for PhoB[8, 9].

Referencias

1. Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563-77.
2. Lozada-Chavez, I., Angarica, V.E., Collado-Vides, J., Contreras-Moreira, B. DNA binding specificity predicts hierarchy in bacterial regulatory networks. (submitted)
3. Angarica, V.E., Perez, A.G., Vasconcelos, A.T., Collado-Vides, J., Contreras-Moreira, B. A structure-based methodology for predicting transcription factor binding sites in complete genomes. (in preparation)
4. Lu, X. J. & Olson, W. K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **31**, 5108-21.
5. Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859-83.
6. Kono, H. & Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **35**, 114-31.
7. Sarai, A. & Kono, H. (2005). Protein-DNA recognition patterns and predictions. *Annu Rev Biophys Biomol Struct* **34**, 379-98.
8. Gonzalez, A.D., Espinosa, V., Vasconcelos, A.T., Perez-Rueda, E., Collado-Vides, J. (2005). TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. *Nucl. Acids Res* **33**, D98-102.
9. Perez, A.G., Angarica, V.E., Vasconcelos, A.T., Collado-Vides, J. (2007). Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes. *Nucl. Acids Res* **35**, D132-6

Universidad de Zaragoza

Facultad de Ciencias