Caracterización y análisis de las interacciones de regulación entre los factores de transcripción y los genes

Joan Maynou, Erola Pairó, Raimon Massanet, Montserrat Vallverdú, Pere Caminal y Alexandre Perera

***Abstract*— El estudio y la comprensión de las redes de interacción entre proteínas es fundamental para entender el funcionamiento de los diferentes procesos biológicos a nivel celular. El conjunto de interacciones entre proteínas, definido como interactoma, es muy complejo debido al número y a los diferentes tipos de interacciones existentes. En este contexto, estudiar las interacciones de regulación entre proteína y ácido desoxirribonucleico (Factor de Transcripción-ADN) es importante para comprender el nivel de expresión de los genes afectados. El principal objetivo de este trabajo es la caracterización desde el punto de vista estadístico de los factores de transcripción que regulan un gen y de los genes regulados por un factor de transcripción. Los datos han sido obtenidos de la base de datos String [1] y de la aplicación de minería de datos de SabioSciences [2]. El trabajo se centra en las interacciones de regulación TF-gen para el organismo *Homo sapiens*.**

I. Motivación

Las proteínas son biomoléculas esenciales para el desarrollo de los procesos biológicos tanto a nivel celular como a nivel de sistemas. Generalmente, las proteínas no actúan de forma independiente, establecen interacciones dinámicas entre más de una proteína para llevar a cabo las funciones biológicas [1], por ejemplo los factores de transcripción. Los factores de transcripción (TF) son proteínas muy específicas vinculadas al proceso de transcripción de un gen. Los factores de transcripción se unen de forma combinatorial a otros factores de modulación y al ácido ribonucleico, RNA, polimerasa. Dicho complejo proteíco se une al promotor, región de ADN que controla la iniciación de la transcripción, para iniciar la transcripción del gen a RNA mensajero, mRNA.

Las interacciones entre los factores de transcripción, pertenecen a un grupo específico de interacciones de co-regulación. A nivel general, las asociaciones entre proteínas (ppi) se clasifican según el tipo de interacción y la actividad que desarrollan: co-interacción, co-regulación y co-localización. La interacción de co-interacción corresponde a una asociación directa entre proteínas. La co-regulación consiste en la asociación entre proteínas que desarrollan una misma actividad biomolecular pero que no interactúan directamente. La co-localización es la asociación entre proteínas que actúan en el mismo compartimiento celular [2]. Cada una de estas interacciones están divididas en diferentes categorías. La asociación física está dividida en interacciones transitorias o permanentes según la duración de la interacción. Las interacciones de co-regulación se dividen en metabólicas o genéticas. Concretamente, las interacciones de regulación genéticas corresponde a las asociaciones entre factores de transcripción. Finalmente, la asociación de co-localización se divide en localización de membrana y localización soluble.

Los datos obtenidos, tanto a nivel experimental como computacional, sobre las interacciones entre proteínas se encuentran almacenados en grandes repositorios o bases de datos. Dichas bases de datos pueden ser clasificadas según la procedencia de los datos [1]: Bases de datos primarias, meta bases de datos y bases de datos de predicción. En las bases de datos primarias, los datos almacenados son datos experimentales. Las principales bases de datos son BioGRID [3], DIP [4], MINT [5] y IntAct [6]. En las meta bases de datos, las ppi provienen de las bases de datos primarias. Finalmente, en las bases de datos de predicción, los datos provienen de las principales bases de datos primarias y, además, hay datos predichos. La principal base de datos de predicción es String [7].

Las bases de datos simplifican, en gran medida, el análisis de los distintos tipos de datos [8]. La información contenida es analizada mediante diferentes aproximaciones computacionales. La teoría de grafos es el método comúnmente usado para visualizar y extraer la información inherente entre la ppi. Las proteínas individuales son modelizadas como vértices y las interacciones directas, obtenidas experimentalmente, corresponden a las aristas del grafo. Diferentes algoritmos, basados en teoría de grafos, han sido desarrollados para estudiar la información que se encuentra en el interactoma [9], [10]. La característica común en todos estos algoritmos es que sólo consideran las interacciones físicas o directas entre proteínas. Añadir a los estudios existentes las interacciones de regulación entre TF-gen, además de las interacciones directas, permitiría ampliar la información inherente que se obtiene del interactoma.

Este trabajo pretende caracterizar desde una perspectiva estadística los genes regulados por un factor de transcripción

Este trabajo se ha realizado con el soporte del Ministerio Español de Educación y Ciencia mediante el programa de la Ramón y Cajal y TEC2010-20886-C02-02 y el CIBER-BBN.

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

J. Maynou, R. Massanet, M. Vallverdú, P. Caminal y A. Perera como miembros de CIBER de Bioingeniería, Biomateriales y Nanomedicina. http://www.isciii.es/htdocs/redes/ciber.jsp

Erola Pairó pertenece al Institut de Bioenginyeria de Catalunya, IBEC,y al Departament d'Electrònica, Universitat de Barcelona, Avinguda Diagonal, 647 08028 Barcelona, España. e-mail: epeiroibec.pcb.ub.es

[1] http://string-db.org/

[2] http://www.sabiosciences.com/

TABLA I

RESUMEN DEL PAQUETE STRINGSABIO

Función	Origen	Definición
idString	String	Extracción ID
intString	String	Extracción PPI
bioString	String	Extracción TF
intSabioSciences	SabioSciences	Extracción TF
SabiotoString	String/Sabio	Homogenización ID
StringSabio	String/Sabio	Eliminación Información Redundante

TABLA II

RESUMEN DE LAS ENTRADAS PAQUETE STRINGSABIO

Función	Entrada
idString	(nombre Gen, taxonomía)
intString	(identificador, umbral significancia
	máximo nodos, taxonomía)
bioString	(Interacciones)
intSabioSciences	(nombre Gen, organismo)
SabiotoString	(Identificador, nombre Gen, taxonomía)
StringSabio	(outString, outSabioSciences)

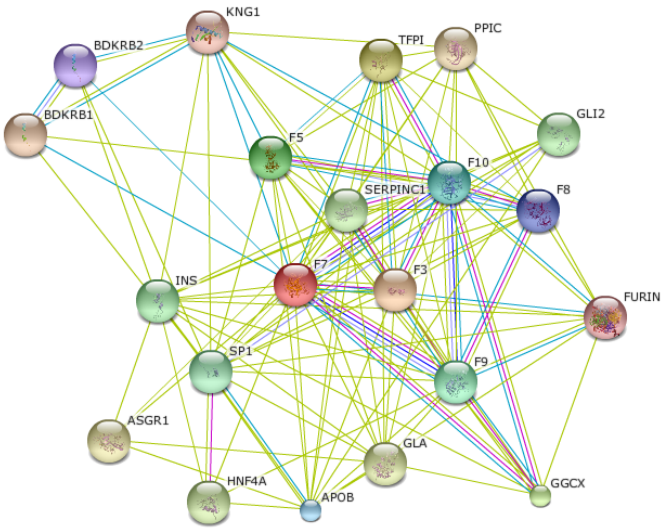

Fig. 1. Grafo de las interacciones directas e indirectas para el F7 según la base de datos String 9.0.

y los factores de transcripción que regulan un gen para el organismo *Homo sapiens*. Concretamente, se ha construido una librería en R [3] [11] StringSabio [4] para extraer las interacciones por regulación genética presentes en las bases de datos String [7] y de la aplicación de búsqueda de factores de transcripción de SabioSciences. La extracción de los datos se realiza de forma automática, remota y no redundante.

II. PAQUETE: STRINGSABIO

El paquete StringSabio es una librería en R para extraer los factores de transcripción presentes en la transcripción de los genes (interacción TF-gen) en la base de datos String y en la aplicación de SabioSciences. Contiene diferentes funciones para extraer los datos de forma independiente, remota y no redundante en cada una de las bases de datos, ver Tabla IV. Cada función tiene sus propias entradas, ver Tabla II. Concretamente, a partir del nombre del gen, la taxonomía del organismo de estudio y concatenando la salida de cada una de las funciones, se obtienen los factores de transcripción vinculados en la regulación de dicho gen. Además, el paquete incluye una amplia documentación y ejemplos de consulta. El paquete StringSabio está disponible en http://sisbio.recerca.upc.edu/R/StringSabio.1.0.tar.gz.

A. Ejemplo

Para ejecutar el paquete StringSabio se requiere el siguiente conjunto de librerías de R.

```
>library(RCurl)
>library(XML)
>library(string)
```

En el siguiente ejemplo, se realiza la búsqueda de los factores de transcripción presentes en las transcripción del FVII para el organismo *Homo sapiens* mediante el paquete StringSabio. Se consideran, solamente, las interacciones de regulación y se amplia el conjunto de interacciones con los datos provenientes de SabioSciences, ver Tabla III. En cambio, la misma búsqueda de forma directa, mediante web, de las interacciones de regulación en la base de datos String considerando 20 interacciones y un umbral de significancia de 0.150, da como resultado el grafo de la Fig. 1.

TABLA III

RESULTADO FVII PARA EL PAQUETE STRINGSABIO

Gen	Factor de Transcripción	Interacción	Origen
FVII	HNF4G	Regulación	String
FVII	SP1	Regulación	String
FVII	HNF4A	Regulación	String/Sabio
FVII	BATF	Regulación	SabioSciences

Mediante este ejemplo, se visualiza como, a través del paquete en R StringSabio, se ha seleccionado solamente los factores de transcripción relacionados con la regulación del gen considerado, obviando las interacciones restantes.

III. DESCRIPCIÓN BASE DE DATOS

Para construir el conjunto completo de interacciones TF-gen para el organismo *Homo sapiens*, se han considerado todos los genes conocidos del *Homo sapiens*, 22812, de la base de datos NCBI [5] [12]. A partir de dichos datos, se han extraído todos los factores de transcripción de las bases de datos String y de la aplicación de minería de datos de SabioSciences.

String es una base de datos de interacciones entre proteínas conocidas y predichas. String Versión 9.0 incluye interacciones directas o físicas e indirectas o funcionales. Las interacciones entre proteínas provienen de bases de datos primarias como BioGRID, DIP, IntAct, MINT,...y de datos predichos. String contiene alrededor de 5 millones de proteínas de 1133 organismos, con más de 100 millones de interacciones reportadas. Todos estos datos provienen de

[3]http://www.r-project.org/

[4]http://sisbio.recerca.upc.edu/R/StringSabio.1.0.tar.gz

[5]http://www.ncbi.nlm.nih.gov/gene

TABLA IV

RESUMEN DE LOS DATOS OBTENIDOS EN STRING Y SABIOSCIENCES

Número Interacciones	String	SabioSciences	String\SabioSciences
193882	103152	89986	744

TABLA V

RESUMEN DEL NÚMERO DE FACTORES DE TRANSCRIPCIÓN POR GEN

	Número Genes
$1 \leq TF < 5$	1809
$5 \leq TF < 10$	7003
$10 \leq TF < 15$	3147
$15 \leq TF < 20$	1245
$20 \leq TF < 30$	1030
$30 \leq TF < 50$	665
$50 \leq TF < 100$	324
$100 \leq TF < 600$	61
$600 \leq TF < \infty$	0

cuatro fuentes diferentes: contexto genómico, experimentos, coexpresión y conocimiento predicho.

SabioSciences, versión del Junio del 2011, es un buscador de factores de transcripción a partir de un gen en cuestión. SabioSciences combina una aplicación de minería de datos, basada en la extracción de relaciones entre proteínas presentes en artículos científicos, con los datos de UCSC Genome para compilar una base de datos de más de 200 factores de transcripción para todos los genes del genoma humano.

El conjunto de datos obtenidos de las bases de datos fueron descargados el 8 de Junio del 2011. En la Tabla IV, se visualiza el número de interacciones que se han obtenido en cada una de las bases de datos.

IV. RESULTADOS

El número de interacciones entre TF-gen reportadas en la base de datos String y de SabioSciences es de 193882, ver tabla IV. Concretamente, 103152 interacciones de regulación están reportadas en la base de datos String y 89986 interacciones de regulación provienen de la aplicación de SabioSciences. Solamente 744 interacciones de regulación están reportadas a la vez en String y SabioSciences.

El grado de interacciones de regulación es diferente según el gen , ver Fig. 2. Se han clasificado los genes por el número de factores de transcripción necesarios para su regulación, ver Tabla V. El grupo comprendido entre 5 y 10 factores de transcripción es el que contiene un mayor número de genes, 7003. Concretamente, el grupo de 8 factores de transcripción es el que tiene más genes. Este grupo está formado por 1773 genes. En cambio, el gen con un mayor número de factores de transcripción es $TAF1$ con 591 factores de transcripción necesarios para su regulación. El inicio de la transcripción del gen $taf1$ está regulada por la unión entre el ácido ribonucleico (ARN) polimerasa II y el resto de polipéptidos, más de 70.

Además, se han clasificado los elementos de regulación, factores de transcripción, por el número de genes que regulan, ver Tabla VI. Se visualiza que la mayoría de factores de transcripción regulan entre 1 y 5 genes, ver Fig. 3.

TABLA VI

RESUMEN DEL NÚMERO DE GENES REGULADOS POR FACTOR DE TRANSCRIPCIÓN

	Número TF
$1 \leq Gen < 5$	467
$5 \leq Gen < 10$	178
$10 \leq Gen < 15$	128
$15 \leq Gen < 20$	112
$20 \leq Gen < 30$	220
$30 \leq Gen < 50$	331
$50 \leq Gen < 100$	404
$100 \leq Gen < 600$	353
$600 \leq Gen < 4300$	51

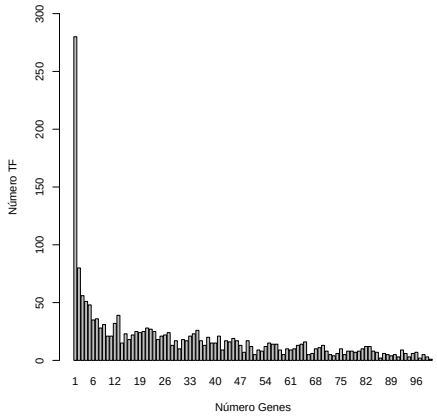

Fig. 2. Histograma del número Genes por Factores de Transcripción.

Concretamente, 280 factores de transcripción regulan un sólo gen. Por otra parte, el elemento que regula más genes es $C17orf64$ (cromosoma 17 open reading frame 64) con 2150.

En la figura 4, se muestra la variabilidad de la distribución de genes respecto a los factores de transcripción. Se obseva que el número de factores de transcripción que regula un gen crece rápidamente hasta alcanzar un valor estacionario, alrededor de 10 factores de transcripción por gen.

V. CONCLUSIÓN

StringSabio es un paquete en R construido para la obtención de las interacciones entre TF-Gen para la base de datos String y de la aplicación SabioSciences. El paquete permite la ejecución de cada una de las funciones de forma independiente, secuencial y no redundante. Permitiendo trabajar solamente con String, con SabioSciences o con las dos aplicaciones a la vez. Dado un organismo y un gen, el paquete extrae el conjunto de factores de transcripción para dicho gen obviando las demás interacciones o proteínas. A partir de todo el conjunto de interacciones entre TF-Gen, se observa que el grado de interacciones de regulación es diferente según el gen. El grupo formado por los genes que interaccionan con 8 factores de transcripción es el más numeroso con 7003 genes. En cambio, el gen con más interacciones de regulación, 591, es $taf1$, gen regulado por la acción de más 70 polipéptidos, juntamente con el ARN

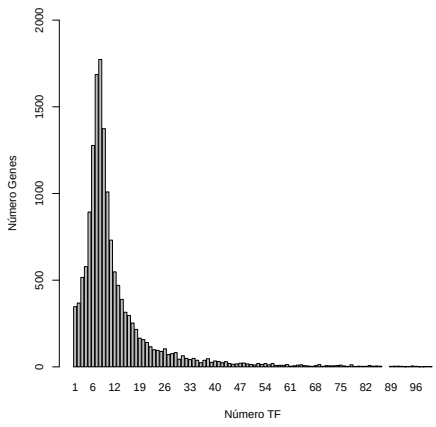

Fig. 3. Histograma del número de Factores de Transcripción por Gen.

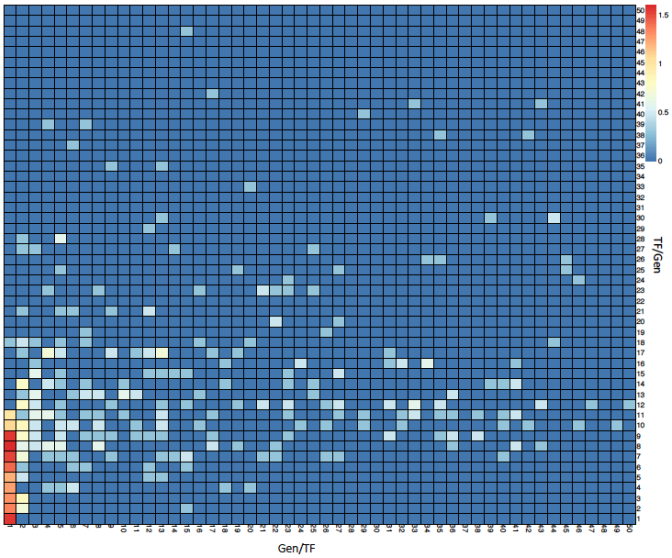

Fig. 4. Heatmap de la Distribución de Genes respecto los Factores de Transcripción

polimerasa II. Por otra parte, la mayoría de factores de transcripción regulan entre 1 a 5 genes. Concretamente, el grupo más numeroso, con 280, es el grupo de factores de transcripción que regulan solamente una gen. En cambio, el factor de transcripción que regula más genes es $C17orf64$ con 2150 genes. Además, se visualiza que el número de factores de transcripción que regulan un gen crece hasta alcanzar un régimen estacionario, próximo a 10 factores de transcripción por gen.

VI. ACKNOWLEDGMENTS

CIBER de Bioingeniería, Biomateriales y Nanomedicina es una iniciativa de ISCIII.

REFERENCES

[1] J. Rivas and C. Fontanillo, "Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks," *Plos Computational Biology*, vol. 6, 2010.

[2] J. Rivas and A. Luis, "Interactome data and databases: different types of protein interaction," *Comp. Funct. Genom.*, vol. 5, pp. 173–178, 2004.

[3] B. Breitkreutz, C. Stark, T. Reguly, L. Boucher, and et al., "The biogrid interaction database." *Nucleic Acids Res*, vol. 36, pp. D637–D640, 2008.

[4] L. Salwinski, C. Miller, A. Smith, F. Pettit, B. J.U., and et al., "The database of interacting proteins," *Nucleic Acids Res*, vol. 32, pp. D449–D451, 2004.

[5] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, and et al., "Mint, the molecular interaction database," *Nucleic Acids Res*, vol. 38, pp. D532–D539, 2009.

[6] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, and et al., "The intact molecular interaction database in 2010," *Nucleic Acids Res*, vol. 38, pp. D525–D531, 2010.

[7] C. Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, "String: a database of predicted functional associations between proteins," *Nucleic Acids Res*, vol. 31, pp. 258–261, 2003.

[8] L. Salwinski and D. Eisenberg, "Computational methods of analysis of protein–protein interactions," *Comp. Funct. Genom.*, vol. 13, pp. 377–382, 2002.

[9] R. Massanet-Vila, P. Caminal, and A. Perera, "Graph theory-based measures as predictors of gene morbidity," in *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2010.

[10] M. Deng, S. Mehta, and F. Sun, "Inferring domain-domain interactions from protein-protein interactions," *Genome Res.*, vol. 12, pp. 1540–1548, 2002.

[11] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[12] D. Maglott, J. Ostell, K. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi." *Nucleic Acids Res*, vol. 35 (Database issue), pp. D26–31, 2007.