Medidas basadas en teoría de grafos y la predicción de la morbosidad de genes

R. Massanet Vila[1,2,3], P. Caminal Magrans[1,2,3], A. Perera Lluna[1,2,3]

[1]Dept. ESAII, Universitat Politècnica de Catalunya (UPC), Barcelona, España;
{raimon.massanet, pere.caminal,alexandre.perera}@upc.edu
[2]Centre de Recerca en Enginyeria Biomèdica (CREB), Barcelona, España;
[3]CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), España;

Resumen

Estudios previos sugieren que las redes de interacción entre proteínas presentan propiedades de la teoría de grafos que pueden tener cierta relación con la morbosidad de los genes. En particular, se ha sugerido que cuando un polimorfismo afecta a un gen, es más probable que se produzca una enfermedad si el grado de ese gen en una red de interacción entre proteínas es elevado. Sin embargo, estos resultados no tienen en cuenta el posible sesgo introducido en los datos por la variación en la cantidad de información que se tiene sobre los diferentes genes. En este trabajo se intenta modelar la morbosidad de genes como una combinación lineal de los grados de los nodos en redes de interacción entre proteínas y la cantidad de información sobre genes disponible en la literatura. Un conjunto de 7461 *genes y* 3665 *enfermedades reportadas en la base de datos* Online Mendelian Inheritance in Man *(OMIM) fue utilizado conjuntamente con una red de interacciones entre proteínas de* 9630 *nodos y* 38756 *interacciones de la* Human Proteome Resource Database *(HPRD). La cantidad de información disponible para cada gen se ha medido minando la base de datos* PubMed. *Los resultados sugieren que la correlación entre el grado de un nodo en la red de interacciones entre proteínas y la morbosidad del gen que el nodo representa es consecuencia, al menos en una parte considerable, de la variación en la cantidad de información disponible para los diferentes genes. Aunque los resultados sugieren una correlación positiva entre el grado de un nodo y su morbosidad, los autores creen que esta correlación debe ser considerada con precaución puesto que podría estar afectada por factores que no se consideraron en este estudio.*

1 Introducción

Los métodos de alto rendimiento de procesamiento para la identificación de proteínas, como *yeast two-hybrid* [1], *high-throughput mass-spectometric protein complex identification (HMS-PCI)* [2], *tandem affinity purification (TAP)* [3], *correlated mRNA expression* y otros, han permitido la construcción, en los últimos años de grandes redes de interacción entre proteínas (RIP) con una fiabilidad relativamente elevada. Aunque los grafos tienen limitaciones importantes a la hora de modelar RIPs, llevan usándose de forma amplia y reiterada para ese fin [4, 5]. Por consiguiente, la teoría de grafos ha sido aplicada al estudio de RIPs para el descubrimiento de sus propiedades de red características. Un esfuerzo particularmente grande ha sido dirigido al hallazago de relaciones entre propiedades de los grafos que representan RIPs y la morbosidad de los genes. Algunos autores han argumentado que la morbosidad está relacionada con la distribución de los grados de los nodos en RIPs. La idea tras esa afirmación es que mutaciones en nodos de alta conectividad podrían causar una disrupción severa en la red.

En [6] los autores afirman que las RIPs, como otras redes reales, tienen una topología *libre de escala*. Este tipo de redes se caracterizan por tener pocos nodos de grado elevado y muchos nodos de grado bajo. Las redes con topología libre de escala son muy robustas frente a errores aleatorios, pero son vulnerables a errores en los nodos centrales (nodos de grado elevado). Estudios realizados sobre organismos simples sugieren que el grado de los nodos en RIPs puede estar asociado con la letalidad de los genes, teniendo los genes letales un grado mayor que los genes no letales [7]. También se ha hallado evidencia de que los genes letales corresponden a genes de grado elevado que además provocan una desconexión en la RIP cuando son eliminados [8]. Estos resultados fortalecen la idea de que la morbosidad de los genes es consecuencia de su rol central en la red proteómica, independientemente de su función biológica.

Por otra parte, la comunidad científica tiende a dedicar un mayor esfuerzo al estudio de genes de morbosidad conocida, así como su entorno, en busca de otros genes que modulen o interaccionen con los genes patológicos. Este hecho podría causar un sesgo en la cantidad de información sobre interacciones entre proteínas disponible para los diferentes genes, teniendo los genes patológicos un número mayor de interacciones reportadas como consecuencia de la mayor atención que la comunidad científica les ha dedicado.

Esto podría contribuir en un efecto causal entre la morbosidad y el grado de un gen, y no al revés.

Este trabajo pretende profundizar en la posible relación entre morbosidad y grado de un gen, teniendo en cuenta la cantidad de información. Para poder estudiar esta relación correctamente, la varianza en los grados de los nodos debería ser ajustada, controlando la variación en la cantidad de información publicada sobre los genes que los nodos representan.

En esta contribución, esto se ha aproximado utilizando un modelo lineal que relaciona de forma estadísticamente significativa la morbosidad de un gen con el grado del nodo correspondiente, aislando la varianza causada por la variación en la cantidad de información disponible.

2 Materiales y métodos

La base de datos *Online Mendelian Inheritance in Man* (OMIM) fue minada para obtener una estimación de la morbosidad de un gen. Los datos de OMIM (*morbid map*) establecen una relación entre fenotipos humanos de origen genético reportados en la literatura y el conjunto de genes que han sido asociados a ellos. La morbosidad de un gen se estimó como el número de enfermedades con las que un gen ha sido asociado. El *morbid map* usado en este trabajo fue descargado de OMIM el 5 de Febrero de 2010. Estos datos relacionan 7461 genes diferentes con 3665 identificadores OMIM (enfermedades o fenotipos).

La base de datos *Human Proteome Resource Database* (HPRD) [9] fue minada con el objetivo de recopilar información de interacciones entre proteínas. Estos datos se obtuvieron a través del sitio web HPRD, versión del 6 de Julio de 2009. Los datos fueron transformados a estructura de grafo no dirigido de 9630 nodos y 38756 aristas.

El servicio web de *PubMed* fue masivamente consultado para obtener una estimación de la cantidad de información que la comunidad científica tiene sobre los diferentes genes. Esta medida se estimó como el número de identificadores de publicaciones diferentes obtenidos al consultar un gen determinado.

De los 9630 nodos del grafo de interacciones entre proteínas, se encontró el correspondiente símbolo de gen para 9374. Para cada uno de los símbolos genéticos se calcularon tres medidas: el grado del nodo correspondiente en el grafo de interacciones, el número de identificadores OMIM (morbosidad), y el número de identificadores PubMed (cantidad de información). Solamente para 1873 nodos del grafo se halló al menos un identificador OMIM asociado.

Para estudiar la relación entre morbosidad y grado se generaron dos muestras. La primera (caso) compuesta por los grados de los 1873 genes con morbosidad mayor que 0. La segunda (control) compuesta por una selección aleatoria del mismo tamaño muestral de los grados de los genes sin morbosidad conocida. La diferencia entre las dos muestras fue medida mediante un test de Mann-Whitney [10].

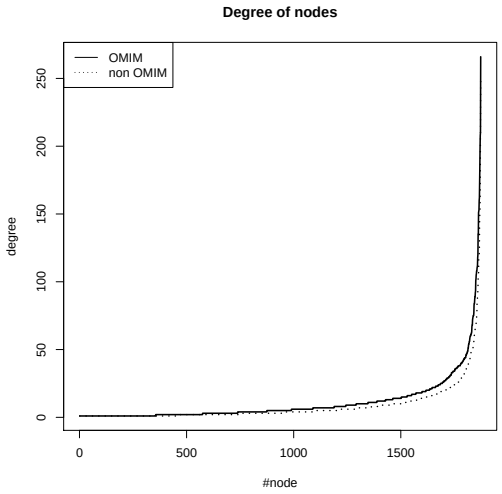


Figura 1: *Distribución de los grados de los nodos correspondientes a genes con morbosidad mayor que* 0 *(linea continua) y genes sin morbosidad conocida (linea de puntos). Las diferencias halladas fueron estadísticamente significativas, con un p-valor máximo de* $6.72e^{-10}$.

Para estudiar más a fondo la influencia en la morbosidad de un gen, el grado medio y la cantidad media de información fueron calculados para cada valor de morbosidad. A continuación se construyó un modelo lineal para cuantificar la influencia ejercida. La morbosidad fue usada como variable de respuesta, mientras que la cantidad media de información y el grado medio fueron usados como variables explicativas.

Todos las tareas de minado de bases de datos y cálculo se realizaron usando el lenguaje de programación estadística R [11].

3 Resultados

Los resultados muestran diferencias estadísticamente significativas entre los grados de genes con morbosidad mayor que 0 y genes sin morobosidad conocida (ver Figura 1), con un p-valor máximo de $6.72e-10$. Este resultado es coherente con estudios previos realizados en otros organismos [7], que sugieren que la morbosidad de un gen puede estar relacionada con el número de interacciones reportadas para la proteína que el gen codifica. A pesar de que este argumento parece lógico e intuitivo, no se ha considerado el efecto ejercido por la variante cantidad de información que la comunidad científica tiene sobre los diferentes genes.

Los genes fueron agrupados en función del número de enfermedades asociadas a ellos. La Figura 2 muestra la distribución de los grados de los nodos para las diferentes categorias. A pesar de que se observa un valor de correlación de Pearson relativamente bajo de 0.20,

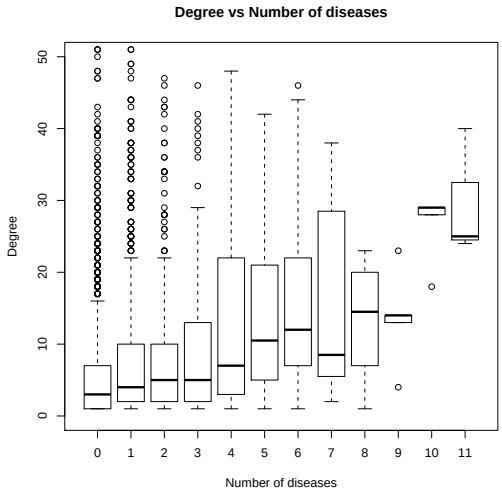


Figura 2: *Grado medio de los nodos en función del número de enfermedades o fenotipos con los que los genes correspondientes han sido relacionados. Los datos sugieren una correlación positiva entre el número de interacciones de un gen y el número de fenotipos asociados a él.*

parece evidente que hay una correlación positiva. Sin embargo, esta correlación podría estar afectada por el hecho de que genes de morbosidad conocida tienden a ser más estudiados, como se ha dicho anteriormente. La Figura 3 muestra la distribución del número de publicaciones para los genes en las diferentes categorías. En este caso también parece evidente que hay una correlación positiva entre las dos variables. El valor de correlación entre ellas es ligeramente más elevado, 0.26.

Con el objetivo de determinar el efecto de la varianza de la cantidad de información se normalizaron los grados de los nodos por el número de publicaciones en las que aparecen los genes correspondientes. La Figura 4 muestra que cuando el número de interacciones es normalizado de esta forma la correlación positiva con el número de enfermedades ya no es tan evidente y el valor de correlación de Pearson cae a -0.12.

Se construyó un modelo lineal, como se explicó en la sección 2, para segregar la varianza introducida por la cantidad de información y el grado de los genes y estudiarlas por separado. Se calculó el modelo descrito por la siguiente ecuación:

$$M(g) = \alpha \cdot I(g) + \beta \cdot D(g) + \gamma \qquad (1)$$

donde $M(g)$ representa la morbosidad del gen g, $I(g)$ es la cantidad media de información disponible y $D(g)$ el grado medio del nodo correspondiente en la red de interacciones entre proteínas.

Las tablas 1 y 2 muestran los valores de regresión obtenidos por el modelo. Para comprobar que los residuos del modelo seguían una distribución normal se aplicó una prueba de Kolmogorov-Smirnov, cuyo p-valor fue de 0.81. La distribución normal de los

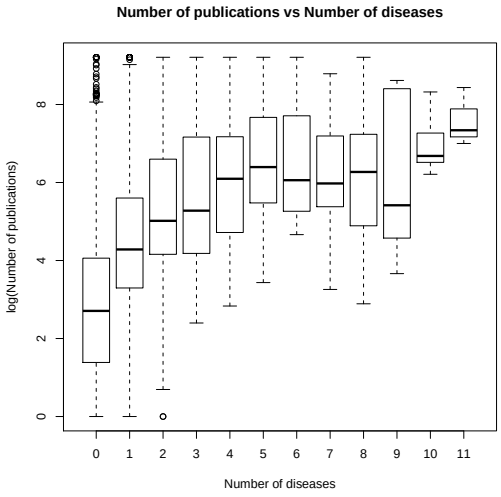


Figura 3: *Número medio de publicaciones (en escala logarítmica) por nodo en función del número de enfermedades asociadas. Los datos sugieren una fuerte correlación positiva entre la cantidad de información disponible sobre un gen y el número de enfermedades asociadas a él.*

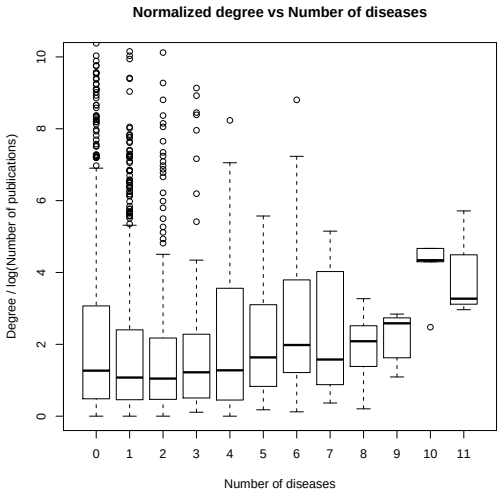


Figura 4: *Grados de los nodos, normalizados por la cantidad de información, en función de la morbosidad de los genes correspondientes. La correlación entre el número de interacciones y la morbosidad cae de forma significativa cuando la variación en la cantidad de información es considerada. Este resultado sugiere que la cantidad de información, modelada aquí por el número de publicaciones, puede tener un efecto importante en esta relación.*

Min	1Q	Mediana	3Q	Max
-2.25	-1.43	-0.03	1.43	2.58

Tabla 1: *Residuos del modelo lineal. Los residuos parecen seguir una distribución normal, con un p-valor de* 0.81 *en una prueba de Kolmogorov-Smirnov de dos colas.*

	Estimado	Err. est.	t-valor	$\Pr(>\|t\|)$
γ	-1.75	1.48	-1.18	0.27
$I(g)$	$3.73e^{-3}$	$9.60e^{-4}$	3.89	$3.67e^{-3}$
$D(g)$	$8.29e^{-2}$	$3.41e^{-2}$	2.43	$3.77e^{-2}$

Tabla 2: *Coeficientes del modelo lineal. Tanto el grado medio como la cantidad media de información por enfermedad son estadísticamente significativos. Sin embargo, la cantidad de información muestra una significación estadística un orden de magnitud mayor que el grado. La regresión lineal tiene una significación de* $1.41e^{-3}$.

residuos indica que el modelo es aplicable a los datos. El bajo p-valor del modelo ($1.41e^{-3}$) sugiere éste que se ajusta satisfactoriamente a los datos. La Figura 5 muestra algunas medidas de calidad que refuerzan la confianza en los resultados del modelo lineal. La Figura 5a muestra que los residuos estandarizados se ajustan a los cuantiles teóricos. Así mismo, la Figura 5b muestra que todos los puntos tienen una distancia de Cook [12] baja, lo cual indica que ningún punto causa un cambio importante en la pendiente de la recta de regresión.

La significación estadística para la cantidad de información es un orden de magnitud mayor (p-valor menor) que para el grado de los genes. Esto sugiere que el efecto producido por la variación en la cantidad de información es más significativo que el producido por la variación en el grado de los genes. Aún así, es interesante notar que el p-valor asociado al grado es significativo independientemente del efecto ejercido por la cantidad de información. Este resultado indica que aún cuando se controla el efecto de la variación en la cantidad de información, se observa un efecto considerable en la variable de respuesta que el modelo atribuye al grado de los genes. Además, el coeficiente obtenido para la variable grado es un orden de magnitud mayor, indicando que dada la misma cantidad de información en dos genes, el número de enfermedades asociadas a ellos crece con relativa celeridad respecto de su grado.

4 CONCLUSIÓN

Los resultados sugieren que la relación entre el grado de un nodo en una red de interacciones entre proteïnas y la morbosidad del gen correspondiente no es tan evidente como puede parecer. Parece haber un sesgo inherente debido a la variación en la cantidad de información disponible en la literatura científica sobre los diferentes genes. Genes relacionados con enfermedades aparecen con más frecuencia en la literatura, puesto que son de mayor interés para la comunidad clínica. Además, se buscan con mayor ahínco proteínas que interaccionen con genes de morbosidad conocida, puesto que son los objetivos más evidentes a la hora de buscar efectos moduladores o nuevos genes candidatos. A pesar de que los resultados sugieren una correlación positiva entre el grado de un nodo y la morbosidad del gen correspondiente, esta relación debería ser considerada con mucha cautela, pues podría estar influenciada por otros factores no considerados en este estudio.

5 AGRADECIMIENTOS

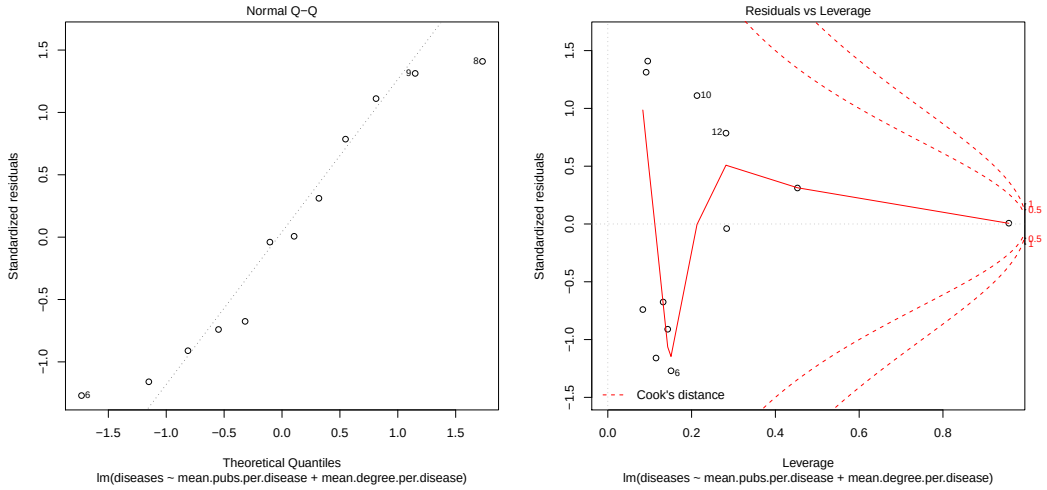
Los autores agradecen el apoyo recibido por parte del Ministerio de Educación y Ciencia a través del programa Ramón y Cajal y TEC2007-63637/TCM así como del Insitituto de Salud Carlos III a través de la iniciativa CIBER-BBN en Bioingeniería , biomateriales y nanomedicina.

Referencias

[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623–627, 02/10 2000.

[2] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers, "Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 01/10 2002.

[3] A.-C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester,



(a) (b)

Figura 5: *(a) Residuos del modelo en función de los valores predichos. (b) Distancia de Cook de los datos ajustados por el modelo.*

P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 01/10 2002.

[4] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 10/20 2005.

[5] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, "Global landscape of protein complexes in the yeast saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 03/30 2006.

[6] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 10/05 2000.

[7] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 05/03 2001.

[8] N. Przulj, D. A. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, no. 3, pp. 340–348, February 12 2004.

[9] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human protein reference database–2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–772, January 1 2009.

[10] H. Mann and D. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[11] R Development Core Team, "R: A language and environment for statistical computing," 2009. [Online]. Available: http://www.R-project.org

[12] R. Cook and S. Weisberg, *Residuals and influence in regression*. New York: Chapman and Hall, 1982.