Conservación de polimorfismos relacionados con enfermedades humanas

H. Brunel[1,2], J.J. Gallardo-Chacón[3], M. Vallverdú[1], P. Caminal[1], A. Perera[1]

[1]Dept. ESAII, Universitat Politècnica de Catalunya (UPC), Barcelona, España;
{helena.brunel, alexandre.perera, joan.josep.gallardo, montserrat.vallverdu, pere.caminal}@upc.edu
[2]Institut de Bioenginyeria de Catalunya (IBEC), España;
[3]CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), España

Resumen

Los estudios de genómica comparativa indican que el análisis de la homología entre secuencias de diferentes especies puede aportar información útil para el estudio de enfermedades humanas. En particular, se conoce que las variantes genéticas biológicamente relevantes se conservan más que otro tipo de secuencias. Las variantes genéticas relacionadas con enfermedades se obtienen a través de estudios de asociación genética. Generalmente, estos estudios se centran en variantes del tipo SNP (Polimorfismo de Nucleótido Simple) pero presentan ciertas dificultades a la hora de la replicación y validación de resultados. La presencia de falsos positivos induce a la búsqueda de criterios de priorización de SNPs, por ejemplo en función de su relevancia biológica. En este trabajo se propone un criterio basado en la conservación de secuencias homólogas entre especies. Se ha comparado la conservación de secuencias para dos muestras de SNPs, formadas por SNPs relacionados con enfermedades, indexados en la base de datos OMIM, y por SNPs neutros seleccionados de forma aleatoria. Se han obtenido diferencias significativas entre los dos grupos, con un p-valor de 0.0016, presentando mayor grado de conservación los SNPs relacionados con enfermedades que los SNPs de control.

1 Introducción

Uno de los mayores retos de la investigación genética es la identificación de factores relacionados con enfermedades humanas mediante el estudio de la relación entre genotipos y enfermedades. Se conoce que la variabilidad genética entre individuos sólo representa un 1 % del genoma humano. Esta proporción corresponde a los diferentes tipos de variantes genéticas responsables de características individuales como los rasgos físicos, la susceptibilidad a contraer ciertas enfermedades o la respuesta a tratamientos médicos [1]. En particular, los polimorfismos de nucleótido simple (SNPs) son la fuente de variación genética más utilizada para el estudio de enfermedades humanas [2]. Un SNP es una posición en el genoma donde se produce una mutación, generalmente la substitución de una base por otra, que se ha conservado de generación en generación, logrando una frecuencia de más de un 1% de la población.En la actualidad, se han validado aproximadamente 3 millones de SNPs (*rs* SNP o reference SNP) y se estima que existen alrededor de 10 millones en el genoma humano (*ss* SNP o Submitted SNP) de los que sólo una pequeña proporción estan relacionados con enfermedades [3].

Los SNPs relacionados con enfermedades se encuentran mediante estudios de asociación genética. Los resultados de estos estudios son difíciles de replicar y/o validar y, a menudo, presentan falsos positivos. Por eso, la búsqueda de criterios para priorizar la relevancia biológica de los SNPs está ganando protagonismo en la investigación genética [4]. La ubicación del SNP respecto al gen, la característica del SNP en cuanto a cambios funcionales en la proteína o la conservación del SNP a lo largo de las especies son posibles criterios de priorización de los SNPs como medida de soporte a los estudios de asociación [5, 6]. En particular, el estudio de la conservación de las secuencias entre especies es de gran interés para la identificación de marcadores genéticos asociados con enfermedades [7]. Los estudios de genómica comparativa han demostrado que gran parte del genoma humano es común con otras especies [8]. Gracias a la disponibilidad de múltiples secuencias genómicas de organismos *modelo*, estos estudios aportan información sobre la relevancia selectiva de las variantes genéticas desde un punto de vista evolutivo [9]. Por otro lado, se cree que las secuencias funcionales tienden a evolucionar más lentamente y por tanto son mas conservadas que secuencias menos relevantes [10, 11, 12, 13].

Los SNPs relacionados con enfermedades y cuya asociación con la enfermedad ha sido confirmada se pueden encontrar en bases de datos públicas como la Online Mendelian Inheritance in Man (OMIM) [14]. La conservación de las secuencias entre especies se mide a partir de herramientas para buscar homologías entre secuencias. Dos secuencias son homólogas si son muy similares y provienen del mismo antecesor [10, 15]. Las medidas de conservación entre secuencias

homólogas se basan en "scores" de similitud [16]. Una de ellas es la entropía de Shannon [17]. La entropía es una medida de desorden introducida por C. Shannon en el campo de las telecomunicaciones [18] y que ha sido aplicada a la bioinformática, sobretodo en estudios de análisis y alineamiento de secuencias [19]. El objetivo de este trabajo es establecer un criterio basado en homologías para validar los resultados obtenidos en estudios de asociación genética. La hipótesis del trabajo establece que los SNPs relacionados con enfermedades presentan mayor conservación entre especies que los SNPs menos relevantes. Para ello se ha realizado un estudio estadístico, comparando la conservación de secuencias de SNPs en una muestra de SNPs indexados en la OMIM y una muestra de SNPs neutros seleccionados de manera aleatoria.

2 Materiales y Métodos

En la figura 1 se muestra de manera esquematica los diferentes estadios del diseño experimental así como las herramientas utilizadas en cada etapa del proceso.

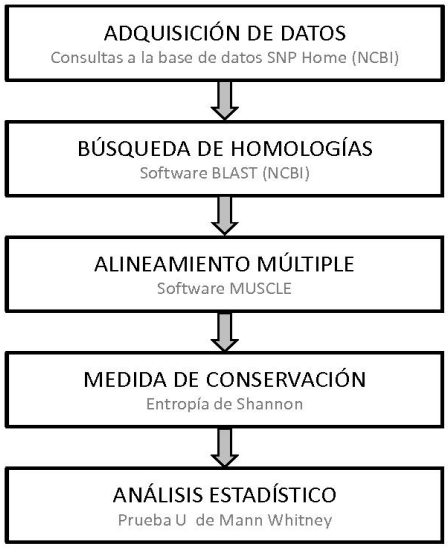

Figura 1. *Etapas de la metodología de trabajo*

Los datos se han obtenido de la base de datos SNP Home de NCBI [20]. Esta base de datos contiene información sobre *rs* SNPs. En particular, para cada SNP, se almacenan datos como el organismo afectado, el tipo de polimorfismo, la mutación o la ubicación del SNP respecto al gen, entre otros. Uno de los campos hace referencia a la relación del SNP con enfermedades humanas, en función de si el SNP aparece en la base de datos OMIM [14]. El motor de búsqueda de la base de datos SNP Home permite hacer consultas de SNPs con diferentes criterios basados en estos atributos. Así pues, se han seleccionado los polimorfismos de tipo SNP, del organismo *Homo sapiens* e indexados en la OMIM. Esta muestra está compuesta por 3547 SNPs, para los que se ha almacenado la secuencia vecina del SNP, una secuencia de tamaño variable conteniendo el SNP, y su ubicación respecto al gen (en un exón, intrón, u otras regiones). Esta muestra se comparará con una muestra de control, formada por 3547 SNPs neutros seleccionados de forma aleatoria mediante consultas a la base de datos SNP Home con la restricción de conservar las proporciones de SNPs de las diferentes clases, factor que puede afectar la conservación de los SNPs y que podría alterar los resultados. En este caso se han clasificado los SNPs según si estan en un exón, en un intrón, o en otras regiones (promotores, UTR (Untranslated Regions), ...). Otra característica de la segunda muestra de SNPs es que está formada por SNPs que no aparecen en el OMIM, para poder aplicar tests estadísticos de comparación de muestras.

En segundo lugar, se han buscado secuencias homólogas a la secuencia de cada SNP. Se ha utilizado un algoritmo de BLAST, un software libre del NCBI [21]. En particular se ha aplicado el algoritmo *blastn* de forma automática para todos los SNPs. El algoritmo *blastn* parte de una secuencia de nucleótidos (la secuencia del SNP) y busca homologías en una base de datos de secuencias de nucleótidos para todas las especies. Este algoritmo depende del parámetro E, también conocido como E-valor, que determina la significación estadística de la homología obtenida. En este caso se ha trabajado con $E = 10^{-3}$, un valor suficientemente significativo y que permite obtener un cierto grado de variabilidad en los datos. El algoritmo de BLAST también permite limitar el número de secuencias homólogas de la salida y seleccionar el tipo de especies *modelo*. En este caso, no se ha especificado ninguna restricción para estos parámetros.

Una vez obtenido el conjunto de secuencias homólogas, se ha aplicado un algoritmo de alineamiento de secuencias. Se ha trabajado con el software MUSCLE [22]. Para optimizar la rendibilidad del algoritmo de alineamiento de múltiples secuencias, se han recortado las secuencias a 100 nucleótidos de longitud, centrados en el SNP. La ventaja de MUSCLE es que permite modificar ciertos parámetros para obtener un mejor rendimiento (precisión y velocidad de cálculo) que otros algoritmos. En este caso, dado que en algunos casos se han obtenido muchas secuencias homólogas, se ha reducido a 2 el número máximo de iteraciones en el alineamiento.

Dada una matriz de secuencias homólogas alineadas, la entropía para la posición correspondiente a un SNP

(S) se define en 1.

$$H = -\sum_{i=0}^{N} p(S_i) \cdot log_2 p(S_i) \quad (1)$$

donde N es el número de símbolos posibles (A, T, C, G) para el SNP S y $p(S_i)$ es la probabilidad de obtener el símbolo i en la columna correspondiente al SNP S de la matriz de secuencias alineadas. La entropía mide el grado de desorden presente en la variable. Valores altos de la entropía corresponden a posiciones muy variables y por tanto poco conservadas mientras que valores bajos de la entropía se asocian con un alto grado de conservación, siendo nula para SNPs totalmente conservados.

Finalmente, se ha efectuado un análisis estadístico sobre las entropías. Se han comparado las entropías para los SNPs del OMIM y para la muestra aleatoria mediante una prueba no paramétrica U de Mann Whitney para muestras independientes. La hipótesis nula establece la no diferencia en la medida de entropía para las dos muestras, mientras que la hipótesis alternativa establece que los SNPs indexados en el OMIM presentan menor entropía y por tanto más conservación que la muestra de SNPs de control. Se ha utilizado una prueba de una cola con un nivel de significación estadística de 0.05.

3 Resultados

En la comparación de la conservación de los SNPs, se han obtenido diferencias significativas en el valor de la entropía para las dos muestras, con un p-valor de 0.0016 en la prueba de Mann-Whitney.

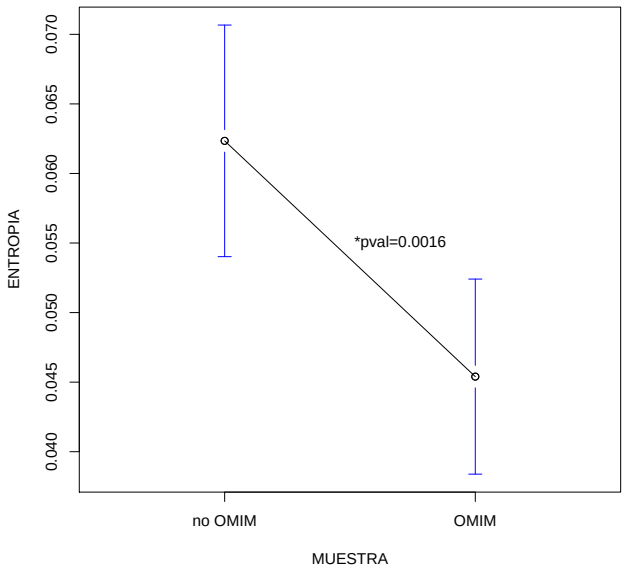

Figura 2. *Comparación de la distribución estadística de la entropía de los SNPs de ambas muestras y p-valor de la prueba U de Mann-Whitney*

Se observa en la figura 2 que la muestra de SNPs relacionados con enfermedades (OMIM) presenta una entropía estadísticamente inferior a la de la muestra aleatoria (noOMIM). Es decir que los SNPs relacionados con enfermedades son más conservados que los SNPs de control. Se puede observar, también, que los valores de entropía son muy bajos, seguramente debido a la presencia de SNPs altamente conservados. En particular, las mutaciones que causan enfermedades suelen encontrarse en regiones exónicas, donde producen modificaciones en la función de la proteína resultante. Estas zonas suelen ser muy conservadas. En la tabla 1 se presentan la media y desviación estándar de la entropía de los SNPs de las diferentes regiones.

Tabla 1. *media y desviación estándar de la entropía de los SNPs de diferentes regiones génicas. N representa el número de SNPs de cada grupo. Diferentes letras representan diferencias significativas entre muestras*

Región	N	OMIM $\mu \pm sd$	noOMIM $\mu \pm sd$
EXÓN	2537	0.0276 ± 0.1504 $_a$	0.0783 ± 0.2621 $_b$
INTRÓN	674	0.1045 ± 0.2997 $_a$	0.0415 ± 0.1875 $_b$
Otras Regiones	336	0.0634 ± 0.2243 $_a$	0.0001 ± 0.0033 $_b$

Se observa que la gran mayoría de SNPs indexados en la OMIM pertenecen a exones. Los SNPs de estas regiones presentan valores de entropía bajos debido a que son más conservados. Las mayores diferencias estadísticas entre muestras se observan para esta región, con un p-valor de $1.29*10^{-13}$. Se observa que los SNPs indexados en la OMIM presentan menor entropía y por tanto mayor conservación que los SNPs de control. Las mutaciones en regiones exónicas pueden implicar un cambio en la secuencia de aminoácidos resultante (mutaciones no sinónimas) o pueden ser "silenciosas" y no producir cambios funcionales en la proteína resultante. Los SNPs exónicos relacionados con enfermedades suelen ser SNPs no sinónimos altamente conservados mientras que en la muestra aleatoria pueden aparecer un mayor número de SNPs silenciosos, que podría hacer subir la media del valor de entropía. Para las otras regiones, se observa que los SNPs relacionados con enfermedades están menos conservados que los SNPs de control. Estas diferencias pueden ser debidas a que se tiene poco conocimiento sobre la funcionalidad de las zonas no codificantes (intrones, promotores, etc.) ya que la mayoría de estudios de genómica evolutiva se centran en secuencias de DNA codificante. Esto implica además que el número de secuencias homólogas obtenidas en el paso 2 de la metodología sea pequeño y facilite la obtención de entropías nulas. Sin embargo, el número de SNPs encontrados en estas regiones es poco representativo. Así pues, los resultados obtenidos permiten aceptar la hipótesis de trabajo con un nivel de confianza del 95% y se puede afirmar que SNPs relacionados con enfermedades genéticas situados en regiones exónicas presentan mayor conservación a lo largo de las especies que SNPs de control.

4 Conclusiones

La metodologia propuesta en este trabajo pretende demostrar que las mutaciones en posiciones que estan relaciondas con patologías genéticas están más conservadas a lo largo de las especies. Mediante una prueba U de Mann- Whitney se ha podido validar la hipótesis de trabajo, que establece que SNPs relacionados con enfermedades, indexados en la base de datos OMIM, presentan más conservación que SNPs neutros selecionados aleatoriamente. La diferencia entre muestras es especialmente importante para SNPs en regiones codificantes. Estos resultados concuerdan con resultados propuestos en la literatura, donde se ha demostrado que los SNPs funcionales, y en particular no sinónimos, se conservan a lo largo de las especies. Esta metodología ha sido propuesta para validar los resultados de estudios de asociación genética, estableciendo un criterio que prioriza los SNPs ubicados en regiones altamente conservadas, ya que tienen más posibilidades de estar relacionados con enfermedades.

5 Agradecimientos

Este trabajo ha sido parcialmente financiado por la CICYT TEC2007-63637/TCM y por el Instituto de Ingeniería Biomédica de Cataluña (IBEC), así como por el programa Ramón y Cajal del ministerio de Educación y Ciencia y las redes temáticas de investigación Cooperativa (RETIC) Cardiovascular (RECAVA) Exp-06/0014/0016RD del Fondo de Investigación Sanitaria (FIS). El CIBER en Bioingeniería, Biomateriales y Nanomedicina es una iniciativa del ISCIII.

References

[1] D. E. Reich, S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler, "Human genome sequence variation and the influence of gene history, mutation and recombination," *Nature genetics*, vol. 32, no. 1, pp. 135–142, Sep 2002.

[2] S.-C. Su, "Single nucleotide polymorphism data analysis - state-of-the-art review on this emerging field from a signal processing viewpoint," pp. 75–82, 2007.

[3] The International HapMap Consortium, "The international hapmap project," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[4] J. Ioannidis, P. Boffetta, J. Little, T. O'Brien, A. Uitterlinden, P. Vineis, D. Balding, A. Chokkalingam, S. Dolan, W. Flanders, J. Higgins, M. McCarthy, D. McDermott, G. Page, T. Rebbeck, D. Seminara, and M. Khoury, "Assessment of cumulative evidence on genetic associations: interim guidelines," *International Journal of Genetic Epidemiology*, vol. 37, pp. 120–132, 2008.

[5] P.Bhatti, D. Church, J. Rutter, J. Struewing, and A. Sigurdson, "Candidate single nucleotide polymorphism selection using publicly available tools: A guide for epidemiologists," *American Journal of Epidemiology*, vol. 164, no. 8, pp. 794–804, 2006.

[6] D. Burke, C. Worth, E. Priego, T. Cheng, L. Smink, T. J.a, and T. Blundell, "Genome bioinformatic analysis of nonsynonymous SNPs," *BMC Bioinformatics*, vol. 8, p. 301, 2007.

[7] J. L. McCauley, S. Kenealy, E. Margulies, N. Schnetz-Boutaud, S. Gregory, S. L. Hauser, J. Oksenberg, M. Pericak-Vance, J. Haines, and D. Mortlock, "SNPs in multi-species conserved sequences (MCS) as useful markers in association studies: a practical approach," *BMC Genomics*, vol. 8, p. 266, 2007.

[8] R. Hardison, "Comparative genomics," *PLoS Biology*, vol. 1, no. 2, p. e58, 2003.

[9] H. Huang, E. Winter, H. Wang, K. Weinstock, H. Xing, L. Goodstadt, P. Stenson, D. Cooper, D. Smith, M. Albà, C. Ponting, and K. Fechtel, "Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes," *Genome Biology*, vol. 5, no. 7, p. R47, 2001.

[10] K. Frazer, L. Elnitski, D.M.Church, I. Dubchak, and R. Hardison, "Cross-species sequence comparisons A review of methods and available resources," *Genome Research*, vol. 13, no. 1, pp. 1–12, 2009.

[11] C. Loots, R. Locksley, C. Blankespoor, Z. Wang, W. Miller, E. Rubin, and K. Frazer, "Identification os a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons," *Science*, vol. 288, pp. 136–140, 2000.

[12] S. Mottagui-Tabar, M. Faghili, Y. Mizuno, P. Engström, B. Lenhard, W. Wasserman, and C. Wahlestedt, "Identification of functional SNPs in the 5-prime flanking sequences of human genes," *BMC Genomics*, vol. 6, p. 18, 2005.

[13] Y. Zhu, M. Spitz, C. Amos, J. Lin, M. Schabath, and X. Wu, "An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology," *Cancer Research*, vol. 64, pp. 2251–2257, 2004.

[14] "The Online Mendelian Inheritance in Man database," http://www.ncbi.nlm.nih.gov/omim.

[15] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[16] N. Stojanovic, L. Florea, C. Riemer, D. Gumucio, J. Slightom, M. Goodman, W. Miller, and R. Hardison, "Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions," *Nucleic Acids Research*, vol. 27, no. 19, p. 3899–3910, 1999.

[17] K.Wang and R. Samudrala, "Incorporating background frequency improves entropy-based residue conservation measures," *BMC Bioinformatics*, vol. 7, p. 385, 2006.

[18] C. Shannon, "A mathematical theory of communication," *The Bell Systems Technical Journal*, vol. 27, pp. 379–423, 1948.

[19] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic acids research*, vol. 18, no. 20, pp. 6097–6100, Oct 25 1990.

[20] "SNP Home, National Center for Biotechnology Information," http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp.

[21] "Basic Local Alignment Search Tool," http://blast.ncbi.nlm.nih.gov/Blast.cgi.

[22] R. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–97, 2004.