Trabajo de Fin de Máster - PEC1 Plan de Trabajo

Jorge Vallejo Ortega

16/03/2021

Endo-Mining: herramienta web para la búsqueda automatizada de genes potencialmente relacionados con la endometriosis a través de minería de textos en PubMed

1. Contexto y Justificación del Trabajo

1.1. Descripción general:

Se desarrollará una aplicación web interactiva para descubrir información contenida en los abstracts de las publicaciones científicas en Pubmed mediante el uso de minería de textos. Se aplicará a la búsqueda de genes relacionados con la endometriosis.

1.2 Justificación del TFG:

La endometriosis es una enfermedad del sistema reproductor femenino caracterizada por crecimiento de tejido endometrial ectópico fuera del útero. Afecta a alrededor del 10% de mujeres en edad reproductora (Malvezzi et al. 2020). Sus principales síntomas son infertilidad y dolor (Rolla 2019), y se ha asociado con depresión y fatiga (Chapron et al., 2019); por lo que supone una importante disminución en su calidad de vida del paciente.

No existe ninguna prueba genética que identifique personas con mayor riesgo de desarrollar la enfermedad, ni biomarcadores que permitan un diagnóstico fiable. Sin embargo, podemos acceder a una amplia literatura científica que explora la relación entre genes y endometriosis (Rolla 2019) y, aunque se estima una heredabilidad de entre 0.27 a 0.51, no se ha demostrado relación directa entre este y ninguno de los genes candidatos. Se le

supone una etiología multifactorial en la que intervendrían diferentes factores genéticos y ambientales (Sapkota, 2017).

PubMed es un recurso en línea de acceso público y gratuito consistente en una base de datos - en continuo crecimiento - que incluye más de 32 millones de citas y abstracts de literatura biomédica, tanto artículos (MEDLINE y PubMed Central) como libros (Bookshelf) (National Library of Medicine, 2021). La inmensa cantidad de información contenida en la base de datos hace que incluso búsquedas restrictivas recuperen muchas veces cientos (o miles) de citas relevantes, un flujo de información difícilmente digerible por el investigador mediante los métodos tradicionales de lectura y análisis de artículos individuales.

Las técnicas de minería de textos permiten a los investigadores de las áreas biomédicas un acceso efectivo y eficiente al conocimiento enterrado en las ingentes cantidades de literatura publicada, además de suplementar la información extraída mediante minería de datos a partir de otras fuentes de datos masivos como la secuenciación de genomas, datos de expresión génica y de estructuras proteicas. Ambas funciones permiten acelerar la investigación biomédica (Aggarwal, 2012).

En resumen, en este trabajo estamos abordando el estudio de una enfermedad - la endometriosis - que afecta negativamente la calidad de vida de un importante porcentaje de la población mundial. Lo hacemos centrándonos en un aspecto todavía muy desconocido como es su relación con la genética, y usando una herramienta analítica y exploratoria - la minería de textos - que permite extraer información de una enorme fuente de datos de acceso público pero poco estructurada como son los abstracts de los artículos científicos. Desarrollando una aplicación web que permita a cualquier persona reproducir el proceso de forma sencilla y automática, estamos contribuyendo dentro de nuestras posibilidades al crecimiento del conocimiento científico y técnico al exponer de forma pública los resultados de nuestro trabajo y el modo en el que hemos llegado hasta ellos.

2. Objetivos

2.1 Objetivo general

- Encontrar genes relacionados con la endometriosis aplicando técnicas de minería de textos.

2.2 Objetivos específicos

1. Desarrollar un script que permita realizar un procedimiento de minería de textos automáticamente, desde la recopilación de datos en bruto hasta la presentación de resultados.

2. Desarrollar una aplicación web implementando el script de minería de textos que resultó del objetivo anterior.

3. Enfoque y método a seguir

Como señalábamos anteriormente, se describe la endometriosis como una enfermedad de etiología poco clara, dependiente de factores ambientales y genéticos. La heredabilidad de la enfermedad no sigue un patrón mendeliano, así que su origen es muy probablemente multifactorial e influenciado por un número desconocido de genes. Existen varias estrategias al alcance de los investigadores que permiten el descubrimiento de genes y regiones genómicas asociadas a enfermedades multifactoriales.

Entre los métodos más antiguos están los tests de desequilibrio de ligamiento (Spielman, 1996), en los que se estudia la asociación entre la enfermedad y un marcador genético cercano a un gen candidato. Tiene la desventaja de que depende de la existencia de genes candidatos y de marcadores muy cercanos a dichos genes.

Otro método consiste en el uso de microarrays para polimorfismos de nucleótido único (SNPs). Éste permite comparar genotipos de pacientes de la enfermedad en una población, frente a sujetos control de la misma población; identificando variantes genómicas asociadas a la enfermedad (Mafra, 2016). Facilita el estudio simultáneo de muchas regiones genómicas, aunque limitadas en número por el espacio físico de la microarray.

Los estudios de asociación de genoma completo (GWAS) se basan en la comparación de genomas secuenciados completos entre individuos que manifiestan la enfermedad e individuos control, al igual que ocurre en el método de microarrays. A diferencia de aquél, el espacio de búsqueda no se limita a las secuencias que pueden incluirse en una microarray,

sino a todo el genoma que ha podido secuenciarse. Tiene la desventaja de que, debido a la enorme cantidad de secuencias que forma un genoma, necesita gran número de muestras para que la potencia de los tests estadísticos sea útil (Sapkota, 2016).

Por último, tenemos la minería de textos. Ésta permite estudiar la asociación entre genes y enfermedades midiendo la cantidad de veces que los genes aparecen nombrados junto a la enfermedad en la literatura biomédica. Tiene la desventaja de que necesita la existencia previa de muchos estudios publicados acerca de la enfermedad; pero a cambio es un método muy rápido y de bajo coste que, al menos en ocasiones, puede producir resultados similares a los de los GWAS (Bouaziz, 2018).

Este último método, la minería de textos, es el que hemos elegido para este trabajo. Es el más adecuado teniendo en cuenta el corto espacio de tiempo en el que tiene que realizarse el trabajo de fin de máster, y que nos encontramos en medio de una emergencia sanitaria que desaconseja o impide el acceso a un grupo de trabajo en laboratorio y favorece el trabajo desde casa. En una situación como esta las ventajas de la minería de textos la hacen muy adecuada: es un método rápido y se basa en el análisis de trabajos ya publicados, por lo que no es necesario acceder a muestras ni a equipos de laboratorio. Tan sólo es necesario tener acceso a Internet y a un ordenador doméstico, ambas cosas al alcance inmediato del autor de este trabajo.

La herramienta resultado de este trabajo se desarrollará utilizando el lenguaje de programación R. En primer lugar, debido a que dispone una amplia variedad de librerías especialmente enfocadas a la minería de textos en general (*lsa, tidytext, tm*) y al acceso a datos de PubMed y el NCBI *(easyPubmed, pubmed.mineR, bibliometrix, rentrez, RISmed*). Y, en segundo lugar, debido a que es el lenguaje mejor conocido por el autor.

El flujo de trabajo se basará en el expuesto en Rani et al. (2015) y Liu (2016), y constará de los siguientes pasos:

1. Reducción de la información. En esta fase se descargará un pequeño subgrupo de abstracts de todos los disponibles en la base de datos PubMed. Para seleccionar dichos abstracts de interés se usarán palabras clave apropiadas al tema de este trabajo y se acotará por fechas. Estos abstracts conformarán el corpus primario de documentos sobre los que se realizará la minería de textos.

2. Preprocesado. Éste consistirá en el desglose de cada abstract del corpus primario en las frases y palabras que lo componen (*tokens*).

3. Reconocimiento y normalización de entidades. Identificaremos aquellos tokens correspondientes a nombres de genes; consolidaremos además sinónimos de genes a un identificador único de cada gen. El resultado de este paso es una lista de genes posiblemente asociados con la endometriosis según la literatura biomédica que compone el corpus primario.

4. Filtrado estadístico. Se realizará un filtrado de la lista mediante un test frente a la distribución hipergeométrica, para distinguir aquellos genes con menor probabilidad de haber sido recuperados por azar.

5. Caracterización de la lista final de genes. La lista de genes candidatos se examinará buscando categorías funcionales enriquecidas según los términos de ontología génica (GO, *gene ontology*). Se tendrán en cuenta las categorías de proceso biológico, componente celular y función molecular.

6. Visualización de la información en forma de tablas y gráficas (nubes de palabras con los genes, diagramas de barras de términos GO).

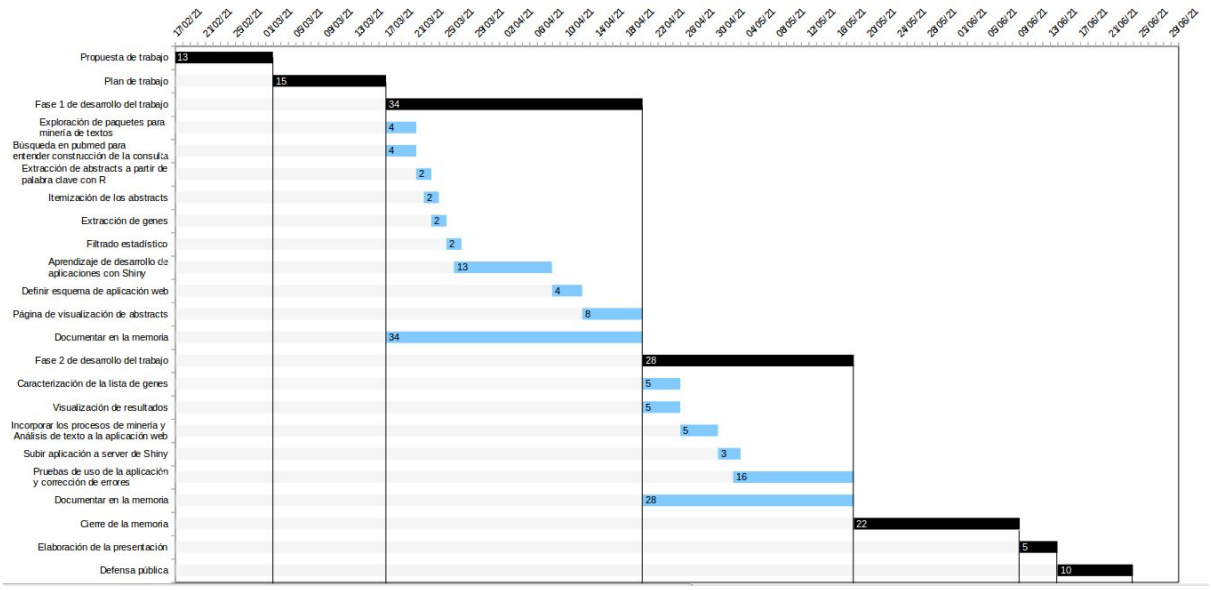
Finalmente, el flujo de trabajo descrito se integrará en el desarrollo de la aplicación web usando el paquete Shiny de R.

4. Planificación con hitos y temporización

4.1 Listado de tareas, plazos e hitos

Tarea	Inicio	Fin	Días	Hito
Propuesta de trabajo	**17/02/21**	**01/03/21**	**13**	**Entrega de PEC0**
Plan de trabajo	**02/03/21**	**16/03/21**	**15**	**Entrega de PEC1**
Fase 1 de desarrollo del trabajo	**17/03/21**	**19/04/21**	**34**	**Entrega de PEC2**
Exploración de paquetes para minería de textos	17/03/21	20/03/21	4	
Búsqueda en Pubmed para entender construcción de la consulta	17/03/21	20/03/21	4	
Extracción de abstracts a partir de palabra clave con R	21/03/21	22/03/21	2	
Preprocesado de los abstracts	22/03/21	23/03/21	2	
Extracción de genes	23/03/21	24/03/21	2	
Filtrado estadístico	25/03/21	26/03/21	2	
Aprendizaje de desarrollo de aplicaciones con Shiny	26/03/21	07/04/21	13	
Definir esquema de aplicación web	08/04/21	11/04/21	4	
Página de visualización de abstracts	12/04/21	19/04/21	8	
Documentar en la memoria	17/03/21	19/04/21	34	
Fase 2 de desarrollo del trabajo	**20/04/21**	**17/05/21**	**28**	**Entrega de PEC3**
Caracterización de la lista de genes	20/04/21	24/04/21	5	
Visualización de resultados	20/04/21	24/04/21	5	
Incorporar los procesos de minería y análisis de texto a la aplicación web	25/04/21	29/04/21	5	
Subir aplicación a server de Shiny	30/04/21	02/05/21	3	
Pruebas de uso de la aplicación y corrección de errores	02/05/21	17/05/21	16	
Documentar en la memoria	20/04/21	17/05/21	28	
Cierre de la memoria	**18/05/21**	**08/06/21**	**22**	**Entrega de PEC4**
Elaboración de la presentación	**09/06/21**	**13/06/21**	**5**	**Entrega de PEC5a**
Defensa pública	**14/06/21**	**23/06/21**	**10**	**Entrega de PEC5b**

4.2 Diagrama de Gantt



4.4 Análisis de riesgos

Descripción del riesgo	**Severidad**	**Probabilidad**	**Mitigación**
No poder desarrollar la aplicación web mediante Shiny	Moderada	Baja	Publicar los resultados en forma de página web estática.
Alguna de las librerías no funciona o es incompatible con el resto	Alta	Moderada	Buscar librerías alternativas. Examinar el código e intentar corregir el problema.
Corpus primario muy extenso provoca un excesivo tiempo de procesamiento.	Modada	Alta	Restringir el periodo de tiempo de publicación de los artículos cuyos abstracts formarán el corpus primario. Utilizar términos de busca menos generales (e.g., en lugar de usar el término “endometriosis”, utilizar términos como “endometriosis diagnosis”, “endometriosis biomarkers”, etcétera.
El corpus usado contiene palabras y siglas que, por casualidad, coinciden con el símbolo o nombre de genes.	Moderada	Baja	Buscar listados ya existentes de tales palabras, y usarlos para filtrar los resultados. Examinar el contexto (frases) en el que aparecen los genes de la lista final, detectar qué genes son ‘falsos’ y generar una lista con los mismos para usarla de filtro.

5. Resultados esperados:

1. **Plan de trabajo**. Documento en donde se enumeran y explican las tareas y sus tiempos estimados de duración.
2. **Memoria**. Documento en formato PDF con el desarrollo del trabajo de fin de máster. Se detallarán las partes más importantes de los scripts programados en R, en forma de pseudocódigo.
3. **Aplicación web interactiva** creada para realizar el proceso de minería de textos y visualizar los resultados. Se proporcionará un enlace a la web de la aplicación.
4. **Código fuente**: Un apéndice de la memoria tendrá el enlace a un repositorio de GitHub en el que se habrá publicado el código completo de la aplicación web.
5. **Presentación virtual**: El autor del trabajo presentará y explicará el mismo a través de una presentación de diapositivas.

6. Estructuración del proyecto

1. Resumen
2. Introducción
    - 2.1 Contexto y justificación del Trabajo
    - 2.2 Objetivos del Trabajo
    - 2.3 Enfoque y método seguido
    - 2.4 Planificación del Trabajo
    - 2.5 Breve sumario de productos obtenidos
    - 2.6 Breve descripción de los otros capítulos de la memoria
3. Estado del arte
4. Metodología
5. Resultados
6. Discusión
7. Valoración económica (Sólo si es necesario)
8. Conclusiones
4. Glosario
5. Bibliografía
6. Anexos

7.Bibliografía

Aggarwal, Charu C., ed. Mining Text Data. New York, NY: Springer, 2012.

Bouaziz, J., R. Mashiach, S. Cohen, A. Kedem, A. Baron, M. Zajicek, I. Feldman, D. Seidman, and D. Soriano. “How Artificial Intelligence Can Improve Our Understanding of the Genes Associated with Endometriosis: Natural Language Processing of the PubMed Database.” BioMed Research International 2018 (March 20, 2018). https://doi.org/10.1155/2018/6217812.

Chapron, Charles, Louis Marcellin, Bruno Borghese, and Pietro Santulli. “Rethinking Mechanisms, Diagnosis and Management of Endometriosis.” Nature Reviews. Endocrinology 15, no. 11 (November 2019): 666–82. https://doi.org/10.1038/s41574-019-0245-z.

Liu, Ji-Long, and Miao Zhao. “A PubMed-Wide Study of Endometriosis.” Genomics 108, no. 3–4 (October 2016): 151–57. https://doi.org/10.1016/j.ygeno.2016.10.003.

Mafra, Fernanda, Diego Mazzotti, Renata Pellegrino, Bianca Bianco, Caio Parente Barbosa, Hakon Hakonarson, and Denise Christofolini. “Copy Number Variation Analysis Reveals Additional Variants Contributing to Endometriosis Development.” Journal of Assisted Reproduction and Genetics 34, no. 1 (January 2017): 117–24. https://doi.org/10.1007/s10815-016-0822-1.

Malvezzi, Helena, Eliana Blini Marengo, Sérgio Podgaec, and Carla de Azevedo Piccinato. “Endometriosis: Current Challenges in Modeling a Multifactorial Disease of Unknown Etiology.” Journal of Translational Medicine 18, no. 1 (August 12, 2020): 311. https://doi.org/10.1186/s12967-020-02471-0.

National Library of Medicine. “About - PubMed.” Accessed March 9, 2021. https://pubmed.ncbi.nlm.nih.gov/about/.

National Library of Medicine. “MEDLINE Overview.” Accessed March 9, 2021. https://www.nlm.nih.gov/medline/medline_overview.html.

Rani, Jyoti, Ab Rauf Shah, and Srinivasan Ramachandran. “Pubmed.MineR: An R Package with Text-Mining Algorithms to Analyse PubMed Abstracts.” Journal of Biosciences 40, no. 4 (October 2015): 671–82. https://doi.org/10.1007/s12038-015-9552-2.

Rolla, Edgardo. “Endometriosis: Advances and Controversies in Classification, Pathogenesis, Diagnosis, and Treatment.” F1000Research 8 (2019). https://doi.org/10.12688/f1000research.14817.1.

Sapkota, Yadav, Valgerdur Steinthorsdottir, Andrew P. Morris, Amelie Fassbender, Nilufer Rahmioglu, Immaculata De Vivo, Julie E. Buring, et al. “Meta-Analysis Identifies Five Novel Loci Associated with Endometriosis Highlighting Key Genes Involved in Hormone Metabolism.” Nature Communications 8, no. 1 (August 2017): 15539. https://doi.org/10.1038/ncomms15539.

Spielman, R. S., and W. J. Ewens. “The TDT and Other Family-Based Tests for Linkage Disequilibrium and Association.” *American Journal of Human Genetics* 59, no. 5 (November 1996): 983–89. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1914831/